SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 1–31

TOPOLOGICAL SENSITIVITY AND SHAPE OPTIMIZATION FOR THE STOKES EQUATIONS*

PH. GUILLAUME† AND K. SID IDRIS†

Abstract. The topological sensitivity analysis provides an asymptotic expansion of a shape function with respect to the insertion of a small hole or obstacle inside a domain. This expansion can then be used for shape optimization. In this paper, such an expansion is obtained for the Stokes equations with general shape functions and arbitrarily shaped holes. A numerical example illustrates the use of the topological sensitivity in a shape optimization problem.

Key words. topological sensitivity, topological derivative, shape optimization, design sensitivity, Stokes equations

AMS subject classifications. 49Q10, 49Q12, 74P05, 74P10, 74P15

DOI. 10.1137/S0363012902411210

1. Introduction.

The topological sensitivity analysis consists of studying the variation of a cost function with respect to a modification of the topology of a domain. It is a basic tool for topological shape optimization, in that it provides a "descent direction" for updating the shape of the domain. In this paper, we consider the case of the Stokes equations. The shape optimization problem consists of minimizing a function $j(\Omega) = J(\Omega, u_\Omega)$ where the solution u_Ω to the Stokes equations (2.1) is defined on a variable open and bounded subset Ω of $\mathbb{R}^n$. For $\varepsilon > 0$, let $\Omega_\varepsilon = \Omega \backslash (\overline{x_0 + \varepsilon\omega})$ be the subset obtained by inserting a small obstacle $\overline{x_0 + \varepsilon\omega}$ into Ω, where $x_0 \in \Omega$ and $\omega \subset \mathbb{R}^n$ is a fixed open and bounded subset containing the origin. Then, an asymptotic expansion of function j is obtained in the following form:

$$j(\Omega_\varepsilon) = j(\Omega) + \varepsilon \delta j(x_0) + o(\varepsilon).$$

The "topological sensitivity" $\delta j(x_0)$ provides information for inserting a small obstacle at x_0: if $\delta j(x_0) < 0$, then $j(\Omega_\varepsilon) < j(\Omega)$ for small ε. More generally, function δj can be used like a descent direction in an optimization process. The step length then consists of choosing the size of the obstacle which is added, located where $\delta j(x)$ is the most negative. For example, in the case of a circular obstacle and $n = 3$, if the cost function involves u_Ω but not Du_Ω, then the first variation of function j reads

$$j(\Omega_\varepsilon) = j(\Omega) + 6\pi\nu\varepsilon\, u_\Omega(x_0).v_\Omega(x_0) + o(\varepsilon),$$

where v_Ω is the adjoint state (see Proposition 4.3). It is interesting to observe that the resulting optimality condition $u_\Omega.v_\Omega \geq 0$ is almost the same as the one given by Buttazzo and Dal Maso [4] in the case of the Laplace equation. When the cost function involves Du_Ω, there is an additional term (see Proposition 4.4).

An asymptotic theory for partial differential equations defined on a singular perturbed domain has been developed by several authors; see, for example, [12]. In the context of shape optimization, the topological sensitivity was introduced by Schumacher [14] followed by Sokolowski and Zochowski [15], who studied the effect of

*Received by the editors July 10, 2002; accepted for publication (in revised form) August 13, 2003; published electronically May 25, 2004.
http://www.siam.org/journals/sicon/43-1/41121.html
†MIP, UMR 5640, INSA, Département de Mathématiques, Complexe scientifique de Rangueil, 31077 Toulouse Cedex 4, France (guillaum@gmm.insa-tlse.fr).

removing a small part of material in structural mechanics. A topological sensitivity framework using an adaptation of the adjoint method [5] and a truncation technique was then introduced by Masmoudi [11] in the case of the Laplace equation with a circular hole and a Dirichlet condition on the boundary of the hole. It was generalized in [9] to the elasticity equations in the case of arbitrarily shaped holes and a Neumann boundary condition. In [10], we have analyzed the case of Poisson's Equation with noncircular holes (with a Dirichlet boundary condition), arbitrary right-hand sides, and cost functions. These results are generalized here to the Stokes equations which have some similar properties: in the three-dimensional case, one observes that the topological sensitivity $\delta j(x_0)$ depends on the shape of the obstacle, whereas it is independent of the shape in the two-dimensional case. This comes from the Stokes paradox, and was already observed by Allaire [1] in the case of periodically distributed holes. There is, however, a difference with Poisson's equation in the three-dimensional case: in the Stokes equations, the topological sensitivity $\delta j(x_0)$ may also depend on the orientation of the obstacle, whereas it was independent of the orientation in the (scalar) Poisson's equation.

First, the formulation of the problem is presented in section 2, and its truncated version is described in section 3. Section 4 presents the main results whose proofs are given in section 6. In the case of a circular obstacle, explicit expressions of the topological sensitivity are given for Dirichlet boundary conditions and for dimensions $n = 2$ or 3. Finally, numerical examples in section 5 illustrate the use of the topological sensitivity in shape optimization.

2. Formulation of the problem. Let Ω be an open and bounded subset of $\mathbb{R}^n$ with boundary Γ, $n = 2$ or 3. The Stokes equations [16] with homogeneous Dirichlet boundary conditions read

$$\left\{\begin{array}{rl} -\nu\Delta u_\Omega + \nabla p_\Omega = f & \text{in } \Omega, \\ \operatorname{div} u_\Omega = 0 & \text{in } \Omega, \\ u_\Omega = 0 & \text{on } \Gamma, \end{array}\right. \tag{2.1}$$

with $\nu > 0$. The case of an inhomogeneous boundary condition on Γ can be treated in a similar way by using a suitable change of the unknown velocity field, of the form $u_\Omega^r = u_\Omega + \mu$, where μ satisfies the given boundary condition. We suppose throughout this paper that $f \in L^q(\Omega)^n$ with $q > n/2$. These equations have a unique solution in $H_0^1(\Omega)^n \times L^2(\Omega)/\mathbb{R}$, and due to the regularity of f, the velocity field u_Ω is continuous in Ω [16, 3]. For a given $x_0 \in \Omega$, consider the modified open subset $\Omega_\varepsilon = \Omega \backslash \overline{\omega_\varepsilon}$, $\omega_\varepsilon = x_0 + \varepsilon\omega$, where ω is a fixed open and bounded subset of $\mathbb{R}^n$ containing the origin ($\omega_\varepsilon = \emptyset$ if $\varepsilon = 0$), whose boundary $\partial\omega$ is connected and piecewise of class $\mathcal{C}^1$ (cf. Figure 1). It is supposed that ε is small enough so that $\overline{\omega_\varepsilon} \subset \Omega$. The modified solution u_{Ω_ε}, p_{Ω_ε} satisfies

$$\left\{\begin{array}{rl} -\nu\Delta u_{\Omega_\varepsilon} + \nabla p_{\Omega_\varepsilon} = f & \text{in } \Omega_\varepsilon, \\ \operatorname{div} u_{\Omega_\varepsilon} = 0 & \text{in } \Omega_\varepsilon, \\ u_{\Omega_\varepsilon} = 0 & \text{on } \Gamma \cup \partial\omega_\varepsilon. \end{array}\right. \tag{2.2}$$

Note that for $\varepsilon = 0$, one has $u_{\Omega_0} = u_\Omega$ and $p_{\Omega_0} = p_\Omega$.

Consider now a cost function $j(\varepsilon)$ of the form

$$j(\varepsilon) = \widetilde{J}_\varepsilon(u_{\Omega_\varepsilon}) \tag{2.3}$$

where $\widetilde{J}_\varepsilon$ is defined on $H_0^1(\Omega_\varepsilon)^n$ for $\varepsilon \geq 0$. We aim to obtain an asymptotic expansion of j with respect to ε. The velocity field u_{Ω_ε} is defined on the variable open subset Ω_ε;

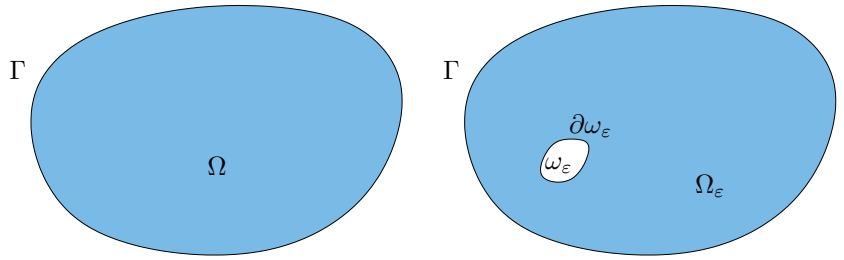


FIG. 1. *The initial domain and the same domain after inclusion of an obstacle.*

thus it belongs to a functional space which depends on ε. In order to obtain an asymptotic expansion of j, we need to work on a fixed functional space (cf. Lemma 6.1). Such a functional space can be constructed by using the domain truncation technique described in the next section (see also [11] and [9]). This truncation is needed only for analysis, and will never be used for practical computations. During the optimization process, the two systems which have to be solved at each step are (2.1) and (4.14).

3. The truncated problem. Let $R > 0$ be such that the closed ball $\overline{B}(x_0, R)$ is included in Ω. The truncated open subset Ω_R and D_ε (cf. Figure 2) are defined by

$$\Omega_R = \Omega \backslash \overline{B}(x_0, R), \quad D_\varepsilon = B(x_0, R) \backslash \overline{\omega_\varepsilon}.$$

Let Γ_R be the boundary of the ball $B(x_0, R)$. We will use the following space of traces on Γ_R:

$$H_V^{1/2}(\Gamma_R)^n = \left\{ \varphi \in H^{1/2}(\Gamma_R)^n ; \int_{\Gamma_R} \varphi . \mathbf{n} \, d\gamma(x) = 0 \right\}, \tag{3.1}$$

where $d\gamma(x)$ denotes the Lebesgue measure on the boundary. The normal $\mathbf{n}$ is chosen outward to D_ε on Γ_R and $\partial\omega_\varepsilon$, regardless of whether D_ε or Ω_R are considered. The dual space of $H_V^{1/2}(\Gamma_R)^n$ is denoted $H_V^{-1/2}(\Gamma_R)^n$. Here $x.y$ denotes the usual dot product of $\mathbb{R}^n$.

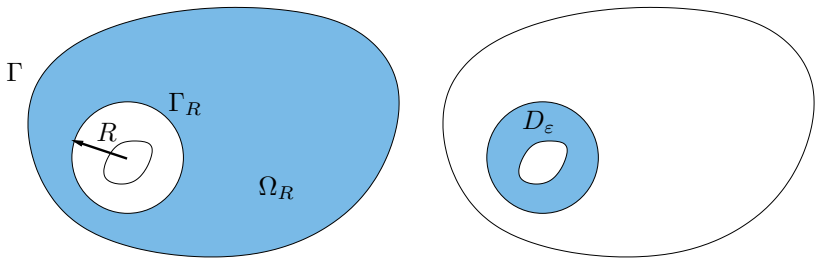


FIG. 2. *The truncated domain.*

For $\varphi \in H_V^{1/2}(\Gamma_R)^n$ and $\varepsilon > 0$, let $(u_\varepsilon^{f,\varphi}, p_\varepsilon^{f,\varphi}) \in H^1(D_\varepsilon)^n \times L_0^2(D_\varepsilon)$ be the solution to the following problem: find $u_\varepsilon^{f,\varphi}$, $p_\varepsilon^{f,\varphi}$ such that

$$\left\{ \begin{array}{rcl} -\nu\Delta u_\varepsilon^{f,\varphi} + \nabla p_\varepsilon^{f,\varphi} = f & \text{in} & D_\varepsilon, \\ \operatorname{div} u_\varepsilon^{f,\varphi} = 0 & \text{in} & D_\varepsilon, \\ u_\varepsilon^{f,\varphi} = \varphi & \text{on} & \Gamma_R, \\ u_\varepsilon^{f,\varphi} = 0 & \text{on} & \partial\omega_\varepsilon. \end{array} \right. \tag{3.2}$$

As usual, the space $L^2(D_\varepsilon)/\mathbb{R}$ is identified to $L^2_0(D_\varepsilon) = \{p \in L^2(D_\varepsilon)\,;\, \int_{D_\varepsilon} p\,dx = 0\}$. This problem has a unique solution [16]. For $\varepsilon = 0$, $(u_0^{f,\varphi}, p_0^{f,\varphi})$ is the solution to

$$\left\{\begin{array}{rcl} -\nu\Delta u_0^{f,\varphi} + \nabla p_0^{f,\varphi} = f & \text{in} & D_0, \\ \operatorname{div} u_0^{f,\varphi} = 0 & \text{in} & D_0, \\ u_0^{f,\varphi} = \varphi & \text{on} & \Gamma_R. \end{array}\right. \tag{3.3}$$

Clearly we have

$$u_\varepsilon^{f,\varphi} = u_\varepsilon^{f,0} + u_\varepsilon^{0,\varphi}, \quad p_\varepsilon^{f,\varphi} = p_\varepsilon^{f,0} + p_\varepsilon^{0,\varphi}. \tag{3.4}$$

For $\varepsilon \geq 0$, the Dirichlet-to-Neumann operator T_ε is defined by

$$\begin{array}{rcl} T_\varepsilon : H_V^{1/2}(\Gamma_R)^n & \longrightarrow & H_V^{-1/2}(\Gamma_R)^n, \\ \varphi & \longmapsto & T_\varepsilon\varphi = (\nu Du_\varepsilon^{0,\varphi} - p_\varepsilon^{0,\varphi} I)\mathbf{n}; \end{array}$$

that is,

$$\langle T_\varepsilon\varphi, \psi\rangle_{-1/2,1/2} = \nu \int_{D_\varepsilon} Du_\varepsilon^{0,\varphi} : Du_\varepsilon^{0,\psi}\,dx. \tag{3.5}$$

Here and in what follows, I denotes the identity matrix, $A : B = \sum_{ij} A_{ij}B_{ij}$ for $A,\ B \in \mathcal{M}_n(\mathbb{R})$, and $Du = (\partial_j u_i)_{i,j=1,n}$ denotes the Jacobian matrix of u. One can observe that $T_\varepsilon\varphi$ is completely determined by $u_\varepsilon^{0,\varphi}$.

Function $f_\varepsilon \in H_V^{-1/2}(\Gamma_R)^n$ is defined by

$$f_\varepsilon = (-\nu Du_\varepsilon^{f,0} + p_\varepsilon^{f,0} I)\mathbf{n};$$

that is,

$$\langle f_\varepsilon, \psi\rangle_{-1/2,1/2} = \int_{D_\varepsilon} f.u_\varepsilon^{0,\psi}\,dx. \tag{3.6}$$

Notice that

$$\int_{D_\varepsilon} Du_\varepsilon^{f,0} : Du_\varepsilon^{0,\psi}\,dx = 0. \tag{3.7}$$

Hence, we have

$$(\nu Du_\varepsilon^{f,\varphi} - p_\varepsilon^{f,\varphi} I)\mathbf{n} = T_\varepsilon\varphi - f_\varepsilon.$$

For $\varepsilon \geq 0$, we can now define the solution $(u_\varepsilon,\, p_\varepsilon) \in H^1(\Omega_R)^n \times L^2(\Omega_R)$ to the truncated problem

$$\left\{\begin{array}{rcl} -\nu\Delta u_\varepsilon + \nabla p_\varepsilon = f & \text{in} & \Omega_R, \\ \operatorname{div} u_\varepsilon = 0 & \text{in} & \Omega_R, \\ u_\varepsilon = 0 & \text{on} & \Gamma, \\ -(\nu Du_\varepsilon - p_\varepsilon I)\mathbf{n} + T_\varepsilon u_\varepsilon = f_\varepsilon & \text{on} & \Gamma_R. \end{array}\right. \tag{3.8}$$

The associated variational formulation is as follows: find $u_\varepsilon \in \mathcal{V}_R$ such that

$$a_\varepsilon(u_\varepsilon, v) = l_\varepsilon(v) \qquad \forall v \in \mathcal{V}_R, \tag{3.9}$$

where the functional space $\mathcal{V}_R$, the bilinear form a_ε, and the linear form l_ε are defined by

$$\mathcal{V}_R = \left\{ u \in H^1(\Omega_R)^n;\ \operatorname{div} u = 0,\ u = 0 \text{ on } \Gamma \right\},$$

$$a_\varepsilon(u,v) = \nu \int_{\Omega_R} Du : Dv\ dx + \int_{\Gamma_R} T_\varepsilon u.v\ d\gamma(x), \tag{3.10}$$

$$l_\varepsilon(v) = \int_{\Omega_R} f.v\ dx + \int_{\Gamma_R} f_\varepsilon.v\ d\gamma(x). \tag{3.11}$$

Symmetry, continuity, and coercivity of a_ε and continuity of l_ε follow directly from (3.5) and (3.6). We suppose that the pressure p_{Ω_ε} solution to (2.2) is chosen in such a way that $\int_{D_\varepsilon} p_{\Omega_\varepsilon}\ dx = 0$. Recall that $f \in L^q(\Omega)^n$ with $q > n/2$.

PROPOSITION 3.1. *Let $\varepsilon \geq 0$. Problems* (2.2) *and* (3.8) *have a unique solution. Moreover, the restriction to Ω_R of the solution u_{Ω_ε}, p_{Ω_ε} to* (2.2) *is the solution u_ε, p_ε to* (3.8)*, and we have on D_ε*

$$(u_{\Omega_\varepsilon})_{|D_\varepsilon} = u_\varepsilon^{f,\varphi}, \quad (p_{\Omega_\varepsilon})_{|D_\varepsilon} = p_\varepsilon^{f,\varphi}, \tag{3.12}$$

where φ is the trace of u_ε on Γ_R.

Proof. Problem (2.2) has a unique solution u_{Ω_ε} [16], and it follows from Lax–Milgram's theorem and [16] that Problem (3.8) has a unique solution u_ε. Let $\varphi = (u_{\Omega_\varepsilon})_{|\Gamma_R}$ and $u_R = (u_{\Omega_\varepsilon})_{|\Omega_R}$. Clearly (3.12) holds for this φ, and it remains to prove that $u_R = u_\varepsilon$. Let $v \in \mathcal{V}_R$ and $\psi = v_{|\Gamma_R}$. Its extension by $u^{0,\psi}$ on D_ε is still denoted by v, and it is divergence free on D_ε. Using (3.5), (3.6), (3.7), and the definition of u_{Ω_ε}, we have

$$\begin{aligned}
\nu \int_{\Omega_R} Du_R\ :\ Dv\ dx &+ \int_{\Gamma_R} (T_\varepsilon u_R - f_\varepsilon).v\ d\gamma(x) \\
&= \nu \int_{\Omega_R} Du_R : Dv\ dx + \int_{D_\varepsilon} \nu Du^{0,\varphi} : Du^{0,\psi} - f.u^{0,\psi}\ dx \\
&= \nu \int_{\Omega_R} Du_R : Dv\ dx + \int_{D_\varepsilon} \nu(Du^{f,\varphi} - Du^{f,0}) : Du^{0,\psi} - f.u^{0,\psi}\ dx \\
&= \nu \int_{\Omega_\varepsilon} Du_{\Omega_\varepsilon} : Dv\ dx - \int_{D_\varepsilon} f.u^{0,\psi}\ dx \\
&= \int_{\Omega_\varepsilon} f.v\ dx - \int_{D_\varepsilon} f.v\ dx = \int_{\Omega_R} f.v\ dx.
\end{aligned}$$

This proves that u_R is a solution to (3.9), and $u_R = u_\varepsilon$ follows from uniqueness of this solution. ∎

We now have at our disposal the fixed functional space $\mathcal{V}_R$ (independent of ε) required by Lemma 6.1. The cost function (2.3) can be redefined in the following way: for $u \in \mathcal{V}_R$, let $\widetilde{u}_\varepsilon \in H^1(\Omega_\varepsilon)^n$ be the extension of u which coincides with u on Ω_R and with $u_\varepsilon^{f,\varphi}$ on D_ε for $\varphi = u_{|\Gamma_R}$. Then, a function J_ε can be defined on $\mathcal{V}_R$ by

$$J_\varepsilon(u) = \widetilde{J}_\varepsilon(\widetilde{u}_\varepsilon). \tag{3.13}$$

Particularly, it follows from the previous proposition that

$$j(\varepsilon) = \widetilde{J}_\varepsilon(u_{\Omega_\varepsilon}) = J_\varepsilon(u_\varepsilon). \tag{3.14}$$

Notice that $J_\varepsilon(u_\varepsilon)$ is independent of the choice of R. For example, for a given target function u_d, if

$$\widetilde{J}_\varepsilon(u_{\Omega_\varepsilon}) = \int_{\Omega_\varepsilon} |u_{\Omega_\varepsilon} - u_\mathrm{d}|^2 dx,$$

then we have

$$J_\varepsilon(u) = \int_{\Omega_R} |u - u_\mathrm{d}|^2 dx + \int_{D_\varepsilon} |u_\varepsilon^{f,\varphi} - u_\mathrm{d}|^2 dx, \quad u \in \mathcal{V}_R \text{ and } \varphi = u_{|\Gamma_R}.$$

4. Asymptotic expansion of the cost function. This section presents the main results of the paper. All asymptotic expansions concern the homogeneous problem (2.1) and involve its solution u_Ω, a cost function J_ε (or $\widetilde{J}_\varepsilon$), and the associated adjoint state v_Ω solution to (4.14). However, they remain valid for inhomogeneous problems (on Γ), provided that all data are written with respect to the inhomogeneous formulation: u_Ω has to be replaced by the inhomogeneous solution u_Ω^r and J_ε by the cost function depending on this solution u_Ω^r. The adjoint state corresponding to the inhomogeneous problem is the same as the one corresponding to the associated homogeneous problem.

A general statement (Theorem 4.1) is followed by applications to two classes of cost functions, the first one involving u, the second one involving Du. Most proofs are reported in section 6. Henceforth we have to distinguish the cases $n = 2$ and $n = 3$. This is due to the fact that the fundamental solutions to the Stokes equations in $\mathbb{R}^2$ and $\mathbb{R}^3$ have an essentially different asymptotic behavior at infinity, and Problem (4.1) generally has no solution if $n = 2$.

4.1. The three-dimensional case. Possibly changing the coordinate system, we can suppose for convenience that $x_0 = 0$. Let v_ω, p_ω be the solution to the exterior problem

$$\left\{\begin{array}{rl} -\nu\Delta v_\omega + \nabla p_\omega = 0 & \text{in} \quad \mathbb{R}^3\backslash\overline{\omega}, \\ \operatorname{div} v_\omega = 0 & \text{in} \quad \mathbb{R}^3\backslash\overline{\omega}, \\ v_\omega = 0 & \text{at infinity}, \\ v_\omega = u_\Omega(x_0) & \text{on} \quad \partial\omega, \end{array}\right. \tag{4.1}$$

where u_Ω is the solution to (2.1). Due to $f \in L^q(\Omega)^n$ with $q > n/2$, u_Ω is continuous in Ω and the above boundary condition is well defined.

Functions v_ω, p_ω can be expressed by a single layer potential on $\partial\omega$. For $y \in \mathbb{R}^3\backslash\{0\}$, let

$$E(y) = \frac{1}{8\pi\nu r}\left(I + \mathbf{e}_r\mathbf{e}_r^T\right), \; P(y) = \frac{y}{4\pi r^3}, \tag{4.2}$$

with $r = \|y\|$ and $\mathbf{e}_r = y/\|y\|$. It is a fundamental solution system to the Stokes equations in $\mathbb{R}^3$; that is

$$-\nu\Delta E_j + \nabla P_j = \delta\mathbf{e}_j, \tag{4.3}$$

where E_j denotes the jth column of E, $(\mathbf{e}_j)_{j=1}^3$ is the canonical basis of $\mathbb{R}^3$, and δ is the Dirac distribution. Then, functions v_ω, p_ω read (recall that $x.y$ denotes the usual dot product of $\mathbb{R}^3$) as follows:

$$\begin{aligned} v_\omega(y) &= \int_{\partial\omega} E(y-x)t_\omega(x)\, d\gamma(x), \quad y \in \mathbb{R}^3\backslash\overline{\omega}, \\ p_\omega(y) &= \int_{\partial\omega} P(y-x).t_\omega(x)\, d\gamma(x), \quad y \in \mathbb{R}^3\backslash\overline{\omega}, \end{aligned} \tag{4.4}$$

where $t_\omega \in H^{-1/2}(\partial\omega)^3$ is a solution to the boundary integral equation (see [8], Chap. XI-B, sect. 5)

$$\int_{\partial\omega} E(y-x)t_\omega(x)\,d\gamma(x) = u_\Omega(x_0) \quad \forall y \in \partial\omega. \tag{4.5}$$

Function t_ω is determined up to a function proportional to the normal; hence it is unique in $H^{-1/2}(\partial\omega)^3/\mathbb{R}\mathbf{n}$, which can be identified to $H_V^{-1/2}(\partial\omega)^3 = (\ker l_\mathbf{n})'$ for $l_\mathbf{n}(\varphi) = \int_{\partial\omega} \varphi.\mathbf{n}\,d\gamma(x)$.

For x bounded (and $r = \|y\|$), we have

$$E(y-x) = E(y) + O\left(\frac{1}{r^2}\right), \quad P(y-x) = P(y) + O\left(\frac{1}{r^3}\right),$$

from which follows the asymptotic expansion at infinity of functions v_ω and p_ω:

$$v_\omega(y) = V_\omega(y) + R_\omega(y),\ p_\omega(y) = P_\omega(y) + S_\omega(y), \tag{4.6}$$

$$V_\omega(y) = E(y)A_\omega(u_\Omega(x_0)), \quad P_\omega(y) = P(y).A_\omega(u_\Omega(x_0)), \tag{4.7}$$

$$A_\omega(u_\Omega(x_0)) = \int_{\partial\omega} t_\omega(x)\,d\gamma(x) \in \mathbb{R}^3, \tag{4.8}$$

$$R_\omega(y) = O\left(\frac{1}{r^2}\right), \quad S_\omega(y) = O\left(\frac{1}{r^3}\right).$$

Notice that $V_\omega \in L^m_{\mathrm{loc}}(\mathbb{R}^3)^3$ for all $m < 3$. Clearly, the function $\alpha \longmapsto A_\omega(\alpha)$ is linear on $\mathbb{R}^3$, and the vector $A_\omega(\alpha)$ depends on the shape of ω. For example, if ω is changed in $k\omega$, $k > 0$, then $v_{k\omega}(ky) = v_\omega(y)$ in (4.1), and it follows from (4.5) that $kt_{k\omega}(kx) = t_\omega(x)$ for $x \in \partial\omega$. Then using (4.8) we obtain $A_{k\omega}(u_\Omega(x_0)) = kA_\omega(u_\Omega(x_0))$. More generally, $A_\omega(\alpha)$ may depend on the orientation of ω, contrary to the scalar case like Poisson's equations [10]. Next we consider W_ω, Q_ω the solution to

$$\begin{cases} -\nu\Delta W_\omega + \nabla Q_\omega = 0 & \text{in} \quad D_0, \\ \operatorname{div} W_\omega = 0 & \text{in} \quad D_0, \\ W_\omega = V_\omega & \text{on} \quad \Gamma_R. \end{cases} \tag{4.9}$$

The main result is the following. It is based on the fact that

$$\varepsilon(W_\omega - V_\omega)_{|D_\varepsilon} \tag{4.10}$$

is the "first order approximation" of $(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi})_{|D_\varepsilon}$ with $\varphi = (u_\Omega)_{|\Gamma_R}$, in a sense which will be stated precisely in section 6. The stronger hypothesis $f \in L^q(\Omega)^n$, $q > n$, is used in the study of the variation of the linear form l_ε (3.11) (cf. Proposition 6.8), which involves the $\mathcal{C}^1$ norm of u_0 around x_0. If l_ε does not depend on ε (which happens, for example, if f vanishes on D_0), then $f \in L^q(\Omega)^n$, $q > n/2$ is sufficient.

THEOREM 4.1. *Let $f \in L^q(\Omega)^n$ with $q > n$ and let J_ε be a function defined on $\mathcal{V}_R$ for all $\varepsilon \geq 0$. Suppose that for all $v \in \mathcal{V}_R$ and $\varepsilon > 0$, one has*

$$J_\varepsilon(v) - J_0(u_0) = DJ_0(u_0)(v-u_0) + \varepsilon\,\delta J(u_0) + o(\varepsilon + \|v-u_0\|_{\mathcal{V}_R}), \tag{4.11}$$

where $DJ_0(u_0)$ is linear and continuous on $\mathcal{V}_R$, and u_ε, $\varepsilon \geq 0$, is the solution to (3.9). *Let $v_0 \in \mathcal{V}_R$ be the solution to the adjoint equation*

$$a_0(w, v_0) = -DJ_0(u_0)w \quad \forall w \in \mathcal{V}_R. \tag{4.12}$$

Let $j(\varepsilon) = J_\varepsilon(u_\varepsilon)$ for $\varepsilon \geq 0$. Then function j has the following asymptotic expansion:

$$j(\varepsilon) = j(0) + \varepsilon\, \delta j(x_0) + o(\varepsilon)$$

with

$$\delta j(x_0) = \int_{\Gamma_R} ((\nu DW_\omega - Q_\omega I) - (\nu DV_\omega - P_\omega I))\, \mathbf{n}.v_0\, d\gamma(x) + \delta J(u_0). \tag{4.13}$$

Function $\delta j(x_0)$ is called the "topological sensitivity" or the "topological gradient." As j is usually independent of R (at least when it is of the form (3.14), which is the "natural" way of posing the problem) and $\delta j(x_0)$ is independent of ε, it follows from the uniqueness of an asymptotic expansion that $\delta j(x_0)$ is also usually independent of R. This is not necessarily true for the terms $\delta a(u_0, v_0)$, $\delta l(v_0)$ (see section 6), or $\delta J(u_0)$ considered separately, because a, l, and J do depend on R.

During an optimization process, what is in fact computed is the solution u_Ω to (2.1) and the adjoint state $v_\Omega \in \mathcal{V}_0$, which is the solution to

$$\nu \int_\Omega Dw : Dv_\Omega\, dx = -D\widetilde{J}_0(u_\Omega)w, \quad \forall w \in \mathcal{V}_0, \tag{4.14}$$

with $\mathcal{V}_0 = \{v \in H_0^1(\Omega)^3;\, \operatorname{div} v = 0 \text{ in } \Omega\}$. As observed in Proposition 3.1, u_0 is the restriction to Ω_R of u_Ω. Similarly, v_0 is the restriction to Ω_R of v_Ω, which can be proved in the same way as in [9]. Hence, the basic property of an adjoint technique is here satisfied, in that function u_Ω (or u_0) and the adjoint state v_Ω (or v_0) do not depend on x_0. Thus *only two systems must be solved* in order to compute the topological sensitivity $\delta j(x)$ for all $x \in \Omega$. Moreover, there exists a unique $q_\Omega \in L^2(\Omega)/\mathbb{R}$ such that

$$\nu \int_\Omega Dw : Dv_\Omega - \int_\Omega q_\Omega \operatorname{div} w = -D\widetilde{J}_0(u_\Omega)\, w, \qquad \forall w \in H_0^1(\Omega)^3, \tag{4.15}$$

and (4.13) can be expressed in the following way (see Corollary 4.2). Proposition 4.3 will show that in fact the two last terms in the right-hand side of (4.16) cancel each other for a large class of cost functions which do not involve Du_Ω. The regularity of $\nu\Delta v_\Omega - \nabla q_\Omega$ depends on u_Ω and on the cost function J; some examples are provided in sections 4.1.1 and 4.1.2.

COROLLARY 4.2. *Under the assumptions of Theorem* 4.1*, if $\nu\Delta v_\Omega - \nabla q_\Omega \in L^q(D_0)^n$ with $q > n/2$, then*

$$\delta j(x_0) = A_\omega(u_\Omega(x_0)).v_\Omega(x_0) + \int_{D_0} (\nu\Delta v_\Omega - \nabla q_\Omega).(V_\omega - W_\omega)\, dx + \delta J(u_0). \tag{4.16}$$

If ω is the unit ball $B(0,1)$, then $v_\omega(y)$, $t_\omega(y)$, and $A_\omega(u_\Omega(x_0))$ are given explicitly by

$$\begin{aligned} v_\omega(y) &= \pi\nu(6E + \Delta E)u_\Omega(x_0), \\ t_\omega(y) &= \frac{3\nu}{2} u_\Omega(x_0), \quad \forall y \in \partial\omega, \\ A_\omega(u_\Omega(x_0)) &= 6\pi\nu u_\Omega(x_0). \end{aligned}$$

Proof. Thanks to Green's formula and (4.9) (with $V_\omega = W_\omega$ on Γ_R), (4.13) also

reads as follows:

$$\begin{aligned}\delta j(x_0) &= \int_{\Gamma_R} ((\nu DW_\omega - Q_\omega I) - (\nu DV_\omega - P_\omega I))\,\mathbf{n}.v_\Omega\, d\gamma(x) + \delta J(u_0)\\ &= \int_{\Gamma_R} (\nu Dv_\Omega - q_\Omega I)\mathbf{n}.V_\omega\, d\gamma(x) - \int_{\Gamma_R} (\nu DV_\omega - P_\omega I)\mathbf{n}.v_\Omega\, d\gamma(x)\\ &\quad - \int_{D_0} (\nu \Delta v_\Omega - \nabla q_\Omega)\mathbf{n}.W_\omega\, dx + \delta J(u_0).\end{aligned}$$

Through a regularization and localization technique, it can be shown that

$$\begin{aligned}&\int_{\Gamma_R} (\nu Dv_\Omega - q_\Omega I)\mathbf{n}.V_\omega\, d\gamma(x) - \int_{\Gamma_R} (\nu DV_\omega - P_\omega I)\mathbf{n}.v_\Omega\, d\gamma(x)\\ &= \int_{D_0} (\nu \Delta v_\Omega - \nabla q_\Omega).V_\omega\, dx - \langle \nu \Delta V_\omega - \nabla P_\omega, \varphi v_\Omega\rangle,\end{aligned}$$

where $\varphi \in \mathcal{D}(D_0)^3$ satisfies $\varphi(x_0) = 1$. It follows from (4.7) and (4.3) that

$$\begin{aligned}\langle -\nu\Delta V_\omega + \nabla P_\omega, \varphi v_\Omega\rangle &= \langle -\nu\Delta(EA_\omega(u_\Omega(x_0))) + \nabla(P.A_\omega(u_\Omega(x_0))), \varphi v_\Omega\rangle\\ &= \sum_j A_\omega(u_\Omega(x_0))_j \,\langle -\nu\Delta E_j + \nabla P_j, \varphi v_\Omega\rangle\\ &= \sum_j A_\omega(u_\Omega(x_0))_j \,\langle \delta \mathbf{e}_j, \varphi v_\Omega\rangle\\ &= A_\omega(u_\Omega(x_0)).\, v_\Omega(x_0).\end{aligned}$$

This proves (4.16). For $\omega = B(0,1)$, function v_ω can be computed explicitly. For $y \neq 0$, we have

$$\begin{aligned}E(y) &= \frac{1}{8\pi\nu r}(I + \mathbf{e}_r\mathbf{e}_r^T),\\ \Delta E(y) &= \frac{1}{4\pi\nu r^3}(I - 3\,\mathbf{e}_r\mathbf{e}_r^T),\\ \pi\nu(6E + \Delta E)(y) &= \frac{1}{4}\left(\frac{3}{r} + \frac{1}{r^3}\right) I + \frac{1}{4}\left(\frac{3}{r} - \frac{3}{r^3}\right)\mathbf{e}_r\mathbf{e}_r^T.\end{aligned}$$

Hence $\pi\nu(6E+\Delta E)(y) = I$ on $\partial B(0,1)$, and it follows from the uniqueness of v_ω that

$$v_\omega(y) = \pi\nu(6E + \Delta E)(y)u_\Omega(x_0).$$

The expressions of $t_\omega(y)$ and $A_\omega(u_\Omega(x_0))$ follow straightforwardly from (4.5) and

$$\int_{\partial B(0,1)} E(y-x)\, d\gamma(x) = \frac{2}{3\nu} I \quad \forall y \in \partial B(0,1). \qquad \square$$

We now examine two classes of cost functions.

4.1.1. First example. It consists of functions of the form

$$\widetilde{J}_\varepsilon(u) = \int_{\Omega_\varepsilon} g(x, u(x))\, dx, \quad u \in H^1(\Omega_\varepsilon)^3. \tag{4.17}$$

The hypotheses on g are the following:

- For all $x \in \Omega$, function $s \mapsto g(x,s)$ is of class $\mathcal{C}^1$ on $\mathbb{R}^3$, its gradient being denoted by $\nabla_s g(x,s)$.
- For all $x \in \Omega$, function $s \mapsto \nabla_s g(x,s)$ is Lipschitz continuous, and there exists a constant M such that

$$|\nabla_s g(x,t) - \nabla_s g(x,s)| \leq M\,|t-s|\,, \quad \forall (x,s,t) \in \Omega \times \mathbb{R}^3 \times \mathbb{R}^3, \tag{4.18}$$

  where $|t|$ denotes the usual norm on $\mathbb{R}^n$.
- Function $x \mapsto \nabla_s g(x,0)$ belongs to $L^2(\Omega)^3$ and $x \mapsto g(x,0)$ belongs to $L^{3/2}(\Omega)$.

These hypotheses imply that for all $(x,s) \in \Omega \times \mathbb{R}^3$

$$|g(x,s)| \leq |g(x,0)| + |\nabla_s g(x,0).s| + \frac{M}{2}\,|s|^2\,, \tag{4.19}$$

$$|\nabla_s g(x,s)| \leq |\nabla_s g(x,0)| + M\,|s|\,, \tag{4.20}$$

and functions $x \mapsto g(x,u(x))$ and $x \mapsto |\nabla_s g(x,u(x))|^2$ are integrable on Ω for all $u \in L^2(\Omega)^3$. If $u \in L^6(\mathcal{O})^3$, $\mathcal{O} \subset\Omega$, then function $x \mapsto g(x,u(x))$ belongs to $L^{3/2}(\mathcal{O})$. The standard example

$$g(x,s) = |s - u_{\rm d}(x)|^2$$

satisfies these hypotheses if $u_{\rm d} \in L^3(\Omega)^3$.

PROPOSITION 4.3. *If these hypotheses are satisfied and if $f \in L^q(\Omega)^n$ with $q > n$, then*

$$\delta J(u_0) = \int_{D_0} \nabla_s g(x,u_\Omega).(W_\omega - V_\omega)\,dx,$$

the adjoint state $(v_\Omega,\, q_\Omega) \in \mathcal{V}_0 \times L^2(\Omega)/\mathbb{R}$ is the solution to

$$-\nu\Delta v_\Omega + \nabla q_\Omega = -\nabla_s g(x,u_\Omega),$$

and function j has the asymptotic expansion

$$j(\varepsilon) = j(0) + \varepsilon A_\omega(u_\Omega(x_0)).v_\Omega(x_0) + o(\varepsilon).$$

If ω is the unit ball $B(0,1)$, then

$$j(\varepsilon) = j(0) + 6\pi\nu\varepsilon\, u_\Omega(x_0).v_\Omega(x_0) + o(\varepsilon).$$

4.1.2. Second example. It consists of functions of the form

$$\widetilde{J}_\varepsilon(u) = \frac{1}{2}\int_{\Omega_\varepsilon} BD(u-u_{\rm d}) : D(u-u_{\rm d})\,dx, \quad u \in H^1(\Omega_\varepsilon)^3, \tag{4.21}$$

where $B \in W^{1,\infty}(\Omega, \mathbb{R}^{9\times 9})$ with $(BM)_{ij}(x) = \sum_{kl} b_{ijkl}(x) M_{kl}$ for $M \in \mathcal{M}_3(\mathbb{R})$, and $u_{\rm d} \in H^1(\Omega)^3$. The operator B is supposed to be symmetric; that is, $b_{ijkl} = b_{klij}$, or equivalently $BM : N = M : BN$ for all M, $N \in \mathcal{M}_3(\mathbb{R})$.

PROPOSITION 4.4. *If these hypotheses are satisfied and if f and $\Delta u_{\rm d}$ belong to $L^q(\Omega)^n$, $q > n$, then the adjoint state $(v_\Omega,\ q_\Omega) \in \mathcal{V}_0 \times L^2(\Omega)/\mathbb{R}$ is the solution to*

$$-\nu\Delta v_\Omega + \nabla q_\Omega = \operatorname{div}\left[BD(u_\Omega - u_{\rm d})\right],$$

the variation of J is given by

$$\delta J(u_0) = \int_{D_0} (\nu\Delta v_\Omega - \nabla q_\Omega).(W_\omega - V_\omega) + \frac{1}{2}\int_{\mathbb{R}^3\setminus\overline{\omega}} B(x_0)Dv_\omega(y) : Dv_\omega(y)\,dy,$$

and function j has the asymptotic expansion

$$j(\varepsilon) = j(0) + \varepsilon A_\omega(u_\Omega(x_0)).v_\Omega(x_0) + \frac{\varepsilon}{2}\int_{\mathbb{R}^3\setminus\overline{\omega}} B(x_0)Dv_\omega(y) : Dv_\omega(y)\,dy + o(\varepsilon).$$

If ω is the unit ball, then

$$v_\omega = \pi\nu(6E + \Delta E)u_\Omega(x_0) = \frac{3r^2+1}{4r^3}u_\Omega(x_0) + \frac{(3r^2-3)u_\Omega(x_0).\mathbf{e}_r}{4r^3}\mathbf{e}_r,$$

and the integral can be computed explicitly.

If $\overline{\omega}$ is the unit ball $\overline{B}(0,1)$, the integral has the form

$$\begin{aligned}&\int_{\mathbb{R}^3\setminus\overline{B}(0,1)} B(x_0)Dv_\omega(y) : Dv_\omega(y)\,dy\\ &= \sum_{ijklmn} (B_{ijkl}u_m u_n)(x_0)\int_{\mathbb{R}^3\setminus\overline{B}(0,1)} (\pi\nu)^2\partial_j(6E+\Delta E)_{im}\partial_l(6E+\Delta E)_{kn}\,dy,\end{aligned}$$

where $u_\Omega = (u_1, u_2, u_3)$, each term in $\partial_j(6E_{im} + \Delta E_{im})\partial_l(6E_{kn} + \Delta E_{kn})$ is of the form y^α/r^s with $s = |\alpha| + p$, $p = 4, 6$, or 8, and each y^α is of the form y_i^2, $y_i^2y_j^2$, or $y_i^2y_j^2y_k^2$ (the other integrals vanish). Then one can use the following formulas:

$$\int_{\mathbb{R}^3\setminus\overline{B}(0,1)} \frac{y_i^2}{r^s}dy = \frac{1}{3}\int_{\mathbb{R}^3\setminus\overline{B}(0,1)} \frac{1}{r^p}dy = \frac{4\pi}{3(p-3)}$$

$$\int_{\mathbb{R}^3\setminus\overline{B}(0,1)} \frac{y_i^2y_j^2}{r^s}dy = \begin{cases} \frac{4\pi}{5(p-3)} & \text{if } i = j,\\ \frac{4\pi}{15(p-3)} & \text{if } i \neq j,\end{cases}$$

$$\int_{\mathbb{R}^3\setminus\overline{B}(0,1)} \frac{y_i^2y_j^2y_k^2}{r^s}dy = \begin{cases} \frac{4\pi}{7(p-3)} & \text{if } i = j = k,\\ \frac{4\pi}{35(p-3)} & \text{if } i = j \neq k,\\ \frac{4\pi}{105(p-3)} & \text{if } i \neq j \neq k \neq i.\end{cases}$$

For example, if $BM(x) = C(x)M$ where $C(x) \in \mathcal{M}_3(\mathbb{R})$ is a symmetric matrix, then

$$j(\varepsilon) = j(0) + \varepsilon[6\pi\nu u_\Omega.v_\Omega + \frac{3\pi}{10}\operatorname{tr} C|u_\Omega|^2 + \frac{21\pi}{10}Cu_\Omega.u_\Omega](x_0) + o(\varepsilon).$$

4.2. The two-dimensional case. We briefly describe the transposition of the previous results to the two-dimensional case. As before, u_Ω and the adjoint state v_Ω, respectively, are the solutions to (2.1) and (4.14). A fundamental solution system to the Stokes equations in $\mathbb{R}^2$ is given here by

$$E(y) = \frac{1}{4\pi\nu}\left(-\log r I + \mathbf{e}_r\mathbf{e}_r^T\right), \quad P(y) = \frac{y}{2\pi r^2}.$$

The exterior problem must now be defined differently than in (4.1), which generally has no solution in the two-dimensional case (Stokes paradox). Let v_ω be the solution to

$$\begin{cases} -\nu\Delta v_\omega + \nabla p_\omega = 0 & \text{in } \mathbb{R}^2\setminus\overline{\omega},\\ \operatorname{div} v_\omega = 0 & \text{in } \mathbb{R}^2\setminus\overline{\omega},\\ v_\omega/\log\|y\| = u_\Omega(x_0) & \text{at } \infty,\\ v_\omega = 0 & \text{on } \partial\omega.\end{cases}$$

Functions v_ω and p_ω have the following asymptotic expansion at infinity:

$$v_\omega(y) = -4\pi\nu E(y)u_\Omega(x_0) + V_\omega + R_\omega(y), \quad R_\omega(y) = O(1/r),$$
$$p_\omega(y) = -4\pi\nu P(y).u_\Omega(x_0) + S_\omega(y), \quad S_\omega(y) = O(1/r^2),$$

where $V_\omega \in \mathbb{R}^2$ is *constant* (it follows from the fact that any tempered distribution solution to the homogeneous Stokes equations in $\mathbb{R}^n$ is a polynomial, which can be proved as in [8, Chap. II, sect. 2, Prop. 3]). The principal part of v_ω is $-4\pi\nu E(y)u_\Omega(x_0)+V_\omega$ and, because V_ω is constant, its derivative is independent of ω. The consequence is that, in contrast to the three-dimensional case, the topological sensitivity does not depend on the shape of ω (cf. Propositions 4.5 and 4.6).

If $\omega = B(0,1)$, then function v_ω can be computed explicitly as follows: for $y \neq 0$, we have

$$E(y) = \frac{1}{4\pi\nu}(-\log r I + \mathbf{e}_r\mathbf{e}_r^T),$$
$$\Delta E(y) = \frac{1}{2\pi\nu r^2}(I - 2\mathbf{e}_r\mathbf{e}_r^T),$$
$$\pi\nu(4E + \Delta E)(y) = \left(-\log r + \frac{1}{2r^2}\right) I + \left(1 - \frac{1}{r^2}\right)\mathbf{e}_r\mathbf{e}_r^T.$$

Hence $\pi\nu(4E + \Delta E)(y) = I/2$ on $\partial B(0,1)$, and it follows from the uniqueness of v_ω that

$$v_\omega(y) = -\pi\nu(4E + \Delta E)(y)u_\Omega(x_0) + \frac{1}{2}u_\Omega(x_0).$$

Next we consider the solution W_ω, Q_ω to the interior problem

$$\begin{cases} -\nu\Delta W_\omega + \nabla Q_\omega = 0 & \text{in } D_0, \\ \operatorname{div} W_\omega = 0 & \text{in } D_0, \\ W_\omega = -4\pi\nu E u_\Omega(x_0) + V_\omega & \text{on } \Gamma_R. \end{cases}$$

The "first order approximation" of $(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi})_{|D_\varepsilon}$ with $\varphi = (u_\Omega)_{|\Gamma_R}$ now becomes (compare with (4.10))

$$\frac{-1}{\log\varepsilon}\left(-4\pi\nu E(x)u_\Omega(x_0) + V_\omega - W_\omega(x)\right)_{|D_\varepsilon}.$$

PROPOSITION 4.5. *The assumptions are the same as in Proposition* 4.3*, with* $\widetilde{J}_\varepsilon$ *of the form*

$$\widetilde{J}_\varepsilon(u) = \int_{\Omega_\varepsilon} g(x, u(x))\,dx, \quad u \in H^1(\Omega_\varepsilon)^2.$$

Then function j *has the following asymptotic expansion:*

$$j(\varepsilon) = j(0) - \frac{4\pi\nu u_\Omega(x_0).v_\Omega(x_0)}{\log\varepsilon} + o\left(\frac{-1}{\log\varepsilon}\right). \tag{4.22}$$

In the next proposition, the first expression of $j(\varepsilon)$ is given for comparison with the three-dimensional case.

PROPOSITION 4.6. *The assumptions are the same as in Proposition* 4.4, *with* $\widetilde{J}_\varepsilon$ *of the form*

$$\widetilde{J}_\varepsilon(u) = \frac{1}{2}\int_{\Omega_\varepsilon} BD(u-u_{\rm d}) : D(u-u_{\rm d})\,dx, \quad u \in H^1(\Omega_\varepsilon)^2.$$

Then function j *has the following asymptotic expansion (with* $D_\varepsilon/\varepsilon = B(0,R/\varepsilon)/\overline{\omega}$*):*

$$\begin{aligned} j(\varepsilon) &= j(0) - \frac{4\pi\nu}{\log\varepsilon} u_\Omega(x_0).v_\Omega(x_0) + \frac{1}{2\log^2\varepsilon}\int_{D_\varepsilon/\varepsilon} B(x_0)Dv_\omega : Dv_\omega\,dy + o\left(\frac{-1}{\log\varepsilon}\right) \\ &= j(0) - \frac{4\pi\nu}{\log\varepsilon} u_\Omega(x_0).v_\Omega(x_0) + \frac{(4\pi\nu)^2}{2\log^2\varepsilon}\int_{D_\varepsilon/\varepsilon} B(x_0)D(Eu_\Omega(x_0)) : D(Eu_\Omega(x_0))\,dy \\ &\quad + o\left(\frac{-1}{\log\varepsilon}\right) \\ &= j(0) - \frac{\pi}{4\log\varepsilon}[16\nu u_\Omega.v_\Omega + (b_{1111}+b_{2222}+b_{1212}+b_{2121}-2b_{2112}-2b_{1122})|u_\Omega|^2 \\ &\quad + 4b_{1212}u_1^2 + 4b_{2121}u_2^2 + 4(-b_{1112}+b_{1222}-b_{2122}+b_{1121})u_1u_2](x_0) + o\left(\frac{-1}{\log\varepsilon}\right), \end{aligned}$$

where $u_\Omega = (u_1, u_2)$.

The proofs use the same tools as for the three-dimensional case (see section 6) and will not be repeated for the two-dimensional case.

5. A numerical example. The example presented in this section shows how topological sensitivity can be used to improve a given criterion. Although the obtained result indicates that the problem may not have a "classical solution," which is quite common in shape optimization, such a result may still have some practical interest for situations where the user does not necessarily need the "best solution," but only an improved design. A classical approach used for obtaining existence of an optimal solution is to use homogeneization [1, 4, 7, 13]. For a review on recent advances in shape optimization methods and existence of an optimal solution (either through penalization or relaxation), we refer the reader to [2] and references therein.

We consider the case of a tank filled with an incompressible fluid, in which some obstacles can be inserted in order to approximate a target flow $u_{\rm d}$. The velocity and the pressure are the solution to

$$-\nu\Delta u_\Omega + \nabla p_\Omega = 0 \ \text{ in } \Omega = [0,1.4]\times[-1.2,1.2], \tag{5.1}$$

where the domain Ω and the boundary conditions are illustrated by Figure 3.

The cost function is defined by (see also Figure 3)

$$\begin{aligned} J(u) &= \int_{\Omega_d} ||u-u_{\rm d}||^2\,dx, \\ u_{\rm d} &= \left(\frac{y+1.2}{2.4^2}, 0\right), \quad \Omega_{\rm d} = \{(x,y)\in\Omega;\ |x|<0.8\}. \end{aligned}$$

The subset $\Omega\backslash\overline{\Omega_{\rm d}}$ is the location where some obstacles can be inserted. For that function, the adjoint equation (4.15) reads

$$\begin{cases} -\nu\Delta v_\Omega + \nabla q_\Omega = -2(u_\Omega - u_{\rm d}) \text{ in } \Omega, \\ v_\Omega = 0 \text{ on } \Gamma, \end{cases}$$

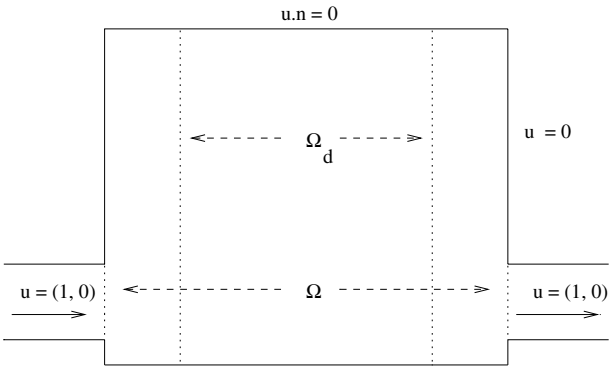


FIG. 3. *The domain Ω and the boundary conditions.*

and we obtain from (4.22) the topological sensitivity

$$j(\varepsilon) = j(0) - \frac{4\pi\nu}{\log \varepsilon} u_\Omega(x_0).v_\Omega(x_0) + o\left(\frac{-1}{\log \varepsilon}\right).$$

For completeness, we recall here the topology optimization algorithm described in [9], which was inspired by the original work of Céa [6]. Let $(m_k)_{k\geq 0}$ be an increasing sequence of volume constraints, with $m_0 = \text{meas}(\Omega_\text{d})$. For example, a geometrical sequence may be chosen. At the kth iteration, the topological sensitivity is denoted by $\delta j_k(x)$, and a small obstacle is added at the point where δj is the most negative. The algorithm is the following [9]:

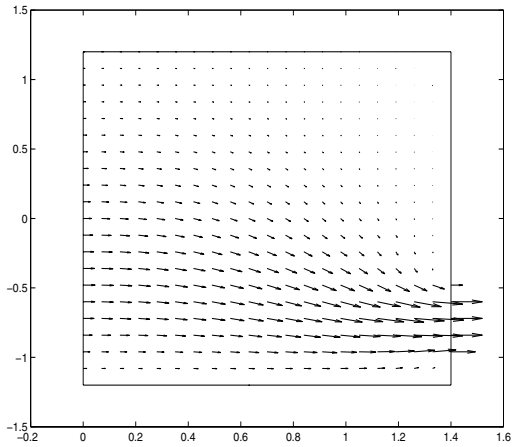
- Initialization: choose $\Omega_0 = \Omega_\text{d}$, and set $k = 0$.
- Repeat until target is reached:
  1. solve (5.1) in Ω_k,
  2. compute the topological sensitivity δj_k,
  3. set $\Omega_{k+1} = \{x \in \Omega_k;\, \delta j_k(x) \geq c_{k+1}\}$, where c_{k+1} is chosen in such a way that $\text{meas}(\Omega_{k+1}) = m_{k+1}$,
  4. $k \leftarrow k + 1$.

Figures 4(a) and 4(b) show the initial flow and the objective flow that we want to obtain after optimization. Only half of the tank is represented. The results after optimization (obstacles and obtained flow) are presented in Figures 4(c) and 4(d). The plotting function of MATLAB normalizes the sizes of the arrows, which is the reason why we show in Figure 4(d) the restriction of the flow (Figure 4(c)) to Ω_d. That allows a better comparison between the objective flow and the obtained flow. The value of the cost function at each iteration is presented in Figure 4(f).

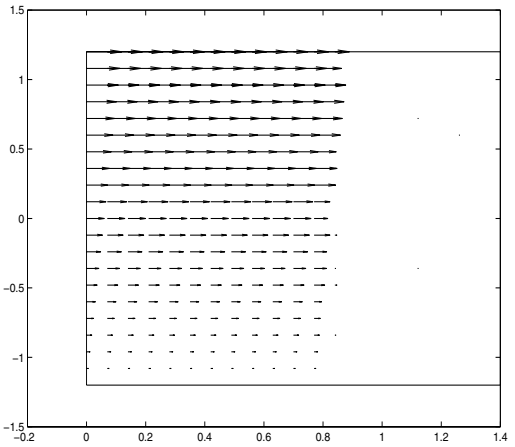
6. Proofs. This section consists of the proofs of Theorem 4.1 and Propositions 4.3 and 4.4. They use the fundamental result from [11], [9] which is recalled here.

LEMMA 6.1. *Let $\mathcal{V}$ be a Hilbert space. For $\varepsilon \geq 0$, let a_ε be a bilinear and symmetric form on $\mathcal{V}$ and l_ε be a linear form on $\mathcal{V}$, such that for all $\varepsilon \geq 0$,*

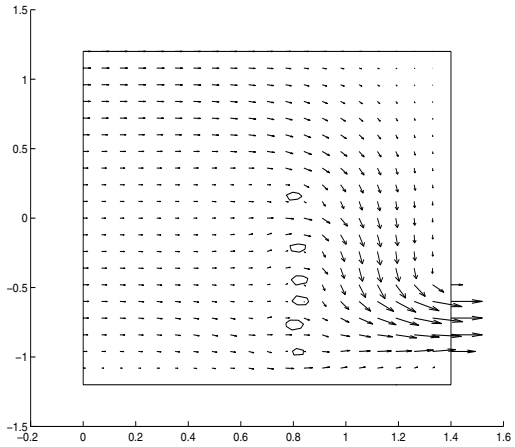
$$\begin{aligned} a_\varepsilon(u,v) &\leq M_1\|u\|\,\|v\|, \quad \forall u,\, v \in \mathcal{V}, \\ a_\varepsilon(u,u) &\geq \alpha\|u\|^2, \quad \forall u \in \mathcal{V}, \\ |l_\varepsilon(v)| &\leq M_2\|v\|, \quad \forall v \in \mathcal{V}. \end{aligned}$$



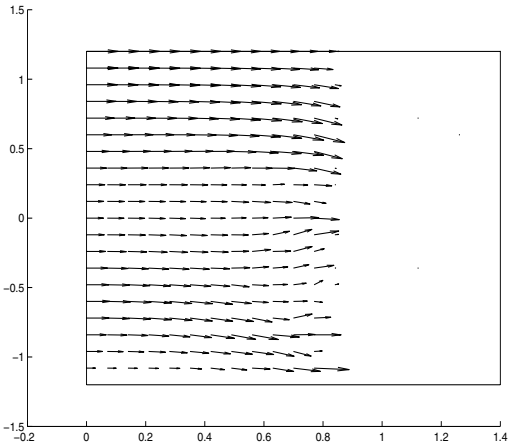
(a) initial flow



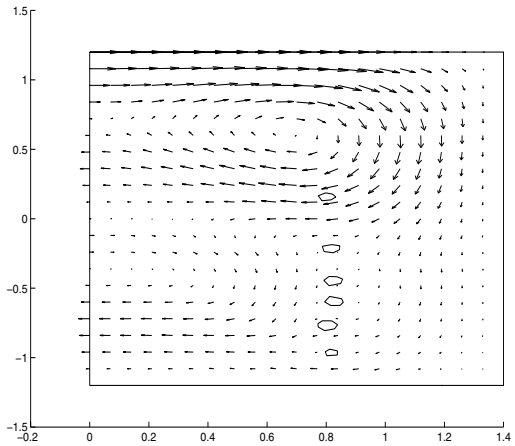
(b) objective flow



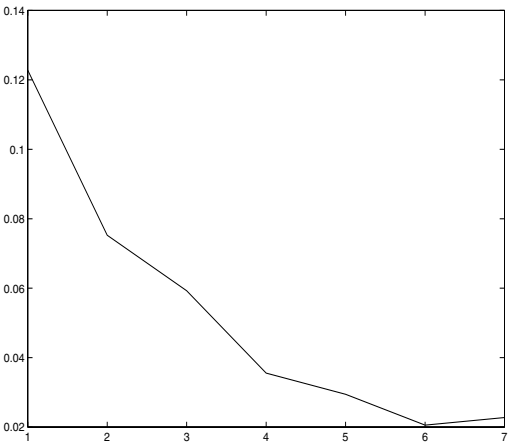
(c) obtained flow on Ω



(d) obtained flow on Ω_d



(e) adjoint flow



(f) cost function

FIG. 4.

We suppose that there exists a bilinear and continuous form δa, a linear and continuous form δl, and a real function $f(\varepsilon) > 0$ defined on $\mathbb{R}_+$ such that

$$\begin{aligned} ||a_\varepsilon - a_0 - f(\varepsilon)\delta a||_{\mathcal{L}_2(\mathcal{V})} &= o(f(\varepsilon)), \\ ||l_\varepsilon - l_0 - f(\varepsilon)\delta l||_{\mathcal{L}(\mathcal{V})} &= o(f(\varepsilon)), \\ \lim_{\varepsilon \to 0} f(\varepsilon) &= 0. \end{aligned}$$

For $\varepsilon \geq 0$, let u_ε be the solution to

$$a_\varepsilon(u_\varepsilon, v) = l_\varepsilon(v) \quad \forall v \in \mathcal{V}.$$

Then

$$\|u_\varepsilon - u_0\|_{\mathcal{V}} = O(f(\varepsilon)).$$

Consider next a cost function of the form $j(\varepsilon) = J_\varepsilon(u_\varepsilon)$, where J_ε is defined on $\mathcal{V}$ and J_0 is differentiable with respect to u, its derivative being denoted by $DJ_0(u)$. Suppose that there exists a function δJ defined on $\mathcal{V}$ such that for all $v \in \mathcal{V}$ and all $\varepsilon > 0$

$$J_\varepsilon(v) - J_0(u) = DJ_0(u)(v-u) + f(\varepsilon)\delta J(u) + o(\|v-u\|_{\mathcal{V}} + f(\varepsilon)).$$

Then j has the asymptotic expansion

$$j(\varepsilon) = j(0) + f(\varepsilon)[\delta a(u_0, v_0) - \delta l(v_0) + \delta J_0(u_0)] + o(f(\varepsilon)),$$

where $v_0 \in \mathcal{V}$ is the solution to the adjoint problem

$$a_0(w, v_0) = -DJ_0(u_0)w \quad \forall w \in \mathcal{V}.$$

Here, the variations of the bilinear form a_ε and the linear form l_ε (see (3.10) and (3.11)) read

$$\begin{aligned} a_\varepsilon(u,v) - a_0(u,v) &= \int_{\Gamma_R} (T_\varepsilon - T_0)u.v \, d\gamma(x), \\ l_\varepsilon(v) - l_0(v) &= \int_{\Gamma_R} (f_\varepsilon - f_0).v \, d\gamma(x). \end{aligned}$$

Hence, the problem reduces to the analysis of $(T_\varepsilon - T_0)\varphi$ for $\varphi \in H_V^{1/2}(\Gamma_R)^3$ and of $f_\varepsilon - f_0$ in $H_V^{-1/2}(\Gamma_R)$. More precisely, it will be shown in sections 6.3 and 6.4 that there exists an operator $\delta T \in \mathcal{L}(H_V^{1/2}(\Gamma_R)^3; H_V^{-1/2}(\Gamma_R)^3)$ and a function $\delta f \in H^{-1/2}(\Gamma_R)$ such that

$$\|T_\varepsilon - T_0 - \varepsilon\delta T\|_{\mathcal{L}(H_V^{1/2}(\Gamma_R)^3;\, H_V^{-1/2}(\Gamma_R)^3)} = O(\varepsilon^2), \tag{6.1}$$

$$\|f_\varepsilon - f_0 - \varepsilon\delta f\|_{H_V^{-1/2}(\Gamma_R)} = O(\varepsilon^2). \tag{6.2}$$

Consequently, defining δa and δl by

$$\begin{aligned} \delta a(u,v) &= \int_{\Gamma_R} \delta T u.v \, d\gamma(x), \quad u,\, v \in \mathcal{V}_R, \\ \delta l(v) &= \int_{\Gamma_R} \delta f.v \, d\gamma(x), \quad v \in \mathcal{V}_R, \end{aligned}$$

will yield straightforwardly

$$\|a_\varepsilon - a_0 - \varepsilon\delta a\|_{\mathcal{L}_2(\mathcal{V}_R)} = O(\varepsilon^2),$$
$$\|l_\varepsilon - l_0 - \varepsilon\delta l\|_{\mathcal{L}(\mathcal{V}_R)} = O(\varepsilon^2),$$

and Lemma 6.1 can be applied. In order to derive (6.1)–(6.2), we need some definitions and preliminary lemmas.

6.1. Definitions. This section describes the functional spaces and norms which will be used in the proofs.

- For a bounded and open subset $\mathcal{O} \subset \mathbb{R}^3$ and $m \geq 0$, the Sobolev space $H^m(\mathcal{O})^3$ is equipped with the norm defined by

$$\|u\|^2_{m,\mathcal{O}} = \sum_{k=0}^{m} |u|^2_{k,\mathcal{O}}\,,$$

where the seminorms $|u|^2_{k,\mathcal{O}}$ are given by

$$|u|^2_{k,\mathcal{O}} := \sum_{|\alpha|=k} ||D^\alpha u||^2_{L^2(\mathcal{O})}. \tag{6.3}$$

The usual space of traces on the boundary of $\mathcal{O}$ is denoted by $H^{1/2}(\partial\mathcal{O})$, and its norm is denoted by $||\cdot||_{1/2,\partial\mathcal{O}}$. The subspace

$$H^{1/2}_V(\partial\mathcal{O})^3 = \left\{\varphi \in H^{1/2}(\partial\mathcal{O})^3;\ \int_{\Gamma_R} \varphi.\mathbf{n}\, d\gamma(x) = 0\right\}$$

is equipped with the norm induced by $H^{1/2}(\partial\mathcal{O})^3$, and $H^{-1/2}_V(\partial\mathcal{O})^3$ denotes its dual space. A special case is when $\psi = (\nu Du - pI)\mathbf{n}$ on Γ_R with $-\nu\Delta u + \nabla p = 0$ in $C(R/2, R)$. For $\varphi \in H^{1/2}_V(\Gamma_R)^3$, let $v \in H^1(C(R/2,R))$ solve the Stokes equations with a null right-hand side, $v = \varphi$ on Γ_R and $v = 0$ on $\Gamma_{R/2}$. Using the well-posedness of the Stokes equations, we have

$$\begin{aligned}\langle(\nu Du - pI)\mathbf{n},\varphi\rangle_{-1/2,1/2} &= \nu \int_{C(R/2,R)} Du : Dv\, dx \\ &\leq \nu|u|_{1,C(R/2,R)}|v|_{1,C(R/2,R)} \\ &\leq c|u|_{1,C(R/2,R)}||\varphi||_{1/2},\end{aligned}$$

which proves that

$$\|(\nu Du - pI)\mathbf{n}\|_{-1/2,\Gamma_R} \leq c|u|_{1,C(R/2,R)}. \tag{6.4}$$

As usual in PDEs, letter c denotes a positive constant independent of the data (e.g., on ε).

- Finally, the space $L^2(\mathcal{O})/\mathbb{R}$ is identified with

$$L^2_0(\mathcal{O}) = \left\{p \in L^2(\mathcal{O});\ \int_{\mathcal{O}} p\, dx = 0\right\}.$$

6.2. Preliminary lemmas. Recall that $x_0 = 0$. We will use extensively the following change of variable. For a given function u defined on a subset $\mathcal{O}$, function $\widetilde{u}$ is defined on $\widetilde{\mathcal{O}} := \mathcal{O}/\varepsilon$ by

$$\widetilde{u}(y) = u(x), \quad y = x/\varepsilon.$$

Due to $Du(x) = D\widetilde{u}(y)/\varepsilon$ and to Definition (6.3), we have

$$|u|^2_{1,\mathcal{O}} = \int_{\mathcal{O}} |Du|^2 \, dx = \frac{1}{\varepsilon^2} \int_{\widetilde{\mathcal{O}}} |D\widetilde{u}|^2 \, \varepsilon^3 dy;$$

hence

$$|u|_{1,\mathcal{O}} = \varepsilon^{1/2} \, |\widetilde{u}|_{1,\widetilde{\mathcal{O}}} \, . \tag{6.5}$$

Similarly, we have

$$||u||_{0,\mathcal{O}} = \varepsilon^{3/2} ||\widetilde{u}||_{0,\widetilde{\mathcal{O}}}. \tag{6.6}$$

LEMMA 6.2. *For $\varphi \in H_V^{1/2}(\partial\omega)$, let v_ω, q_ω be the solution to the problem*

$$\left\{ \begin{array}{rl} -\nu\Delta v_\omega + \nabla q_\omega = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega}, \\ \operatorname{div} v_\omega = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega}, \\ v_\omega = 0 & \text{at } \infty, \\ v_\omega = \varphi & \text{on } \partial\omega. \end{array} \right. \tag{6.7}$$

Function v_ω is split into

$$\begin{aligned} v_\omega(y) &= V_\omega(y) + R_\omega(y), \\ V_\omega(y) &= E(y) \int_{\partial\omega} t_\omega(x) \, d\gamma(x), \end{aligned}$$

where $E(y)$ is defined in (4.2) and $t_\omega \in H_V^{-1/2}(\partial\omega)^3$ is the unique solution to

$$\int_{\partial\omega} E(y-x) t_\omega(x) d\gamma(x) = \varphi(y) \quad \forall y \in \partial\omega. \tag{6.8}$$

There exists a constant $c > 0$ (independent of φ and ε) such that

$$\begin{aligned} \|V_\omega\|_{0,C(R/(2\varepsilon),R/\varepsilon)} &\leq c\varepsilon^{-1/2} \|\varphi\|_{1/2,\partial\omega}, \\ |V_\omega|_{1,C(R/(2\varepsilon),R/\varepsilon)} &\leq c\varepsilon^{1/2} \|\varphi\|_{1/2,\partial\omega}, \\ \|V_\omega\|_{0,D_\varepsilon/\varepsilon} &\leq c\varepsilon^{-1/2} \|\varphi\|_{1/2,\partial\omega}, \\ |V_\omega|_{1,D_\varepsilon/\varepsilon} &\leq c||\varphi||_{1/2,\partial\omega}, \\ \|R_\omega\|_{0,C(R/(2\varepsilon),R/\varepsilon)} &\leq c\varepsilon^{1/2} \|\varphi\|_{1/2,\partial\omega}, \\ |R_\omega|_{1,C(R/(2\varepsilon),R/\varepsilon)} &\leq c\varepsilon^{3/2} \|\varphi\|_{1/2,\partial\omega}, \\ \|R_\omega\|_{1,D_\varepsilon/\varepsilon} &\leq c\|\varphi\|_{1/2,\partial\omega}. \end{aligned}$$

Proof. Function v_ω reads

$$v_\omega(y) = \int_{\partial\omega} E(y-x) t_\omega(x) \, d\gamma(x), \quad y \in \mathbb{R}^3\backslash\overline{\omega}.$$

Using a Taylor expansion of E computed at the point y and the well-posedness of (6.8) we have for large $r = ||y||$

$$|V_\omega(y)| \leq \frac{c}{r}\|\varphi\|_{1/2,\partial\omega},\ |R_\omega(y)| \leq \frac{c}{r^2}\|\varphi\|_{1/2,\partial\omega},$$
$$|DV_\omega(y)| \leq \frac{c}{r^2}\|\varphi\|_{1/2,\partial\omega},\ |DR_\omega(y)| \leq \frac{c}{r^3}\|\varphi\|_{1/2,\partial\omega},$$

from which the above inequalities follow straightforwardly. □

LEMMA 6.3. *For $\varepsilon > 0$ and $\varphi \in H_V^{1/2}(\Gamma_R)^3$, let v_ε, q_ε be the solution to the problem*

$$\left\{\begin{array}{rl} -\nu\Delta v_\varepsilon + \nabla q_\varepsilon = 0 & \text{in } D_\varepsilon, \\ \operatorname{div} v_\varepsilon = 0 & \text{in } D_\varepsilon, \\ v_\varepsilon = \varphi & \text{on } \Gamma_R, \\ v_\varepsilon = 0 & \text{on } \partial\omega_\varepsilon. \end{array}\right. \tag{6.9}$$

There exist a constant $c > 0$ (independent of φ and ε) and $\varepsilon_1 > 0$ such that for all $0 < \varepsilon < \varepsilon_1$,

$$\|v_\varepsilon\|_{1,D_\varepsilon} \leq c\,\|\varphi\|_{1/2,\Gamma_R}.$$

Proof. Let $\varepsilon_0 > 0$. Problem (6.9) is well-posed; hence there exists a constant c such that

$$|v_{\varepsilon_0}|_{1,D_{\varepsilon_0}} \leq c\|\varphi\|_{1/2,\Gamma_R}.$$

Let $\varepsilon_1 \leq \varepsilon_0$ be such that $D_{\varepsilon_0} \subset D_\varepsilon$ for all $\varepsilon < \varepsilon_1$. Let $\widehat{v}_{\varepsilon_0}$ be the extension of v_{ε_0} to D_ε by 0. Function v_ε minimizes the energy $|v|_{1,D_\varepsilon}$ over the affine space

$$\left\{v \in H^1(D_\varepsilon)^3;\ v = \varphi \text{ on } \Gamma_R,\ \operatorname{div} v = 0 \text{ and } v = 0 \text{ on } \partial\omega\right\};$$

hence, for all $\varepsilon \leq \varepsilon_1$ we have

$$|v_\varepsilon|_{1,D_\varepsilon} \leq |\widehat{v}_{\varepsilon_0}|_{1,D_\varepsilon} = |v_{\varepsilon_0}|_{1,D_{\varepsilon_0}} \leq c\|\varphi\|_{1/2,\Gamma_R}.$$

We also have

$$\|v_0\|_{0,D_0} \leq c\,\|\varphi\|_{1/2,\Gamma_R}.$$

Then, denoting by $\widehat{v}_\varepsilon$ the extension by 0 of v_ε to D_0 and using Poincaré's inequality on D_0 yields

$$\begin{aligned}\|v_\varepsilon\|_{0,D_\varepsilon} = \|\widehat{v}_\varepsilon\|_{0,D_0} &\leq \|\widehat{v}_\varepsilon - v_0\|_{0,D_0} + \|v_0\|_{0,D_0} \\ &\leq c\,|\widehat{v}_\varepsilon - v_0|_{1,D_0} + \|v_0\|_{0,D_0} \\ &\leq c\,|\widehat{v}_\varepsilon|_{1,D_0} + c\,\|v_0\|_{1,D_0} = c\,|v_\varepsilon|_{1,D_\varepsilon} + c\,\|v_0\|_{1,D_0} \\ &\leq c\,\|\varphi\|_{1/2,\Gamma_R}.\end{aligned}$$ □

LEMMA 6.4. *For $\varepsilon > 0$ and $\psi \in H^1(D_0)^3$ such that $\operatorname{div}\psi = 0$, let u_ε, p_ε be the solution to the problem*

$$\left\{\begin{array}{rl} -\nu\Delta u_\varepsilon + \nabla p_\varepsilon &= 0 \quad \text{in } D_\varepsilon, \\ \operatorname{div} u_\varepsilon &= 0 \quad \text{in } D_\varepsilon, \\ u_\varepsilon &= 0 \quad \text{on } \Gamma_R, \\ u_\varepsilon &= \psi \quad \text{on } \partial\omega_\varepsilon. \end{array}\right. \tag{6.10}$$

There exist a constant $c > 0$ (independent of ψ and ε) and $\varepsilon_1 > 0$ such that for all $0 < \varepsilon < \varepsilon_1$,

$$\begin{aligned} |u_\varepsilon|_{1,C(R/2,R)} &\le c\varepsilon \|\psi(\varepsilon y)\|_{1/2,\partial\omega}, \\ \|u_\varepsilon\|_{0,D_\varepsilon} &\le c\varepsilon \left\|\psi(\varepsilon y)\right\|_{1/2,\partial\omega}, \\ |u_\varepsilon|_{1,D_\varepsilon} &\le c\varepsilon^{1/2} \left\|\psi(\varepsilon y)\right\|_{1/2,\partial\omega}. \end{aligned}$$

Proof. Let $\widetilde{v}_\varepsilon$, $\widetilde{q}_\varepsilon$ be the solution to the exterior problem

$$\left\{ \begin{array}{rl} -\nu\Delta\widetilde{v}_\varepsilon + \nabla\widetilde{q}_\varepsilon = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega}, \\ \operatorname{div}\widetilde{v}_\varepsilon = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega}, \\ \widetilde{v}_\varepsilon = 0 & \text{at infinity}, \\ \widetilde{v}_\varepsilon = \psi(\varepsilon y) & \text{on } \partial\omega. \end{array} \right.$$

Function u_ε can be written

$$u_\varepsilon = v_\varepsilon - w_\varepsilon$$

where $v_\varepsilon(x) = \widetilde{v}_\varepsilon(x/\varepsilon)$. Function w_ε itself is the solution to

$$\left\{ \begin{array}{rl} -\nu\Delta w_\varepsilon + \nabla s_\varepsilon = 0 & \text{in } D_\varepsilon, \\ \operatorname{div} w_\varepsilon = 0 & \text{in } D_\varepsilon, \\ w_\varepsilon = v_\varepsilon & \text{on } \Gamma_R, \\ w_\varepsilon = 0 & \text{on } \partial\omega_\varepsilon. \end{array} \right.$$

It follows from Lemma 6.3, (6.5), (6.6), and Lemma 6.2 that there exist $c > 0$ and $\varepsilon_1 > 0$ such that for all $0 < \varepsilon < \varepsilon_1$,

$$\begin{aligned} \|w_\varepsilon\|_{1,D_\varepsilon} &\le c \|v_\varepsilon\|_{1/2,\Gamma_R} \\ &\le c \|v_\varepsilon\|_{1,C(R/2,R)} \\ &\le c(|v_\varepsilon|_{0,C(R/2,R)} + |v_\varepsilon|_{1,C(R/2,R)}) \\ &= c(\varepsilon^{3/2}|\widetilde{v}_\varepsilon|_{0,C(R/2\varepsilon,R/\varepsilon)} + \varepsilon^{1/2}|\widetilde{v}_\varepsilon|_{1,C(R/2\varepsilon,R/\varepsilon)}) \\ &\le c\varepsilon \left\|\psi(\varepsilon y)\right\|_{1/2,\partial\omega}. \end{aligned} \tag{6.11}$$

Hence

$$\begin{aligned} |u_\varepsilon|_{1,C(R/2,R)} = |v_\varepsilon - w_\varepsilon|_{1,C(R/2,R)} &\le |v_\varepsilon|_{1,C(R/2,R)} + |w_\varepsilon|_{1,D_\varepsilon} \\ &\le c\varepsilon \left\|\psi(\varepsilon y)\right\|_{1/2,\partial\omega}. \end{aligned}$$

Similarly we have

$$\begin{aligned} \|v_\varepsilon\|_{0,D_\varepsilon} &= \varepsilon^{3/2} \left\|\tilde{v}_\varepsilon\right\|_{0,D_\varepsilon/\varepsilon} \le c\varepsilon \left\|\psi(\varepsilon y)\right\|_{1/2,\partial\omega}, \\ |v_\varepsilon|_{1,D_\varepsilon} &= \varepsilon^{1/2} \left|\tilde{v}_\varepsilon\right|_{1,D_\varepsilon/\varepsilon} \le c\varepsilon^{1/2} \left\|\psi(\varepsilon y)\right\|_{1/2,\partial\omega} \end{aligned}$$

and

$$\begin{aligned} \|u_\varepsilon\|_{0,D_\varepsilon} &\le c\varepsilon \left\|\psi(\varepsilon y)\right\|_{1/2,\partial\omega}, \\ |u_\varepsilon|_{1,D_\varepsilon} &\le c\varepsilon^{1/2} \left\|\psi(\varepsilon y)\right\|_{1/2,\partial\omega}. \end{aligned}$$

□

Lemmas 6.3 and 6.4 are summarized in the following lemma.

LEMMA 6.5. *For $\varepsilon > 0$, $\varphi \in H_V^{1/2}(\Gamma_R)^3$, and $\psi \in H^1(D_0)^3$ such that* $\operatorname{div}\psi = 0$, *let v_ε, q_ε be the solution to the problem*

$$\begin{cases} -\nu\Delta v_\varepsilon + \nabla q_\varepsilon = 0 & \text{in } D_\varepsilon, \\ \operatorname{div} v_\varepsilon = 0 & \text{in } D_\varepsilon, \\ v_\varepsilon = \varphi & \text{on } \Gamma_R, \\ v_\varepsilon = \psi & \text{on } \partial\omega_\varepsilon. \end{cases}$$

There exist a constant $c > 0$ (independent of φ, ψ and ε) and $\varepsilon_1 > 0$ such that for all $0 < \varepsilon < \varepsilon_1$,

$$\begin{aligned} |v_\varepsilon|_{1,C(R/2,R)} &\le c\left(\|\varphi\|_{1/2,\Gamma_R} + \varepsilon\,\|\psi(\varepsilon y)\|_{1/2,\partial\omega}\right), \\ ||v_\varepsilon||_{0,D_\varepsilon} &\le c\left(\|\varphi\|_{1/2,\Gamma_R} + \varepsilon\,\|\psi(\varepsilon y)\|_{1/2,\partial\omega}\right), \\ |v_\varepsilon|_{1,D_\varepsilon} &\le c\left(\|\varphi\|_{1/2,\Gamma_R} + \varepsilon^{1/2}\,\|\psi(\varepsilon y)\|_{1/2,\partial\omega}\right). \end{aligned}$$

6.3. Variation of the bilinear form. The variation of the bilinear form a_ε reads

$$a_\varepsilon(u,v) - a_0(u,v) = \int_{\Gamma_R} (T_\varepsilon - T_0)u.v\, d\gamma(x).$$

For $\varphi \in H_V^{1/2}(\Gamma_R)^3$, recall that $u_\varepsilon^{0,\varphi}$ is the solution to (3.2) or (3.3) if $\varepsilon = 0$. Let $v_\omega^{0,\varphi}$, $p_\omega^{0,\varphi}$ be the solution to

$$\begin{cases} -\nu\Delta v_\omega^{0,\varphi} + \nabla p_\omega^{0,\varphi} = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega}, \\ \operatorname{div} v_\omega^{0,\varphi} = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega}, \\ v_\omega^{0,\varphi} = 0 & \text{at } \infty, \\ v_\omega^{0,\varphi} = u_0^{0,\varphi}(x_0) & \text{on } \partial\omega. \end{cases} \tag{6.12}$$

As in (4.6), let $V_\omega^{0,\varphi}(y) = E(y)A_\omega(u_0^{0,\varphi}(x_0))$ be the dominant part of $v_\omega^{0,\varphi}$, and let $W_\omega^{0,\varphi}$, $Q_\omega^{0,\varphi}$ be the associated solution to (4.9) with $W_\omega^{0,\varphi} = V_\omega^{0,\varphi}$ on Γ_R. The linear operator δT (independent of ε) is defined as

(6.13)
$$\begin{aligned} \delta T : H_V^{1/2}(\Gamma_R)^3 &\longrightarrow H_V^{-1/2}(\Gamma_R)^3, \\ \varphi &\longmapsto \delta T\varphi := [(\nu DW_\omega^{0,\varphi} - Q_\omega^{0,\varphi} I) - (\nu DV_\omega^{0,\varphi} - P_\omega^{0,\varphi} I)]\mathbf{n}. \end{aligned}$$

PROPOSITION 6.6. *The asymptotic expansion of T_ε is*

$$\|T_\varepsilon - T_0 - \varepsilon\delta T\|_{\mathcal{L}(H_V^{1/2}(\Gamma_R)^3;\, H_V^{-1/2}(\Gamma_R)^3)} = O(\varepsilon^2). \tag{6.14}$$

Proof. Let $\varphi \in H_V^{1/2}(\Gamma_R)^3$. For simplicity we may drop the subscripts $(.)^{0,\varphi}$. For $y = x/\varepsilon$, we have $v_\omega(y) = V_\omega(y) + R_\omega(y)$ with $V_\omega(x/\varepsilon) = \varepsilon V_\omega(x)$; $p_\omega(y) = P_\omega(y) + S_\omega(y)$ with $P_\omega(x/\varepsilon) = \varepsilon^2 P_\omega(x)$; and $R_\omega(y) = O(1/||y||^2)$, $S_\omega(y) = O(1/||y||^3)$ (see (4.2), (4.6), (4.7)). Let

$$\psi_\varepsilon(x) = (T_\varepsilon - T_0 - \varepsilon\delta T)\varphi(x).$$

We have

$$\begin{aligned} \psi_\varepsilon(x) &= (\nu Du_\varepsilon - p_\varepsilon I)\mathbf{n} - (\nu Du_0 - p_0 I)\mathbf{n} - \varepsilon[(\nu DW_\omega - Q_\omega I)\mathbf{n} - (\nu DV_\omega - P_\omega I)\mathbf{n}] \\ &= \nu D(w_\varepsilon(x) - R_\omega(x/\varepsilon))\mathbf{n} - \left(s_\varepsilon(x) - \frac{1}{\varepsilon}S_\omega(x/\varepsilon)\right)\mathbf{n}, \end{aligned}$$

where w_ε, s_ε is defined by

$$w_\varepsilon(x) = u_\varepsilon(x) - u_0(x) + v_\omega(x/\varepsilon) - \varepsilon W_\omega(x),$$
$$s_\varepsilon(x) = p_\varepsilon(x) - p_0(x) + \frac{1}{\varepsilon} p_\omega(x/\varepsilon) - \varepsilon\, Q_\omega(x).$$

Functions w_ε, s_ε are solutions to

$$\left\{\begin{array}{rcll} -\nu\Delta w_\varepsilon + \nabla s_\varepsilon = & 0 & & \text{in } D_\varepsilon, \\ \operatorname{div} w_\varepsilon = & 0 & & \text{in } D_\varepsilon, \\ w_\varepsilon = & v_\omega(x/\varepsilon) - \varepsilon W_\omega(x) & & \text{on } \Gamma_R, \\ w_\varepsilon = & [-u_0(x) + u_0(x_0) - \varepsilon W_\omega(x)] & & \text{on } \partial\omega_\varepsilon. \end{array}\right.$$

In order to apply Lemma 6.5, we have to estimate the two right-hand sides.

On Γ_R, due to $W_\omega(x) = V_\omega(x)$, we have

$$v_\omega(x/\varepsilon) - \varepsilon W_\omega(x) = R_\omega(x/\varepsilon).$$

Due to the definition of v_ω and W_ω, we have

$$\int_{\Gamma_R} (v_\omega(x/\varepsilon) - \varepsilon W_\omega(x))\mathbf{n}\, d\gamma(x) = 0.$$

Using (6.5), (6.6), Lemma 6.2, and elliptic regularity we obtain

$$\begin{aligned} \|v_\omega(x/\varepsilon) - \varepsilon W_\omega\|_{1/2,\Gamma_R} &= \|R_\omega(x/\varepsilon)\|_{1/2,\Gamma_R} \\ &\le c\|R_\omega(x/\varepsilon)\|_{1,C(R/2,R)} \\ &\le c(\|R_\omega(x/\varepsilon)\|_{0,C(R/2,R)} + |R_\omega(x/\varepsilon)|_{1,C(R/2,R)}) \\ &= c(\varepsilon^{3/2}|R_\omega(y)|_{0,C(R/2\varepsilon,R/\varepsilon)} + \varepsilon^{1/2}|R_\omega(y)|_{1,C(R/2\varepsilon,R/\varepsilon)}) \\ &\le c\varepsilon^2\, \|u_0(x_0)\|_{1/2,\partial\omega} \\ &\le c\varepsilon^2 ||\varphi||_{1/2,\Gamma_R}. \end{aligned}$$

On $\partial\omega_\varepsilon$, putting $\theta_\varepsilon(x) := (-u_0(x) + u(x_0) - \varepsilon W_\omega(x))/\varepsilon$, we have $\operatorname{div}\theta_\varepsilon = 0$ in D_0, and for small ε,

$$\begin{aligned} \|\theta_\varepsilon(\varepsilon y)\|_{1/2,\partial\omega} &\le c\|\theta_\varepsilon(\varepsilon y)\|_{1,\omega} \\ &= c\left\|\frac{u_0(\varepsilon y) - u_0(x_0)}{\varepsilon} + W_\omega(\varepsilon y)\right\|_{1,\omega} \\ &\le c(||u_0||_{\mathcal{C}^2(B(0,R/2))} + ||W_\omega||_{\mathcal{C}^1(B(0,R/2))}) \\ &\le c(||\varphi||_{1/2,\Gamma_R} + ||V_\omega||_{1/2,\Gamma_R}) \\ &\le c||\varphi||_{1/2,\Gamma_R}. \end{aligned}$$

We can now apply Lemma 6.5, which gives

$$\begin{aligned} |w_\varepsilon(x)|_{1,C(R/2,R)} &\le c(\varepsilon^2||\varphi||_{1/2,\Gamma_R} + \varepsilon||\varepsilon\theta_\varepsilon(\varepsilon y)||_{1/2,\partial\omega}) \\ &\le c\varepsilon^2\, ||\varphi||_{1/2,\Gamma_R}. \end{aligned}$$

Finally, observing that $-\nu\Delta V_\omega + \nabla P_\omega = 0$ in $C(R/2, R)$, which implies that $-\nu\Delta(w_\varepsilon(x) - R(x/\varepsilon)) + \nabla(s_\varepsilon(x) - S_\omega(x/\varepsilon)/\varepsilon) = 0$ in $C(R/2, R)$, it follows from (6.4), (6.5), and

Lemma 6.2 that

$$\begin{aligned}||\psi||_{-1/2,\Gamma_R} &= \|\nu D(w_\varepsilon(x) - R(x/\varepsilon))\mathbf{n} - (s_\varepsilon(x) - S_\omega(x/\varepsilon)/\varepsilon)\mathbf{n}\|_{-1/2,\Gamma_R} \\ &\le c(|w_\varepsilon|_{1,C(R/2,R)} + |R_\omega(x/\varepsilon)|_{1,C(R/2,R)}) \\ &= c(|w_\varepsilon|_{1,C(R/2,R)} + \varepsilon^{1/2}|R_\omega(y)|_{1,C(R/2\varepsilon,R/\varepsilon)}) \\ &\le c\varepsilon^2\|\varphi\|_{1/2,\Gamma_R}.\end{aligned}$$

Hence

$$\|(T_\varepsilon - T_0 - \varepsilon\delta T)\varphi\|_{-1/2,\Gamma_R} = O(\varepsilon^2). \qquad \square$$

The asymptotic expansion of the bilinear form a_ε now follows straightforwardly.

PROPOSITION 6.7. *Let*

$$\delta a(u,v) = \int_{\Gamma_R} \delta T u.v \, d\gamma(x), \quad u,\, v \in \mathcal{V}_R.$$

Then the asymptotic expansion of the bilinear form a_ε is given by

$$\|a_\varepsilon - a_0 - \varepsilon\delta a\|_{\mathcal{L}_2(\mathcal{V}_R)} = O(\varepsilon^2).$$

6.4. Variation of the linear form. The technique is the same as in section 6.3. The difference comes from the boundary condition imposed on $\partial\omega$ to the solution of the exterior problem: $v_\omega^{0,\varphi} = u_0^{0,\varphi}(x_0)$ in (6.12) for the study of the bilinear form; $v_\omega^{f,0} = u_0^{f,0}(x_0)$ in (6.15) for the study of the linear form. Hence estimations involving $\|\varphi\|_{1/2,\Gamma_R}$ are replaced by estimations involving $\|f\|_{L^q}$. The variation of the linear form l_ε reads

$$l_\varepsilon(v) - l_0(v) = \int_{\Gamma_R} (f_\varepsilon - f_0).v \, d\gamma(x).$$

Recall that $u_\varepsilon^{f,0}$ is the solution to (3.2) or (3.3) if $\varepsilon = 0$. Let $v_\omega^{f,0}$, $p_\omega^{f,0}$ be the solution to

$$\left\{\begin{array}{rl} -\nu\Delta v_\omega^{f,0} + \nabla p_\omega^{f,0} = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega}, \\ \operatorname{div} v_\omega^{f,0} = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega}, \\ v_\omega^{f,0} = 0 & \text{at } \infty, \\ v_\omega^{f,0} = u_0^{f,0}(x_0) & \text{on } \partial\omega. \end{array}\right. \tag{6.15}$$

As in (4.6), let $V_\omega^{f,0}(y) = E(y)A_\omega(u_0^{f,0}(x_0))$ be the dominant part of $v_\omega^{f,0}$, and let $W_\omega^{f,0}$, $Q_\omega^{f,0}$ be the associated solution to (4.9) with $W_\omega^{f,0} = V_\omega^{f,0}$ on Γ_R. Function $\delta f \in H_V^{-1/2}(\Gamma_R)^3$ (independent of ε) is defined by

$$\delta f = [-(\nu D W_\omega^{f,0} - Q_\omega^{f,0} I) + (\nu D V_\omega^{f,0} - P_\omega^{f,0} I)]\mathbf{n}. \tag{6.16}$$

PROPOSITION 6.8. *Let $f \in L^q(\Omega)^n$, $q > n$. The asymptotic expansion of f_ε is*

$$\|f_\varepsilon - f_0 - \varepsilon\delta f\|_{-1/2,\Gamma_R} = O(\varepsilon^2).$$

Proof. The proof runs as in Proposition 6.6 (we drop the subscripts $(.)^{f,0}$) with w_ε and θ_ε defined by

$$\begin{aligned} w_\varepsilon(x) &= u_\varepsilon(x) - u_0(x) + v_\omega(x/\varepsilon) - \varepsilon W_\omega(x), \\ \theta_\varepsilon(x) &= (-u_0(x) + u_0(x_0) - \varepsilon W_\omega(x))/\varepsilon. \end{aligned}$$

The only difference lies in the elliptic regularity estimates

$$|u_0(x_0)| \leq \|u_0\|_{\mathcal{C}^0(D_0)} \leq c\,\|f\|_{L^q},$$

and for small ε

$$\begin{aligned}
\|\theta_\varepsilon(\varepsilon y)\|_{1/2,\partial\omega} &\leq c\,\|\theta_\varepsilon(\varepsilon y)\|_{1,\omega} \\
&\leq c\,\left\|\frac{u_0(\varepsilon y)-u_0(x_0)}{\varepsilon}+W_\omega(\varepsilon y)\right\|_{1,\omega} \\
&\leq c\,(\|u_0\|_{\mathcal{C}^1(B(0,R/2))}+||W_\omega||_{\mathcal{C}^1(B(0,R/2))}) \\
&\leq c\,(\|f\|_{L^q}+||V_\omega||_{1/2,\Gamma_R}) \\
&\leq c\,\|f\|_{L^q}. \qquad \square
\end{aligned}$$

The asymptotic expansion of the linear form l_ε now follows straightforwardly.

PROPOSITION 6.9. *Let*

$$\delta l(v)=\int_{\Gamma_R}\delta f.v\,d\gamma(x),\quad v\in\mathcal{V}_R.$$

Then the asymptotic expansion of linear form l_ε is given by

$$\|l_\varepsilon-l_0-\varepsilon\delta l\|_{\mathcal{L}(\mathcal{V}_R)}=O(\varepsilon^2).$$

6.5. Proof of Theorem 4.1. It follows from Propositions 6.7 and 6.8 and Lemma 6.1 that

$$j(\varepsilon)=j(0)+(\delta a(u_0,v_0)-\delta l(v_0)+\delta J(u_0))\varepsilon+o(\varepsilon).$$

With $\varphi=u_{0|\Gamma_R}$, it follows from (3.4), (3.12), (4.1), (6.12), and (6.15) that

$$v_\omega=v_\omega^{f,0}+v_\omega^{0,\varphi},$$

which implies

$$\begin{aligned}
V_\omega &= V_\omega^{f,0}+V_\omega^{0,\varphi}, \\
W_\omega &= W_\omega^{f,0}+W_\omega^{0,\varphi}.
\end{aligned}$$

Then, using (6.13), (6.16), and Propositions 6.7 and 6.9, we obtain

$$\begin{aligned}
\delta a(u_0,v_0)-\delta l(v_0) &= \int_{\Gamma_R}[(\nu DW_\omega^{0,\varphi}-Q_\omega^{0,\varphi}I)-(\nu DV_\omega^{0,\varphi}-P_\omega^{0,\varphi}I)]\mathbf{n}.v_0\;d\gamma(x) \\
&\quad+\int_{\Gamma_R}[(\nu DW_\omega^{f,0}-Q_\omega^{f,0}I)-(\nu DV_\omega^{f,0}-P_\omega^{f,0}I)]\mathbf{n}.v_0\;d\gamma(x) \\
&= \int_{\Gamma_R}[(\nu DW_\omega^{f,\varphi}-Q_\omega^{f,\varphi}I)-(\nu DV_\omega^{f,\varphi}-P_\omega^{f,\varphi}I)]\mathbf{n}.v_0\;d\gamma(x),
\end{aligned}$$

which achieves the proof of Theorem 4.1.

6.6. Proof of Proposition 4.3. This section describes the variations of $J_\varepsilon(u) = \widetilde{J}_\varepsilon(\widetilde{u}_\varepsilon)$ (see (3.13)) when $\widetilde{J}_\varepsilon$ is of the form (4.17)

$$\widetilde{J}_\varepsilon(u) = \int_{\Omega_\varepsilon} g(x, u(x))\, dx, \quad u \in H^1(\Omega_\varepsilon)^3.$$

The hypotheses on g (4.18)–(4.20) described in section 4 are supposed to be satisfied. Throughout the two next sections, $\widetilde{u}_\varepsilon \in H^1(\Omega_\varepsilon)^3$, $\varepsilon \geq 0$, denotes the extension of $u \in \mathcal{V}_R$ which coincides with u on Ω_R and with $u_\varepsilon^{f,\varphi}$ on D_ε for $\varphi = u_{|\Gamma_R}$.

LEMMA 6.10. *Let $\varphi \in H_V^{1/2}(\Gamma_R)^3$ and $f \in L^q(\Omega)^3$, $q > n$. Let $u_\varepsilon^{f,\varphi}$ and $u_0^{f,\varphi}$ be, respectively, the solutions to* (3.2) *and* (3.3)*. Then*

$$||u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon(W_\omega^{f,\varphi} - V_\omega^{f,\varphi})||_{0,D_\varepsilon} = O(\varepsilon^{3/2}), \tag{6.17}$$

$$||u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon W_\omega^{f,\varphi} + v_\omega^{f,\varphi}(x/\varepsilon)||_{1,D_\varepsilon} = O(\varepsilon^{3/2}), \tag{6.18}$$

where $V_\omega^{f,\varphi}$ is the dominant part (4.6) *of the solution $v_\omega^{f,\varphi}$ to the exterior problem* (4.1) *with $u_0^{f,\varphi}(x_0)$ substituted for $u_\Omega(x_0)$, and $W_\omega^{f,\varphi}$ is the associated solution to* (4.9)*.*

Proof. Recall that $v_\omega = V_\omega + R_\omega$ (4.6) with $V_\omega(x/\varepsilon) = \varepsilon V_\omega(x)$ and $R_\omega(y) = O(1/||y||^2)$ (we drop the subscripts $(.)^{f,\varphi}$). Let

$$\begin{aligned} w_\varepsilon(x) &= \left(u_\varepsilon - u_0 - \varepsilon(W_\omega - V_\omega)\right)(x) + R_\omega(x/\varepsilon) \\ &= u_\varepsilon(x) - u_0(x) + v_\omega(x/\varepsilon) - \varepsilon W_\omega(x). \end{aligned} \tag{6.19}$$

Function w_ε (with the appropriate s_ε) is the solution to

$$\left\{ \begin{array}{rcll} -\nu\Delta w_\varepsilon + \nabla s_\varepsilon = & 0 & & \text{in } D_\varepsilon, \\ \operatorname{div} w_\varepsilon = & 0 & & \text{in } D_\varepsilon, \\ w_\varepsilon = & v_\omega(x/\varepsilon) - \varepsilon W_\omega(x) & & \text{on } \Gamma_R, \\ w_\varepsilon = & -u_0(x) + u_0(x_0) - \varepsilon W_\omega(x) & & \text{on } \partial\omega_\varepsilon. \end{array} \right.$$

Using the same arguments as in the proofs of Proposition 6.6 and 6.8 we obtain

$$\begin{aligned} \|v_\omega(x/\varepsilon) - \varepsilon Q_\omega\|_{1/2,\Gamma_R} &\leq c\varepsilon^2 \left(\|\varphi\|_{1/2,\Gamma_R} + \|f\|_{L^q} \right), \\ \| - u_0(\varepsilon y) + u_0(x_0) - \varepsilon Q_\omega)\|_{1/2,\partial\omega} &\leq c\varepsilon \left(\|\varphi\|_{1/2,\Gamma_R} + \|f\|_{L^q} \right). \end{aligned}$$

It follows from Lemma 6.5 that

$$\begin{aligned} ||w_\varepsilon||_{0,D_\varepsilon} &\leq c\varepsilon^2 \left(\|\varphi\|_{1/2,\Gamma_R} + \|f\|_{L^q} \right), \\ ||w_\varepsilon||_{1,D_\varepsilon} &\leq c\varepsilon^{3/2} \left(\|\varphi\|_{1/2,\Gamma_R} + \|f\|_{L^q} \right). \end{aligned}$$

The second equation proves (6.18). Due to (6.6), Lemma 6.2, and elliptic regularity we also have

$$\|R_\omega(x/\varepsilon)\|_{0,D_\varepsilon} = \varepsilon^{3/2} \|R_\omega\|_{0,D_\varepsilon/\varepsilon} \leq c\varepsilon^{3/2} \left(\|\varphi\|_{1/2,\Gamma_R} + \|f\|_{L^q} \right).$$

We conclude by using $u_\varepsilon - u_0 - \varepsilon(W_\omega - V_\omega) = w_\varepsilon(x) - R_\omega(x/\varepsilon)$. ∎

The variation $J_\varepsilon(u) - J_0(u)$ is given by the next lemma.

LEMMA 6.11. *For $u \in \mathcal{V}_R$ we have*

$$\begin{aligned} J_\varepsilon(u) &= J_0(u) + \varepsilon\delta J(u) + o(\varepsilon), \\ \delta J(u) &= \int_{D_0} \nabla_s g(x, \widetilde{u}_0(x)).(W_\omega - V_\omega)\, dx, \end{aligned}$$

where W_ω and V_ω are defined as in Lemma 6.10, with $\varphi = u_{|\Gamma_R}$.

Proof. Let

$$I_\varepsilon = J_\varepsilon(u) - J_0(u) - \varepsilon \int_{D_0} \nabla_s g(x, \widetilde{u}_0).(W_\omega - V_\omega)\, dx.$$

On D_ε we have $\widetilde{u}_\varepsilon = u_\varepsilon^{f,\varphi}$ for $\varepsilon \geq 0$, $\varphi = u$ on Γ_R, and on Ω_R we have $\widetilde{u}_\varepsilon = \widetilde{u}_0$. Hence

$$\begin{aligned} I_\varepsilon &= \widetilde{J}_\varepsilon(\widetilde{u}_\varepsilon) - \widetilde{J}_0(\widetilde{u}_0) - \varepsilon \int_{D_0} \nabla_s g(x, \widetilde{u}_0).(W_\omega - V_\omega)\, dx \\ &= \int_{\Omega\varepsilon} g(x, \widetilde{u}_\varepsilon)\, dx - \int_\Omega g(x, \widetilde{u}_0)\, dx - \varepsilon \int_{D_0} \nabla_s g(x, \widetilde{u}_0).(W_\omega - V_\omega)\, dx \\ &= \int_{D_\varepsilon} g(x, u_\varepsilon^{f,\varphi}) - g(x, u_0^{f,\varphi})\, dx - \int_{\omega_\varepsilon} g(x, u_0^{f,\varphi})\, dx - \varepsilon \int_{D_0} \nabla_s g(x, u_0^{f,\varphi}).(W_\omega - V_\omega)\, dx. \end{aligned}$$

Due to the hypotheses on g (4.18) and (4.19), we have, for all $(x, s, t) \in \Omega \times \mathbb{R}^3 \times \mathbb{R}^3$,

$$\begin{aligned} &g(x,t) - g(x,s) = \nabla_s g(x,s).(t-s) + \theta(x,s,t)(t-s).(t-s), \\ &\|\theta(x,s,t)\|_{\mathcal{L}(\mathbb{R}^3)} \leq \frac{M}{2}. \end{aligned}$$

Then

$$\begin{aligned} I_\varepsilon = &\int_{D_\varepsilon} \nabla_s g(x, u_0^{f,\varphi}).\left(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon(W_\omega - V_\omega)\right)\, dx \\ &-\varepsilon \int_{\omega_\varepsilon} \nabla_s g(x, u_0^{f,\varphi}).(W_\omega - V_\omega)\, dx - \int_{\omega_\varepsilon} g(x, u_0^{f,\varphi})\, dx \\ &+ \int_{D_\varepsilon} \theta(x, u_0^{f,\varphi}, u_\varepsilon^{f,\varphi})(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi}).(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi})\, dx, \end{aligned}$$

and

$$\begin{aligned} |I_\varepsilon| \leq &\int_{D_\varepsilon} \left|\nabla_s g(x, u_0^{f,\varphi}).\left(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon(W_\omega - V_\omega)\right)\right|\, dx \\ &+\varepsilon \int_{\omega_\varepsilon} \left|\nabla_s g(x, u_0^{f,\varphi}).(W_\omega - V_\omega)\right|\, dx + \int_{\omega_\varepsilon} \left|g(x, u_0^{f,\varphi})\right|\, dx + \int_{D_\varepsilon} \frac{M}{2}|u_\varepsilon^{f,\varphi} - u_0^{f,\varphi}|^2\, dx. \end{aligned}$$

It follows from the hypotheses on g (4.18)–(4.20), Lemma 6.10, the regularity of $u_0^{f,\varphi}$ (which implies that $x \mapsto g(x, u_0^{f,\varphi}(x))$ is in $L^{3/2}(D_{R/2})$), $||V_\omega||_{0,\omega_\varepsilon} = c\varepsilon^{1/2}$, and $||\nabla_s g(., u_0^{f,\varphi}(.))||_{0,\omega_\varepsilon} = o(1)$ that

$$\begin{aligned} \int_{D_\varepsilon} \left|\nabla_s g(x, u_0^{f,\varphi}).\left(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon(W_\omega - V_\omega)\right)\right|\, dx &\leq c||u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon(W_\omega - V_\omega)||_{0,D_\varepsilon} \\ &= O(\varepsilon^{3/2}), \\ \int_{D_\varepsilon} \frac{M}{2}|u_\varepsilon^{f,\varphi} - u_0^{f,\varphi}|^2\, dx &= O(\varepsilon^2), \\ \varepsilon \int_{\omega_\varepsilon} \left|\nabla_s g(x, u_0^{f,\varphi}).(W_\omega - V_\omega)\right|\, dx &\leq \varepsilon||\nabla_s g(., u_0^{f,\varphi}(.))||_{0,\omega_\varepsilon}||W_\omega - V_\omega||_{0,\omega_\varepsilon} \\ &= o(\varepsilon^{3/2}), \end{aligned}$$

$$\int_{\omega_\varepsilon} \left| g(x, u_0^{f,\varphi}) \right| \, dx \leq \left(\int_{\omega_\varepsilon} \left| g(x, u_0^{f,\varphi}) \right|^{3/2} \, dx \right)^{2/3} \left(\int_{\omega_\varepsilon} dx \right)^{1/3} = o(\varepsilon). \tag{6.20}$$

Hence

$$I_\varepsilon = o(\varepsilon). \qquad \square$$

We can now check hypothesis (4.11) involved in Theorem 4.1.

PROPOSITION 6.12. *Function J_0 is differentiable on $\mathcal{V}_R$ and we have, for all $u, v \in \mathcal{V}_R$,*

$$J_\varepsilon(v) - J_0(u) = \varepsilon \delta J(u) + DJ_0(u)(v-u) + o(\varepsilon + \|v-u\|_{\mathcal{V}_R}).$$

Proof. We have

$$J_0(u) = \widetilde{J}_0(\widetilde{u}) = \int_\Omega g(x, \widetilde{u}(x)) \, dx.$$

It follows from the hypotheses on g (4.18) that function $\widetilde{J}_0$ is differentiable on $H^1(\Omega)^3$ with

$$D\widetilde{J}_0(\widetilde{u}_0)w = \int_\Omega \nabla_s g(x, \widetilde{u}_0).w \, dx, \quad w \in H^1(\Omega)^3.$$

Thus J_0 is differentiable on $\mathcal{V}_R$, and for $w \in \mathcal{V}_R$ extended by $\widehat{w} \in H^1(\Omega)^3$ with $-\nu\Delta\widehat{w} + \nabla\widehat{q} = 0$ in D_0 and $\operatorname{div}\widehat{w} = 0$, we have

$$DJ_0(u)w = D\widetilde{J}_0(\widetilde{u}_0)\widehat{w}.$$

Hence, applying Lemma 6.11 yields

$$\begin{aligned} J_\varepsilon(v) - J_0(u) &= J_\varepsilon(v) - J_0(v) + J_0(v) - J_0(u) \\ &= \varepsilon\delta J(v) + o(\varepsilon) + DJ_0(u)(v-u) + o(\|v-u\|_{\mathcal{V}_R}) \\ &= \varepsilon\delta J(u) + DJ_0(u)(v-u) + o(\varepsilon + \|v-u\|_{\mathcal{V}_R}) \\ &\quad + \varepsilon(\delta J(v) - \delta J(u)). \end{aligned}$$

It remains to prove that $\varepsilon(\delta J(v) - \delta J(u)) = o(\varepsilon + \|v-u\|_{\mathcal{V}_R})$. For this it is sufficient to prove that $\delta J(v) - \delta J(u) = O(\|v-u\|_{\mathcal{V}_R})$. With the notation defined below in (6.21) and (6.22), it follows from Lemma 6.11 that

$$\begin{aligned} \delta J(v) - \delta J(u) &= \int_{D_0} \nabla_s g(x, \widetilde{v}_0).(W_\omega^v - V_\omega^v) - \nabla_s g(x, \widetilde{u}_0).(W_\omega^u - V_\omega^u) \, dx \\ &= \int_{D_0} \left[\nabla_s g(x, \widetilde{v}_0) - \nabla_s g(x, \widetilde{u}_0) \right].(W_\omega^v - V_\omega^v) \, dx \\ &\quad + \int_{D_0} \nabla_s g(x, \widetilde{u}_0).\left[(W_\omega^v - V_\omega^v) - (W_\omega^u - V_\omega^u) \right] \, dx. \end{aligned}$$

Hence, using the hypotheses on g (4.18)–(4.19) we obtain

$$\begin{aligned} |\delta J(v) - \delta J(u)| &\leq \int_{D_0} M|\widetilde{v}_0 - \widetilde{u}_0||W_\omega^v - V_\omega^v| \, dx \\ &\quad + \int_{D_0} \left(|\nabla_s g(x,0)| + M|\widetilde{u}_0| \right) \left(|W_\omega^v - W_\omega^u| + |V_\omega^v - V_\omega^u| \right) \, dx. \end{aligned}$$

We conclude by using linearity and continuity of

$$\begin{array}{ccccccc} \mathcal{V}_R & \to & H_V^{1/2}(\Gamma_R)^3 & \to & H^1(D_0)^3 & \to & L^2(D_0)^3, \\ u & \mapsto & \varphi := u_{|\Gamma_R} & \mapsto & u_0^{f,\varphi} & \mapsto & V_\omega^u := EA_\omega(u_0^{f,\varphi}(x_0)) \end{array} \tag{6.21}$$

and

$$\begin{array}{ccccccc} \mathcal{V}_R & \to & H^{1/2}(\Gamma_R)^3 & \to & \mathbb{R}, \\ u & \mapsto & (V_\omega^u)_{|\Gamma_R} & \mapsto & W_\omega^u. \end{array} \tag{6.22}$$

□

Hence, hypothesis (4.11) is fulfilled and we can apply Theorem 4.1. The adjoint equation (4.14) reads

$$\nu \int_\Omega Dw : Dv_\Omega \, dx = -\int_\Omega \nabla_s g(x, u_\Omega).w \, dx;$$

hence

$$-\nu\Delta v_\Omega + \nabla q_\Omega = -\nabla_s g(x, u_\Omega). \tag{6.23}$$

It follows from (4.16), Proposition 6.11, and (6.23) that

$$\begin{aligned} \delta j(x_0) &= A_\omega(u_\Omega(x_0)).v_\Omega(x_0) + \int_{D_0} (\nu\Delta v_\Omega - \nabla q_\Omega).(V_\omega - W_\omega) \, dx + \delta J(u_0) \\ &= A_\omega(u_\Omega(x_0)).v_\Omega(x_0) + \int_{D_0} (\nu\Delta v_\Omega - \nabla q_\Omega)(V_\omega - W_\omega) \, dx \\ &\quad + \int_{D_0} \nabla_s g(x, u_\Omega)(W_\omega - V_\omega) \, dx \\ &= A_\omega(u_\Omega(x_0)).v_\Omega(x_0), \end{aligned}$$

which achieves the proof of Proposition 4.3.

6.7. Proof of Proposition 4.4. Here $\widetilde{J}_\varepsilon$ is of the form (4.21)

$$\widetilde{J}_\varepsilon(u) = \frac{1}{2} \int_{\Omega_\varepsilon} BD(u - u_{\mathrm{d}}) : D(u - u_{\mathrm{d}}) \, dx, \quad u \in H^1(\Omega_\varepsilon)^3.$$

The notation is the same as in section 6.6. For $u \in \mathcal{V}_R$, we have

$$J_\varepsilon(u) = \widetilde{J}_\varepsilon(\widetilde{u}_\varepsilon) = \frac{1}{2} \int_{\Omega_\varepsilon} BD(\widetilde{u}_\varepsilon - u_{\mathrm{d}}) : D(\widetilde{u}_\varepsilon - u_{\mathrm{d}}) \, dx.$$

Due to the assumptions Δu_{d}, $f \in L^q(\Omega)^3$ with $q > n$, we have Du_{d}, $D\,\widetilde{u}_0 \in \mathcal{C}^0(\overline{B}(0, R/2))^9$ (see [8, Chap. II, sect. 3, Prop. 6]); hence

$$\int_{\omega_\varepsilon} BD(\widetilde{u}_0 - u_{\mathrm{d}}) : D(\widetilde{u}_0 - u_{\mathrm{d}}) \, dx = O(\varepsilon^3).$$

This and the fact that $b_{ijkl}(x) = b_{klij}(x)$ yield

$$J_\varepsilon(u) - J_0(u) = \frac{1}{2} \int_{D_\varepsilon} 2\, BD(\widetilde{u}_0 - u_{\mathrm{d}}) : D(\widetilde{u}_\varepsilon - \widetilde{u}_0) + BD(\widetilde{u}_\varepsilon - \widetilde{u}_0) : D(\widetilde{u}_\varepsilon - \widetilde{u}_0) \, dx + o(\varepsilon).$$

Equation (6.19) reads here as

$$w_\varepsilon(x) = \widetilde{u}_\varepsilon(x) - \widetilde{u}_0(x) + v_\omega(x/\varepsilon) - \varepsilon W_\omega(x)$$

and

$$\begin{aligned} J_\varepsilon(u) - J_0(u) = &\int_{D_\varepsilon} BD(\widetilde{u}_0 - u_{\mathrm{d}}) : D(\varepsilon W_\omega - v_\omega(x/\varepsilon) + w_\varepsilon)\,dx \\ &+\frac{1}{2}\int_{D_\varepsilon} BD(\varepsilon W_\omega - v_\omega(x/\varepsilon) + w_\varepsilon) : D(\varepsilon W_\omega - v_\omega(x/\varepsilon) + w_\varepsilon)\,dx + o(\varepsilon). \end{aligned}$$

Recall that $v_\omega = V_\omega + R_\omega$ (4.6) with $V_\omega(x/\varepsilon) = \varepsilon V_\omega(x)$ and $R_\omega(y) = O(1/||y||^2)$. Then

$$\begin{aligned} J_\varepsilon(u) - J_0(u) = &\int_{D_\varepsilon} BD(\widetilde{u}_0 - u_{\mathrm{d}}) : D(\varepsilon W_\omega(x) - \varepsilon V_\omega(x))\,dx \\ &-\int_{D_\varepsilon} BD(\widetilde{u}_0 - u_{\mathrm{d}}) : D_x R_\omega(x/\varepsilon)\,dx \\ &+\int_{D_\varepsilon} BD(\widetilde{u}_0 - u_{\mathrm{d}}) : Dw_\varepsilon(x)\,dx + \frac{\varepsilon^2}{2}\int_{D_\varepsilon} BDW_\omega : DW_\omega\,dx \\ &+\frac{1}{2}\int_{D_\varepsilon} BD_x v_\omega(x/\varepsilon) : D_x v_\omega(x/\varepsilon)\,dx - \varepsilon\int_{D_\varepsilon} BD_x v_\omega(x/\varepsilon) : DW_\omega\,dx \\ &+\int_{D_\varepsilon} BD(\varepsilon W_\omega - v_\omega(x/\varepsilon)) : Dw_\varepsilon\,dx + \frac{1}{2}\int_{D_\varepsilon} BDw_\varepsilon : Dw_\varepsilon\,dx + o(\varepsilon). \end{aligned}$$

Here D_x denotes the derivative with respect to x, and particularly $D(v(x/\varepsilon)) = D_x v(x/\varepsilon) = Dv(x/\varepsilon)/\varepsilon$. We have $||D_x R_\omega(x/\varepsilon)||_{L^1(D_\varepsilon)} = \varepsilon^2 ||DR_\omega||_{L^1(D_\varepsilon/\varepsilon)} = O(\varepsilon^2|\log\varepsilon|)$; hence

$$\begin{aligned} \left|\int_{D_\varepsilon} BD(\widetilde{u}_0 - u_{\mathrm{d}}) : D_x R_\omega(x/\varepsilon)\,dx\right| &\le ||BD(\widetilde{u}_0 - u_{\mathrm{d}})||_\infty ||D_x R_\omega(x/\varepsilon)||_{L^1(D_\varepsilon)} \\ &\le (||\widetilde{u}_0||_{3,D_0} + ||u_{\mathrm{d}}||_{1,\infty,D_0})||D_x R_\omega(x/\varepsilon)||_{L^1(D_\varepsilon)} \\ &= o(\varepsilon). \end{aligned}$$

It follows from the regularity of W_ω and Lemmas 6.2 and 6.10 that

$$\int_{D_\varepsilon} BD(\widetilde{u}_0 - u_{\mathrm{d}}) : Dw_\varepsilon(x)\,dx = O(\varepsilon^{3/2}),$$

$$\int_{D_\varepsilon} \varepsilon^2 BDW_\omega : DW_\omega\,dx = O(\varepsilon^2),$$

$$\varepsilon\int_{D_\varepsilon} BD_x v_\omega(x/\varepsilon) : DW_\omega\,dx = O(\varepsilon^{3/2}),$$

$$\int_{D_\varepsilon} BD(\varepsilon W_\omega - v_\omega(x/\varepsilon)) : Dw_\varepsilon\,dx = O(\varepsilon^2),$$

$$\int_{D_\varepsilon} BDw_\varepsilon : Dw_\varepsilon\,dx = O(\varepsilon^3).$$

Hence, using $DV_\omega = O(1/r^2)$, which implies that $\int_{\omega_\varepsilon} BD(\widetilde{u}_0 - u_\mathrm{d}) : D(W_\omega - V_\omega)\,dx = O(\varepsilon)$, we obtain

$$J_\varepsilon(u) - J_0(u) = \varepsilon \int_{D_0} BD(\widetilde{u}_0 - u_\mathrm{d}) : D(W_\omega - V_\omega)\,dx + \frac{1}{2}\int_{D_\varepsilon} BD_x v_\omega(x/\varepsilon) : D_x v_\omega(x/\varepsilon)\,dx + o(\varepsilon).$$

The adjoint equation reads for $\varphi \in \mathcal{D}^3(D_0)$ as

$$\int_{D_0} (-\nu\Delta v_\Omega + \nabla q_\Omega).\varphi\,dx = -\int_{D_0} BD(u_\Omega - u_\mathrm{d}) : D\varphi\,dx. \tag{6.24}$$

Due to $B \in W^{1,\infty}(\Omega)^{9\times 9}$ and Δu_d, $f \in L^q(\Omega)^3$ (thus $D^2 u_\mathrm{d}$, $D^2 u_\Omega \in L^q(D_0)$; cf. the Calderon–Zygmund theorem [8, Chap. II, sect. 3, Prop. 8]), we have $-\nu\Delta v_\Omega + \nabla q_\Omega = \mathrm{div}\,[BD(u_\Omega - u_\mathrm{d})] \in L^q(D_0)^3$. Moreover, $q > n/2$ and $W_\omega - V_\omega \in L^m(D_0)^3$ for all $m < 3$; thus $(-\nu\Delta v_\Omega + \nabla q_\Omega).(P_\omega - Q_\omega) \in L^1(D_0)$. Hence, as $W_\omega - V_\omega$ vanishes on Γ_R, (6.24) still holds for $\varphi = W_\omega - V_\omega$, and

$$J_\varepsilon(u_0) - J_0(u_0) = \int_{D_0} \varepsilon(\nu\Delta v_\Omega - \nabla q_\Omega).(W_\omega - V_\omega) + \frac{1}{2}\int_{D_\varepsilon} BD_x v_\omega(x/\varepsilon) : D_x v_\omega(x/\varepsilon)\,dx + o(\varepsilon).$$

Then the proof can be achieved as in section 6.6. It follows from Corollary 4.2 that

$$j(\varepsilon) = j(0) + \varepsilon A_\omega(u_\Omega(x_0)).v_\Omega(x_0) + \frac{1}{2}\int_{D_\varepsilon} BD_x v_\omega(x/\varepsilon) : D_x v_\omega(x/\varepsilon)\,dx + o(\varepsilon).$$

Using Lebesgue's convergence theorem, we deduce that

$$\int_{D_\varepsilon} B(x) D_x v_\omega(x/\varepsilon) : D_x v_\omega(x/\varepsilon)\,dx = \varepsilon \int_{D_\varepsilon/\varepsilon} B(\varepsilon y) Dv_\omega(y) : Dv_\omega(y)\,dy = \varepsilon \int_{\mathbb{R}^3\setminus\overline{\omega}} B(x_0) Dv_\omega(y) : Dv_\omega(y)\,dy + o(\varepsilon),$$

which proves that

$$j(\varepsilon) = j(0) + \varepsilon A_\omega(u_\Omega(x_0)).v_\Omega(x_0) + \frac{\varepsilon}{2}\int_{\mathbb{R}^3\setminus\overline{\omega}} B(x_0) Dv_\omega(y) : Dv_\omega(y)\,dy + o(\varepsilon)$$

and completes the proof of Proposition 4.4.

Acknowledgment. The authors are grateful to the referees for their careful reading and their valuable suggestions concerning the presentation.

REFERENCES

[1] G. Allaire, *Homogenization of the Navier-Stokes equations in open sets perforated with tiny holes. I. Abstract framework, a volume distribution of holes*, Arch. Ration. Mech. Anal., 113 (1990), pp. 209–259.

[2] G. Allaire and A. Henrot, *On some recent advances in shape optimization*, C. R. Acad. Sci. Paris, Sér. IIb, 329 (2001), pp. 383–396.

[3] H. Brezis, *Analyse fonctionnelle*, in Théorie et applications, Masson, Paris, 1993.
[4] G. Buttazzo and G. Dal Maso, *Shape optimization for Dirichlet problems: Relaxed formulation and optimality conditions*, Appl. Math. Optim. 23 (1991), pp. 17–49.
[5] J. Céa, *Conception optimale ou identification de forme: Calcul rapide de la dérivée directionnelle de la fonction coût*, RAIRO Modél Math. Anal. Numer., 20 (1986), pp. 371–402 (in French).
[6] J. Céa, A. Gioan, and J. Michel, *Quelques résultats sur l'identification de domaines*, Calcolo, 10 (1973), pp. 207–232 (in French).
[7] M. Chipot and G. Dal Maso, *Relaxed shape optimization: The case of nonnegative data for the Dirichlet problem*, Adv. Math. Sci. Appl., 1 (1992), pp. 47–81.
[8] R. Dautray and J. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques, Vol.* 1, in INSTN: Collection Enseignement, Masson, Paris 1987 (in French).
[9] S. Garreau, Ph. Guillaume, and M. Masmoudi, *The topological asymptotic for PDE systems: The elasticity case*, SIAM J. Control Optim., 39 (2001), pp. 1756–1778.
[10] Ph. Guillaume and K. Sid Idris, *The topological asymptotic expansion for the Dirichlet problem*, SIAM J. Control Optim., 41 (2002), pp. 1042–1072.
[11] M. Masmoudi, *The topological asymptotic expansion*, in Computational Methods for Control Applications, GAKUTO Internat. Ser. Math. Sci. Appl. 16, H. Kawarada and J. Periaux, eds., Gakkotosho, Tokyo, 2001, pp. 53–72.
[12] W. G. Mazja, S. A. Nazarov, and B. A. Plamennevskii, *Asymptotic Theory of Elliptic Boundary Value Problems in Singular Perturbed Domains*, Birkhaüser Verlag, Basel, 2000.
[13] F. Murat and L. Tartar, *Calcul des variations et homogénéisation*, in Les méthodes de l'homogénéisation: Théorie et Applications en Physique, Eyrolles, Paris, 1985, pp. 319–369 (in French).
[14] A. Schumacher, *Topologieoptimierung von Bauteilstrukturen unter Verwendung von Lochpositionierungkrieterien*, Thesis, Universität-Gesamthochschule-Siegen, Siegen, Germany, 1995.
[15] J. Sokolowski and A. Zochowski, *On the topological derivative in shape optimization*, SIAM J. Control Optim., 37 (1999), pp. 1251–1272.
[16] R. Temam, *Navier Stokes Equations: Theory and Numerical Analysis*, 3rd ed., Stud. Math. Appl. 2, North–Holland, Amsterdam, 1984.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 32–57

PRACTICAL AND ASYMPTOTIC STABILIZATION OF CHAINED SYSTEMS BY THE TRANSVERSE FUNCTION CONTROL APPROACH*

PASCAL MORIN† AND CLAUDE SAMSON†

Abstract. A control approach for the practical and asymptotic stabilization of nonlinear driftless systems subjected to additive perturbations is proposed. Such perturbations arise naturally, for instance, in the modeling of trajectory stabilization problems for controllable driftless systems on Lie groups. The objective of the approach is to provide practical stability of an arbitrary given point in the state space, whatever the perturbations, and asymptotic stability (resp., convergence to the point) when the perturbations are absent (resp., tend to zero). A general framework is presented in this paper, and a control solution is proposed for the class of the chained systems.

Key words. nonlinear systems, stabilization, feedback control, Lie groups

AMS subject classifications. 93B05, 93B29, 93C10

DOI. 10.1137/S0363012903421868

1. Introduction.

The development of the transverse function (t.f.) approach [18] finds its original motivation in the problem of *practical* stabilization of the origin of a control system in the form

$$\mathcal{S}: \qquad \dot{x} = \sum_{i=1}^{m} u_i X_i(x) + P(x,t), \tag{1}$$

with $x \in \mathbb{R}^n$, $n > m$, $\{X_1, \ldots, X_m\}$ a set of smooth vector fields (v.f.) that satisfy the Lie algebra rank condition (LARC) on an open ball centered at $x = 0$, and P an additive perturbation, continuous in x and t but otherwise *arbitrary*. Note that such a perturbation may well forbid the existence of any equilibrium point for the controlled system. The t.f. approach provides a general solution to this problem. Up to now, and to our knowledge, this solution is unique in its class, even though several other methods and many control laws have been devised during the last decade to address the stabilization problem when $P \equiv 0$. These studies were motivated in the first place by Brockett's theorem [5] according to which, if $m < n$ and the control v.f. evaluated at $x = 0$ are linearly independent, no smooth or even continuous pure state feedback can make the origin of the system asymptotically stable. Different types of feedback laws have been considered to circumvent this difficulty, although not all of them guarantee Lyapunov stability. Discontinuous feedback [1, 3, 6, 11] and hybrid feedback [2, 15, 23] are two possibilities. Another one, more related to the present approach, consists of using continuous time-varying feedback [21, 7, 20, 27, 22, 13, 19, 16, 14]. An early survey on the control of nonholonomic systems, whose kinematic models are nonlinear driftless systems, can also be found in [4]. The importance of considering the perturbed case in association with the objective of practical stabilization is well illustrated when $\mathcal{S}$ is a system on a Lie group and the control objective consists of tracking a trajectory. Indeed, it is shown in [18] that the *error system* associated with

*Received by the editors January 29, 2003; accepted for publication (in revised form) October 7, 2003; published electronically May 25, 2004.
http://www.siam.org/journals/sicon/43-1/42186.html
†INRIA, 2004 Route des Lucioles, B. P. 93, 06902 Sophia-Antipolis Cedex, France (Pascal.Morin@inria.fr, Claude.Samson@inria.fr)

this problem is in the same form as the original system, except for the presence of a perturbation P. Moreover, when the trajectory is not a solution of the control system, asymptotic stabilization is not possible. Other reasons for considering practical stabilization as a reasonable control objective, in the case of nonlinear driftless systems, are also pointed out in [18]: lack of robustness of exponential (continuous/time-varying or discontinuous) stabilizers, nonexistence of feedback controllers capable of stabilizing asymptotically every feasible trajectory [12], and incapacity of most existing asymptotic stabilizers to ensure ε-ultimate boundedness of the closed-loop trajectories when a destabilizing perturbation P is present. However, it is important to realize that practical stabilization is by no means opposed to asymptotic stabilization. It is merely a weaker requirement, whose interest resides precisely in the fact that it is weaker and thus applicable to more numerous situations. Once practical stabilization is granted, it may still be possible, and desirable in some cases, to achieve asymptotic stabilization, or at least convergence to zero—when, for instance, P vanishes after some time. For the same reasons, feedback controllers derived with the t.f. approach should not be considered as antagonistic to other controllers proposed for nonlinear driftless systems—asymptotic stabilizers, in particular. A more pertinent issue is the possibility of deriving a practical stabilizer which also ensures asymptotic stabilization when the perturbation P allows for it. For instance, can the t.f. approach be used for this purpose?

This question is addressed in the present paper, and a partial positive answer is obtained. More precisely, an extension of the approach in [18] is proposed in order to achieve asymptotic stabilization of the origin of $\mathcal{S}$ when $P \equiv 0$, and asymptotic convergence to the origin when P tends to zero as time tends to infinity. The main ingredient of this extension is the concept of a *generalized* t.f. introduced in section 2. The principles of the t.f. approach and design of stabilizers are also laid out in this section. A solution to the problem of practical and asymptotic stabilization for the popular class of the chained systems is proposed in section 3, and illustrated by simulation results in section 4. The practical relevance of this case comes from the widespread use of chained systems to model the kinematic equations of various mechanical systems subjected to nonholonomic constraints (unicycle and car-like mobile robots, for instance) and also the possibility of using them as homogeneous approximations of dynamics involved in several other physical systems (ships, induction motors, etc.). Finding a more general solution, which applies to a broader class of systems, remains an open subject of research. In order to facilitate the reading of the paper, we have distributed the proofs of our results into two sections: the cores are given in section 5, whereas intermediate technical results of lesser conceptual significance are regrouped in the appendix.

Since the t.f. approach finds its most natural exposition in the context of systems which are invariant on Lie groups, we have chosen to recast the systems and control problems evoked above in this framework. Let us recall the prominent role played by Lie groups in control theory [26, 10]. In particular, controllable driftless systems can always be approximated by controllable driftless homogeneous systems which are, after a possible dynamic extension, systems on Lie groups. The chained systems, which are more specifically addressed here, are systems on Lie groups.

The following notation is used throughout the paper. The tangent space of a manifold M at a point p is denoted as T_pM. The differential of a smooth mapping f between manifolds, at a point p, is denoted as $df(p)$. The torus of dimension $p \in \mathbb{N}$ is $\mathbb{T}^p$ with $\mathbb{T} \triangleq \mathbb{R}/2\pi\mathbb{Z}$. An element $\theta \in \mathbb{T}$ is identified with the real number in $(-\pi, +\pi]$ which belongs to the class of equivalence of θ. Addition of angles makes $\mathbb{T}^p$

a Lie group. The ith component of $\sigma \in \mathbb{T}^p$ is denoted as σ_i, i.e., $\sigma = (\sigma_1, \ldots, \sigma_p)$. The canonical basis of $\mathbb{R}^p$ is the set of unitary vectors $\{e_i\}_{i=1,\ldots,p}$. Since this set is also the natural basis of the Lie algebra of $\mathbb{T}^p$, a vector field v on $\mathbb{T}^p$ is identified with its vector of coordinates in this basis, i.e., $v = (v_1, \ldots, v_p)'$ if $v = \sum_{i=1}^p v_i e_i$. If $\sigma(.)$ is a smooth curve on $\mathbb{T}^p$, this identification allows us to view $\dot{\sigma}(t)$ as a vector in $\mathbb{R}^p$. Consider a differentiable mapping f from $\mathbb{T}^p$ to a manifold M. By a slight abuse of notation, and for the sake of simplifying the writing of several forthcoming equations, we write the Lie derivative of f along e_i at σ as $\frac{\partial f}{\partial \sigma_i}(\sigma)$, or $\frac{\partial f}{\partial \sigma}(\sigma)e_i$, instead of $df(\sigma)(e_i)$ (or $L_{e_i} f(\sigma)$), even though the normal use of the partial derivative symbol refers to a system of coordinates on M. Accordingly, along an arbitrary v.f. v on $\mathbb{T}^p$, we write $\frac{\partial f}{\partial \sigma}(\sigma)v \triangleq df(\sigma)(v)$. We also use standard notation for Lie groups—see, e.g., [8] for more details on this topic. G denotes a Lie group of dimension n, with Lie algebra (of left-invariant v.f.) $\mathfrak{g}$. For simplicity, we assume that G is connected so that there exists a globally defined left-invariant distance d_G on G. The identity element of G is denoted by e. Left and right translations are denoted by l and r, respectively, i.e., $l_\sigma(\tau) = r_\tau(\sigma) = \sigma\tau$. As usual, if $X \in \mathfrak{g}$, then $\exp tX$ is the solution at time t of $\dot{g} = X(g)$ with initial condition $g(0) = e$. The adjoint representation of G is Ad; i.e., for $\sigma \in G$, $\mathrm{Ad}(\sigma) = dI_\sigma(e)$ with $I_\sigma : G \longrightarrow G$ defined by $I_\sigma(g) = \sigma g \sigma^{-1}$. By extension we define the v.f. $\mathrm{Ad}(\sigma)X$ on G by $\mathrm{Ad}(\sigma)X(g) = dl_g(e)(\mathrm{Ad}(\sigma)X(e))$. The differential of Ad is ad, and $(\mathrm{ad}X, Y) = [X, Y]$, the Lie bracket of X and Y.

2. Control of perturbed driftless systems by the t.f. approach. Consider a control system

$$\mathcal{S}(g): \qquad \dot{g} = \sum_{i=1}^m u_i X_i(g) + P(g,t) \tag{2}$$

on a Lie group G, with $X_1, \ldots, X_m$ independent left-invariant smooth v.f. that satisfy the LARC. We assume that the drift term $P(g,t)$ is a continuous function of g and t, and that[1]

$$\forall (g,t) \in G \times \mathbb{R}, \qquad P(g,t) \in \mathrm{span}\{X_1(g), \ldots, X_m(g)\}^\perp, \tag{3}$$

where orthogonality refers to an arbitrary Riemannian metric on G. The definition of a *transverse function*, as originally given in [17] for v.f. on an arbitrary manifold—i.e., not necessarily on a Lie group—is now recalled.

DEFINITION 1. *Let $X_1, \ldots, X_m$ denote smooth v.f. on a manifold M. A function $f \in \mathcal{C}^\infty(\mathbb{T}^p; M)$ is called a* transverse function *(for the v.f. $X_1, \ldots, X_m$) if*

$$\forall \sigma \in \mathbb{T}^p, \quad \mathrm{span}\{X_1(f(\sigma)), \ldots, X_m(f(\sigma))\} + df(\sigma)(T_\sigma \mathbb{T}^p) = T_{f(\sigma)} M. \tag{4}$$

Another way of writing the above relation (with the notation explained before) is

$$\forall \sigma \in \mathbb{T}^p, \quad \mathrm{span}\left\{X_1(f(\sigma)), \ldots, X_m(f(\sigma)), \frac{\partial f}{\partial \sigma_1}(\sigma), \ldots, \frac{\partial f}{\partial \sigma_p}(\sigma)\right\} = T_{f(\sigma)} M. \tag{5}$$

Note that, by this definition, the image set $\mathrm{Im}(f) = f(\mathbb{T}^p)$ is compact. The main contribution of [17] was to show that if a set of v.f. $X_1, \ldots, X_m$ satisfies the LARC at some point $q \in M$, then for any neighborhood $\mathcal{U}$ of q there exists a transverse function with values in $\mathcal{U}$.

[1] Note that (3) can always be obtained after the application of a suitable preliminary feedback.

In the context of stabilization, transverse functions allow to use $\dot{\sigma}$ as a new—virtual—control input vector. This leads us to introduce the following dynamic extension of $\mathcal{S}(g)$, which evolves on $G \times \mathbb{T}^p$:

$$\mathcal{S}(g,\sigma): \quad \begin{cases} \dot{g} = \sum_{i=1}^{m} u_i X_i(g) + P(g,t), \\ \dot{\sigma} = u_\sigma, \end{cases} \tag{6}$$

where (u, u_σ) is viewed as an extended control vector. In the following subsection, the practical stabilization of $g = e$ for $\mathcal{S}(g)$, based on the t.f. approach, is addressed. More details on the approach, as well as several examples with explicit derivations of t.f., can be found in [18].

2.1. Practical stabilization. The formulation of a general practical stabilization problem which can be associated with $\mathcal{S}(g)$ is as follows: given an arbitrary neighborhood $\mathcal{U}_G(e)$ of e, determine a (smooth, or at least continuous) feedback control (which depends on g and, eventually, on other variables) which asymptotically stabilizes some compact set $\mathcal{D}_G \subset \mathcal{U}_G(e)$. The t.f. control approach provides a solution to this problem. This solution is now recalled.

Consider the change of variables on $G \times \mathbb{T}^p$ defined by $\Psi_f(g,\sigma) = (f(\sigma)g^{-1}, \sigma)$, with f a t.f. such that $f(\mathbb{T}^p) \subset \mathcal{U}_G(e)$. From now on, in order to ease the notation, the element $f(\sigma)g^{-1} \in G$ associated with g and $f(\sigma)$ will be abbreviated as z, i.e., $z \triangleq f(\sigma)g^{-1}$. By differentiating both members of the equality $zg = f(\sigma)$, one easily verifies that, along any trajectory $(g,\sigma)(.)$ of $\mathcal{S}(g,\sigma)$,

$$\dot{z} = -dr_{g^{-1}}(f(\sigma))\left(\sum_{i=1}^{m} u_i X_i(f(\sigma)) - \frac{\partial f}{\partial \sigma}(\sigma)\dot{\sigma} + dl_z(g)P(g,t)\right). \tag{7}$$

Therefore, $\mathcal{S}(g,\sigma)$ is equivalent to the control system

$$\bar{\mathcal{S}}(z,\sigma): \begin{cases} \dot{z} = -dr_{g^{-1}}(f(\sigma))\left(\sum_{i=1}^{m} u_i X_i(f(\sigma)) - \frac{\partial f}{\partial \sigma}(\sigma)u_\sigma + dl_z(g)P(g,t)\right), \\ \dot{\sigma} = u_\sigma. \end{cases} \tag{8}$$

From the definition of Ψ_f, the asymptotic stability of $\{e\} \times \mathbb{T}^p$ for $\bar{\mathcal{S}}(z,\sigma)$ is equivalent to the asymptotic stability of $\{(f(\sigma),\sigma) : \sigma \in \mathbb{T}^p\}$ for $\mathcal{S}(g,\sigma)$. It is also equivalent to the asymptotic stability of $f(\mathbb{T}^p)$ for $\mathcal{S}(g)$, provided that, for some left-invariant distance on G, the initial value $\sigma(0)$ of σ is chosen so as to minimize the distance between $z(0) = f(\sigma(0))g(0)^{-1}$ and e.

Now, for any v.f. Z on G, the property of transversality of f ensures that the equation

$$\sum_{i=1}^{m} u_i X_i(f(\sigma)) - \frac{\partial f}{\partial \sigma}(\sigma)u_\sigma = -dl_z(g)P(g,t) - dr_g(z)Z(z) \tag{9}$$

admits a feedback solution $(u, u_\sigma)(g,\sigma,t)$. Applying any[2] such feedback law to $\bar{\mathcal{S}}(z,\sigma)$, and using the fact that $(dr_g(z))^{-1} = dr_{g^{-1}}(f)$, it follows from (7) that

$$\dot{z} = Z(z). \tag{10}$$

[2]The only (weak) requirement is that the solutions of $\mathcal{S}(g,\sigma)$ must be well defined for $t \in [0,\infty)$.

Therefore, provided that Z is chosen so as to asymptotically stabilize e for system (10), the feedback law (u, u_σ) defined by (9) makes the set $\{e\} \times \mathbb{T}^p$ asymptotically stable for $\bar{\mathcal{S}}(z, \sigma)$.

In general, the solution (u, u_σ) of (9) is not unique. It is shown in [18], however, that one can always find[3] t.f. $f \in \mathcal{C}^\infty(\mathbb{T}^{n-m}; G)$, i.e., such that $p = n - m$ with the notation of Definition 1. It is clear from the transversality condition (4) that this value of p is minimal and that the solution (u, u_σ) of (9), given f, is unique in this case. Allowing the t.f. f to depend on a larger number of variables provides complementary control inputs which can be used to guarantee complementary control objectives. The asymptotic stabilization of e for $\mathcal{S}(g)$ when $P \equiv 0$ will, for instance, be addressed in this way.

2.2. A framework for asymptotic stabilization. Let us introduce, in the framework of Lie groups, the following specific class of transverse functions.

DEFINITION 2. *Consider a function $f \in \mathcal{C}^\infty(\mathbb{T}^{n-m} \times \mathbb{T}^{n-m}; G)$ and the associated family of functions $\{f_\beta\}_{\beta \in \mathbb{T}^{n-m}}$ defined by $f_\beta(\theta) = f(\theta, \beta)$. The function f is called a* generalized *t.f. for the v.f. $X_1, \ldots, X_m$ on the Lie group G if*

$$\begin{aligned} &\forall \sigma = (\theta, \beta) \in \mathbb{T}^{n-m} \times \mathbb{T}^{n-m}, \\ &\operatorname{span}\{X_1(f(\sigma)), \ldots, X_m(f(\sigma))\} + df_\beta(\theta)(T_\theta \mathbb{T}^{n-m}) = T_{f(\sigma)} G \end{aligned} \tag{11}$$

and

$$\forall \beta \in \mathbb{T}^{n-m}, \quad f(0, \beta) = e. \tag{12}$$

From now on, variables in $\mathbb{T}^{n-m}$ will be indexed starting from $m + 1$; i.e., if $\theta \in \mathbb{T}^{n-m}$, then $\theta = (\theta_{m+1}, \ldots, \theta_n)$. With the notation specified in the introduction, another way of writing relation (11) is

$$\begin{aligned} &\forall \sigma = (\theta, \beta) \in \mathbb{T}^{n-m} \times \mathbb{T}^{n-m}, \\ &\operatorname{span}\left\{X_1(f(\sigma)), \ldots, X_m(f(\sigma)), \frac{\partial f}{\partial \theta_{m+1}}(\sigma), \ldots, \frac{\partial f}{\partial \theta_n}(\sigma)\right\} = T_{f(\sigma)} G. \end{aligned} \tag{13}$$

It is clear that any generalized t.f. is a t.f. It is also quite simple to build a generalized t.f. $f \in \mathcal{C}^\infty(\mathbb{T}^{n-m} \times \mathbb{T}^{n-m}; G)$ from a t.f. $\bar{f} \in \mathcal{C}^\infty(\mathbb{T}^{n-m}; G)$. For example, define

$$\forall (\theta, \beta) \in \mathbb{T}^{n-m} \times \mathbb{T}^{n-m}, \quad f(\theta, \beta) = (\bar{f}(\beta))^{-1} \bar{f}(\theta + \beta).$$

Let us now consider any generalized t.f. We let

$$\dot{\theta} = v, \qquad \dot{\beta} = w, \tag{14}$$

so that $\dot{\sigma} = u_\sigma = (v, w)$. With this notation, (9)—whose satisfaction yields $\dot{z} = Z(z)$—is equivalent to

$$\sum_{i=1}^{m} u_i X_i(f(\sigma)) - \frac{\partial f}{\partial \theta}(\sigma) v = \frac{\partial f}{\partial \beta}(\sigma) w - dl_z(g) P(g, t) - dr_g(z) Z(z). \tag{15}$$

From (11), this equation has a unique feedback solution $(u, v)(g, \sigma, t)$ for any function w. The v.f. Z is again chosen so as to make $z = e$ asymptotically stable. Now the

[3] Expressions of such functions are given in that paper—see also the next subsection.

objective is to determine w in order to make θ tend to zero. Indeed, this latter property implies, in view of (12), that f tends to e so that, from the fact that $z = f(\sigma)g^{-1}$ tends to e, the asymptotic convergence of g to e follows. Note that such a convergence cannot be obtained without the drift term P satisfying some extra conditions. For instance, if $P(e,t)$ is periodically different from zero, then it follows from (3) that e cannot be an equilibrium for system (2), whatever the control u. Moreover, under mild complementary regularity conditions upon the function P, convergence of $P(g,t)$ to zero when g tends to e and t tends to infinity is necessary to the convergence of the system's solutions to e.

The feedback law (u,v) defined by (15) ensures the convergence of z to e independently of w. Hence, the asymptotic behavior of $\theta(t)$ and $\beta(t)$, for the controlled system, is described by the *zero-dynamics* obtained by setting $z = e$ in (15), i.e.,

$$\sum_{i=1}^{m} u_i(g,\sigma,t)X_i(f(\sigma)) - \frac{\partial f}{\partial \theta}(\sigma)v(g,\sigma,t) = \frac{\partial f}{\partial \beta}(\sigma)w - P(f(\sigma),t). \tag{16}$$

From the initial assumption that the v.f. $X_1,\ldots,X_m$ are independent, there exist v.f. $X_{m+1},\ldots,X_n$ such that $\operatorname{span}\{X_1,\ldots,X_n\} = \mathfrak{g}$. For any such set of v.f., there exist smooth functions $a_{i,j}$ and $b_{i,j}$ such that

(17)
$$\forall j = m+1,\ldots,n\,, \quad \frac{\partial f}{\partial \theta_j}(\sigma) = \sum_{i=1}^{n} a_{i,j}(\sigma)X_i(f(\sigma))\,, \quad \frac{\partial f}{\partial \beta_j}(\sigma) = \sum_{i=1}^{n} b_{i,j}(\sigma)X_i(f(\sigma)).$$

With d_i $(i = m+1,\ldots,n)$ denoting the one-forms defined by $\langle d_i, X_k\rangle = \delta_{i,k}$ (the Kronecker delta), the application of d_i to each side of (16) yields, since $\dot\theta = v$,

$$A(\sigma)\dot\theta = -B(\sigma)w + \sum_{i=m+1}^{n} \langle d_i(f(\sigma)), P(f(\sigma),t)\rangle e_i \tag{18}$$

with

$$A(\sigma) \triangleq (a_{i,j}(\sigma))_{i,j=m+1,\ldots,n}, \qquad B(\sigma) \triangleq (b_{i,j}(\sigma))_{i,j=m+1,\ldots,n}, \tag{19}$$

and e_i the $(i-m)$th unit vector in $\mathbb{R}^{n-m}$. Note that the transversality condition (11) is equivalent to the matrix $A(\sigma)$ being invertible for any σ.

Equation (18) is important because it explicitly relates the control w (the time-derivative of β) to the variation of θ. In particular, the simplification obtained when $P \equiv 0$, i.e.,

$$\dot\theta = -A^{-1}(\sigma)B(\sigma)w, \tag{20}$$

suggests some ways of choosing w to make $|\theta(t)|$ nonincreasing on the zero-dynamics. However, a difficulty arising at this stage, to ensure the convergence of $\theta(t)$ to zero, comes from the fact that $B(\sigma)$ tends to the null matrix when θ tends to zero, since $f(0,\beta) = e\ \forall\beta \Rightarrow \frac{\partial f}{\partial \beta}(0,\beta) = 0,\ \forall\beta$. This difficulty is itself related to the well-known impossibility of ensuring *exponential* stabilization of e by means of a *smooth* feedback [13, Thm. 3]. The matter would still be easily settled if $B(\sigma)$ were invertible everywhere except at $\theta = 0$. Unfortunately, this is not true in general, and further inspection of this matrix, in relation to the way the structure of f combines with the

structure of the Lie algebra $\mathfrak{g}$, is required. Although we do not know whether a solution always exists, we were able to use the specific structure of the Lie algebra associated with the chained systems and derive a solution in this case. Prior to reporting it in the next section, we propose below a formulation of the problem which, whereas it is restricted to the zero-dynamics (20), simplifies the search for a solution for the complete system.

PROBLEM 1. *Given a neighborhood $\mathcal{U}_G(e)$ of e, determine a triplet (f, w, V) consisting of*

(i) *a generalized t.f. $f \in \mathcal{C}^\infty(\mathbb{T}^{n-m} \times \mathbb{T}^{n-m}; \mathcal{U}_G(e))$,*

(ii) *a function $w \in \mathcal{C}^1(\mathcal{U}_{\mathbb{T}^{n-m}}(0); \mathbb{R}^{n-m})$,*

(iii) *a function $V \in \mathcal{C}^1(\mathcal{U}_{\mathbb{T}^{n-m}}(0); \mathbb{R})$ with bounded first-order partial derivatives such that*

1. *$\forall \theta \in V^{-1}([0, V_{\max}))$, $h_{V_m}(|\theta|) \leq V(\theta) \leq h_{V_M}(|\theta|)$ with h_{V_m} and h_{V_M} two $\mathcal{K}$-functions, and $V_{\max} > 0$ a real number such that $V^{-1}([0, V_{\max})) \subset \mathcal{U}_{\mathbb{T}^{n-m}}(0)$;*
2. *the following proposition is true:*

$$\forall \beta \in \mathbb{T}^{n-m},\ \forall \theta \in V^{-1}([0, V_{\max})), \quad L_F(\pi^* V)(\sigma) \leq -\gamma V(\theta)^l, \quad \gamma, l > 0, \tag{21}$$

with π and F defined by

$$\forall \sigma = (\theta, \beta), \quad \pi(\sigma) = \theta,\ F(\sigma) = -A^{-1}(\sigma) B(\sigma) w(\theta). \tag{22}$$

Note that (21) clearly implies that $\theta = 0$ is locally asymptotically stable for the system (20). Note also that the inclusion $\mathcal{U}_{\mathbb{T}^{n-m}}(0) \subset \mathbb{T}^{n-m}$ has to be strict since (21) would otherwise contradict the known nonexistence of global asymptotic stabilizers on $\mathbb{T}^{n-m}$. Once the above problem is solved, it is not difficult to infer a solution to the problem of asymptotic stabilization of e for system $\mathcal{S}(g)$ when $P \equiv 0$. Such a solution is pointed out in the following proposition.

PROPOSITION 1. *Let Z denote a smooth v.f. which asymptotically stabilizes e for the system $\dot{z} = Z(z)$. Assume that Problem 1 is solved by a triplet $(f, w^\star, V)$, and consider for $\mathcal{S}(g, \sigma)$ the feedback control (u, v, w) with (u, v) defined by (15) and w defined by*

$$w(\theta) = k\left(\frac{1}{V_{\max} - V(\theta)}\right) w^\star(\theta), \tag{23}$$

with k denoting any $\mathcal{K}_\infty$-function. Assume also that the initial condition $\theta(0)$ is chosen in $V^{-1}([0, V_{\max}))$. Then

1. *whatever P, the above-defined feedback control asymptotically stabilizes the set $\{(f(\sigma), \sigma) : \sigma \in V^{-1}([0, V_{\max})) \times \mathbb{T}^{n-m}\}$ for $\mathcal{S}(g, \sigma)$;*
2. *if $P \equiv 0$, then this control asymptotically stabilizes the set $\{e\} \times \{0\} \times \mathbb{T}^{n-m}$ for $\mathcal{S}(g, \sigma)$;*
3. *if $P(g, t)$ tends to zero as $t \longrightarrow +\infty$, uniformly w.r.t. g in compact sets, then $(g, \theta)(t) \longrightarrow (e, 0)$ as $t \longrightarrow +\infty$.*

Note that when P, Z, and k are differentiable, the stabilizing feedback control (u, v, w) so obtained is also differentiable. When $P \equiv 0$ and $\theta(0) = 0$, this control asymptotically stabilizes e for $\mathcal{S}(g)$. However, as in the case of a time-periodic Lipschitz-continuous asymptotic stabilizer of $\mathcal{S}(g)$, the control's differentiability rules out the possibility of a uniform convergence rate as fast as exponential. On the other hand, while the frequency of a time-periodic stabilizer is constant, the time-derivatives of θ and β, which may be interpreted as self-adapting frequencies in the case of a stabilizer derived with the t.f. approach, asymptotically tend to zero.

2.3. A class of generalized t.f. In this section, we introduce a class of generalized t.f. which is instrumental in solving Problem 1 for the class of the chained systems. First, we need to recall the definition of a graded basis of $\mathfrak{g}$ (see [18]). This definition is similar to the one of a *basis adapted to the control filtration* [9, 25]; a complementary requirement is that some elements of the basis be expressed as Lie brackets of other elements of the basis.

DEFINITION 3. *Let* $X_1, \ldots, X_m \in \mathfrak{g}$ *denote independent v.f. such that* $\mathrm{Lie}(X_1, \ldots, X_m) = \mathfrak{g}$. *Let* $\mathfrak{u} = \mathrm{span}\{X_1, \ldots, X_m\}$, *and define inductively, for* $k = 2, \ldots, K$, $\mathfrak{u}^k = \mathfrak{u}^{k-1} + [\mathfrak{u}, \mathfrak{u}^{k-1}]$ *with* $K = \min\{k : \mathfrak{u}^k = \mathfrak{g}\}$. *A* graded basis *of* $\mathfrak{g}$ *associated with* $X_1, \ldots, X_m$ *is an ordered basis* $\{X_1, \ldots, X_n\}$ *of* $\mathfrak{g}$ *associated with two mappings* $\lambda, \rho : \{m+1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$ *such that*

1. *for any* $k \in \{1, \ldots, K\}$, $\mathfrak{u}^k = \mathrm{span}\{X_1, X_2, \ldots, X_{\dim \mathfrak{u}^k}\}$;
2. *for* $k \geq 2$ *and* $\dim \mathfrak{u}^{k-1} < i \leq \dim \mathfrak{u}^k$, $X_i = [X_{\lambda(i)}, X_{\rho(i)}]$ *with* $X_{\lambda(i)} \in \mathfrak{u}^a$, $X_{\rho(i)} \in \mathfrak{u}^b$, *and* $a + b = k$.

With any graded basis of $\mathfrak{g}$, one can associate a *weight-vector* $(r_1, \ldots, r_n)$ defined by

$$r_i = k \Longleftrightarrow X_i \in \mathfrak{u}^k \setminus \mathfrak{u}^{k-1} \Longleftrightarrow \dim \mathfrak{u}^{k-1} < i \leq \dim \mathfrak{u}^k.$$

Note that $1 = r_1 \leq r_2 \leq \cdots \leq r_n = K$, and, from Definition 3, $\forall i > m$, $r_i = r_{\lambda(i)} + r_{\rho(i)}$.

With $\{X_1, \ldots, X_n\}$ any graded basis of $\mathfrak{g}$, let us define $f \in \mathcal{C}^\infty(\mathbb{T}^{n-m} \times \mathbb{T}^{n-m}; G)$ by

$$\forall \sigma = (\theta, \beta) \in \mathbb{T}^{n-m} \times \mathbb{T}^{n-m}, \quad f(\sigma) = f_n(\sigma_n) \cdots f_{m+1}(\sigma_{m+1}), \tag{24}$$

with $f_j : \mathbb{T} \times \mathbb{T} \longrightarrow G$ defined by

$$\forall \sigma_j = (\theta_j, \beta_j), \quad f_j(\sigma_j) = \exp\left(\alpha_j(\sigma_j) X_j\right) \exp\left(\alpha_{j,\lambda}(\sigma_j) X_{\lambda(j)} + \alpha_{j,\rho}(\sigma_j) X_{\rho(j)}\right), \tag{25}$$

where

$$\alpha_{j,\lambda}(\sigma_j) = \varepsilon_j^{r_{\lambda(j)}} (\sin(\theta_j + \beta_j) - \sin \beta_j), \qquad \alpha_{j,\rho}(\sigma_j) = \varepsilon_j^{r_{\rho(j)}} (\cos(\theta_j + \beta_j) - \cos \beta_j),$$

$$\alpha_j(\sigma_j) = \frac{\varepsilon_j^{r_j}}{2} \sin \theta_j, \tag{26}$$

and the ε_j's are positive real numbers. This function obviously satisfies (12). As for the transversality condition (11), we have the following result.

PROPOSITION 2. *Let* $X_1, \ldots, X_m$ *denote independent v.f. on a Lie group* G *of dimension* n. *Assume that* $\mathrm{Lie}(X_1, \ldots, X_m) = \mathfrak{g}$. *Let* $f \in \mathcal{C}^\infty(\mathbb{T}^{n-m} \times \mathbb{T}^{n-m}; G)$ *be defined by* (24), (25), (26), *with* $\{X_1, \ldots, X_n\}$ *a graded basis of* $\mathfrak{g}$. *Then there exist real positive numbers* $\eta_{m+1}, \ldots, \eta_n$ *and* ε_0 *such that, for* $(\varepsilon_{m+1}, \ldots, \varepsilon_n) = \varepsilon(\eta_{m+1}, \ldots, \eta_n)$ *with* $\varepsilon \in (0, \varepsilon_0)$, f *satisfies* (11). *More precisely, the* η_k*'s can be defined recursively by choosing any* $\eta_{m+1} > 0$ *and for* $k = m+2, \ldots, n$, *choosing* η_k *large enough w.r.t.* $\eta_{m+1}, \ldots, \eta_{k-1}$.

3. Asymptotic stabilization of chained systems. A solution to Problem 1 is provided in the case where $G = \mathbb{R}^n$, $m = 2$, and the control v.f. X_1, X_2 are defined by

$$X_1(x) = (1, 0, x_2, \ldots, x_{n-1})', \quad X_2 = (0, 1, 0, \ldots, 0)' \tag{27}$$

with $g = x = (x_1, \ldots, x_n)'$ and $e = 0$.

The v.f. X_1 and X_2 defined by (27) are left-invariant w.r.t. the group operation

$$(xy)_i = \begin{cases} x_i + y_i & \text{if } i = 1, 2, \\ x_i + y_i + \displaystyle\sum_{j=2}^{i-1} \frac{y_1^{i-j}}{(i-j)!} x_j & \text{otherwise,} \end{cases}$$

with $x, y \in \mathbb{R}^n$ (see [24], for instance). Furthermore, $\mathrm{Lie}(X_1, X_2) = \mathfrak{g}$, so that chained systems (with $P \equiv 0$) are controllable, and the v.f.

$$X_1, X_2, X_k \triangleq [X_1, X_{k-1}] \quad (k = 3, \ldots, n) \tag{28}$$

define a graded basis. The associated weight-vector r is given by

$$r_1 = r_2 = 1, \qquad r_k = k - 1 \quad (k = 3, \ldots, n). \tag{29}$$

Since the underlying Lie group G is $\mathbb{R}^n$, a simple example of v.f. which globally exponentially stabilizes the origin of $\dot{z} = Z(z)$ on $\mathbb{R}^n$ is defined by $Z(z) = Kz$, with K denoting any $n \times n$ Hurwitz-stable matrix. The main result is stated next.

THEOREM 1. *When $m = 2$ and the v.f. X_1, X_2 are given by* (27)*, there exist real positive numbers $\eta_{m+1}, \ldots, \eta_n$ such that a solution to Problem* 1 *is the triplet (f, w, V) consisting of*

1. *the candidate generalized t.f. defined by* (24)–(26) *with $(\varepsilon_{m+1}, \ldots, \varepsilon_n) = \varepsilon(\eta_{m+1}, \ldots, \eta_n)$ and $\varepsilon > 0$ chosen small enough so that f ranges in $\mathcal{U}_{\mathbb{R}^n}(0)$,*
2. *the function $w \in \mathcal{C}^1((-\pi, \pi)^{n-2}; \mathbb{R}^{n-2})$ defined by*

$$w_i(\theta_i) = \frac{1}{\eta_i^{i-2}} |\theta_i|^{(i-3)} \theta_i \qquad (i = 3, \ldots, n), \tag{30}$$

3. *the function $V \in \mathcal{C}^1((-\pi, \pi)^{n-2}; \mathbb{R})$ defined by*

$$V(\theta) \triangleq \sum_{i=3}^{n} \eta_i^{i-3/2} |\theta_i|^{n+2-i} \quad \text{with} \quad V_{\max} = \min_{i=3,\ldots,n} \{\eta_i^{i-3/2} \pi^{n+2-i}\}.$$

Remark 1. The proof of this theorem in section 5.3 involves a recursive procedure for the determination of the numbers $\eta_{m+1}, \ldots, \eta_n$, which is similar to the one indicated in Proposition 2.

Remark 2. The solution to Problem 1 given in Theorem 1 applies also to a unicycle-like mobile robot without having to transform its kinematic equations into the chain form—the only restriction is that ε must be smaller than some finite upper bound $\varepsilon_0 > 0$, whatever $\mathcal{U}_G(e)$, whereas, in the case of a chained system, $\varepsilon_0 = +\infty$. One only has to check that the proof of Theorem 1 works as well in this case with $n = 3$, $G = \mathbb{R}^2 \times S^1$, $g = (x, y, \alpha)'$, and the system's control v.f. defined by

$$X_1(g) = (\cos\alpha, \sin\alpha, 0)', \quad X_2(g) = (0, 0, 1)'. \tag{31}$$

These v.f. are left-invariant w.r.t. the group operation

$$g_1 g_2 = \begin{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + R(\alpha_1) \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \\ \alpha_1 + \alpha_2 \end{pmatrix}$$

with $g_i = (x_i, y_i, \alpha_i)'$ and $R(\alpha_1)$ the rotation matrix of angle α_1. Also, $\mathrm{Lie}(X_1, X_2) = \mathfrak{g}$ and $\{X_1, X_2, X_3 = [X_1, X_2]\}$ constitutes a graded basis of $\mathfrak{g}$ with weight vector $(r_1 = r_2 = 1, r_3 = 2)$.

Let us comment on the rate of convergence provided by a feedback control derived according to Proposition 1 and Theorem 1 when $P \equiv 0$. This will be the starting point of a more general discussion about what the t.f. approach can offer in comparison with other control design methods, its limitations and assets. Assuming that the v.f. Z used in the expression of (u, v) is chosen so as to stabilize the origin of $\dot{z} = Z(z)$ exponentially, the rate of convergence of $g(t)$ to e coincides with the slower rate of convergence of $\theta(t)$ to zero on the zero-dynamics. This latter rate is itself given by the rate of convergence of $V(\theta(t))$ to zero, and is thus related to the integer l in relation (21). From (47) in the proof of Theorem 1, we have $l = \frac{n+1}{2}$, and one deduces that $V(\theta(t))$ tends to zero as quickly as $t^{-\frac{2}{n-1}}$. In fact, a complementary analysis would show that $V(\theta(t))$ cannot tend to zero faster. Now, since $k_1|\theta|^{n-1} \leq V(\theta) \leq k_2|\theta|^2$ in the neighborhood of $\theta = 0$, this in turn implies that $|\theta(t)|$ may (and will usually) not tend to zero faster than $t^{-\frac{2}{(n-1)^2}}$. The same rate holds for the convergence of $|g(t)|$ towards e. This polynomial rate of convergence is similar to the one which can be obtained by applying a smooth time-periodic stabilizer to $\mathcal{S}(g)$. Therefore, one can conclude that, as far as asymptotic stabilization is concerned, no clear advantage results from designing a stabilizer with the t.f. approach. In the authors' opinion this conclusion is correct, but it conveys only a partial picture of the properties granted by the approach. Indeed, the primary feature of such a controller, which motivated the development of the t.f. approach in the first place, is the capacity of ensuring practical stabilization, with easily tunable arbitrary small ultimate bound of the state error, independently of the "perturbation" P acting on the system. As shown in [18], this allows, for example, the tracking of *any* trajectory in the state space (it does not have to be a solution to the system's equations) with arbitrarily good precision, in the sense that tracking errors are ultimately bounded by a prespecified (nonzero, but otherwise as small as desired) threshold. To our knowledge, no other controller proposed so far in the literature has this capacity. Our motivation for the present paper was to show that such a controller can also be endowed with the extra property of ensuring asymptotic point-stabilization when such a feature is desirable. This is achieved via the concept of a generalized t.f. depending upon two sets of variables whose time-derivatives are used as extra control inputs. Transversality is maintained with respect to the first set θ, while the second set β is used to enforce some type of "phase-tuning," which allows us to reduce the size of the t.f. when the perturbation P vanishes.

4. Simulation results. The control law proposed in the previous section has been tested by simulation on the four-dimensional (4d) chained system. The following parameters for the definition of the transverse function have been used: $\varepsilon = 0.2$, $\eta_3 = 1, \eta_4 = 8$. The v.f. $Z(z)$ in (15) has been chosen as $Z(z) = -0.3z$. Finally, the $\mathcal{K}_\infty$-function k in (23) has been defined by $k(s) = 10V_{\max}s$, with $V_{\max}$ as specified in Theorem 1. The initial condition for the simulation was $x(0) = (0, 0, 0, 10)'$, and $\sigma(0) = 0$. Figure 1 displays the state variables versus time. As discussed in the previous section, the convergence rate to zero is slow. For comparison, Figure 2 displays the same variables when no attempt is made to achieve convergence to zero, i.e., with $w = 0$ and $\beta = 0$ in the control law defined by (15). In this case $\theta(t)$ exponentially converges to some $\theta_{\lim} \in \mathbb{T}^{n-m}$, and $x(t)$ exponentially converges to $f(\theta_{\lim}, 0)$. Note that the solution to Problem 1 given by Theorem 1 is only one of its kind, and that much room is left for improving the proposed stabilization method.

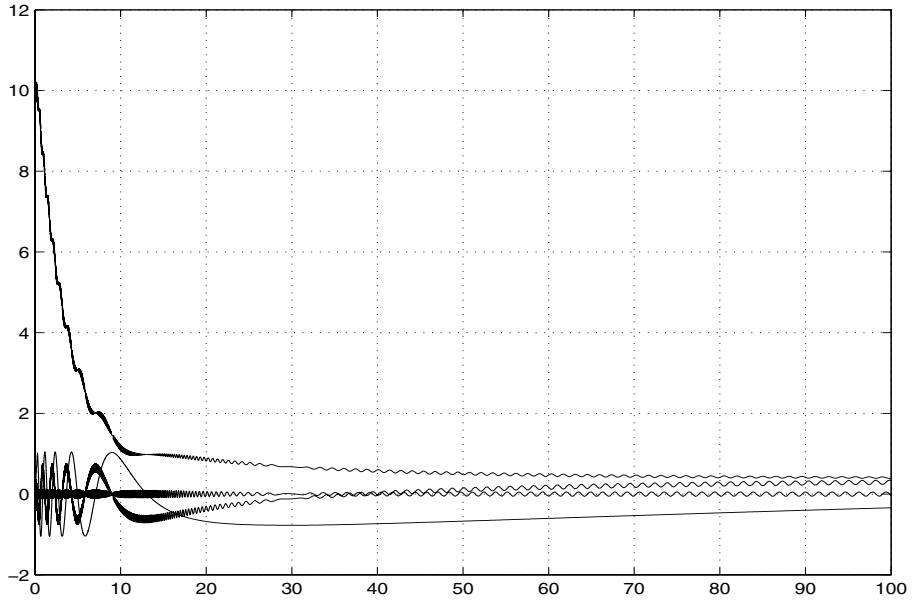


FIG. 1. *State variables for the* 4*d chained system, asymptotic stabilization.*

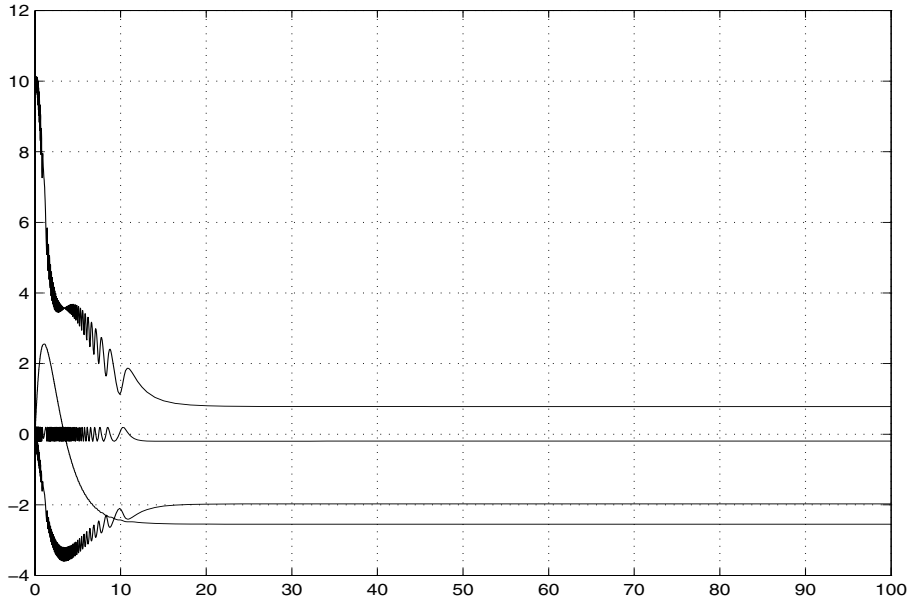


FIG. 2. *State variables for the* 4*d chained system, practical stabilization.*

5. Proofs.

5.1. Proof of Proposition 1. Let us first recall, as shown in section 2.2, that the control $(u, v)(g, \sigma, t)$ defined by (15) yields

$$\dot{z} = Z(z), \tag{32}$$

with $z = fg^{-1}$ and Z chosen so as to ensure the asymptotic stability of e for the above system. Now, applying the one-forms d_i $(i = m+1, \ldots, n)$ to each side of the equality (15) yields (compare with (18))

$$\dot{\theta} = -A^{-1}(\sigma)B(\sigma)w(\theta) + A^{-1}(\sigma) \sum_{i=m+1}^{n} \langle d_i(f), dl_z(g)P(g,t) + dr_g(z)Z(z)\rangle e_i, \tag{33}$$

where we have used the same notation as in section 2.2. Using the fact that $z = fg^{-1}$, we rewrite this equation as

$$\begin{aligned} \dot{\theta} = -A^{-1}(\sigma)B(\sigma)w(\theta) + A^{-1}(\sigma) \sum_{i=m+1}^{n} \langle d_i(f), dl_z(z^{-1}f)P(z^{-1}f,t) \\ + dr_{z^{-1}f}(z)Z(z)\rangle e_i. \end{aligned} \tag{34}$$

By using (23), this in turn implies that, along a solution of the controlled system,

$$\frac{d}{dt}V(\theta) = k\left(\frac{1}{V_{\max} - V(\theta)}\right) L_F(\pi^* V)(\sigma) + Q(g,\sigma,t),$$

with

$$F(\sigma) \triangleq -A^{-1}(\sigma)B(\sigma)w^\star(\theta),$$

$$Q(g,\sigma,t) \triangleq \frac{\partial V}{\partial \theta}(\theta)A^{-1}(\sigma) \sum_{i=m+1}^{n} \langle d_i(f), dl_z(z^{-1}f)P(z^{-1}f,t) + dr_{z^{-1}f}(z)Z(z)\rangle e_i.$$

Therefore, in view of (21),

$$\frac{d}{dt}V(\theta) \le -\xi(V(\theta)) + Q(g,\sigma,t), \qquad \xi(V) \triangleq k\left(\frac{1}{V_{\max} - V}\right)\gamma V^l. \tag{35}$$

Let us show that $\theta(t)$ cannot leave the set $V^{-1}([0, V_{\max}))$. We first remark that, on any time-interval $[0,T)$ such that $\theta(t)$ stays in this set, there exists a constant M_T, independent of the trajectory $\theta(.)$, such that $|Q(g(t),\sigma(t),t)| \le M_T$ because (i) by assumption, $\frac{\partial V}{\partial \theta}$ is bounded on $\mathcal{U}_{\mathbb{T}^{n-m}}(0) \supset V^{-1}([0, V_{\max}))$; (ii) z, and subsequently z^{-1}, are bounded due to the asymptotic stability of e for the system (32); (iii) P is continuous. Since, by (35), ξ is a bijective increasing function from $[0, V_{\max})$ to $[0, +\infty)$, we deduce from (35) that on any such interval $[0,T)$

$$V(\theta(t)) \le \max\{\xi^{-1}(M_T), V(\theta(0))\} < V_{\max}.$$

This implies that $V(\theta(t))$ cannot tend to $V_{\max}$ in finite time, so that $\theta(t)$ remains in the set $V^{-1}([0, V_{\max}))$. This in turn implies that the control law is well defined along any trajectory of the closed-loop system with initial conditions $(g(0), \theta(0), \beta(0))$ such that $z(0) = f(\sigma(0))g(0)^{-1}$ is in the stability domain of e for the system $\dot{z} = Z(z)$ and $\theta(0) \in V^{-1}([0, V_{\max}))$, and that such a trajectory is complete. Point 1 of Proposition 1 then follows directly from the asymptotic stability of $z = e$, as ensured by (32), and the invariance of the set $V^{-1}([0, V_{\max}))$ for the variable θ.

As for point 2, which assumes that $P \equiv 0$, it is sufficient to consider trajectories with initial conditions $(z(0), \theta(0))$ in a small neighborhood of the point $(e, 0)$. From the definition of Q, the asymptotic stability of $z = e$, combined with the invariance

of the set $V^{-1}([0, V_{\max}))$ for the variable θ and the fact that V has bounded partial derivatives on this set, yields the existence of a $\mathcal{K}$-function h_z such that

$$\forall t \geq 0, \qquad |Q(g(t), \sigma(t), t)| \leq h_z(d_G(z(0), e)).$$

Therefore, in view of (35),

$$\forall t \geq 0, \qquad \frac{d}{dt} V(\theta(t)) \leq -k_V V(\theta(t))^l + h_z(d_G(z(0), e)),$$

with $k_V = k(\frac{1}{V_{\max}})\gamma(> 0)$. This in turn implies

$$\forall t \geq 0, \qquad V(\theta(t)) \leq \left(\frac{h_z(d_G(z(0), e))}{k_V} \right)^{1/l} + V(\theta(0)). \tag{36}$$

In view of (36) and property 1 in Problem 1,

$$\forall t \geq 0, \qquad |\theta(t)| \leq h_{V_m}^{-1} \left(\left(\frac{h_z(d_G(z(0), e))}{k_V} \right)^{1/l} + h_{V_M}(|\theta(0)|) \right).$$

This relation, combined with the asymptotic stability of $z = e$, implies the stability of the set $\{e\} \times \{0\} \times \mathbb{T}^{n-m}$ for $\mathcal{S}(g, \sigma)$. The convergence of the closed-loop trajectories to this set simply results from the convergence of $Q(g(t), \sigma(t), t)$ to zero when $z(t)$ tends to e, since $Z(z(t))$ then converges to zero. In view of (35), this yields the convergence of $V(\theta(t))$ to zero.

When it is assumed only that $P(g, t)$ tends to zero when t tends to infinity—uniformly w.r.t. g in compact sets—the term $Q(g(t), \sigma(t), t)$ in (35) still converges to zero, because the asymptotic stability of $z = e$ implies that $Z(z(t))$ converges to $Z(e) = 0$. Hence, the convergence of $V(\theta(t))$ to zero is still ensured, so that $\theta(t)$ tends to zero and $f(\sigma(t))$ tends to e (using the property (12) of a generalized t.f.). Therefore $(g, \theta)(t)$ tends to the point $(e, 0)$, as announced in point 3 of the proposition.

5.2. Proof of Proposition 2. The following notation is used in the forthcoming proofs. With v denoting a smooth function of the real variables x and y—possibly vector-valued—we write $v = \mathrm{o}(x^k)$ (resp., $v = \mathrm{O}(x^k)$) if $(|v(x, y)|/|x|^k) \longrightarrow 0$ as $|x| \longrightarrow 0$ (resp., if $(|v(x, y)|/|x|^k) \leq K < \infty$ in some neighborhood of $x = 0$) uniformly w.r.t. the y variable which takes values in a compact set. Finally, for indexed variables x_i with $i = k, \ldots, n$, we define the set of indexed vectors $\{\bar{x}_p\}_{p \in \{k, \ldots, n\}}$ by setting $\bar{x}_p = (x_k, \ldots, x_p)$.

Remark 3. Various results in the paper, starting with Proposition 2, refer to t.f. which depend on a vector of parameters $\varepsilon \in \mathbb{R}^{n-m}$, used as a means to monitor the "size" of the functions. Relations (24)–(26) define such a family of transverse functions. A member of this family could have been denoted as f_ε or $f(\varepsilon, .)$ in order to point out the functional dependence upon ε explicitly. However, for the sake of simplifying the (already cumbersome) notation used in the paper, we have chosen to systematically omit the argument ε when referring to t.f. It is nonetheless important to keep this dependence in mind when reading the forthcoming proofs. In particular, several functions associated with an arbitrary member of the family of t.f. defined by (24)–(26) will be introduced in Lemmas 1 and 2. Each of them is thus also a function of ε. For the sake of keeping the notation coherent throughout the paper, the index is again omitted when referring to such a function.

The proof of Proposition 2 consists of three steps summarized in the form of three lemmas, which are proved in the appendix.

LEMMA 1. *Assume that the assumptions of Proposition* 2 *are satisfied. Then, for each* $j \in \{m+1, \dots, n\}$ *and* $i \in \{1, \dots, n\}$, *there exist analytic functions* $v_{i,j}$ *and* $w_{i,j}$ *of* $\varepsilon_j \in \mathbb{R}$ *and* $\sigma_j \in \mathbb{T} \times \mathbb{T}$ *such that*

$$\frac{\partial f_j}{\partial \theta_j}(\sigma_j) = \sum_{i=1}^{n} v_{i,j}(\sigma_j) X_i(f_j(\sigma_j)), \quad \frac{\partial f_j}{\partial \beta_j}(\sigma_j) = \sum_{i=1}^{n} w_{i,j}(\sigma_j) X_i(f_j(\sigma_j)), \tag{37}$$

with

$$v_{i,j} = \begin{cases} \mathrm{O}(\varepsilon_j^{r_i}) & \forall i, \\ \mathrm{o}(\varepsilon_j^{r_i}) & \text{if } i < j \text{ and } r_i = r_j, \\ \dfrac{\varepsilon_j^{r_j}}{2} + \mathrm{o}(\varepsilon_j^{r_j}) & \text{if } i = j, \end{cases} \tag{38}$$

and

$$w_{i,j} = \begin{cases} \mathrm{O}(\varepsilon_j^{r_i})\mathrm{O}(\theta_j) & \forall i, \\ \varepsilon_j^{r_j}(1 - \cos\theta_j) + \mathrm{o}(\varepsilon_j^{r_j})\mathrm{o}(\theta_j^2) & \text{if } i = j. \end{cases} \tag{39}$$

In the following lemma, $\mathrm{O}(\bar{\varepsilon}_m)$ formally appears when setting $j = m+1$ in $\mathrm{O}(\bar{\varepsilon}_{j-1})$, although ε_m has not been defined previously. The lemma's statement is nonetheless valid, provided that $\mathrm{O}(\bar{\varepsilon}_m)$ is identified with the null function.

LEMMA 2. *Assume that the assumptions of Proposition* 2 *are satisfied. Then, for each* $j \in \{m+1, \dots, n\}$ *and* $i \in \{1, \dots, n\}$, *there exist analytic functions* $a_{i,j}$ *and* $b_{i,j}$ *of* $\bar{\varepsilon}_j \in \mathbb{R}^{j-m}$ *and* $\sigma \in \mathbb{T}^{n-m} \times \mathbb{T}^{n-m}$ *such that*

$$\frac{\partial f}{\partial \theta_j}(\sigma) = \sum_{i=1}^{n} a_{i,j}(\sigma) X_i(f(\sigma)), \quad \frac{\partial f}{\partial \beta_j}(\sigma) = \sum_{i=1}^{n} b_{i,j}(\sigma) X_i(f(\sigma)), \tag{40}$$

with

$$a_{i,j} = \begin{cases} \mathrm{O}(\bar{\varepsilon}_j^{r_i}) & \forall i, \\ \mathrm{O}(\bar{\varepsilon}_{j-1})\mathrm{O}(\bar{\varepsilon}_j^{r_i - 1}) + \mathrm{o}(\bar{\varepsilon}_j^{r_i}) & \text{if } i < j \text{ and } r_i = r_j, \\ \dfrac{\varepsilon_j^{r_j}}{2} + \mathrm{O}(\bar{\varepsilon}_{j-1})\mathrm{O}(\bar{\varepsilon}_j^{r_j - 1}) + \mathrm{o}(\bar{\varepsilon}_j^{r_j}) & \text{if } i = j, \end{cases} \tag{41}$$

and

$$b_{i,j} = \begin{cases} \mathrm{O}(\bar{\varepsilon}_j^{r_i})\mathrm{O}(\bar{\theta}_j) & \forall i, \\ \varepsilon_j^{r_j}(1 - \cos\theta_j) + \mathrm{O}(\bar{\varepsilon}_{j-1})\mathrm{O}(\bar{\varepsilon}_j^{r_j - 1})\mathrm{O}(\theta_j)\mathrm{O}(\bar{\theta}_{j-1}) + \mathrm{o}(\bar{\varepsilon}_j^{r_j})\mathrm{o}(\bar{\theta}_j^2) & \text{if } i = j. \end{cases} \tag{42}$$

Note that, if all O and o terms in the above expressions were equal to zero, then the transversality property would simply follow from (40)–(41) and the fact that $\{X_1, \dots, X_n\}$ is a basis of $\mathfrak{g}$. Although this is not the case, one can show that these terms can be neglected, provided that the ε_j's are adequately chosen.

LEMMA 3. *Assume that the assumptions of Proposition* 2 *are satisfied. Then there exist* $n - m$ *numbers* $\eta_{m+1}, \dots, \eta_n$ *and* $\varepsilon_0 > 0$ *such that choosing*

$$(\varepsilon_{m+1}, \dots, \varepsilon_n) = \varepsilon(\eta_{m+1}, \dots, \eta_n)$$

with $\varepsilon \in (0, \varepsilon_0)$ *yields*

$$\forall \sigma \in \mathbb{T}^{n-m} \times \mathbb{T}^{n-m}, \quad \mathrm{Det}\, A(\sigma) \neq 0 \quad \textit{with } A(\sigma) = (a_{i,j}(\sigma))_{i,j=m+1,\ldots,n}. \tag{43}$$

5.3. Proof of Theorem 1. One easily verifies that for any positive real numbers $\eta_3, \ldots, \eta_n$ the functions w and V satisfy (ii) and (iii) of Problem 1 with $\mathcal{U}_{\mathbb{T}^{n-m}}(0) = (-\pi, \pi)^{n-m}$. It is also clear that property 1 of Problem 1 is verified. We show below that, for an adequate choice of positive $\eta_3, \ldots, \eta_n$, properties (i) and 2 are also satisfied. The proof relies on the following lemma, proved in the appendix, which points out complementary properties of the functions $a_{i,j}$ and $b_{i,j}$ in Lemma 2 in the case of the chained systems.

LEMMA 4. *In the case of chained systems, the functions* $a_{i,j}$ *and* $b_{i,j}$ $(i = 1, \ldots, n,$ $j = 3, \ldots, n)$ *are homogeneous polynomials of degree* r_i *in* $\varepsilon_3, \ldots, \varepsilon_j$. *Furthermore,*

$$a_{i,j} = \mathrm{O}(\bar{\theta}_j^{r_i - r_j}), \quad b_{i,j} = \mathrm{O}(\bar{\theta}_j^{\max(1, r_i - r_j + 2)}). \tag{44}$$

Let

$$A_p(\sigma) \triangleq (a_{i,j}(\sigma))_{i,j=3,\ldots,p}, \quad B_p(\sigma) \triangleq (b_{i,j}(\sigma))_{i,j=3,\ldots,p}, \tag{45}$$

and note that $A_n = A$ and $B_n = B$, with A and B defined by (19).

PROPOSITION 3. *For any* $p = 3, \ldots, n$, *there exists a set of positive numbers* $\{\eta_3, \ldots, \eta_p\}$ *such that setting* $(\varepsilon_3, \ldots, \varepsilon_p) = \varepsilon(\eta_3, \ldots, \eta_p)$ *with* $\varepsilon > 0$ *implies that*

(i) *the matrix* $A_p(\sigma)$ *is invertible for any* σ, *and*

$$\forall i, j = 3, \ldots, p, \quad (A_p^{-1}(\sigma))_{i,j} = \mathrm{O}(\bar{\theta}_p^{r_i - r_j}); \tag{46}$$

(ii) *the following is true:*

$$V_p(\bar{\theta}_p) < V_{p,max} \Longrightarrow L_{F_p} V_p(\sigma) \leq -\alpha_p |\bar{\theta}_p|^{n+1} \quad (\alpha_p > 0) \tag{47}$$

with

$$V_p(\bar{\theta}_p) \triangleq \sum_{i=3}^{p} \eta_i^{i-3/2} |\theta_i|^{n+2-i}, \qquad V_{p,max} = \min_{i=3,\ldots,p} \{\eta_i^{i-3/2} \pi^{n+2-i}\} \tag{48}$$

and

$$F_p(\sigma) = -A_p^{-1}(\sigma) B_p(\sigma) \bar{w}_p(\bar{\theta}_p). \tag{49}$$

With $p = n$, property (i) of this proposition implies that the function f satisfies the transversality condition (11). Since (12) is trivially verified from (24), (25), (26), property (i) of Problem 1 follows. As for property 2 in Problem 1, it is true by (ii) in the above proposition. Note that, to be fully precise, in (47) one should write $L_{F_p}(\pi_p^* V_p)(\sigma)$ instead of $L_{F_p} V_p(\sigma)$, with $\pi_p(\sigma) = \bar{\theta}_p$ (compare with (21)). For the sake of simplifying the notation in the forthcoming proof, we have chosen to keep this small abuse of notation.

Proof of Proposition 3. We proceed by induction. For $p = 3$, it follows from (29) and Lemma 2 that

$$a_{3,3}(\sigma) = \frac{\varepsilon_3^2}{2} + \mathrm{o}(\varepsilon_3^2), \quad b_{3,3}(\sigma) = \varepsilon_3^2 (1 - \cos\theta_3) + \mathrm{o}(\varepsilon_3^2)\mathrm{o}({\theta_3}^2). \tag{50}$$

Lemma 4 implies that the o terms in the above equation are identically equal to zero, since $a_{3,3}$ and $b_{3,3}$ are homogeneous polynomials of degree $r_3 = 2$ in ε_3. Therefore, $a_{3,3}(\sigma) > 0$ for any $\varepsilon_3 > 0$, and the point (i) of the proposition is verified.

Take $\eta_3 = 1$. From (30), (49), (50), and the fact that the o terms in (50) are equal to zero,

$$F_3(\sigma) = -a_{3,3}^{-1}(\sigma)\left(\varepsilon_3^2(1-\cos\theta_3)\right)\theta_3 = -2(1-\cos\theta_3)\theta_3. \tag{51}$$

From (48), one easily checks that

$$L_{F_3}V_3(\sigma) = -2(n-1)(1-\cos\theta_3)|\theta_3|^{n-1}. \tag{52}$$

Since $V_3(\theta_3) = |\theta_3|^{n-1}$ and $V_{3,max} = \pi^{n-1}$, we deduce from (52) that

$$V_3(\theta_3) < \pi^{n-1} \Longrightarrow \qquad L_{F_3}V_3(\sigma) \leq -\alpha_3|\theta_3|^{n+1}$$

for some $\alpha_3 > 0$. Point (ii) of the proposition is thus verified with $\eta_3 = 1$ and $\varepsilon = \varepsilon_3 > 0$, and this concludes the proof of Proposition 3 for $p = 3$.

Let us now assume that points (i) and (ii) of the proposition hold true up to some $p < n$, with $\bar{\varepsilon}_p = \bar{\eta}_p$, and show that they are also true for $p+1$, with $\bar{\varepsilon}_{p+1} = \bar{\eta}_{p+1}$. This will in turn imply that they are true when $\bar{\varepsilon}_{p+1} = \varepsilon\bar{\eta}_{p+1}$ with $\varepsilon > 0$, thanks to the homogeneity properties of the $a_{i,j}$'s and $b_{i,j}$'s—see Lemma 4. Indeed, when $\bar{\eta}_{p+1}$ is multiplied by ε, then A_{p+1} and B_{p+1} are just premultiplied by the diagonal matrix $\mathrm{Diag}(\varepsilon^{r_3}, \ldots, \varepsilon^{r_{p+1}})$, thus leaving $F_{p+1}(\sigma)$ and the subsequent analysis unchanged.

From (45), A_{p+1} and B_{p+1} can be written as

$$A_{p+1} = \begin{pmatrix} A_p & a_{*,p+1} \\ a_{p+1,*} & a_{p+1,p+1} \end{pmatrix}, \quad B_{p+1} = \begin{pmatrix} B_p & b_{*,p+1} \\ b_{p+1,*} & b_{p+1,p+1} \end{pmatrix}, \tag{53}$$

with the star denoting the indexes from 1 to p, i.e., $a_{p+1,*} = (a_{p+1,1}, \ldots, a_{p+1,p})$ and $a_{*,p+1} = (a_{1,p+1}, \ldots, a_{p,p+1})'$. Let us recall (see, e.g., [29, Chap. 2]) that if A_{11} and A_{22} are square matrices with A_{11} nonsingular, the matrix

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

is invertible if and only if the Schur complement $S \triangleq A_{22} - A_{21}A_{11}^{-1}A_{12}$ of A_{11} in A is invertible. Then

$$A^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{pmatrix}. \tag{54}$$

From (53), the Schur complement of A_p in A_{p+1} is $S = a_{p+1,p+1} - a_{p+1,*}A_p^{-1}a_{*,p+1}$, and, in view of (29) and Lemmas 2 and 4,

$$S = \frac{\varepsilon_{p+1}^p}{2} + q^{p-1}(\varepsilon_{p+1}), \tag{55}$$

with $q^{p-1}(\varepsilon_{p+1})$ a polynomial of degree $p-1$ in ε_{p+1}. (Note, from the domain definition of the functions $a_{i,j}$ in Lemma 2, that the term $a_{p+1,*}A_p^{-1}a_{*,p+1}$ depends on ε_{p+1} only through $a_{*,p+1}$ so that, by Lemma 4, it is a polynomial of degree $r_p = p-1$ in ε_{p+1}.) This implies that S, and thus A_{p+1}, are invertible for ε_{p+1} large enough. In order to prove (i), it remains to show that (46) holds true for $p+1$. Since (46) is true for p,

for any $p = 3, \dots, n-1$ and $\varepsilon_3, \dots, \varepsilon_{p+1}$ such that A_p and A_{p+1} are invertible, let us use (54) to decompose A_{p+1}^{-1} as follows:

$$A_{p+1}^{-1} = \begin{pmatrix} A_p^{-1} & 0 \\ 0 & S^{-1} \end{pmatrix} + \begin{pmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{21} & 0 \end{pmatrix} \triangleq \begin{pmatrix} A_p^{-1} & 0 \\ 0 & S^{-1} \end{pmatrix} + \Xi, \tag{56}$$

with

$$\begin{aligned} \Xi_{11} &= A_p^{-1} a_{*,p+1} S^{-1} a_{p+1,*} A_p^{-1}, \\ \Xi_{12} &= -A_p^{-1} a_{*,p+1} S^{-1}, \\ \Xi_{21} &= -S^{-1} a_{p+1,*} A_p^{-1}. \end{aligned}$$

By Lemma 4, $a_{i,j}$ is a homogeneous polynomial in $\varepsilon_3, \dots, \varepsilon_j$ of degree r_i and satisfies (44). Therefore, there exists a constant C such that

$$\forall \sigma, \quad |a_{i,j}(\sigma)| \leq C |\bar{\varepsilon}_j|^{r_i} |\bar{\theta}_j|^{r_i - r_j}. \tag{57}$$

Then, from Lemma 4, relations (46), (55), and (57), and using the fact that neither A_p nor $a_{p+1,*}$ depend on ε_{p+1}, one infers that

$$\varepsilon_{p+1} \geq 1 \Longrightarrow \begin{cases} |\xi_{i,j}| \leq C \varepsilon_{p+1}^{-1} |\bar{\theta}_{p+1}|^{r_i - r_j} & \text{for } i \leq p, \\ |\xi_{i,j}| \leq C \varepsilon_{p+1}^{-p} |\bar{\theta}_{p+1}|^{r_i - r_j} & \text{for } i = p+1, \end{cases} \tag{58}$$

with $\Xi = \{\xi_{i,j}\}_{i,j=3,\dots,p+1}$. The fact that (46) holds true for $p+1$, provided that it is true up to p, directly follows from (56) and (58). Note that a relation similar to (57) holds for $b_{i,j}$, i.e.,

$$\forall \sigma, \quad |b_{i,j}(\sigma)| \leq C |\bar{\varepsilon}_j|^{r_i} |\bar{\theta}_j|^{\max(1, r_i - r_j + 2)}. \tag{59}$$

This relation will be used later on.

Let us now examine the case of (ii). Throughout the rest of the proof, we assume that $\varepsilon_{p+1} \geq 1$. From (53) and (56),

$$A_{p+1}^{-1} B_{p+1} \bar{w}_{p+1} = \begin{pmatrix} A_p^{-1} B_p \bar{w}_p \\ S^{-1} b_{p+1,p+1} w_{p+1} \end{pmatrix} + D_2, \tag{60}$$

with

$$D_2 = \Xi B_{p+1} \bar{w}_{p+1} + \begin{pmatrix} A_p^{-1} b_{*,p+1} w_{p+1} \\ S^{-1} b_{p+1,*} \bar{w}_p \end{pmatrix}. \tag{61}$$

From (42) and Lemma 4, it is not difficult to deduce that

$$b_{p+1,p+1} = \varepsilon_{p+1}^p (1 - \cos \theta_{p+1}) + R_b, \tag{62}$$

with

$$|R_b| \leq C \varepsilon_{p+1}^{p-1} |\bar{\theta}_{p+1}|^2 \tag{63}$$

for some constant C—recall that $\varepsilon_{p+1} \geq 1$. From the definition of $F_{p+1}(\sigma)$ in Proposition 3 and from relations (60) and (62),

$$F_{p+1}(\sigma) = \underbrace{\begin{pmatrix} F_p(\sigma) \\ -S^{-1} \varepsilon_{p+1}^p (1 - \cos \theta_{p+1}) w_{p+1}(\theta) \end{pmatrix}}_{D_0} - \underbrace{\begin{pmatrix} 0 \\ S^{-1} R_b w_{p+1}(\theta) \end{pmatrix}}_{D_1} - D_2. \tag{64}$$

We claim that the Lie derivative $L_{D_0}V_{p+1}$ of V_{p+1} along D_0 defined by (64) satisfies

$$L_{D_0}V_{p+1}(\sigma) \leq -\alpha_p|\bar{\theta}_p|^{n+1} - \alpha_1\varepsilon_{p+1}^{1/2}|\theta_{p+1}|^{n+1} \quad (\alpha_p, \alpha_1 > 0). \tag{65}$$

Indeed, by (48), $V_{p+1} = V_p + \varepsilon_{p+1}^{p-1/2}|\theta_{p+1}|^{n-p+1}$ (recall that $\bar{\varepsilon}_{p+1} = \bar{\eta}_{p+1}$), and it follows from (64) that

$$L_{D_0}V_{p+1}(\sigma) = L_{F_p}V_p(\sigma) - (n-p+1)S^{-1}\varepsilon_{p+1}^{2p-1/2}(1-\cos\theta_{p+1})w_{p+1}(\theta)\theta_{p+1}^{\{n-p\}}, \tag{66}$$

with the notation $x^{\{n\}} = |x|^{n-1}x$, also used in subsequent relations. From (47),

$$L_{F_p}V_p(\sigma) \leq -\alpha_p|\bar{\theta}_p|^{n+1}, \tag{67}$$

and, proceeding as for $a_{3,3}$, it is simple to verify, by using (30), (55), and the fact that $\varepsilon_{p+1} = \eta_{p+1} \geq 1$, that

$$-(n-p+1)S^{-1}\varepsilon_{p+1}^{2p-1/2}(1-\cos\theta_{p+1})w_{p+1}(\theta)\theta_{p+1}^{\{n-p\}} \leq -\alpha_1\varepsilon_{p+1}^{1/2}|\theta_{p+1}|^{n+1}. \tag{68}$$

Then, (65) follows from (66), (67), and (68).

From (30), (55), (63), and (64), it is straightforward to verify—by again using the condition $\varepsilon_{p+1} \geq 1$—that

$$|L_{D_1}V_{p+1}(\sigma)| \leq \alpha_2\varepsilon_{p+1}^{-1/2}|\bar{\theta}_p|^{n+1} + \alpha_2|\theta_{p+1}|^{n+1}. \tag{69}$$

Finally, we claim that

$$|L_{D_2}V_{p+1}(\sigma)| \leq \left(\frac{\alpha_p}{2} + \alpha_3\varepsilon_{p+1}^{-1/2}\right)|\bar{\theta}_p|^{n+1} + \alpha_4|\theta_{p+1}|^{n+1}. \tag{70}$$

Indeed, from (53), (56), and (61),

$$D_2 = \begin{pmatrix} (\Xi_{11}B_p + \Xi_{12}b_{p+1,*})\bar{w}_p + (\Xi_{11}b_{*,p+1} + \Xi_{12}b_{p+1,p+1})w_{p+1} + A_p^{-1}b_{*,p+1}w_{p+1} \\ \Xi_{21}B_p\bar{w}_p + \Xi_{21}b_{*,p+1}w_{p+1} + S^{-1}b_{p+1,*}\bar{w}_p \end{pmatrix}.$$

By using (29), (30), (46), (57), (58), and (59), it is tedious but not difficult to show that

$$\begin{cases} |(D_2)_i| \leq C\varepsilon_{p+1}^{-1}|\bar{\theta}_{p+1}|^i + C|\bar{\theta}_{p+1}|^{i-p+1}|\theta_{p+1}|^{p-1} & \text{for } i = 3, \ldots, p, \\ |(D_2)_{p+1}| \leq C\varepsilon_{p+1}^{-p}|\bar{\theta}_{p+1}|^{p+1}. \end{cases} \tag{71}$$

We infer from (48) and (71) that

$$|L_{D_2}V_p(\sigma)| \leq \alpha_5\varepsilon_{p+1}^{-1}|\bar{\theta}_{p+1}|^{n+1} + \alpha_6|\bar{\theta}_{p+1}|^{n-p+2}|\theta_{p+1}|^{p-1}. \tag{72}$$

By using Young's inequality, one shows that

$$\begin{aligned} \alpha_6|\bar{\theta}_{p+1}|^{n-p+2}|\theta_{p+1}|^{p-1} &\leq \frac{\alpha_p}{2}|\bar{\theta}_{p+1}|^{n+1} + \alpha_7|\theta_{p+1}|^{n+1} \\ &\leq \frac{\alpha_p}{2}|\bar{\theta}_p|^{n+1} + \alpha_8|\theta_{p+1}|^{n+1} \end{aligned} \tag{73}$$

for other constants α_7, α_8. We deduce from (72) and (73) that

$$|L_{D_2}V_p(\sigma)| \leq \left(\frac{\alpha_p}{2} + \alpha_9\varepsilon_{p+1}^{-1}\right)|\bar{\theta}_p|^{n+1} + \alpha_{10}|\theta_{p+1}|^{n+1}. \tag{74}$$

We also deduce from (71) that

$$|L_{D_2}(V_{p+1} - V_p)(\sigma)| \leq \alpha_{11}\varepsilon_{p+1}^{-1/2}|\bar{\theta}_{p+1}|^{n+1}, \tag{75}$$

and (70) then follows from (74), (75), and the condition $\varepsilon_{p+1} \geq 1$.

Let us now use (65), (69), and (70) to get an upper bound for $L_{F_{p+1}}V_{p+1}$. We obtain

$$\begin{aligned} L_{F_{p+1}}V_{p+1}(\sigma) &= (L_{D_0}V_{p+1} - L_{D_1}V_{p+1} - L_{D_2}V_{p+1})(\sigma) \\ &\leq -\left(\frac{\alpha_p}{2} - \alpha_{12}\varepsilon_{p+1}^{-1/2}\right)|\bar{\theta}_p|^{n+1} - \left(\alpha_1\varepsilon_{p+1}^{1/2} - \alpha_{13}\right)|\theta_{p+1}|^{n+1}. \end{aligned}$$

Since, by (65), α_p and α_1 are strictly positive, for ε_{p+1} large enough there exists $\alpha_{p+1} > 0$ such that

$$L_{F_{p+1}}V_{p+1}(\sigma) \leq -\alpha_{p+1}|\bar{\theta}_{p+1}|^{n+1}.$$

This concludes the proofs of Proposition 3 and Theorem 1.

Appendix: Proofs of Lemmas 1–4. The proofs of these lemmas rely on the following two properties.

Claim 1. *Let Y and Z denote two time-dependent left-invariant v.f. on G, and g, h solutions of $\dot{g} = Y(g,t)$ and $\dot{h} = Z(h,t)$, respectively. Then $\nu \triangleq gh$ is a solution of $\dot{\nu} = \mathrm{Ad}(h^{-1})Y(\nu,t) + Z(\nu,t)$.*

This is simple to verify. Indeed one has

$$\begin{aligned} \tfrac{d}{dt}(gh) &= dl_g(h)\dot{h} + dr_h(g)\dot{g} \\ &= dl_g(h)Z(h,t) + dr_h(g)Y(g,t) \\ &= Z(gh,t) + dr_h(g)dl_g(e)Y(e,t), \end{aligned}$$

so that one has to show only that $dr_h(g)dl_g(e) = dl_{gh}(e)\mathrm{Ad}(h^{-1})$. For this purpose, it suffices to use the definition of the Ad operator, i.e.,

$$\begin{aligned} \mathrm{Ad}(h) &= d(l_h \circ r_{h^{-1}})(e) \\ &= dl_h(r_{h^{-1}}(e))dr_{h^{-1}}(e) \\ &= dl_h(h^{-1})dr_{h^{-1}}(e), \end{aligned}$$

and well-known relations obtained by differentiating both members of the identities $l_g \circ r_h = r_h \circ l_g$, $l_g \circ l_{g^{-1}} = i_d$, and $l_{gh} = l_g \circ l_h$. The desired result is then obtained as follows:

$$\begin{aligned} dr_h(g)dl_g(e) &= dl_g(h)dr_h(e) \\ &= dl_g(h)dl_h(e)dl_{h^{-1}}(h)dr_h(e) \\ &= dl_{gh}(e)dl_{h^{-1}}(h)dr_h(e) \\ &= dl_{gh}(e)\mathrm{Ad}(h^{-1}). \end{aligned}$$

Claim 2. *Let $\{X_1, \ldots, X_n\}$ denote a graded basis of the Lie algebra $\mathfrak{g}$ of a Lie group G. Let $\lambda, \rho, q \in \{1, \ldots, n\}$, $\alpha_p \in \mathbb{R}$, and $s \in \mathbb{N}$. Then, there exist analytic functions $g_1, \ldots, g_n$ such that, for any $\alpha_\lambda, \alpha_\rho \in \mathbb{R}$,*

$$\sum_{j=s}^{\infty} \frac{1}{j!}(\mathrm{ad}^j(\alpha_\lambda X_\lambda + \alpha_\rho X_\rho), X_q) = \sum_{k=1}^{n} g_k(\alpha_\lambda, \alpha_\rho)X_k.$$

Furthermore, if $\alpha_\lambda, \alpha_\rho$ are analytic functions of x and y such that $\alpha_\lambda = \mathrm{O}(x^{r_\lambda})$ and $\alpha_\rho = \mathrm{O}(x^{r_\rho})$, then $g_k(\alpha_\lambda, \alpha_\rho)$ is an analytic function of x and y and $g_k(\alpha_\lambda, \alpha_\rho) = \mathrm{O}(x^{\max\{s\ \min\{r_\lambda, r_\rho\}, r_k - r_q\}})$.

The proof of this claim, which can be viewed as a direct adaptation of [28, sect. 2], is given in [18, App. A, Claim 2].

Proof of Lemma 1. In order to simplify the notation, let

$$X_\lambda = X_{\lambda(j)}\,,\ X_\rho = X_{\rho(j)}\,,\ \ \alpha_\lambda = \alpha_{j,\lambda}\,,\ \alpha_\rho = \alpha_{j,\rho}. \tag{76}$$

With this notation, it follows from (25) that $f_j(\sigma_j) = g_j(\sigma_j)h_j(\sigma_j)$, with $g_j(\sigma_j) = \exp(\alpha_j(\sigma_j)X_j)$ and $h_j(\sigma_j) = \exp(\alpha_\lambda(\sigma_j)X_\lambda + \alpha_\rho(\sigma_j)X_\rho)$.

Let $\sigma_j(.)$ denote an arbitrary smooth curve on $\mathbb{T}^2$. By using the fact that $\frac{d}{dt}\exp X(t) = \frac{d}{ds}\exp(X(t) + s\frac{d}{dt}X(t))_{|s=0}$ and that (see, e.g., [8, p. 105])

$$\frac{d}{ds}\exp(X + sY)_{|s=0} = (\phi(\mathrm{ad}X), Y)(\exp X)\,, \quad \phi(z) \triangleq \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)!} z^k,$$

one infers that

$$\begin{aligned}
\dot{h}_j &\triangleq \frac{d}{dt} h_j(\sigma_j(t)) \\
&= \frac{d}{ds} \exp\left(\alpha_\lambda X_\lambda + \alpha_\rho X_\rho + s\frac{d}{dt}(\alpha_\lambda X_\lambda + \alpha_\rho X_\rho)\right)_{|s=0} \\
&= (\phi(\mathrm{ad}(\alpha_\lambda X_\lambda + \alpha_\rho X_\rho)), \dot{\alpha}_\lambda X_\lambda + \dot{\alpha}_\rho X_\rho)(h_j).
\end{aligned}$$

One has also $\dot{g}_j = \dot{\alpha}_j X_j(g_j)$. The application of Claim 1 then yields

$$\dot{f}_j = (\phi(\mathrm{ad}(\alpha_\lambda X_\lambda + \alpha_\rho X_\rho)), \dot{\alpha}_\lambda X_\lambda + \dot{\alpha}_\rho X_\rho)(f_j) + \dot{\alpha}_j \mathrm{Ad}(\exp(-\alpha_\lambda X_\lambda - \alpha_\rho X_\rho))X_j(f_j). \tag{77}$$

Let us now use the fact (see, e.g., [8, p. 128]) that $\mathrm{Ad}(\exp Y)Z = (\exp \mathrm{ad}Y, Z)$. From (77),

$$\begin{aligned}
\dot{f}_j \ &= (\phi(\mathrm{ad}(\alpha_\lambda X_\lambda + \alpha_\rho X_\rho)), \dot{\alpha}_\lambda X_\lambda + \dot{\alpha}_\rho X_\rho)(f_j) \\
&\quad + \dot{\alpha}_j(\exp \mathrm{ad}(-\alpha_\lambda X_\lambda - \alpha_\rho X_\rho), X_j)(f_j) \\
&= \dot{\alpha}_\lambda X_\lambda(f_j) + \dot{\alpha}_\rho X_\rho(f_j) - \frac{1}{2}[\alpha_\lambda X_\lambda + \alpha_\rho X_\rho, \dot{\alpha}_\lambda X_\lambda + \dot{\alpha}_\rho X_\rho](f_j) \\
&\quad + \sum_{k=2}^{\infty} \frac{(-1)^k}{(k+1)!}(\mathrm{ad}^k(\alpha_\lambda X_\lambda + \alpha_\rho X_\rho), \dot{\alpha}_\lambda X_\lambda + \dot{\alpha}_\rho X_\rho)(f_j) \\
&\quad + \dot{\alpha}_j X_j(f_j) + \dot{\alpha}_j \sum_{k=1}^{\infty} \frac{1}{k!}(\mathrm{ad}^k(-\alpha_\lambda X_\lambda - \alpha_\rho X_\rho), X_j)(f_j).
\end{aligned} \tag{78}$$

Since $X_j = [X_\lambda, X_\rho]$ by Definition 3, it comes from (78) that

$$\begin{aligned}
\dot{f}_j = \dot{\alpha}_\lambda X_\lambda(f_j) + \dot{\alpha}_\rho X_\rho(f_j) + \left(\dot{\alpha}_j - \frac{1}{2}(\alpha_\lambda \dot{\alpha}_\rho - \alpha_\rho \dot{\alpha}_\lambda)\right) X_j(f_j) \\
+ (\alpha_\lambda \dot{\alpha}_\rho - \alpha_\rho \dot{\alpha}_\lambda) \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{(k+2)!}(\mathrm{ad}^k(\alpha_\lambda X_\lambda + \alpha_\rho X_\rho), X_j)(f_j) \\
+ \dot{\alpha}_j \sum_{k=1}^{\infty} \frac{1}{k!}(\mathrm{ad}^k(-\alpha_\lambda X_\lambda - \alpha_\rho X_\rho), X_j)(f_j).
\end{aligned} \tag{79}$$

It follows from (26) that

$$\alpha_\lambda, \frac{\partial \alpha_\lambda}{\partial \theta_j}, \frac{\partial \alpha_\lambda}{\partial \beta_j} = \mathrm{O}(\varepsilon_j^{r_\lambda}); \quad \alpha_\rho, \frac{\partial \alpha_\rho}{\partial \theta_j}, \frac{\partial \alpha_\rho}{\partial \beta_j} = \mathrm{O}(\varepsilon_j^{r_\rho}); \quad \alpha_j, \frac{\partial \alpha_j}{\partial \theta_j}, \frac{\partial \alpha_j}{\partial \beta_j} = \mathrm{O}(\varepsilon_j^{r_j}). \tag{80}$$

Therefore, by application of Claim 2 (with $x = \varepsilon_j$ and $y = \theta_j$),

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{(k+2)!} (\mathrm{ad}^k(\alpha_\lambda X_\lambda + \alpha_\rho X_\rho), X_j)(f_j) = \sum_{k=1}^{n} g_k(\alpha_\lambda, \alpha_\rho) X_k(f_j) \tag{81}$$

for some analytic functions $g_1, \ldots, g_n$ which verify

$$g_k(\alpha_\lambda, \alpha_\rho) = \mathrm{O}(\varepsilon_j^{\max\{1, r_k - r_j\}}). \tag{82}$$

Similarly, by applying Claim 2 again,

$$\sum_{k=1}^{\infty} \frac{1}{k!} (\mathrm{ad}^k(-\alpha_\lambda X_\lambda - \alpha_\rho X_\rho), X_j)(f_j) = \sum_{k=1}^{n} h_k(\alpha_\lambda, \alpha_\rho) X_k, \tag{83}$$

with

$$h_k(\alpha_\lambda, \alpha_\rho) = \mathrm{O}(\varepsilon_j^{\max\{1, r_k - r_j\}}). \tag{84}$$

From (79), (81), and (83), we get

(85)

$$\begin{aligned}
\dot{f}_j = &\left(\dot{\alpha}_\lambda + (\alpha_\lambda \dot{\alpha}_\rho - \alpha_\rho \dot{\alpha}_\lambda)\, g_\lambda(\alpha_\lambda, \alpha_\rho) + \dot{\alpha}_j\, h_\lambda(\alpha_\lambda, \alpha_\rho)\right) X_\lambda(f_j) \\
&+ \left(\dot{\alpha}_\rho + (\alpha_\lambda \dot{\alpha}_\rho - \alpha_\rho \dot{\alpha}_\lambda)\, g_\rho(\alpha_\lambda, \alpha_\rho) + \dot{\alpha}_j\, h_\rho(\alpha_\lambda, \alpha_\rho)\right) X_\rho(f_j) \\
&+ \left(\dot{\alpha}_j - \frac{1}{2}(\alpha_\lambda \dot{\alpha}_\rho - \alpha_\rho \dot{\alpha}_\lambda) + (\alpha_\lambda \dot{\alpha}_\rho - \alpha_\rho \dot{\alpha}_\lambda)\, g_j(\alpha_\lambda, \alpha_\rho) + \dot{\alpha}_j\, h_j(\alpha_\lambda, \alpha_\rho)\right) X_j(f_j) \\
&+ \sum_{k \notin \{\lambda, \rho, j\}} \left((\alpha_\lambda \dot{\alpha}_\rho - \alpha_\rho \dot{\alpha}_\lambda)\, g_k(\alpha_\lambda, \alpha_\rho) + \dot{\alpha}_j\, h_k(\alpha_\lambda, \alpha_\rho)\right) X_k(f_j).
\end{aligned}$$

Since this equality holds along any smooth curve $\sigma_j(.)$ on $\mathbb{T}^2$, it is also true when the time-derivatives are replaced by partial derivatives w.r.t. either θ_j or β_j.

Now, it follows from (26) that

$$d\alpha_j - \frac{1}{2}(\alpha_\lambda d\alpha_\rho - \alpha_\rho d\alpha_\lambda) = \frac{\varepsilon_j^{r_j}}{2} d\theta_j + \varepsilon_j^{r_j}(1 - \cos\theta_j) d\beta_j \tag{86}$$

and

$$\alpha_\lambda, \ \frac{\partial \alpha_\lambda}{\partial \beta_j}, \ \alpha_\rho, \ \frac{\partial \alpha_\rho}{\partial \beta_j}, \ \alpha_j = \mathrm{O}(\theta_j); \quad \frac{\partial \alpha_j}{\partial \beta_j} = 0. \tag{87}$$

Furthermore, if f is an analytic function of ε and θ such that $f = \mathrm{O}(|\varepsilon|^p)$ and $f = \mathrm{O}(|\theta|^q)$, then $f = \mathrm{O}(|\varepsilon|^p)\mathrm{O}(|\theta|^q)$. Therefore, by using (80), (82), (84), (86), and (87) in (85), it is tedious but simple to recover all relations in Lemma 1. (For the last relation of (39), note that g_j in (85) is an $\mathrm{O}(\theta_j)$ because it is a function of α_λ and α_ρ, which vanishes when $\alpha_\lambda = \alpha_\rho = 0$).

Proof of Lemma 2. From Claim 1 and relations (24) and (37),

(88)
$$\frac{\partial f}{\partial \theta_j} = \sum_{k=1}^{n} v_{k,j} \mathrm{Ad}(f_{m+1}^{-1} \cdots f_{j-1}^{-1}) X_k(f)\,, \quad \frac{\partial f}{\partial \beta_j} = \sum_{k=1}^{n} w_{k,j} \mathrm{Ad}(f_{m+1}^{-1} \cdots f_{j-1}^{-1}) X_k(f).$$

From the fact that $\mathrm{Ad}(g_1 g_2) = \mathrm{Ad}(g_1)\mathrm{Ad}(g_2)$ and (25),

(89)
$$\begin{aligned} \mathrm{Ad}(f_{m+1}^{-1} \cdots f_{j-1}^{-1}) &= \prod_{p=m+1}^{j-1} \mathrm{Ad}(f_p^{-1}) \\ &= \prod_{p=m+1}^{j-1} \mathrm{Ad}(\exp(-\alpha_{p,\lambda} X_{\lambda(p)} - \alpha_{p,\rho} X_{\rho(p)}))\mathrm{Ad}(\exp -\alpha_p X_p). \end{aligned}$$

By application of Claim 2, for any $p, q, k = 1, \ldots, n$ and $(\alpha_p, \alpha_q) \in \mathbb{R}^2$,

$$\mathrm{Ad}(\exp -\alpha_p X_p - \alpha_q X_q) X_k = X_k + \sum_{i=1}^{n} h_{p,q}^i(\alpha_p, \alpha_q) X_i$$

for some analytic functions $h_{p,q}^i$. Moreover, if $\alpha_p = \mathrm{O}(\varepsilon^{r_p})$ and $\alpha_q = \mathrm{O}(\varepsilon^{r_q})$ are analytic functions, then $h_{p,q}^i(\alpha_p, \alpha_q) = \mathrm{O}(\varepsilon^{\max(1, r_i - r_k)})$. This is used to infer from (80) and (89) that

$$\mathrm{Ad}(f_{m+1}^{-1} \cdots f_{j-1}^{-1}) X_k = X_k + \sum_i g_{j,k}^i X_i \quad \text{with } g_{j,k}^i = \mathrm{O}(\bar{\varepsilon}_{j-1}^{\max(1, r_i - r_k)}). \tag{90}$$

From (90),

$$\sum_{k=1}^{n} v_{k,j} \mathrm{Ad}(f_{m+1}^{-1} \cdots f_{j-1}^{-1}) X_k(f) = \sum_{i=1}^{n} \left(v_{i,j} + \sum_{k=1}^{n} v_{k,j} g_{j,k}^i \right) X_i(f),$$

and a similar expression holds when replacing v by w. Therefore, in view of (88), equation (40) holds with

$$a_{i,j} \triangleq v_{i,j} + \sum_{k=1}^{n} v_{k,j}\, g_{j,k}^i = A + B + C\,, \tag{91}$$
$$A = \sum_{r_k \le r_i} v_{k,j}\, g_{j,k}^i\,,\ B = v_{i,j}\,,\ C = \sum_{r_k > r_i} v_{k,j}\, g_{j,k}^i,$$

and

$$b_{i,j} \triangleq w_{i,j} + \sum_{k=1}^{n} w_{k,j}\, g_{j,k}^i = D + E + F\,, \tag{92}$$
$$D = \sum_{r_k \le r_i} w_{k,j}\, g_{j,k}^i\,,\ E = w_{i,j}\,,\ F = \sum_{r_k > r_i} w_{k,j}\, g_{j,k}^i.$$

Lemma 2 follows from this decomposition. Let us first show how (41) is obtained. From (38) and (90), A, B, and C in (91) are $\mathrm{O}(\bar{\varepsilon}_j^{r_i})$. This gives the first relation of (41).

For $i < j$ and $r_i = r_j$, A vanishes at $\bar{\varepsilon}_{j-1} = 0$ because of (90), and in view of (38), $B = \mathrm{o}(\bar{\varepsilon}_j^{r_i})$ and $C = \mathrm{o}(\bar{\varepsilon}_j^{r_i})$. This gives the second relation of (41).

For $i = j$, the only difference with the previous case comes from the B term, which, in view of (38), is equal to $\varepsilon_j^{r_j}/2 + \mathrm{o}(\varepsilon_j^{r_j})$. This gives the third relation of (41).

Let us now show how (42) is obtained. From (90),

$$g^i_{j,k} = \mathrm{O}(\bar{\theta}_{j-1}) \tag{93}$$

because, by (25) and (26),

$$\bar{\theta}_{j-1} = 0 \Longrightarrow f_{m+1} = \cdots = f_{j-1} = e \Longrightarrow \mathrm{Ad}(f_{m+1}^{-1} \cdots f_{j-1}^{-1}) X_k = X_k.$$

The first relation of (42) is then simply obtained from (39), (90), (92), and (93).

For $i = j$, E in (92) accounts for the term $\varepsilon_j^{r_j}(1 - \cos\theta_j)$—up to higher order terms—in the second relation of (42), whereas D and F account for the remaining term by inspection of (39), (90), and (93).

Proof of Lemma 3. The lemma is a direct consequence of the following property, which can be proved by induction exactly as in the proof of [18, Lem. 3]:

$$\begin{array}{l} \forall k = m+1, \ldots, n,\ \exists \bar{\eta}_k \in \mathbb{R}^{k-m},\ \exists \alpha_k > 0\ : \\ \bar{\varepsilon}_k = \varepsilon_k \bar{\eta}_k \Rightarrow D_k \geq \alpha_k \varepsilon_k^{\bar{r}_k} + o(|\varepsilon_k|^{\bar{r}_k}), \end{array} \tag{94}$$

with $\bar{r}_k = r_{m+1} + \cdots + r_k$ and D_k the function defined by

$$D_k(\sigma) \triangleq \mathrm{Det}(a_{i,j}(\sigma))_{i,j=m+1,\ldots,k}.$$

The first step consists of showing that (94) holds for $k = m+1$. From relation (41) in Lemma 2, $a_{m+1,m+1} = \frac{1}{2}\varepsilon_{m+1}^{r_{m+1}} + o(\varepsilon_{m+1}^{r_{m+1}})$. Since $D_{m+1} = a_{m+1,m+1}$ and $\bar{r}_{m+1} = r_{m+1}$, (94) is verified with $\eta_{m+1} = 1$ and $\alpha_{m+1} = \frac{1}{2}$. For the subsequent steps of the proof, the reader is referred to [18].

Proof of Lemma 4. Let us first show how Lemma 4—relation (44), in particular—is obtained from the following two claims.

Claim 3. *For any i, j,*

$$\left\{ \begin{array}{ll} v_{i,j} = \varepsilon_j^{r_i} \tilde{v}_{i,j} & \textit{with}\ \ \tilde{v}_{i,j} = \mathrm{O}(\theta_j^{r_i - r_j}), \\ w_{i,j} = \varepsilon_j^{r_i} \tilde{w}_{i,j} & \textit{with}\ \ \tilde{w}_{i,j} = \mathrm{O}(\theta_j^{\max(1, r_i - r_j + 2)}), \end{array} \right. \tag{95}$$

where the functions $\tilde{v}_{i,j}$ and $\tilde{w}_{i,j}$ do not depend on the ε_k's.

Claim 4. *Each function $g^i_{j,k}$ in* (90) *is a polynomial in $\varepsilon_3, \ldots, \varepsilon_{j-1}$ homogeneous of degree $r_i - r_k$. Furthermore,*

$$g^i_{j,k} = \left\{ \begin{array}{ll} \mathrm{O}(\bar{\theta}_{j-1}^{r_i - r_k}) & \textit{if}\ r_j \leq r_k < r_i, \\ \mathrm{O}(\bar{\theta}_{j-1}^{r_i - r_j + 1}) & \textit{if}\ r_k < r_j < r_i. \end{array} \right. \tag{96}$$

From these claims, and from (91) and (92), it is straightforward to show that $a_{i,j}$ and $b_{i,j}$ are polynomials homogeneous of degree r_i in $\varepsilon_3, \ldots, \varepsilon_j$. Then, by (95), E and F in (92) are $\mathrm{O}(\bar{\theta}_j^{\max(1, r_i - r_j + 2)})$. As for the term D, it can be decomposed as

$$D = \sum_{r_k < r_i} w_{k,j}\, g^i_{j,k} + \sum_{r_k = r_i} w_{k,j}\, g^i_{j,k}. \tag{97}$$

From (95) and (96), the first sum in (97) is an $\mathrm{O}(\theta_j^{\max(1,r_k-r_j+2)})\mathrm{O}(\bar\theta_{j-1}^{r_i-r_k})$ if $r_j \le r_k < r_i$, and an $\mathrm{O}(\theta_j)\mathrm{O}(\bar\theta_{j-1}^{r_i-r_j+1})$ if $r_k < r_j < r_i$. Therefore, in both cases, it is an $\mathrm{O}(\bar\theta_j^{\max(1,r_i-r_j+2)})$. As for the second sum in (97), it follows from (95) that it is an $\mathrm{O}(\bar\theta_j^{\max(1,r_i-r_j+2)})$. This proves (44) for the term $b_{i,j}$. The proof for $a_{i,j}$ is similar.

It remains to prove Claims 3 and 4. In the case of a chained system, each element X_j of the graded basis, for $j = 3, \dots, n$, is equal to $[X_{\lambda(j)}, X_{\rho(j)}]$ with $\lambda(j) = 1$ and $\rho(j) = j-1$. It is also a constant v.f. With the notation used in the proof of Lemma 1, these two facts imply that

$$(\mathrm{ad}(\alpha_\lambda X_\lambda + \alpha_\rho X_\rho), X_j) = \begin{cases} \alpha_\lambda X_{j+1} & \text{if } j < n, \\ 0 & \text{if } j = n. \end{cases}$$

Relation (79) in Lemma 1 then becomes

$$\dot f_j = \dot\alpha_\lambda X_\lambda(f_j) + \dot\alpha_\rho X_\rho(f_j) + \left(\dot\alpha_j - \frac{1}{2}(\alpha_\lambda\dot\alpha_\rho - \alpha_\rho\dot\alpha_\lambda)\right) X_j(f_j)$$
$$+ (\alpha_\lambda\dot\alpha_\rho - \alpha_\rho\dot\alpha_\lambda)\sum_{k=1}^{n-j}\frac{(-1)^{k+1}}{(k+2)!}\alpha_\lambda^k X_{j+k}(f_j) + \dot\alpha_j\sum_{k=1}^{n-j}\frac{(-\alpha_\lambda)^k}{k!}X_{j+k}(f_j).$$

Claim 3 is easily obtained by identifying this equality with (37), and by using (26) and (29).

Let us now prove Claim 4 by showing how relation (96) is obtained. The first step involves the evaluation of $\mathrm{Ad}(f_p^{-1})X_k$, for $p \in \{3, \dots, n-1\}$ and $k \in \{1, \dots, n\}$. We distinguish two cases.

Case 1. $k \ne 1$. From the definition (28) of $X_1, \dots, X_n$ and from (25),

$$\begin{aligned} \mathrm{Ad}(f_p^{-1})X_k &= \mathrm{Ad}(\exp(-\alpha_{p,\lambda}X_1 - \alpha_{p,\rho}X_{p-1}))\mathrm{Ad}(\exp -\alpha_p X_p)X_k \\ &= \mathrm{Ad}(\exp(-\alpha_{p,\lambda}X_1 - \alpha_{p,\rho}X_{p-1}))X_k \\ &= X_k + \sum_{j=1}^{n-k}\frac{(-\alpha_{p,\lambda})^j}{j!}X_{k+j} \\ &= X_k + \sum_{j=1}^{n-k}\varepsilon_p^j\, h_{p,k}^{k+j}X_{k+j} \quad \text{with } h_{p,k}^{k+j} = \mathrm{O}(\theta_p^j), \end{aligned} \tag{98}$$

where the last equality comes from (26) and (29), and $h_{p,k}^{k+j}$ is a function which does not depend on ε_p. From (29), $r_{k+j} = r_k + j$ for $k > 1$ and $0 \le j \le n-k$. Therefore, from (26) and (98),

$$\mathrm{Ad}(f_p^{-1})X_k = X_k + \sum_{i>k}\varepsilon_p^{r_i-r_k}h_{p,k}^i X_i \quad \text{with } h_{p,k}^i = \mathrm{O}(\theta_p^{r_i-r_k}). \tag{99}$$

By applying (99) recursively, it follows that, for any $k \ne 1$,

$$\mathrm{Ad}(f_3^{-1}\cdots f_{j-1}^{-1})X_k = X_k + \sum_{i>k} g_{j,k}^i X_i \quad \text{with } g_{j,k}^i = \mathrm{O}(\bar\theta_{j-1}^{r_i-r_k}), \tag{100}$$

where each $g_{j,k}^i$ is a polynomial homogeneous of degree $r_i - r_k$ in $\varepsilon_3, \dots, \varepsilon_{j-1}$. This yields (96) for $r_j \le r_k \le r_i$, and also for $r_k < r_j < r_i$ (and $k \ne 1$) after noticing that, in this case, $r_i - r_k \ge r_i - r_j + 1$.

Case 2. $k = 1$. We have

(101)
$$\begin{aligned}
\mathrm{Ad}(f_p^{-1})X_1 &= \mathrm{Ad}(\exp(-\alpha_{p,\lambda}X_1 - \alpha_{p,\rho}X_{p-1}))\mathrm{Ad}(\exp -\alpha_p X_p)X_1 \\
&= \mathrm{Ad}(\exp(-\alpha_{p,\lambda}X_1 - \alpha_{p,\rho}X_{p-1}))(X_1 + \alpha_p X_{p+1}) \\
&= X_1 + \alpha_p X_{p+1} + [-\alpha_{p,\lambda}X_1 - \alpha_{p,\rho}X_{p-1}, X_1 + \alpha_p X_{p+1}] \\
&\quad + \sum_{k=2}^{\infty} \frac{1}{k!}\Big(\mathrm{ad}^{k-1}(-\alpha_{p,\lambda}X_1 - \alpha_{p,\rho}X_{p-1}), \\
&\qquad\qquad [-\alpha_{p,\lambda}X_1 - \alpha_{p,\rho}X_{p-1}, X_1 + \alpha_p X_{p+1}]\Big) \\
&= X_1 + \alpha_p X_{p+1} - \alpha_{p,\lambda}\alpha_p X_{p+2} + \alpha_{p,\rho}X_p \\
&\quad + \sum_{k=2}^{\infty} \frac{1}{k!}(\mathrm{ad}^{k-1}(-\alpha_{p,\lambda}X_1 - \alpha_{p,\rho}X_{p-1}), -\alpha_{p,\lambda}\alpha_p X_{p+2} + \alpha_{p,\rho}X_p) \\
&= X_1 + \alpha_p X_{p+1} - \alpha_{p,\lambda}\alpha_p X_{p+2} + \alpha_{p,\rho}X_p \\
&\quad - \alpha_{p,\lambda}\alpha_p \sum_{k=2}^{\infty} \frac{(-\alpha_{p,\lambda})^{k-1}}{k!} X_{p+2+k-1} + \alpha_{p,\rho} \sum_{k=2}^{\infty} \frac{(-\alpha_{p,\lambda})^{k-1}}{k!} X_{p+k-1}.
\end{aligned}$$

It follows from (26) and (101) that

$$\mathrm{Ad}(f_p^{-1})X_1 = X_1 + \sum_{i>1} \varepsilon_p^{r_i - r_1} h_p^i X_i \quad \text{with } h_p^i = \mathrm{O}(\theta_p^{r_i - r_p}), \tag{102}$$

and h_p^i does not depend on ε_p. By applying (102) recursively and by using (100), it follows that

$$\mathrm{Ad}(f_3^{-1} \cdots f_{j-1}^{-1})X_1 = X_1 + \sum_{i>1} g_{j,1}^i X_i \quad \text{with } g_{j,1}^i = \mathrm{O}(\bar{\theta}_{j-1}^{r_i - r_{j-1}}) = \mathrm{O}(\bar{\theta}_{j-1}^{r_i - r_j + 1}),$$

where each $g_{j,1}^i$ is polynomial homogeneous of degree $r_i - r_1$ in $\varepsilon_3, \ldots, \varepsilon_{j-1}$. This concludes the proof of Claim 4.

Acknowledgments. The authors wish to thank the anonymous reviewers who read the first version of the paper in depth and provided numerous relevant remarks and suggestions to improve the presentation.

REFERENCES

[1] A. Astolfi, *Discontinuous control of nonholonomic systems*, Systems Control Lett., 27 (1996), pp. 37–45.

[2] M. K. Bennani and P. Rouchon, *Robust stabilization of flat and chained systems*, in Proceedings of the European Control Conference (ECC), Rome, 1995, pp. 2642–2646.

[3] A. Bloch and S. Drakunov, *Stabilization and tracking in the nonholonomic integrator via sliding modes*, Systems Control Lett., 29 (1996), pp. 91–99.

[4] A. Bloch, M. Reyhanoglu, and N. McClamroch, *Control and stabilization of nonholonomic dynamic systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1746–1757.

[5] R. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. Sussmann, eds., Birkhäuser Boston, Cambridge, MA, 1983, pp. 181–191.

[6] C. Canudas de Wit and O. J. Sørdalen, *Exponential stabilization of mobile robots with nonholonomic constraints*, IEEE Trans. Automat. Control, 37 (1992), pp. 1791–1797.

[7] J.-M. Coron, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.
[8] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces*, Academic Press, New York, 1978.
[9] H. Hermes, *Nilpotent and high-order approximations of vector field systems*, SIAM Rev., 33 (1991), pp. 238–264.
[10] M. Kawski and H. Sussmann, *Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory*, in Operators, Systems, and Linear Algebra, D. Prätzel-Wolters, U. Helmke, and E. Zerz, eds., Teubner, Leipzig, 1997, pp. 111–128.
[11] W. Lin and P. Radom, *Recursive design of discontinuous controllers for uncertain driftless systems in power chained form*, IEEE Trans. Automat. Control, 45 (2000), pp. 1886–1892.
[12] D. Lizárraga, *Obstructions to the existence of universal stabilizers for smooth control systems*, Math. Control Signals Systems, 16 (2004), pp. 255–277.
[13] R. M'Closkey and R. Murray, *Exponential stabilization of driftless nonlinear control systems using homogeneous feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 614–628.
[14] P. Morin, J.-B. Pomet, and C. Samson, *Design of homogeneous time-varying stabilizing control laws for driftless controllable systems via oscillatory approximation of Lie brackets in closed loop*, SIAM J. Control. Optim., 38 (1999), pp. 22–49.
[15] P. Morin and C. Samson, *Exponential stabilization of nonlinear driftless systems with robustness to unmodeled dynamics*, ESAIM Control. Optim. Calc. Var. 4 (1999), pp. 1–36.
[16] P. Morin and C. Samson, *Control of non-linear chained systems. from the Routh-Hurwitz stability criterion to time-varying exponential stabilizers*, IEEE Trans. Automat. Control, 45 (2000), pp. 141–146.
[17] P. Morin and C. Samson, *A characterization of the Lie algebra rank condition by transverse periodic functions*, SIAM J. Control. Optim., 40 (2001), pp. 1227–1249.
[18] P. Morin and C. Samson, *Practical stabilization of driftless systems on Lie groups: The transverse function approach*, IEEE Trans. Automat. Control, 48 (2003), pp. 1496–1508.
[19] J.-B. Pomet and C. Samson, *Exponential stabilization of nonholonomic systems in power form*, in Proceedings of the IFAC Symposium on Robust Control Design, Rio de Janeiro, Brazil, 1994, pp. 447–452.
[20] J.-B. Pomet, *Explicit design of time-varying stabilizing control laws for a class of controllable systems without drift*, Systems Control Lett., 18 (1992), pp. 467–473.
[21] C. Samson, *Velocity and torque feedback control of a nonholonomic cart*, Int. Workshop in Adaptive and Nonlinear Control: Issues in Robotics (1990); also in Lecture Notes in Control and Inform. Sci. 162, Springer-Verlag, Berlin, 1991, pp. 125–151.
[22] C. Samson, *Control of chained systems. Application to path following and time-varying point-stabilization*, IEEE Trans. Automat. Control, 40 (1995), pp. 64–77.
[23] O. J. Sørdalen and O. Egeland, *Exponential stabilization of nonholonomic chained systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 35–49.
[24] H. Struemper and P. Krishnaprasad, *Tracking and Stabilization for Control Systems on Matrix Lie Groups*, Tech. report 97-34, Institute for Systems Research (ISR), University of Maryland, College Park, MD, 1997.
[25] H. Struemper, *Nilpotent approximation and nilpotentization for under-actuated systems on matrix Lie groups*, in Proceedings of the IEEE Conference on Decision and Control (CDC), Phoenix, AZ, 1998, pp. 4188–4193.
[26] H. J. Sussmann, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.
[27] A. Teel, R. Murray, and G. Walsh, *Nonholonomic control systems: From steering to stabilization with sinusoids*, Internat. J. Control, 62 (1995), pp. 849–870.
[28] J. Wei and E. Norman, *On global representations of the solutions of linear differential equations as a product of exponentials*, Proc. Amer. Math. Soc., 15 (1964), pp. 327–334.
[29] K. Zhou and J. Doyle, *Essentials of Robust Control*, Prentice–Hall, Englewood Cliffs, NJ, 1998.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 58–85

PERPETUAL CONVERTIBLE BONDS*

MIHAI SÎRBU†, IGOR PIKOVSKY‡, AND STEVEN E. SHREVE†

Abstract. A firm issues a convertible bond. At each subsequent time, the bondholder must decide whether to continue to hold the bond, thereby collecting coupons, or to convert it to stock. The firm may at any time call the bond. Because calls and conversions often occur far from maturity, it is not unreasonable to model this situation with a perpetual convertible bond, i.e., a convertible coupon-paying bond without maturity. This model admits a relatively simple solution, under which the value of the perpetual convertible bond, as a function of the value of the underlying firm, is determined by a nonlinear ordinary differential equation.

Key words. convertible bonds, stochastic calculus, viscosity solutions

AMS subject classifications. 90A09, 60H30, 60G44

DOI. 10.1137/S0363012902412458

1. Introduction.

Firms raise capital by issuing debt (bonds) and equity (shares of stock). The convertible bond is intermediate between these two instruments. A convertible is a bond in the sense that it entitles its owner to receive coupons plus the return of the principle at maturity. However, prior to maturity, the holder may "convert" the bond, surrendering it for a preset number of shares of stock. The price of the bond is thus dependent on the price of the firm's stock. Finally, prior to maturity, the firm may "call" the bond, forcing the bondholder to either surrender it to the firm for a previously agreed price or else convert it for stock as above.

After issuing a convertible bond, the firm's objective is to exercise its call option in order to maximize the value of shareholder equity. The bondholder's objective is to exercise his conversion option in order to maximize the value of the bond. If stock and convertible bonds are the only assets issued by a firm, then the value of the firm is the sum of the value of these two types of assets. In idealized markets where the Miller–Modigliani [17], [18] assumptions hold, changes in corporate capital structure do not affect firm value. In particular, the value of the firm does not change at the time of conversion, and the only change in the value of the firm at the time of call is a reduction by the call price paid to the bondholder if the bondholder surrenders rather than converts the bond. By acting to maximize the value of equity, the firm is in fact minimizing the value of the convertible bond. By acting to maximize the value of the bond, the bondholder is in fact minimizing the value of equity. This creates a two-person, zero-sum game.

Brennan and Schwartz [5] and Ingersoll [11] address the convertible bond pricing problem via the arbitrage pricing theory developed by Merton [16] and underlying the option pricing formula of Black and Scholes [4]. This leads to the conclusion that the firm should call as soon as the conversion value of the bond (the value the bondholder would receive if he converts the bond to stock) rises to the call price. There has been

*Received by the editors July 30, 2002; accepted for publication (in revised form) October 16, 2003; published electronically May 25, 2004. This material is based upon work supported by the National Science Foundation under grants DMS-0103814 and DMS-0139911.
http://www.siam.org/journals/sicon/43-1/41245.html
†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, 15213 (msirbu@andrew.cmu.edu, shreve@cmu.edu).
‡Global Modeling and Analytics, Credit Suisse First Boston, One Cabot Square, London, E14 4QJ, UK (igor.pikovsky@csfb.com).

considerable discussion whether firms call bonds at this time; see, e.g., [1], [2], [8], [12].

In the Brennan and Schwartz [5] model, dividends and coupons are paid at discrete dates. Between these dates the value of the firm is a geometric Brownian motion and the price of the convertible bond is governed by the linear second-order partial differential equation developed by Black and Scholes [4]. Brennan and Schwartz [6] generalize that model to allow random interest rates and debt senior to the convertible bond. In Ingersoll [11], coupons are paid out continuously, and for most of the results obtained, dividends are zero. Again, the bond price is governed by a linear second-order partial differential equation. In [5] the bond should not be converted except possibly immediately prior to a dividend payment; in [11] the bond should not be converted except possibly at maturity. Therefore, neither of these papers needs to address the free boundary problem which would arise if early conversion were optimal.

The present paper assumes that a firm's value comprises equity and convertible bonds. To simplify the discussion, we assume the equity is in the form of a single share of stock, and there is a single convertible bond. We assume the value of the issuing firm has constant volatility, the bond continuously pays a coupon at a fixed rate, and the firm equity pays a dividend at a rate which is a fixed fraction of the equity value. In particular, payments are always up to date and there is no issue of accrued interest at the time of a call, default, or conversion. Default occurs if the coupon payments cause the firm value to fall to zero, in which case the bond has zero recovery. In this model, both the bond price and the stock price are functions of the underlying firm value. As pointed out by [3], this means that the stock price does not have constant volatility. Furthermore, because the stock price is the difference between firm value and bond price and because dividends are paid proportionally to the stock price, the differential equation characterizing the bond price as a function of the firm value is *nonlinear*. The development of a mathematical methodology to treat this nonlinearity is the rationale for this paper.

To simplify the analysis, we assume the bond is *perpetual*, i.e., it never matures. This removes the time parameter from the problem, and the free boundary problems associated with optimal call and optimal conversion become "free point" problems. Perpetual bonds are the asymptotic case of finite-maturity bonds; work along the lines of this paper on these bonds is forthcoming. Also, as noted by Ingersoll [11], perpetual convertible bonds are unknown in the market, but they are close relatives of preferred stock, which does trade. Preferred stock does not mature, it can often be called by the issuing firm, and it can be converted to common stock by its owner.

In the time-independent setting of this paper, it is possible to place the convertible bond pricing problem on a firm theoretical foundation. Indeed, the price we obtain is shown to be the only arbitrage-free price in a perfectly liquid market in which the bond, the stock, and a constant-interest-rate money market can be traded. To establish this we first make the assumption that the respective parties adopt not necessarily optimal call and conversion strategies and derive the corresponding no-arbitrage bond price (Theorem 2.1). We then pose the determination of optimal call and conversion strategies as a two-person, zero-sum game and show that the game has a value (Theorem 2.4). We give a full description of the bond price as a function of the firm value in Theorem 2.5. One of the conclusions of that theorem is that it can be optimal to call the bond before the conversion price has reached the call price.

2. The model. We consider a firm whose value at time $t \geq 0$ is denoted by $X(t)$. We assume that prior to call or conversion of the convertible bond, the evolution of

$X(t)$ is governed by the stochastic differential equation

$$dX(t) = h(X(t))\,dt - c\,dt + \sigma X(t)\,dW(t), \tag{2.1}$$

where W is a one-dimensional Brownian motion on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, h is a Lipschitz continuous function satisfying $h(0) = 0$, and c and σ are positive constants. We denote by $\{\mathcal{F}(t); t \geq 0\}$ the filtration generated by the Brownian motion W, augmented by the null sets in $\mathcal{F}$.

At time t, the firm has a debt $D(t)$, and so the equity value is

$$S(t) = X(t) - D(t). \tag{2.2}$$

The debt is in the nature of a convertible bond, which pays coupons at the constant rate c. The bond never matures. The firm's dividend policy is to pay continuously to shareholders at a rate δ times the equity, where $\delta > 0$.

At any time, the owner of the convertible bond may *convert* it for stock. Upon conversion the bondholder will be issued new stock so that his share of the total equity of the company is the *conversion factor* γ, where $0 < \gamma < 1$. To simplify the discussion, let us assume that before conversion the firm has one share of stock outstanding. We are denoting by $X(t)$ the value of the firm and by $D(t)$ the size of the debt before conversion. Therefore, (2.2) gives the price of firm's single share of stock before conversion. Upon conversion, the firm issues new stock and the former bondholder becomes a stockholder. The total value of the firm's outstanding stock is $X(t)$, and the value of the stock owned by the former bondholder is $\gamma X(t)$. Therefore, the price of the share of the stock outstanding before conversion is now $(1-\gamma)X(t)$.

At any time, the firm may *call* the bond, which requires the bondholder to either immediately surrender it for the fixed *conversion price* $K > 0$ or else immediately convert it as described above. If the bond is surrendered, no new stock is issued and the price of the firm's single outstanding share becomes $X(t) - K$. In this model the firm may not call the bond if $X(t) < K$; i.e., there is no provision for reissuing debt.

Equation (2.1) describes the evolution of $X(t)$ only before call or conversion. Prior to conversion or call, the firm value $X(t)$ may drop to zero, in which case the firm declares bankruptcy and coupons and dividends cease.

There is a constant interest rate r, and we assume $\delta < r$. Prior to call or conversion of the bond, there are three tradable assets: the firm's stock, the convertible bond, and a money market paying interest r. We assume that all these are infinitely divisible and there are no transaction costs. Thus, the value $V(t)$ of a portfolio which holds $\Delta_1(t)$ shares of stock and $\Delta_2(t)$ convertible bonds at time t and finances this by investing or borrowing at interest rate r evolves according to the stochastic differential equation

$$\begin{aligned} dV(t) = {} & \Delta_1(t)\big(dS(t) + \delta S(t)\big) + \Delta_2(t)\big(dD(t) + c\,dt\big) \\ & + r\big(V(t) - \Delta_1(t)S(t) - \Delta_2(t)D(t)\big)\,dt. \end{aligned} \tag{2.3}$$

An arbitrage arises if one can begin with $V(0) = 0$ and use $\{\mathcal{F}(t)\}$-adapted processes Δ_1 and Δ_2 so that at some bounded stopping time τ at or before the minimum of the time of call, the time of conversion and the time of bankruptcy, $V(\tau) \geq 0$ almost surely and $V(\tau) > 0$ with positive probability. We restrict ourselves to trading strategies $\Delta_1(t)$, $\Delta_2(t)$ which cause $V(t)$ to be uniformly bounded from below for $0 \leq t \leq \tau$. Our goal is to price the convertible bond, under the assumption that the firm issuing the bond and the bondholder behave optimally, in a way which precludes arbitrage.

If the bond is called, the bondholder surrenders it for the call price K if K exceeds the conversion value $\gamma X(t)$ and converts it if $\gamma X(t) > K$. If $\gamma X(t) = K$, the bondholder is indifferent between surrender and conversion. Thus, if the bond is called when the firm value is $X(t)$, then the value of the bond is $\max\{K, \gamma X(t)\}$. If the bond has not been called, we assume the bondholder adopts a rule of the form: "convert as soon as the value of the firm equals or exceeds C_o." For the firm, we consider call strategies of the form "call the first time the value of the firm equals or exceeds C_a." The firm must choose $C_a \geq K$; if $C_a < K$, the firm would call when the firm value was insufficient to pay the call price. The firm and bondholder each choose a strategy, characterized by positive constants $C_a \geq K$ and $C_o > 0$. Once C_a, C_o are chosen, we want to find the price of the bond as a smooth function of the value of the firm such that no arbitrage can occur.

To set the notation, for an arbitrary number $a > 0$ we define the *nonlinear* differential operator acting on functions $f \in C[0,a] \cap C^2(0,a)$ by

$$\mathcal{N}f(x) \triangleq rf(x) - (rx - c)f'(x) + \delta(x - f(x))f'(x) - \frac{1}{2}\sigma^2 x^2 f''(x). \tag{2.4}$$

We shall see that this differential operator corresponds to the stochastic differential equation for the firm value

$$dX(t) = \big(rX(t) - c\big)\,dt - \delta\big(X(t) - f(X(t))\big)\,dt + \sigma X(t)\,dW(t), \tag{2.5}$$

rather than (2.1) posited above. This turns out to be the so-called *risk-neutral* evolution of the value of the firm. Under the risk-neutral evolution, the firm value has mean rate of change r reduced by the coupon and dividend payments. The volatility σ is the same as in (2.1). An interesting feature of this model is that the function f appearing in (2.5), which determines the evolution of the "state" under the risk neutral measure for this problem, must be determined by optimality considerations. It is not known a priori.

THEOREM 2.1. *Suppose $C_a \geq K$ and $C_o > 0$ are chosen (not necessarily optimally) by the firm and bondholder, respectively, and set*

$$a_* \triangleq \min\{C_a, C_o\}, \quad \tau_* \triangleq \inf\{t \geq 0; X(t) \notin (0, a_*)\}. \tag{2.6}$$

Assume $X(0) \in (0, a_*)$ and

$$D(t) = f(X(t)), \qquad 0 \leq t \leq \tau_*, \tag{2.7}$$

for a function $f \in C[0,a_*] \cap C^2(0,a_*)$ satisfying the boundary conditions

$$f(0) = 0, \quad f(a_*) = \begin{cases} \gamma a_* & \text{if } 0 < C_o < C_a, \\ \max\{K, \gamma a_*\} & \text{if } K \leq C_a \leq C_o. \end{cases} \tag{2.8}$$

If there is no arbitrage, then

$$\mathcal{N}f(x) = c \text{ for } 0 < x < a_*. \tag{2.9}$$

Conversely, if the function f satisfies (2.8) *and* (2.9) *and the derivative f' is bounded on $(0, a_*)$, then there is no arbitrage.*

Proof. Assume that the price of the bond is $D(t) = f(X(t))$ for a function $f \in C[0,a_*] \cap C^2(0,a_*)$ satisfying (2.8). In particular, the value of the equity is $S(t) = X(t) - f(X(t))$ for $0 \leq t \leq \tau_*$. Taking (2.3) into account, we see that the value

$V(t)$ of a self-financing portfolio starting with initial capital $V(0) = 0$ and containing $\Delta_1(t)$ shares of stock and $\Delta_2(t)$ units of convertible bond evolves according to

$$\begin{aligned} dV(t) = {} & \Delta_1(t)\big[d\big(X(t) - f(X(t))\big) + \delta\big(X(t) - f(X(t))\big)dt\big] \\ & + \Delta_2(t)\big[df(X(t)) + c\,dt\big] \\ & + r\big[V(t) - \Delta_1(t)\big(X(t) - f(X(t))\big) - \Delta_2(t)f(X(t))\big]dt. \end{aligned}$$

Therefore,

$$\begin{aligned} & d(e^{-rt}V(t)) \\ & = e^{-rt}\big[\Delta_1(t)\big(1 - f'(X(t))\big) + \Delta_2(t)f'(X(t))\big]dX(t) \\ & \quad + e^{-rt}\Delta_1(t)\left[-\frac{1}{2}\sigma^2X^2(t)f''(X(t)) - (r-\delta)X(t) + (r-\delta)f(X(t))\right]dt \\ & \quad + e^{-rt}\Delta_2(t)\left[\frac{1}{2}\sigma^2X^2(t)f''(X(t)) + c - rf(X(t))\right]dt. \end{aligned} \tag{2.10}$$

We choose $\Delta_1(t) = f'(X(t))\text{sgn}(\mathcal{N}f(X(t)) - c)$ and $\Delta_2(t) = -(1 - f'(X(t)))$ sgn $(\mathcal{N}f(X(t)) - c)$, so that $\Delta_1(t)(1 - f'(X(t))) + \Delta_2(t)f'(X(t)) = 0$. With these choices (2.10) becomes $d(e^{-rt}V(t)) = |\mathcal{N}f(X(t)) - c|\,dt$. This equation shows that the portfolio value $V(t)$ is bounded from below by $V(0) = 0$ and provides an arbitrage unless $\mathcal{N}f(x) = c$ for $0 < x < a_*$.

We now prove the converse. Assume $D(t) = f(X(t))$ for $0 \le t \le \tau_*$, and f satisfies (2.8) and (2.9). Let $\tau \le \tau_*$ be a bounded stopping time. Since $\frac{h(X(t))}{X(t)}$ and $\frac{f(X(t))}{X(t)}$ are bounded for $0 \le t \le \tau_*$, we can use Girsanov's theorem to construct an equivalent probability measure $\widetilde{\mathbb{P}}$ such that

$$\int_0^t \frac{h(X(s))}{X(s)}ds + \sigma W(t) = rt - \delta\int_0^t \left(1 - \frac{f(X(s))}{X(s)}\right)ds + \sigma\widetilde{W}(t) \tag{2.11}$$

for $0 \le t \le \tau$, where $\widetilde{W}$ is a Brownian motion under $\widetilde{\mathbb{P}}$. The differential of the value of the firm may be rewritten as

$$dX(t) = rX(t)dt - \delta(X(t) - f(X(t)))dt - cdt + \sigma X(t)d\widetilde{W}(t), \qquad 0 \le t \le \tau. \tag{2.12}$$

Let us consider the value $V(t)$ starting with initial capital $V(0) = 0$ corresponding to a self-financing trading strategy $\Delta_1(t)$, $\Delta_2(t)$ for $0 \le t \le \tau$. We can write the evolution of $V(t)$ as

$$\begin{aligned} d(e^{-rt}V(t)) = {} & \Delta_1(t)\big(d(e^{-rt}S(t)) + \delta e^{-rt}S(t)dt\big) \\ & + \Delta_2(t)\big(d(e^{-rt}D(t)) + ce^{-rt}dt\big). \end{aligned} \tag{2.13}$$

Since $D(t) = f(X(t))$, $S(t) = X(t) - f(X(t))$, and the function f is smooth, we can apply Itô's formula to obtain

$$\begin{aligned} & d(e^{-rt}S(t)) + \delta e^{-rt}S(t)dt \\ & \qquad = e^{-rt}\left(\mathcal{N}f(X(t)) - c\right)dt + e^{-rt}(1 - f'(X(t)))\sigma X(t)d\widetilde{W}(t), \end{aligned} \tag{2.14}$$

$$\begin{aligned} & d(e^{-rt}D(t)) + ce^{-rt}dt \\ & \qquad = -e^{-rt}\left(\mathcal{N}f(X(t)) - c\right) + e^{-rt}f'(X(t))\sigma X(t)d\widetilde{W}(t). \end{aligned} \tag{2.15}$$

We assume $\mathcal{N}f(x) - c = 0$ for $0 < x < a_*$, and taking into account (2.14), (2.15), and (2.13), we conclude that $e^{-r(t\wedge\tau)}V(t\wedge\tau)$ is a local martingale under $\widetilde{\mathbb{P}}$. But V is uniformly bounded below, and Fatou's lemma implies $\widetilde{\mathbb{E}}[e^{-r\tau}V(\tau)] \le V(0) = 0$. This means it is impossible to have $\widetilde{\mathbb{P}}\{V(\tau) \ge 0\} = 1$ and $\widetilde{\mathbb{P}}\{V(\tau) > 0\} > 0$. Since $\widetilde{\mathbb{P}}$ is equivalent to the probability measure $\mathbb{P}$, no arbitrage exists. □

In the remainder of this section we state the principal results of the paper. Their proofs are provided in section 7.

To compute the "no arbitrage" price of the convertible bond for some (not necessarily optimal) call and conversion levels, we need an existence and uniqueness result for boundary value problems associated with (2.9).

THEOREM 2.2. *Let y_1 be a positive number and $0 < y_1 \le x_1$. Then there exists a unique solution $f \in C[0,x_1] \cap C^2(0,x_1)$ of the boundary value problem*

$$\begin{cases} \mathcal{N}f(x) = c \text{ for } x \in (0,x_1), \\ f(0) = 0,\ f(x_1) = y_1. \end{cases} \tag{2.16}$$

Furthermore, the derivative f' is bounded on $(0,x_1)$. If $y_1 < x_1$, then $f'(x) < 1$ for all $x \in (0,x_1)$.

Taking into account Theorem 2.1, Theorem 2.2, and the discussion regarding the price of the bond at call or conversion time, we see that once the call and conversion levels have been set, the "no-arbitrage" price of the convertible bond is

$$D(t) = f(X(t), C_a, C_o), \tag{2.17}$$

where the function $f(x, C_a, C_o)$ is given in the next definition.

DEFINITION 2.3.

(i) *If $0 < C_o < C_a$, define $f(x, C_a, C_o)$ for $0 \le x \le C_o$ to be the unique solution of the equation $\mathcal{N}f = c$ on $(0, C_o)$ satisfying the boundary conditions $f(0) = 0$, $f(C_o) = \gamma C_o$. For $x \ge C_o$, define*

$$f(x, C_a, C_o) = \begin{cases} \gamma x, & C_o \le x < C_a, \\ \max\{K, \gamma x\}, & x \ge C_a. \end{cases}$$

(ii) *If $K \le C_a \le C_o$, define $f(x, C_a, C_o)$ for $0 \le x \le C_a$ to be the unique solution of the equation $\mathcal{N}f = c$ on $(0, C_a)$ satisfying the boundary conditions $f(0) = 0$, $f(C_a) = \max\{K, \gamma C_a\}$. For $x \ge C_a$, define $f(x, C_a, C_o) = \max\{K, \gamma x\}$.*

Equation (2.17) provides a bond price once the call and conversion levels C_a and C_o have been chosen. The firm wishes to minimize the value of the bond (in order to maximize the value of equity), and the bondholder wishes to maximize the value of the bond. This creates a two-person game, and according to the next theorem, this game has a value.

THEOREM 2.4. *There exist $C_a^* \ge K$ and $C_o^* > 0$ such that for each $x \ge 0$, we have*

$$f(x, C_a^*, C_o^*) = \inf_{C_a \ge K} f(x, C_a, C_o^*) = \sup_{C_o > 0} f(x, C_a^*, C_o). \tag{2.18}$$

Equation (2.18) implies the following equalities, so we can define

$$f_*(x) \triangleq f(x, C_a^*, C_o^*) = \sup_{C_o>0} \inf_{C_a \ge K} f(x, C_a, C_o) = \inf_{C_a \ge K} \sup_{C_o > 0} f(x, C_a, C_o). \tag{2.19}$$

This is the price of the bond as a function of the underlying firm value x, and C_a^* and C_o^* are the optimal call and optimal conversion levels, respectively.

THEOREM 2.5. *The function f_* is in $C[0,\infty)$ and is described by one of three cases. There are two constants $0 \le K_1 < K_2$ depending on r, δ, σ, c, and γ.*

(i) *If $K > K_2$, then $f_* \in C^1(0,\infty)$ and satisfies*

$$0 < f_*'(x) < 1 \text{ for } x > 0. \tag{2.20}$$

In this case,

$$C_o^* = \min\{x > 0; f_*(x) = \gamma x\} = \frac{K_2}{\gamma},$$

f_* restricted to $(0, C_o^*)$ is the unique classical solution of $\mathcal{N}f_* = c$ on $(0, C_o^*)$ with boundary conditions $f_*(0) = 0$ and $f_*(C_o^*) = \gamma C_o^*$,

$$f_*(x) = \gamma x \text{ for } x \ge C_o^*, \tag{2.21}$$

and $C_a^* = \frac{K}{\gamma} > C_o^* = \frac{K_2}{\gamma}$.

(ii) *If $K_1 \le K \le K_2$, then f_* restricted to $(0, K/\gamma)$ is the unique classical solution of $\mathcal{N}f_* = c$ on $(0, K/\gamma)$ with the boundary conditions $f_*(0) = 0$ and $f_*(K/\gamma) = K$. We have*

$$0 < f_*'(x) < 1 \text{ for } 0 < x < \frac{K}{\gamma}, \tag{2.22}$$

$$f_*(x) = \gamma x \text{ for } x \ge \frac{K}{\gamma}. \tag{2.23}$$

In this case, $C_o^* = C_a^* = \frac{K}{\gamma}$.

(iii) *If $K_1 > 0$, there is a third case. A sufficient condition for $K_1 > 0$ is $0 < \gamma < \frac{1}{2}$. In the third case, $0 < K < K_1$, f_* restricted to $(0, K/\gamma)$ is continuously differentiable, $C_a^* \in (K, K/\gamma)$, and f_* restricted to $(0, C_a^*)$ is the unique solution of $\mathcal{N}f_* = c$ on $(0, C_a^*)$ with the boundary conditions $f_*(0) = 0$, $f_*(C_a) = K$. We have*

$$0 < f_*'(x) < 1 \text{ for } 0 < x < C_a^*, \tag{2.24}$$

$$f_*(x) = \begin{cases} K, & C_a^* \le x \le \frac{K}{\gamma}, \\ \gamma x, & x \ge \frac{K}{\gamma}. \end{cases} \tag{2.25}$$

In particular, $f_*'(C_a^*-) = 0$ and $K < C_a^* < C_o^* = \frac{K}{\gamma}$.

From Theorem 2.5 we see that the firm debt at time t is $D(t) = f_*(X(t))$, and (2.2) becomes

$$S(t) = X(t) - f_*(X(t)). \tag{2.26}$$

So long as $x \in (0, C_a^* \wedge C_o^*)$, the function $F(x) \triangleq x - f_*(x)$ is strictly increasing because of (2.20), (2.22), and (2.24) and hence has an inverse F^{-1}. We may invert (2.26) to obtain $X(t) = F^{-1}(S(t))$, and thereby obtain a formula for the market price of the convertible bond in terms of the equity of the firm: $D(t) = f_*(F^{-1}(S(t)))$. In all three cases of Theorem 2.5, the firm should call as soon as $D(t)$ rises to the call price K. In cases (i) and (ii), this is the first time the conversion value of the bond rises to the call price. In case (iii), the call should occur before the conversion value rises to the call price. The owner of the bond should convert as soon as $D(t) - \gamma F^{-1}(S(t))$ falls to zero, i.e., as soon as the difference between the bond price and the bond's conversion value falls to zero.

3. Generation of candidate functions. Theorem 2.5 asserts that for small values of x, the value $f_*(x)$ of the convertible bond satisfies the second-order ordinary differential equation $\mathcal{N}f(x) = c$. Not only is this equation nonlinear, it is also singular at $x = 0$. Rather than solving the differential equation $\mathcal{N}f(x) = c$ directly, we generate a one-parameter family of solutions to the variational inequality

$$\min\{\mathcal{N}f(x) - c, f(x) - \gamma x\} = 0. \tag{3.1}$$

To do this, we first construct for a fixed function $g \in C[0,a]$, a solution to the variational inequality

$$\min\{\mathcal{L}_g f(x) - c, f(x) - \gamma x\} = 0, \tag{3.2}$$

subject to boundary conditions $f(0) = 0,\ f(a) = \gamma a$. Here, the *linear* differential operator $\mathcal{L}_g$ is defined by

$$\mathcal{L}_g f(x) \triangleq rf(x) - (rx - c)f'(x) + \delta(x - g(x))f'(x) - \frac{1}{2}\sigma^2 x^2 f''(x). \tag{3.3}$$

In section 6 we prove existence of a function g for which the solution to this equation is g itself.

DEFINITION 3.1. *Let $a \in (0,\infty)$ be given. Denote $\mathcal{D}_a = [0,a]$ and let $\mathcal{G}_a$ be the set of continuous functions $g\colon \mathcal{D}_a \to \mathbb{R}$ which are continuously differentiable on $(0,a)$ and satisfy*

$$\begin{aligned} &g(0) = 0, \quad g(a) = \gamma a, \\ &g(x) \geq \gamma x, \quad -M_a \leq g'(x) < 1 \quad \forall x \in (0,a), \end{aligned}$$

where M_a will be defined in Proposition 5.6. We denote by $\overline{\mathcal{G}}_a$ the closure of $\mathcal{G}_a$ with respect to the supremum norm in $C[0,a]$.

Denote $\mathcal{D}_\infty = [0,\infty)$ and let $\mathcal{G}_\infty$ be the set of continuous functions $g\colon \mathcal{D}_\infty \to \mathbb{R}$ which are continuously differentiable on $(0,\infty)$ and satisfy

$$\begin{aligned} &g(0) = 0, \quad g(x) = \gamma x \quad \forall x \in [b_g,\infty), \\ &g(x) \geq \gamma x, \quad 0 \leq g'(x) < 1 \quad \forall x \in (0,\infty), \end{aligned}$$

where b_g is a finite number depending on the function g. Let (C_γ, d) be the complete metric space of continuous functions on $\mathcal{D}_\infty$ which satisfy $\lim_{x\to\infty}[g(x) - \gamma x] = 0$, and d is the supremum metric. We denote by $\overline{\mathcal{G}}_\infty$ the closure of $\mathcal{G}_\infty$ in (C_γ, d).

For $a \in (0,\infty]$, $g \in \overline{\mathcal{G}}_a$, and $x \in \mathcal{D}_a$, we define $X^x(t)$ by $X^x(0) = x$ and

$$dX^x(t) = rX^x(t)\,dt - \delta\left(X^x(t) - g(X^x(t))\right)\,dt - c\,dt + \sigma X^x(t)\,dW(t) \tag{3.4}$$

for $0 \leq t \leq \tau_0^x \wedge \tau_a^x$, where $\tau_y^x \triangleq \inf\{t \geq 0; X^x(t) = y\}$. We then set

$$T_a g(x) \triangleq \sup_{0\leq\tau\leq\tau_0^x\wedge\tau_a^x} \mathbb{E}\left[\int_0^\tau e^{-ru}c\,du + \mathbb{I}_{\{\tau<\infty\}}e^{-r\tau}\gamma X^x(\tau)\right], \tag{3.5}$$

where the supremum is over stopping times τ which satisfy $0 \leq \tau \leq \tau_0^x \wedge \tau_a^x$.

We interpret the objects in Definition 3.1 as follows. Suppose we have a function g which maps the value of the firm into the value of convertible bond. Then $S(t)$ in (2.2) is given by $S(t) = X(t) - g(X(t))$. As we have already seen in the proof of

Theorem 2.1 (see (2.12)), under a "risk-neutral" measure, we expect the value of the firm to have mean rate of growth equal to the interest rate r, reduced by the dividend and coupon payments. In other words, if $g(x)$ is the value of the bond when x is the value of the firm, then the evolution of the value of the firm should be given by (3.4).

The fortunes of the firm, which depend on the function g and the initial condition x, may result in bankruptcy at time τ_0^x. If bankruptcy never occurs, then $\tau_0^x = \infty$. The bondholder collects dividends at rate c until bankruptcy occurs or until he converts the bond to stock. He may convert at any stopping time $\tau \le \tau_a^x$; if he has not converted by the time τ_a, he must do so at this time. The parameter a in this restriction on the stopping time τ will allow us to construct a one-parameter family of solutions to (2.4), and we shall later see that the correct choice of the parameter a depends on the call price K. However, in this interpretation of the function $T_a g$, we do not permit the firm to call. Since the conversion option is worthless after bankruptcy, we assume without loss of generality that $0 \le \tau \le \tau_0^x$. Upon conversion, the bondholder receives stock valued at $\gamma X^x(\tau)$. It follows that the risk-neutral value of a conversion strategy τ is $\mathbb{E}\left[\int_0^\tau e^{-ru} c\, du + \mathbb{I}_{\{\tau<\infty\}} e^{-r\tau}\gamma X^x(\tau)\right]$, and $T_a g(x)$ is the value of the optimal conversion strategy, if it exists.

We began this discussion with the supposition that $g(x)$ is the value of the convertible bond when x is the value of the firm. But the value of the convertible bond should be the risk-neutral discounted value of coupons collected plus the risk-neutral discounted value of the stock received upon conversion. In other words, we seek a function $f \in \overline{\mathcal{G}}_a$ such that $T_a f = f$. Such a function will satisfy (2.9), at least for small values of x.

In section 4 we prove continuity of the function $T_a g$. In section 5 we show that, like g, the function $T_a g$ is in $\overline{\mathcal{G}}_a$, and we state the *Hamilton–Jacobi–Bellman equation* (3.2) satisfied by $T_a g$. In section 6, we show that the mapping $T_a \colon \overline{\mathcal{G}}_a \to \overline{\mathcal{G}}_a$ has a unique fixed point, which we call f_a. Section 7 shows that for each call price K, there is a value of a so that f_a is a part of the function described in Theorem 2.5. This enables us to prove Theorems 2.4 and 2.5. Finally, at the end of section 7 we also prove Theorem 2.2.

4. Continuity of candidate functions. Let $a \in (0,\infty]$ and $g \in \overline{\mathcal{G}}_a$ be given, and define $T_a g$ by (3.5). If a is finite, we extend g to be constant on $(-\infty, 0]$ and on $[a,\infty)$. Since the extended g is Lipschitz, we may use (3.4) to define $X^x(t)$ for all $t \ge 0$. The assumptions on g ensure that for some $\eta > 0$, $\delta(x - g(x)) + c \ge \eta x$ for all $x \ge 0$. We now set $Z(t) = \exp\left\{-\sigma W(t) - \frac{1}{2}\sigma^2 t\right\}$, so that $d(Z(t)X^x(t)) \le (r - \sigma^2 - \eta) Z(t) X^x(t)\, dt$ for all $0 \le t \le \tau_0^x$. Integration yields

$$Z(t)X^x(t) \le x + \left(r - \sigma^2 - \eta\right) \int_0^t Z(u)X^x(u)\, du, \quad 0 \le t \le \tau_0^x,$$

and an application of Gronwall's inequality gives the bound

$$X^x(t) \le \frac{x}{Z(t)} e^{\left(r-\sigma^2-\eta\right)t} = x \exp\left\{\sigma W(t) + \left(r - \frac{1}{2}\sigma^2 - \eta\right) t\right\}, \ 0 \le t \le \tau_0^x. \tag{4.1}$$

LEMMA 4.1. *The function $T_a g$ satisfies the bounds*

$$\gamma x \le T_a g(x) \le \frac{c}{r} + \gamma x \quad \forall x \in \mathcal{D}_a. \tag{4.2}$$

Proof. The lower bound in (4.2) follows from taking $\tau \equiv 0$ in (3.5). For the upper bound, we apply the optional sampling theorem and Fatou's lemma to the martingale

$\exp\left\{\sigma W(t) - \frac{1}{2}\sigma^2 t\right\}$ and use (4.1) to obtain for any stopping time τ satisfying $0 \leq \tau \leq \tau_0^x$

$$\begin{aligned}(4.3)\qquad \mathbb{E}e^{-r\tau}X^x(\tau) &\leq x\mathbb{E}\exp\left\{\sigma W(\tau) - \frac{1}{2}\sigma^2\tau\right\} \\ &\leq x\liminf_{t\to\infty}\mathbb{E}\exp\left\{\sigma W(t\wedge\tau) - \frac{1}{2}\sigma^2(t\wedge\tau)\right\} \;=\; x.\end{aligned}$$

Therefore,

$$T_a g(x) \leq \int_0^\infty e^{-ru}c\,du + \gamma \sup_{0\leq\tau\leq\tau_0^x}\mathbb{E}e^{-r\tau}X^x(\tau) \leq \frac{c}{r} + \gamma x.$$

LEMMA 4.2. *For all $y \geq 0$, τ_y^x is almost surely continuous in x at all $x \geq 0$.*

Proof. It is possible to choose for each initial condition a version of the process $X^x(t)$, $t \geq 0$, such that $X^x(t)$ is jointly continuous in (t,x), almost surely (see [15, Theorem 4.2.5]). Because of the uniqueness of the solution to (3.4), we have for $0 \leq \xi < x \leq y$ that $X^\xi(t) \leq X^x(t)$, $0 \leq t < \infty$, almost surely; if these processes ever coalesce, they would henceforth coincide. This implies that $\lim_{\xi\uparrow x}\tau_y^\xi \geq \tau_y^x$. On the other hand, $\tau_y^x = \inf\{t \geq 0; X^x(t) > y\}$, which implies that $\lim_{\xi\uparrow x}\tau_y^\xi \leq \tau_y^x$. Therefore,

$$(4.4)\qquad \lim_{\xi\uparrow x}\tau_y^\xi = \tau_y^x.$$

By a similar argument, we conclude

$$(4.5)\qquad \lim_{\xi\downarrow x}\tau_y^\xi = \tau_y^x.$$

Combining (4.4) and (4.5), we see that, almost surely, $\lim_{\xi\to x}\tau_y^\xi = \tau_y^x$, $0 \leq x < y$, and (4.4) holds for $x = y$. A similar argument shows that $\lim_{\xi\to x}\tau_y^\xi = \tau_y^x$, $0 \leq y < x$, and (4.5) holds for $x = y$. □

Using (4.1) to bound $e^{-rt}X^x(t)$, $\lim_{t\to\infty}\exp\{\sigma W(t) - \frac{1}{2}\sigma^2 t\} = 0$, joint continuity of $X^x(t)$ in (t,x), and Lemma 4.2, we conclude that the process

$$(4.6)\qquad Y^x(t) \triangleq \int_0^{t\wedge\tau_0^x\wedge\tau_a^x} e^{-ru}c\,du + \mathbb{I}_{\{t\wedge\tau_0^x\wedge\tau_a^x<\infty\}}e^{-r(t\wedge\tau_0^x\wedge\tau_a^x)}\gamma X^x(t\wedge\tau_0^x\wedge\tau_a^x)$$

is jointly continuous in $(t,x) \in [0,\infty] \times \mathcal{D}_a$, almost surely. In particular, we have continuity at time $t = \infty$, where

$$Y^x(\infty) \triangleq \int_0^{\tau_0^x\wedge\tau_a^x} e^{-ru}c\,du + \mathbb{I}_{\{\tau_0^x\wedge\tau_a^x<\infty\}}e^{-r(\tau_0^x\wedge\tau_a^x)}\gamma X^x(\tau_0^x\wedge\tau_a^x).$$

LEMMA 4.3. *The function $T_a g$ is lower semicontinuous on $\mathcal{D}_a$.*

Let τ be any nonnegative stopping time. Lemma 4.2 implies that $\tau\wedge\tau_0^x\wedge\tau_a^x$ is almost surely continuous in x. The function

$$h_{\tau,a} = \mathbb{E}\left[\int_0^{\tau\wedge\tau_0^x\wedge\tau_a^x} e^{-ru}c\,du + \mathbb{I}_{\{\tau\wedge\tau_0^x\wedge\tau_a^x<\infty\}}e^{-r(\tau\wedge\tau_0^x\wedge\tau_a^x)}\gamma X^x(\tau\wedge\tau_0^x\wedge\tau_a^x)\right]$$

is thus lower semicontinuous (Fatou's lemma), and $T_a g(x) = \sup_\tau h_{\tau,a}(x)$, the supremum of lower semicontinuous functions, is lower semicontinuous. □

We know from inequality (4.1) that

$$\sup_{0\leq t\leq\infty} Y^x(t) \leq \frac{c}{r} + \gamma x \sup_{t\geq 0} \exp\left\{\sigma W(t) - \left(\eta + \frac{1}{2}\sigma^2\right)t\right\} = \frac{c}{r} + \gamma x e^{\sigma W^*}, \tag{4.7}$$

where $W^* = \sup_{t\geq 0}\left[W(t) - \left(\frac{\sigma}{2} + \frac{\eta}{\sigma}\right)t\right]$. According to [13, Exercise 5.9, Chapter 3], W^* has density

$$\mathbb{P}\{W^* \in db\} = 2\left(\frac{\sigma}{2} + \frac{\eta}{\sigma}\right)\exp\left\{-2\left(\frac{\sigma}{2} + \frac{\eta}{\sigma}\right)b\right\} db, \qquad b > 0. \tag{4.8}$$

This means that $\mathbb{E}e^{\sigma W^*} < \infty$, so we obtain

$$\mathbb{E}\sup_{0\leq t\leq\infty} Y^x(t) < \infty. \tag{4.9}$$

In light of Lemmas 4.1 and 4.3, the set

$$\mathcal{S}_g \triangleq \{x \in \mathcal{D}_a; T_a g(x) = \gamma x\} = \{x \in \mathcal{D}_a : T_a g(x) \leq \gamma x\}$$

is closed, contains the origin, and contains a if a is finite. We define

$$\tau_*^x \triangleq \inf\{t \geq 0; X^x(t) \in \mathcal{S}_g\}, \tag{4.10}$$

a stopping time satisfying $\tau_*^x \leq \tau_0^x \wedge \tau_a^x$. Since inequality (4.9) holds, it is known from the general theory of optimal stopping that the process

$$Z^x(t) \triangleq \int_0^{t\wedge\tau_0^x\wedge\tau_a^x} e^{-ru} c\, du + \mathbb{I}_{\{t\wedge\tau_0^x\wedge\tau_a^x<\infty\}} e^{-r(t\wedge\tau_0^x\wedge\tau_a^x)} T_a g(X^x(t\wedge\tau_0^x\wedge\tau_a^x)) \tag{4.11}$$

is a supermartingale for $0 \leq t \leq \infty$; the stopped process $Z^x(t\wedge\tau_*^x)$, $0 \leq t \leq \infty$, is a martingale; and τ_*^x is an optimal stopping time, i.e.,

$$T_a g(x) = \mathbb{E}\left[\int_0^{\tau_*^x} e^{-ru} c du + \mathbb{I}_{\{\tau_*^x<\infty\}} e^{-r\tau_*^x}\gamma X^x(\tau_*^x)\right] = \mathbb{E}Y^x(\tau_*^x). \tag{4.12}$$

To prove this, one can first show, using the Markov property, that the process $\{Z^x(t)\}_{0\leq t\leq\infty}$ is the *Snell envelope* of $\{Y^x(t)\}_{0\leq t\leq\infty}$, i.e.,

$$Z^x(t) = \operatorname{ess\,sup}_{\tau\geq t}\mathbb{E}\left[Y^x(\tau)|\mathcal{F}_t\right], \tag{4.13}$$

and then appeal to [14, Appendix D] . Another way to prove it is to combine Theorem 1, page 124, and Theorem 3, page 127, from [20].

LEMMA 4.4. *Assume $a = \infty$. We have*

$$\gamma x \leq T_\infty g(x) \leq x \quad \forall x \in \mathcal{D}_\infty, \tag{4.14}$$

and there is a number $b > 0$ such that

$$T_\infty g(x) = \gamma x \quad \forall x \in [b, \infty). \tag{4.15}$$

If $a \in (0, \infty)$, we have

$$\gamma x \leq T_a g(x) \leq x \quad \forall x \in \mathcal{D}_a. \tag{4.16}$$

Proof. We shall construct a number $b > 0$ and a function $\varphi\colon [0,\infty) \mapsto \mathbb{R}$ such that

$$\gamma x \le \varphi(x) \le x \ \forall x \in [0,b], \quad \varphi(x) = \gamma x \ \forall x \in [b,\infty); \tag{4.17}$$

φ'' is defined and continuous on $[0,\infty)$, except at $\sqrt{b}$ and b, but has one-sided derivatives at these points; φ' is defined, bounded, and continuous on $[0,\infty)$ except at $\sqrt{b}$, but has one-sided derivatives at this point which satisfy

$$D^-\varphi(\sqrt{b}) - D^+\varphi(\sqrt{b}) > 0, \tag{4.18}$$

$$\mathcal{L}_g\varphi(x) \ge c \quad \forall x \in [0,\infty) \setminus \{\sqrt{b}, b\}. \tag{4.19}$$

Once b and φ are constructed, we choose an arbitrary $x \ge 0$. With $X(t) = X^x(t)$, the extension of Itô's rule to continuous, piecewise C^2 functions [13, Chapter 3, Theorem 7.1 and Corollary 7.2] implies that

$$\begin{aligned} d\left(e^{-rt}\varphi(X(t))\right) = -e^{-rt}\mathcal{L}_g\varphi(X(t))\,dt - e^{-rt}\left(D^-\varphi(\sqrt{b}) - D^+\varphi(\sqrt{b})\right)d\Lambda(t) \\ + e^{-rt}\sigma X(t)\varphi'(X(t))\,dW(t), \end{aligned}$$

where $\Lambda(t)$ is the (nondecreasing) local time of X at $\sqrt{b}$. From (4.18) and (4.19), we see that

$$d\left(e^{-rt}\varphi(X(t))\right) \le -e^{-rt}c\,dt + e^{-rt}\sigma X(t)\varphi'(X(t))\,dW(t).$$

Hence, for any stopping time $\tau \le \tau_0^x$ and any deterministic time T, we have

$$\mathbb{E}e^{-r(\tau\wedge T)}\varphi(X(\tau\wedge T)) \le \varphi(x) - \mathbb{E}\int_0^{\tau\wedge T} e^{-rt}c\,dt,$$

where we have used the boundedness of φ' and (4.1) to ensure that the expectation of the Itô integral is zero. This last inequality implies

$$\varphi(x) \ge \mathbb{E}\left[\int_0^{\tau\wedge T} e^{-rt}c\,dt + \mathbb{I}_{\{\tau<\infty\}}e^{-r(\tau\wedge T)}\gamma X(\tau\wedge T)\right].$$

Letting $T \to \infty$ and using Fatou's lemma, then maximizing over τ, we obtain $\varphi(x) \ge T_a g(x)$. Relations (4.14), (4.15) follow from (4.2) and (4.17).

If $a \in (0,\infty)$, then the function $h(x) = x$ on $[0,a]$ is two times continuously differentiable on $(0,a)$ and satisfies $\mathcal{L}_g h(x) \ge c$. Since $h(x) \ge \gamma x$ for each $0 \le x \le a$, we can do the same computation as above for the function h instead of φ and obtain (4.16).

The remainder of the proof is the construction of b and φ. For $b > e^2$, define the positive function $\eta(b) \triangleq (1-\gamma)/\left(\frac{1}{2}\log b + \frac{1}{\sqrt{b}} - 1\right)$. Consider the function

$$k(b) \triangleq c\,[\gamma - \eta(b) - 1] + \frac{1}{2}\delta\sqrt{b}(1-\gamma)\,(\gamma - \eta(b)) - \frac{1}{2}\sigma^2\eta(b)\sqrt{b}.$$

Since $\lim_{b\to\infty}\eta(b) = 0$, we have $\lim_{b\to\infty}k(b) = \infty$. We fix a value $b > e^2$ for which $k(b) > 0,\ \eta(b) < \gamma$. For any $g \in \overline{\mathcal{G}}_a$ we know that $\lim_{x\to\infty}[g(x) - \gamma x] = 0$, so for b

sufficiently large, we also have

$$x - g(x) \geq \frac{1}{2}(1-\gamma)\sqrt{b} \quad \forall x \in [\sqrt{b}, \infty), \tag{4.20}$$

$$\delta(x - g(x)) \geq \frac{(1-\gamma)c}{\gamma} \quad \forall x \in [b, \infty). \tag{4.21}$$

With b chosen to satisfy all the above properties, we set

$$\varphi(x) = \begin{cases} x, & 0 \leq x \leq \sqrt{b}, \\ \gamma x + \eta(b)\sqrt{b}\left(\frac{x}{b} - \log\frac{x}{b} - 1\right), & \sqrt{b} < x < b, \\ \gamma x, & x \geq b. \end{cases} \tag{4.22}$$

A straightforward computation verifies that φ has the desired properties.

COROLLARY 4.5. *The function $T_a g$ is continuous on $\mathcal{D}_a$.*

Proof. Recall from the proof of Lemma 4.3 that for each $y \geq 0$, the stopping time τ_y^x is a continuous function of x. The complement of the closed set $\mathcal{S}_g$,

$$\mathcal{C}_g \triangleq \{x \in \mathcal{D}_a; T_a g(x) > \gamma x\},$$

is a countable union of disjoint open intervals, and on each of these intervals (α, β), we have $\tau_*^x = \tau_\alpha^x \wedge \tau_\beta^x$, which is a continuous function of $x \in [\alpha, \beta]$. On the set $\mathcal{S}_g$, $\tau_*^x \equiv 0$. Hence, τ_*^x is continuous on both $\mathcal{S}_g$ and its complement $\mathcal{C}_g$. To show that τ_*^x is continuous on $\mathcal{D}_a = \mathcal{C}_g \cup \mathcal{S}_g$, it remains only to show that if $\{x_n\}_{n=1}^\infty$ is a sequence in $\mathcal{C}_g$ converging to $x \in \mathcal{S}_g$, then $\tau_*^{x_n} \to \tau_*^x = 0$. But $\tau_*^{x_n} \leq \tau_x^{x_n}$ and $\tau_x^{x_n} \to \tau_x^x = 0$, almost surely (Lemma 4.2), so the desired result holds.

For $a < \infty$ we have $0 \leq X^x(t \wedge \tau_*^x) \leq a$. For $a = \infty$, Lemma 4.4 implies there exists $b > 0$ such that $[b, \infty) \subset \mathcal{S}_g$. In this case, $0 \leq X^x(t \wedge \tau_*^x) \leq \max\{x, b\}$. The continuity of $T_a g$ follows from the representation (4.12), the continuity of τ_*^x, the joint continuity of $Y^x(t)$ on $[0, \infty] \times \mathcal{D}_a$, and the dominated convergence theorem. ∎

PROPOSITION 4.6. *The function $T_a g$ is twice continuously differentiable on $\mathcal{C}_g$ and satisfies the equation*

$$\mathcal{L}_g T_a g = c \text{ on } \mathcal{C}_g. \tag{4.23}$$

If $g \in \mathcal{G}_a$, then $T_a g$ is three times continuously differentiable on $\mathcal{C}_g$.

Proof. Let $x \in \mathcal{C}_g$ be given, and choose $0 < \alpha < x < \beta$ such that $(\alpha, \beta) \subset \mathcal{C}_g$. Consider the linear, second-order ordinary differential equation

$$\mathcal{L}_g h(x) = c \quad \forall x \in (\alpha, \beta), \tag{4.24}$$

with the boundary conditions $h(\alpha) = T_a g(\alpha)$, $h(\beta) = T_a g(\beta)$. Because the coefficients of (4.24) are continuous, the equation has a twice continuously differentiable solution h satisfying these boundary conditions. If $g \in \mathcal{G}_a$, so that the coefficients of (4.24) are continuously differentiable, then h is three times continuously differentiable. Itô's formula implies that

$$\begin{aligned} d\left[e^{-rt}h(X^x(t))\right] &= e^{-rt}\left[-\mathcal{L}_g h(X^x(t))\,dt + \sigma X^x(t) h'(X^x(t))\,dW(t)\right] \\ &= -e^{-rt}c\,dt + e^{-rt}\sigma X^x(t)h'(X^x(t))\,dW(t). \end{aligned}$$

Integrating this equation from $t = 0$ to $t = \tau_\alpha^x \wedge \tau_\beta^x$ and taking expectations, we obtain

$$\begin{aligned} h(x) &= \mathbb{E}\left[\int_0^{\tau_\alpha^x \wedge \tau_\beta^x} e^{-rt} c\, dt + e^{-r(\tau_\alpha^x \wedge \tau_\beta^x)} h(X^x(\tau_\alpha^x \wedge \tau_\beta^x))\right] \\ &= \mathbb{E}\left[\int_0^{\tau_\alpha^x \wedge \tau_\beta^x} e^{-rt} c\, dt + e^{-r(\tau_\alpha^x \wedge \tau_\beta^x)} T_a g(X^x(\tau_\alpha^x \wedge \tau_\beta^x))\right] \\ &= \mathbb{E} Z^x(\tau_\alpha^x \wedge \tau_\beta^x) \;=\; Z^x(0) \;=\; T_a g(x), \end{aligned}$$

where we have used the fact that $Z^x(t \wedge \tau_\alpha^x \wedge \tau_\beta^x)$ is a bounded martingale, since $\tau_\alpha^x \wedge \tau_\beta^x \le \tau_*^x$. □

Remark 4.7. Let us denote by $D^{\pm} T_a g$ the derivatives from the right and left of $T_a g$, when these one-sided derivatives exist. We likewise denote by $DT_a g$ the derivative of $T_a g$, when the derivative exists. Because it is open, the set $\mathcal{C}_g$ is a countable union of disjoint open intervals, which we call the *components* of $\mathcal{C}_g$. Let (α, β) be one of these components. The second-order differential operator $\mathcal{L}_g$ does not degenerate to a first-order operator at any point in $[\alpha, \beta]$, except at α when $\alpha = 0$. Therefore, the function h in the proof of Proposition 4.6 is twice continuously differentiable at the endpoint β and also at α provided that $\alpha > 0$. We conclude that $D^- T_a g(\beta) = \lim_{x \uparrow \beta} DT_a g(x)$ exists. If $\alpha > 0$, then $D^+ T_a g(\alpha) = \lim_{x \downarrow \alpha} DT_a g(x)$ also exists.

5. The invariance property of T_a. As in the previous section, let $a \in (0, \infty]$ and $g \in \overline{\mathcal{G}}_a$ be given, and define $T_a g$ by (3.5). In this section we show that T_a maps $\mathcal{G}_a$ into itself. For this we use the theory of viscosity solutions of Hamilton–Jacobi–Bellman equations developed by Crandall and Lions (see [7], [9]). The proof of Proposition 5.1 below is standard, so we omit it. See [19] for a similar proof.

PROPOSITION 5.1. *The function $T_a g$ is a viscosity solution of the Hamilton–Jacobi–Bellman equation*

$$\min\{\mathcal{L}_g h(x) - c, h(x) - \gamma x\} = 0 \quad \forall x \in \mathcal{D}_a. \tag{5.1}$$

We use Proposition 5.1 to deduce other information about $T_a g$.

COROLLARY 5.2. *Given any $b \in (0, a)$, the set $\mathcal{C}_g \cap (0, b)$ is nonempty.*

Proof. Suppose $T_a g(x) = \gamma x$ for all $x \in [0, b]$. Then

$$\mathcal{L}_g T_a g(x) - c = (\gamma - 1)c + \delta\gamma(x - g(x)),$$

which is strictly negative for $x > 0$ sufficiently small. This violates the viscosity supersolution property of $T_a g$. □

LEMMA 5.3. *If $(0, a) \cap \mathcal{S}_g$ contains a point b, then $[b, \infty) \cap \mathcal{D}_a \subset \mathcal{S}_g$.*

Proof. Assume $b \in (0, a) \cap \mathcal{S}_g$ and denote $\overline{\varphi}(x) = \gamma x$. Because $T_a g(b) = \overline{\varphi}(b)$ and $T_a g \ge \overline{\varphi}$, the viscosity supersolution property for $T_a g$ implies

$$c \le \mathcal{L}_g \overline{\varphi}(b) = c\gamma + \delta\gamma(b - g(b)).$$

But the function $x \to x - g(x)$ is nondecreasing on $\mathcal{D}_a$. Therefore

$$c \le c\gamma + \delta\gamma(x - g(x)) \quad \forall x \in [b, \infty) \cap \mathcal{D}_a.$$

We must show that $T_a g(x) \le \overline{\varphi}(x)$ for all $x \in [b, \infty) \cap \mathcal{D}_a$. Assume on the contrary that $\eta \triangleq \sup\{T_a g(x) - \overline{\varphi}(x); x \in [b, \infty) \cap \mathcal{D}_a\}$ is positive and let x_0 attain

the supremum in the definition of η. (The supremum is attained because both $T_a g$ and $\overline{\varphi}$ are continuous, and if $a = \infty$, then $T_a g(x) = \overline{\varphi}(x)$ for all sufficiently large x.)

We take $\varphi(x) = \overline{\varphi}(x) + \eta$ for $x \in [b, \infty) \cap \mathcal{D}_a$, so that $\varphi(x) \geq T_a g(x)$ for $x \in [b, \infty) \cap \mathcal{D}_a$ and $\varphi(x_0) = T_a g(x_0)$. We have $\varphi(b) > T_a g(b)$ and can choose φ on $(0, b)$ so that it is twice continuously differentiable and dominates $T_a g$ on all of $(0, a)$. Because $T_a g$ is a viscosity subsolution of (5.1) and $\varphi(x_0) = T_a g(x_0) > \gamma x_0$, we obtain

$$\mathcal{L}_g \varphi(x_0) = r T_a g(x_0) - \gamma(r x_0 - c) + \delta\gamma(x_0 - g(x_0)) \leq c \leq \gamma c + \delta\gamma(x_0 - g(x_0)),$$

and hence $T_a g(x_0) \leq \gamma x_0$, a contradiction to the choice of x_0. We conclude that $T_a g(x) \leq \overline{\varphi}(x)$ for $x \in [b, \infty) \cap \mathcal{D}_a$. □

From Corollary 5.2 and Lemmas 5.3 and 4.4, we have the following.

PROPOSITION 5.4. *If a is finite, then $\mathcal{C}_g = (0, b)$ for some $b \in (0, a]$ and $\mathcal{S}_g = \{0\} \cup [b, a]$. If $a = \infty$, then $\mathcal{C}_g = (0, b)$ for some $b \in (0, \infty)$ and $\mathcal{S}_g = \{0\} \cup [b, \infty)$.*

Let b be as in Proposition 5.4. We have already seen that $T_a g$ is twice continuously differentiable on $\mathcal{C}_g = (0, b)$ with a one-sided derivative $D^- T_a g(b)$ at b. Since $T_a g(x) = \gamma x$ on $\mathcal{S}_g$, this function is clearly differentiable on the set (b, a) if $b < a$, with one-sided derivative $D^+ T_a g(b) = \gamma$. It remains to examine the differentiability of $T_a g$ at the point b.

PROPOSITION 5.5 (smooth pasting). *The function $T_a g$ is continuously differentiable on $(0, a)$.*

Proof. It suffices to show in the case that $b < a$ that $D^- T_a g(b) = D^+ T_a g(b)$. Because $T_a g(x) \geq \gamma x$ for all $x \in \mathcal{D}_a$ and $T_a g(b) = \gamma b$, we must have $D^- T_a g(b) \leq \gamma$. If $D^- T_a g(b) < \gamma$, we choose $m \in (D^- T_a g(b), D^+ T_a g(b))$, $k > 0$, and define $\varphi(x) = \gamma b + m(x - b) + k(x - b)^2$ for x in an open interval containing b. Note that $\varphi(b) = T_a g(b)$ and $\varphi'(b) = m$. Therefore, $\varphi(x) < T_a g(x)$ for $x \neq b$ in a sufficiently small neighborhood of b (whose size depends on k). We construct φ outside this neighborhood so that φ is twice continuously differentiable on $(0, a)$ and $\varphi(x) \leq T_a g(x)$ for all $x \in (0, a)$. Because $T_a g$ is a viscosity supersolution of (5.1), the inequality

$$0 \leq \mathcal{L}_g \varphi(b) - c = r\gamma b - (rb - c)m + \delta(b - g(b))m - \sigma^2 b^2 k - c$$

must hold. Since $k > 0$ is arbitrary, this is impossible. □

We have proved so far the following properties of the value function $T_a g$: for any $g \in \overline{\mathcal{G}}_a$, $T_a g$ is a continuous function on $\mathcal{D}_a$ and it has a continuous derivative on $(0, a)$, $T_a g(0) = 0$ and $T_a g(x) \geq \gamma x$ for all $x \in \mathcal{D}_a$. If a is finite, then $T_a g(x) = \gamma a$; if $a = \infty$, then $T_\infty g(x) = \gamma x$ for x sufficiently large.

We now need to prove an invariance property for the operator T_a. Up to this point, we have taken g to be an arbitrary function in $\overline{\mathcal{G}}_a$. For the next proposition, we must restrict our attention to $g \in \mathcal{G}_a$.

PROPOSITION 5.6. *Let $a \in (0, \infty]$ be given. Then T_a maps $\mathcal{G}_a$ into $\mathcal{G}_a$.*

Proof. Assume that $g \in \mathcal{G}_a$. By the above remark, it remains only to show that $-M_a \leq D T_a g < 1$ on $(0, a)$ if a is finite and $0 \leq D T_a g < 1$ if $a = \infty$.

First we claim that the function $\psi = D T_a g$ (defined on $(0, a)$) cannot attain a positive local maximum or a negative local minimum in $\mathcal{C}_g$. By Proposition 4.6, ψ is C^2 on $\mathcal{C}_g$. Assume that ψ has a positive local maximum at $x_* \in \mathcal{C}_g$. Thus, we have $\psi'(x_*) = 0$. In particular,

$$\frac{d}{dx}(T_a g(x) - x\psi(x))\Big|_{x = x_*} = -x_* \psi'(x_*) = 0.$$

Equation (4.23) implies for $x \in \mathcal{C}_g$ that

$$\begin{aligned} c &= \mathcal{L}_g T_a g(x) \\ &= r(T_a g(x) - x\psi(x)) + c\psi(x) + \delta(x - g(x))\psi(x) - \frac{1}{2}\sigma^2 x^2 \psi'(x), \end{aligned}$$

and thus

$$0 = \frac{d}{dx}\mathcal{L}_g T_a g(x)\Big|_{x=x_*} = \delta(1 - g'(x_*))\psi(x_*) - \frac{1}{2}\sigma^2 x_*^2 \psi''(x_*).$$

Because ψ has a local maximum at x_*, $\psi''(x_*) \le 0$. But $1 - g'(x_*)$ is positive, and $\psi(x_*) > 0$. We have a contradiction, and hence ψ cannot have a positive local maximum in $\mathcal{C}_g$. If ψ has a negative local minimum at x_*, we likewise have a contradiction.

We consider now the case that $a = \infty$. For $x < y$ we have $X^x(t) \le X^y(t)$ almost surely and $\tau_0^x \le \tau_0^y$ almost surely. It follows from (3.5) that $T_\infty g$ is nondecreasing. The lower bound $DT_\infty g \ge 0$ is established. For the upper bound, $DT_\infty g(x) < 1$, we recall that $\mathcal{C}_g = (0, b)$ for some $b \in (0, \infty)$. Assume there were a point $x_0 \in (0, b)$ where $DT_\infty g(x_0) \ge 1$. We know that $DT_\infty g(b) = \gamma < 1$. Now consider a point $x_1 \in (0, x_0)$. If $DT_\infty g(x_1) < 1$, then $DT_\infty g$ would have a positive local maximum in the interval (x_1, b), which is impossible. We conclude that $DT_\infty g(x_1) \ge 1$. In other words, if there were a point $x_0 \in (0, b)$ where $DT_\infty g(x_0) \ge 1$, then $DT_\infty g \ge 1$ on the whole interval $(0, x_0)$. The upper bound in (4.14) would immediately imply that $T_\infty g(x) = x$ for $0 \le x \le x_0$, and once again $DT_\infty g$ would have a positive local maximum in $(0, b)$. We conclude that $DT_\infty g(x_0) < 1$ for all $x_0 \in (0, b)$.

If a is finite, we can modify the above argument, using (4.16) in place of (4.14) and $D^- T_a g(b) \le \gamma$ (in case $\mathcal{C}_g = (0, a)$), to obtain the upper bound $DT_a g < 1$ on $(0, a)$.

The proof of the lower bound $DT_a g(x) \ge -M_a$ for the case $a < \infty$ is more involved. Again using the notation $\mathcal{C}_g = (0, b)$, we assume there is $x_0 \in (0, b)$ such that $DT_a g(x_0) < 0$. Let $x_1 \in (0, x_0)$. The continuous function $DT_a g$ attains its minimum on $[x_1, b]$ at x_1 or b, since it cannot attain a negative interior minimum. In case the minimum is attained at x_1, this means that $DT_a g(x_1) < DT_a(x_0) < 0$. For any $0 < x_2 < x_1$, DT_g cannot attain a negative interior minimum on $[x_2, x_0]$, so we can conclude that $DT_a g(x_2) < DT_a g(x_1) < 0$. This should hold for any $0 < x_2 < x_1$, which is in contradiction to $T_a g(0) = 0$, $T_a g(x) \ge \gamma x$. So if $DT_a g(x_0) < 0$, then for any $x_1 \in (0, x_0)$, $DT_a g$ attains its negative minimum on $[x_1, b]$ at b. This means that

$$D^- T_a g(b) \le \inf_{0 < x \le b} D_a T g(x) < 0. \tag{5.2}$$

In other words, the derivative $D_a T g$ either is nonnegative or, if it has negative values, is bounded below by $DT_a^- g(b)$. Of course, the latter case can only happen for $b = a$. The first case satisfies the conclusion, so we assume

$$D^- T_a g(a) = \min_{x \in (0, a]} DT_a g(x) < 0. \tag{5.3}$$

This means that $\mathcal{C}_g = (0, a)$ and hence $\mathcal{L}_g T_a g(x) = c$ for all $x \in (0, a)$. Let h satisfy $\mathcal{L}_g h(x) = c$ for $x \in (\gamma a, a)$ and $h(\gamma a) = \gamma a$, $h(a) = \gamma a$. Since $T_a g(\gamma a) \le \gamma a = h(\gamma a)$, $T_a g(a) = \gamma a = h(a)$ and $\mathcal{L}_g T_a g(x) = \mathcal{L}_g h(x)$ for all $x \in (0, a)$, the usual comparison argument based on the maximum principle yields $T_a g(x) \le h(x)$ for all $x \in [\gamma a, a]$. But $T_a g(a) = h(a)$, and this implies

$$D^- T_a g(a) \ge D^- h(a). \tag{5.4}$$

It suffices to find a lower bound on $D^- h(a)$. We have $0 \le \gamma x \le T_a g(x) \le h(x)$ for $x \in [\gamma a, a]$. In order to find an upper bound on h, we let $x^* \in [\gamma a, a]$ be such that $h(x^*) = \max_{x\in[\gamma a,a]} h(x)$. If x^* is an interior point of $[\gamma a, a]$, then $h'(x^*) = 0$ and $h''(x^*) \le 0$. But $\mathcal{L}_g h(x^*) = c$ from which we conclude that $\max_{x\in[\gamma a,a]} h(x) = h(x^*) \le \frac{c}{r}$. If x^* is not an interior point of $[\gamma a, a]$, then $\max_{x\in[\gamma a,a]} h(x) \le h(\gamma a) = h(a) = \gamma a$. In either case, we have

$$0 \le h(x) \le \max\left\{\gamma a, \frac{c}{r}\right\} \quad \forall x \in [\gamma a, a]. \tag{5.5}$$

We know that

$$0 \le g(x) \le a \quad \forall\ x \in [\gamma a, a]. \tag{5.6}$$

Neither (5.5) nor (5.6) depends on the lower bound $-M_a \le g'(x)$ satisfied by functions g in $\mathcal{C}_a$ when a is finite.

Since $h(\gamma a) = h(a)$, there exists $x_0 \in (\gamma a, a)$ such that $h'(x_0) = 0$. We solve the equation $\mathcal{L}_g h = c$ on $(\gamma a, a)$ for h'' and then integrate to obtain

$$h'(x) = \int_{x_0}^{x} \frac{2}{\sigma^2 y^2}\big[rh(y) - (ry - c)h'(y) + \delta(y - g(y))h'(y) - c\big]\, dy \tag{5.7}$$

for all $x \in [x_0, a]$. Taking into account the bounds (5.5) and (5.6), we may use Gronwall's inequality to obtain $|h'(a)| \le M_a$ for some constant M_a depending only on the bounds $\max\{\gamma a, \frac{c}{r}\}$ and a appearing in (5.5) and (5.6) and also depending on the interval $[\gamma a, a]$. From (5.3), (5.4) we conclude that $DT_a g(x) \ge -M_a$ for all $x \in (0, a)$. □

Remark 5.7. M_a is bounded in a as long as a is bounded away from 0.

6. The fixed point property. For $a = \infty$ we recall that $\overline{\mathcal{G}}_\infty$ is a closed subset of the complete metric space (C_γ, d) (see Definition 3.1). For $a < \infty$, the set $\overline{\mathcal{G}}_a$ is a closed convex subset of the Banach space $C[0, a]$ endowed with the supremum norm. We denote by $d(f, g)$ the metric associated with the supremum norm. We have proved that $T_\infty(\overline{\mathcal{G}}_\infty) \subset C_\gamma$ and $T_a(\overline{\mathcal{G}}_a) \subset C[0, a]$ for $a < \infty$. We also know (in both cases $a = \infty$ and $a < \infty$) that $T_a(\mathcal{G}_a) \subset \mathcal{G}_a$. In this section we prove that $T_a(\overline{\mathcal{G}}_a) \subset \overline{\mathcal{G}}_a$ and the operator T_a has a unique fixed point in $\overline{\mathcal{G}}_a$. Many of the arguments in the rest of the paper are based on the following lemma.

LEMMA 6.1 (comparison). *Let $0 \le \alpha < \beta$ and $f, g \in C(\alpha, \beta)$ be given. Consider $\varphi \in C^1(\alpha, \beta)$ a viscosity subsolution of $\mathcal{L}_f \varphi(x) \le c$ on (α, β) and $\psi \in C^1(\alpha, \beta)$ a viscosity supersolution of $\mathcal{L}_g \psi(x) \ge c$ on (α, β). Assume that at least one of the functions is a classical ($C^2(\alpha, \beta)$) solution of the corresponding differential inequality and that the function $\varphi - \psi$ attains a local maximum at $x_* \in (\alpha, \beta)$. Then*

$$r(\varphi(x_*) - \psi(x_*)) \le \delta(f(x_*) - g(x_*))\varphi'(x_*) = \delta(f(x_*) - g(x_*))\psi'(x_*).$$

Proof. Let us assume that $\varphi \in C^2(\alpha, \beta)$ is a classical solution of $\mathcal{L}_f \varphi(x) \le c$. (The argument in the other case is identical.) This means that

$$r\varphi(x_*) - (rx_* - c)\varphi'(x_*) + \delta(x_* - f(x_*))\varphi'(x_*) - \frac{1}{2}\sigma^2 x_*^2 \varphi''(x_*) \le c.$$

The function $\psi - \varphi$ attains a local minimum at x_*, and since φ is C^2 in a neighborhood of x_*, we can consider φ as a test function when we apply the definition of the viscosity supersolution ψ. We obtain the inequality

$$r\psi(x_*) - (rx_* - c)\varphi'(x_*) + \delta(x_* - g(x_*))\varphi'(x_*) - \frac{1}{2}\sigma^2 x_*^2 \varphi''(x_*) \ge c.$$

Comparing the above results, we conclude that

$$r(\varphi(x_*) - \psi(x_*)) \le \delta(f(x_*) - g(x_*))\varphi'(x_*).$$

Since x_* is a point of interior maximum for $\varphi - \psi$, and both φ and ψ have continuous derivatives on $(0, \alpha)$, we have that $\varphi'(x_*) = \psi'(x_*)$. □

PROPOSITION 6.2. *For $0 < a \le \infty$, we have $T_a(\overline{\mathcal{G}}_a) \subset \overline{\mathcal{G}}_a$, and the mapping T_a has a unique fixed point in $\overline{\mathcal{G}}_a$.*

Proof. Let $f, g \in \overline{\mathcal{G}}_a$ be given. We denote $\varphi = T_a f$ and $\psi = T_a g$. Since $\varphi(0) = \psi(0) = 0$, we know that $\sup_{x\in\mathcal{D}_a}(\varphi(x) - \psi(x)) \ge 0$. We recall that φ, ψ are continuous on $[0, a]$ for finite a (or they are continuous on $[0, \infty)$ and equal to γx for x large enough if $a = \infty$). Thus there exists x_* such that $\varphi(x_*) - \psi(x_*) = \max_{x\in[0,a]}(\varphi(x) - \psi(x))$. If $\varphi(x_*) - \psi(x_*) = 0$, then

$$\sup_{x\in\mathcal{D}_a}(\varphi(x) - \psi(x)) \le \varphi(x_*) - \psi(x_*) = 0 \le \frac{\delta}{r}\max\{M_a, 1\}d(f, g).$$

Assume that $\varphi(x_*) - \psi(x_*) > 0$. Since $\varphi(0) = \psi(0)$ and $\varphi(a) = \psi(a)$ (or $\varphi(x) = \psi(x)$ for all x large enough if $a = \infty$), we see that $0 < x_* < a$. Moreover, since $\varphi(x_*) > \psi(x_*) \ge \gamma x_*$, we know $x_* \in \mathcal{C}_f = \{x : \ \varphi(x) > \gamma x\}$.

We remember that φ is a C^2 function on the open set $\mathcal{C}_f$, it is a classical solution of $\mathcal{L}_f\varphi = c$ on $\mathcal{C}_f$, and ψ is a viscosity supersolution of $\mathcal{L}_g\psi \ge c$. Lemma 6.1 implies $r(\varphi(x_*) - \psi(x_*)) \le \delta(f(x_*) - g(x_*))\varphi'(x_*)$. Therefore,

$$\sup_{x\in[0,a]}\big(\varphi(x) - \psi(x)\big) \le \varphi(x_*) - \psi(x_*) \le \frac{\delta}{r}|f(x_*) - g(x_*)||\varphi'(x_*)|.$$

Since $\varphi'(x_*) = \psi'(x_*)$, it is enough to assume that at least one of the functions f and g is an element of $\mathcal{G}_a$ to conclude that $|\varphi'(x_*)| \le \max\{M_a, 1\}$, where $M_a = 0$ for $a = \infty$. Consequently, we obtain

$$\sup_{x\in[0,a]}\big(\varphi(x) - \psi(x)\big) \le \frac{\delta}{r}\max\{M_a, 1\}d(f, g).$$

We can switch φ and ψ in the argument above and obtain a similar inequality for $\psi - \varphi$. In other words, we have proved that

$$d(T_a f, T_a g) \le \frac{\delta}{r}\max\{M_a, 1\}d(f, g), \tag{6.1}$$

provided that at least one of the functions f and g is an element of $\mathcal{G}_a$.

We now choose $f \in \overline{\mathcal{G}}_a$, and let $f_n \in \mathcal{G}_a$ be such that $d(f_n, f) \to 0$ as $n \to \infty$. Using (6.1) we immediately obtain $d(T_a f_n, T_a f) \to 0$ as $n \to \infty$, and since $T_a f_n \in \mathcal{G}_a$ for all n, we conclude that $T_a f \in \overline{\mathcal{G}}_a$.

A similar approximation argument ($f_n \to f,\ g_n \to g,\ f_n, g_n \in \mathcal{G}_a$), together with

$$d(T_a f, T_a g) \le d(T_a f, T_a f_n) + d(T_a f_n, T_a g_n) + d(T_a g_n, T_a g)$$

yields

$$d(T_a f, T_a g) \le \frac{\delta}{r}\max\{M_a, 1\}d(f, g) \quad \forall f, g \in \overline{\mathcal{G}}_a.$$

We consider separately the two cases $a = \infty$ and $a < \infty$. If $a = \infty$, then $M_\infty = 0$. Since $\delta < r$, T_a is a contraction on the complete metric space $(\overline{\mathcal{G}}_\infty, d)$. Applying the Banach fixed point theorem, we conclude that T_a has a unique fixed point in $\overline{\mathcal{G}}_\infty$.

If $a < \infty$, the Arzela–Ascoli theorem implies that $\overline{\mathcal{G}}_a$ is a convex and compact subset of the Banach space $C[0, a]$. Since $T_a : \overline{\mathcal{G}}_a \to \overline{\mathcal{G}}_a$ is a continuous mapping with respect to the norm of $C[0, a]$, Schauder's fixed point theorem implies that there exists a fixed point of T_a in $\overline{\mathcal{G}}_a$. Suppose there were two fixed points of T_a, namely f and g. Assume without loss of generality that

$$f(x_*) - g(x_*) = \max_{x\in[0,a]} \big(f(x) - g(x)\big) > 0,$$

so $x_* \in \mathcal{C}_f$. We apply Lemma 6.1 to conclude

$$r(f(x_*) - g(x_*)) \le \delta f'(x_*)(f(x_*) - g(x_*)),$$

which is impossible since $f(x_*) - g(x_*) > 0$, $\delta < r$, and $f'(x_*) \le 1$. (We use here the fact that f has a continuous derivative on $(0, a)$ and $f \in \overline{\mathcal{G}}_a$ to conclude $f'(x_*) \le 1$.) This means that $f \le g$ on $[0, a]$. Interchanging f and g, we obtain $f = g$, so the fixed point is unique. □

We denote by f_a the unique fixed point of T_a in $\overline{\mathcal{G}}_a$. The function f_a is continuous on $\mathcal{D}_a$ and continuously differentiable on $(0, a)$. Associated with the function f_a is a number $b_a \in (0, a]$ such that

$$\mathcal{L}_{f_a} f_a(x) = c, \quad f_a(x) > \gamma x, \quad 0 < x < b_a, \tag{6.2}$$

$$\mathcal{L}_{f_a} f_a(x) \ge c, \quad f_a(x) = \gamma x, \quad b_a < x \le a. \tag{6.3}$$

Even if $a = \infty$, b_∞ is finite.

PROPOSITION 6.3. *The number b_a is given by*

$$b_a = \begin{cases} a & \text{if } a \le b_\infty, \\ b_\infty & \text{if } a \ge b_\infty. \end{cases} \tag{6.4}$$

Proof. The proof is based on the same comparison argument for viscosity solutions that allowed us to conclude that the fixed point is unique in Proposition 6.2, namely an application of Lemma 6.1.

Consider first the case $a \le b_\infty$, and suppose $b_a < a$. The function f_a is defined only on $[0, a]$, but we may extend it by the formula

$$\overline{f}_a(x) = \begin{cases} f_a(x) & \text{if } 0 \le x \le a, \\ \gamma x & \text{if } x \ge a. \end{cases} \tag{6.5}$$

It is apparent from (6.3) that $\overline{f}_a$ is continuous on $[0, \infty)$ and continuously differentiable on $(0, \infty)$. Furthermore, for $x \ge a$ we have

$$c \le \mathcal{L}_{\overline{f}_a} \overline{f}_a(a) = c\gamma + \delta\gamma(1 - \gamma)a \le c\gamma + \delta\gamma(1 - \gamma)x = \mathcal{L}_{\overline{f}_a} \overline{f}_a(x). \tag{6.6}$$

Using (6.6) we conclude that $\overline{f}_a$ is a viscosity solution of the equation $\min\{\mathcal{L}_f f(x) - c, f(x) - \gamma x\} = 0$ on $(0, \infty)$. Furthermore, $\overline{f}_a$ has a continuous derivative on $(0, \infty)$ and $\overline{f}_a(x) = \gamma x$ for large x. We can now compare $\overline{f}_a$ and f_∞. We know that either

$$\sup_{x\in[0,\infty)} (\overline{f}_a(x) - f_\infty(x)) \le 0$$

or there exists $x_* \in (0, b_\infty)$ such that

$$\overline{f}_a(x_*) - f_\infty(x_*) = \sup_{x\in[0,\infty)} (\overline{f}_a(x) - f_\infty(x)) > 0.$$

In the latter case, $x_* \in \mathcal{C}_{\overline{f}_a}$ and Lemma 6.1 implies

$$r(\overline{f}_a(x_*) - f_\infty(x_*)) \le \delta f'_\infty(x_*)(\overline{f}_a(x_*) - f_\infty(x_*)),$$

which is impossible since $r < \delta$ and $f'_\infty(x_*) \le 1$. This means that the only possibility is $\overline{f}_a \le f_\infty$. In the same way we prove that $f_\infty \le \overline{f}_a$, so $\overline{f}_a = f_\infty$. This implies that $b_a = b_\infty$, which contradicts the hypothesis $b_a < a \le b_\infty$.

The case $a > b_\infty$ is similar since $f_a(a) = \gamma a = f_\infty(a)$, and the restriction of f_∞ to $[0, a]$ is a viscosity solution of (3.1) on $(0, a)$. We can use the same comparison argument to conclude that $f_\infty|_{[0,a]} = f_a$, which implies $b_a = b_\infty$.

COROLLARY 6.4. *For $0 < a \le \infty$, the function f_a is in $\mathcal{G}_a$.*

Proof. We have already seen that f_a is continuously differentiable, and since $f_a \in \overline{\mathcal{G}}_a$, we conclude that $-M_a \le f'_a(x) \le 1$ for $0 < x < a$. It remains only to prove that the derivative f'_a cannot attain the value 1.

Assume, by contradiction, that $f'_a(x_0) = 1$ for some $x_0 \in (0, a)$. This means f'_a has a maximum at x_0 and $x_0 \in \mathcal{C}_{f_a}$, where f_a is two times continuously differentiable. Hence, $f''_a(x_0) = 0$. Moreover, $\mathcal{L}_{f_a} f_a(x_0) = c$, so $(r - \delta)(x_0 - f_a(x_0)) = 0$. Since $r - \delta > 0$ we see that $f_a(x_0) = x_0$. The function f_a is thus a solution of the ordinary differential equation $\mathcal{L}_f f(x) = c$ with initial conditions $f(x_0) = x_0$, $f'(x_0) = 1$ on the interval $[x_0, b_a]$. However, the only such solution to this equation is $f(x) = x$, and we conclude that $f_a(x) = x$ for $x_0 \le x \le b_a$. This contradicts the fact that $f_a(b_a) = \gamma b_a < b_a$.

COROLLARY 6.5. *For every $0 < a \le \infty$, the function f_a is concave for small values of x, it has a right derivative at $x = 0$, and $D^+ f_a(0) \le 1$.*

Proof. Since $f_a = T_a f_a$ and we just proved that $f_a \in \mathcal{G}_a$, we know from the first part of the proof of Proposition 5.6 that the derivative $f'_a = DT_a f_a$ cannot attain a positive local maximum in $(0, b_a)$. Since $f_a(0) = 0$, $f_a(b_a) = \gamma b_a$, and f_a is differentiable on $(0, b_a)$, we can conclude from the mean-value theorem that there exists $x_\gamma \in (0, b_a)$ with $f'_a(x_\gamma) = \gamma$. Since $D^- f_a(b_a) \le \gamma$, we can argue that for any $x_1 < x_2 \le x_\gamma$ we have $f'_a(x_1) > f'_a(x_2)$. To do this, we first use the fact that f'_a cannot attain a positive interior maximum on $[x_2, b_a]$ to conclude that $f'_a(x_2) > f'_a(x_\gamma) = \gamma$ and then use the fact that f'_a cannot attain a positive interior maximum on $[x_1, x_\gamma]$ to further conclude that $f'_a(x_1) > f'_a(x_2)$. In other words, the derivative f'_a is strictly decreasing on $(0, x_\gamma)$. This means that the function f_a is concave on $[0, x_\gamma]$ and

$$D^+ f_a(0) \triangleq \lim_{x\to 0} \frac{f_a(x) - 0}{x - 0} = \lim_{x\to 0} f'_a(x) \tag{6.7}$$

is well defined. It is obvious that $D^+ f_a(0) \le 1$.

7. Proofs of Theorems 2.2, 2.4, and 2.5. For each call price K we construct a function f^* so that $f^*(x)$ is the value of the convertible bond when the value of the firm is x. For small values of x, the function $f^*(x)$ agrees with $f_a(x)$ for an appropriately chosen a, depending on K. In order to proceed, we must first understand the dependence of f_a on the parameter a. For this purpose, we define $m\colon (0,\infty) \to (0,\infty)$ by

$$m(a) = \max_{x\in[0,a]} f_a(x). \tag{7.1}$$

Because $f_a = f_\infty|_{[0,a]}$ for $a \geq b_\infty$ and f_∞ is nondecreasing by virtue of its membership in $\overline{\mathcal{G}}_\infty$, we have

$$m(a) = \gamma a \quad \forall a \geq b_\infty. \tag{7.2}$$

For $a < b_\infty$ and $x \in (0, a)$, we have $f_a(x) > \gamma x$ (Proposition 6.3 and the inequality in (6.2)), and so it is possible that $m(a) > \gamma a$ for $0 < a < b_\infty$. We shall in fact discover that there is a number $b_0 \in [0, b_\infty)$ such that $m(a) > \gamma a$ for $0 < a < b_0$, whereas $m(a) = \gamma a$ for $a \geq b_0$ (see Remark 7.3).

LEMMA 7.1. *The function $m\colon (0,\infty) \to (0,\infty)$ is strictly increasing and continuous and satisfies $\lim_{a \downarrow 0} m(a) = 0$.*

Proof. It is clear from (7.2) that m is strictly increasing on $[b_\infty, \infty)$. We first show that m is nondecreasing on $(0, b_\infty]$. Let $0 < a_1 < a_2 \leq b_\infty$ be given. Since $f_{a_1}(0) = 0 = f_{a_2}(0)$ and $f_{a_1}(a_1) = \gamma a_1 < f_{a_2}(a_1)$, if the function $f_{a_1} - f_{a_2}$ attains a positive maximum over $[0, a_1]$ it must be at an interior point $x_* \in (0, a_1)$. But $\mathcal{L}_{f_{a_1}} f_{a_1}(x) = c = \mathcal{L}_{f_{a_2}} f_{a_2}(x)$ for $0 < x < a_1$, and $x \in \mathcal{C}_{f_{a_1}}$, where f_{a_1} is C^2. Lemma 6.1 implies that

$$r\big(f_{a_1}(x_*) - f_{a_2}(x_*)\big) \leq \delta\big(f_{a_1}(x_*) - f_{a_2}(x_*)\big) f'_{a_1}(x_*),$$

which is impossible because $\delta < r$ and $f'_{a_1}(x_*) \leq 1$. We conclude that $f_{a_1}(x) \leq f_{a_2}(x)$ for all $x \in [0, a_1]$. Therefore m is nondecreasing on $(0, b_\infty]$.

By the same comparison argument, the function $f_{a_2} - f_{a_1}$ cannot attain a positive maximum in $(0, a_1)$, so $f_{a_2}(x) - f_{a_1}(x) \leq f_{a_2}(a_1) - \gamma a_1$ for $0 \leq x \leq a_1$. It follows that

$$\begin{aligned} m(a_2) - m(a_1) &= \max\Big\{ \max_{x\in[0,a_1]} f_{a_2}(x), \max_{x\in[a_1,a_2]} f_{a_2}(x)\Big\} - m(a_1) \\ &\leq \max\Big\{ \max_{x\in[0,a_1]} \big(f_{a_2}(x) - f_{a_1}(x)\big), \max_{x\in[a_1,a_2]} \big(f_{a_2}(x) - \gamma a_1\big)\Big\} \\ &= \max\Big\{ f_{a_2}(a_1) - \gamma a_1, \max_{x\in[a_1,a_2]} \big(f_{a_2}(x) - \gamma a_1\big)\Big\} \\ &= -\gamma a_1 + \max_{x\in[a_1,a_2]} f_{a_2}(x). \end{aligned} \tag{7.3}$$

By virtue of its membership in $\overline{\mathcal{G}}_{a_2}$ and Remark 5.7, the function f_{a_2} satisfies $f'_{a_2}(x) \geq -C$ for all $x \in (0, a_2)$ and some positive constant C which is bounded away from zero so long as a_2 is bounded away from zero. Thus, for $x \in [a_1, a_2]$,

$$f_{a_2}(x) = f_{a_2}(a_2) - \int_x^{a_2} f'_{a_2}(y)\, dy \leq \gamma a_2 + C(a_2 - x) \leq \gamma a_2 + C(a_2 - a_1).$$

Substituting this into (7.3), we conclude that

$$0 \leq m(a_2) - m(a_1) \leq (C + \gamma)(a_2 - a_1), \tag{7.4}$$

so long as a_2 is bounded away from zero. The function m is thus continuous.

We now prove that $m(a_1) < m(a_2)$. Assume, by contradiction, that $m(a_1) = m(a_2)$. Let $x_0 \in [0, a_1]$ be such that $f_{a_1}(x_0) = m(a_1)$. We must actually have $x_0 \in (0, a_1)$ because $m(a_1) = m(a_2) \geq \gamma a_2 > \gamma a_1 = f_{a_1}(a_1) > 0 = f_{a_1}(0)$. We have already shown that f_{a_2} dominates f_{a_1} on $[0, a_1]$, and hence we must have $f_{a_1}(x_0) = f_{a_2}(x_0)$. The comparison argument using Lemma 6.1 shows that neither $f_{a_2} - f_{a_1}$ nor $f_{a_1} - f_{a_2}$ can have a positive maximum in the open interval $(0, x_0)$; we conclude that

$$f_{a_1}(x) = f_{a_2}(x) \quad \forall x \in [0, x_0]. \tag{7.5}$$

Both f_{a_1} and f_{a_2} are solutions of the ordinary differential equation $\mathcal{L}_f f(x) = c$ on $[x_0, a_1]$ and have the same initial conditions $f_{a_1}(x_0) = f_{a_2}(x_0)$, $f'_{a_1}(x_0) = f'_{a_2}(x_0)$. It follows that

$$f_{a_1}(x) = f_{a_2}(x) \quad \forall x \in [x_0, a_1]. \tag{7.6}$$

This implies that $f_{a_2}(a_1) = f_{a_1}(a_1) = \gamma a_1$, which contradicts Proposition 6.3. We conclude that m is strictly increasing on $(0, b_\infty]$.

Finally, since $f_a(x) \le x$ for $0 \le x \le a$, we see that $0 \le m(a) \le a$, and consequently $\lim_{a \downarrow 0} m(a) = 0$. □

LEMMA 7.2.

(i) *Assume $m(\overline{a}) > \gamma \overline{a}$ for some $\overline{a} > 0$. Then $\overline{a} < \frac{c}{r\gamma}$ and $m(a) > \gamma a$ for all $a \in (0, \overline{a})$.*

(ii) *If $m(a) > \gamma a$, the function f_a attains its maximum over $[0, a]$ at a unique point $x_a \in (0, a)$.*

Proof. (i) If $a \ge \frac{c}{r\gamma}$, we define $h(x) = \gamma a \ge \frac{c}{r}$ for $x \in [0, a]$. Then $\mathcal{L}_{f_a} h(x) \ge c$ for $0 < x < a$. Lemma 6.1 shows that $f_a - h$ cannot have a positive maximum in $(0, a)$, and since $f_a(0) = 0 \le h(0)$ and $f_a(a) = \gamma a = h(x)$, we conclude that $f_a(x) \le h(a)$ for all $0 \le x \le a$. Consequently, the maximum of f_a is $m(a) = \gamma a$.

Assume now that $m(\overline{a}) > \gamma \overline{a}$ for some $\overline{a} > 0$. We have just seen that $\overline{a} < \frac{c}{r\gamma}$. Let $a \in (0, \overline{a})$ be given. Define $\ell = a/\overline{a} < 1$ and rescale the function $f_{\overline{a}}$ by setting $f(x) = \ell f_{\overline{a}}(\frac{x}{\ell})$ for all $x \in [0, a]$. We compute $f'(x) = f'_{\overline{a}}(\frac{x}{\ell})$ and $f''(x) = \frac{1}{\ell} f''_{\overline{a}}(\frac{x}{l})$, from which we conclude that

$$\mathcal{L}_f f(x) = \ell \mathcal{L}_{f_{\overline{a}}} f_{\overline{a}} \left(\frac{x}{l}\right) + c(1-l) f'_{\overline{a}} \left(\frac{x}{l}\right) \le \ell c + c(1-\ell) = c \quad \forall x \in (0, a).$$

Lemma 6.1 shows that $f - f_a$ cannot have a positive maximum over $[0, a]$ at a point in $(0, a)$. But $f(0) = f_a(0) = 0$ and $f(a) = f_a(a) = \gamma a$, and therefore $f_a(x) \ge f(x)$ for all $x \in [0, a]$. In particular,

$$m(a) = \max_{x \in [0,a]} f_a(x) \ge \max_{x \in [0,a]} f(x) = \ell m(\overline{a}) > \ell \gamma \overline{a} = \gamma a. \tag{7.7}$$

(ii) Let us assume now that $m(a) > \gamma a$ and there exist $0 < x_0 < y_0 < a$ such that $f_a(x_0) = f_a(y_0) = m(a)$. Since $f_a(x) \le m(a)$ for $x_0 \le x \le y_0$ we see that f_a has a local minimum at some point $x_1 \in (x_0, y_0)$. Then $f'_a(x_1) = 0$, $f''_a(x_1) \ge 0$, and we may use the equation $\mathcal{L}_{f_a} f_a(x_1) = c$ to obtain $r f_a(x_1) \ge c$. This is impossible because $f_a(x_1) \le m(a) < m(\frac{c}{\gamma r}) = \frac{c}{r}$.

Remark 7.3. We define $b_0 \triangleq \sup\{a > 0, m(a) > \gamma a\}$, where we set $b_0 = 0$ if $m(a) = \gamma a$ for all $a > 0$. Lemma 7.2 shows that $m(a) > \gamma a$ for all $a \in (0, b_0)$. This lemma further shows that $b_0 \le \frac{c}{\gamma r}$. Since for $x \ge b_\infty$ we have $f_\infty(x) = \gamma x$ and $\mathcal{L}_{f_\infty} f_\infty(x) \ge c$, we conclude that

$$r\gamma b_\infty - (r b_\infty - c)\gamma + \delta(b_\infty - \gamma b_\infty)\gamma \ge c,$$

which implies $\delta(1-\gamma) b_\infty \gamma \ge c(1-\gamma)$, and consequently $b_\infty \ge \frac{c}{\gamma \delta}$. In summary,

$$0 \le b_0 \le \frac{c}{\gamma r} < \frac{c}{\gamma \delta} \le b_\infty. \tag{7.8}$$

LEMMA 7.4. *If $0 < \gamma < \frac{1}{2}$, then $b_0 > 0$.*

Proof. For small values of a, we construct a quadratic subsolution of

$$\begin{cases} \mathcal{L}_g g(x) \le c \text{ for } 0 < x < a, \\ g(0) = 0,\ g(a) = \gamma a, \end{cases} \tag{7.9}$$

which satisfies $\max_{x\in[0,a]} g(x) > \gamma a$. According to Lemma 6.1, $g - f_a$ cannot have a positive maximum over $[0,a]$ in $(0,a)$, and since $g(0) = f_a(0) = 0$, $g(a) = f_a(a) = \gamma a$, we see that $f_a \ge g$ on $[0,a]$. It follows that $m(a) > \gamma a$.

The remainder of the proof is the construction of g. We define

$$g(x) = -\frac{x^2}{2a} + \left(\gamma + \frac{1}{2}\right) x,$$

so that $g(0) = 0$ and $g(a) = \gamma a$. Direct computation results in

$$\begin{aligned} &\mathcal{L}_g g(x) \\ &= \frac{rx^2}{2a} - \frac{cx}{a} + \left(\gamma + \frac{1}{2}\right) c - \frac{\delta x^3}{2a^2} + \frac{3\delta\gamma x^2}{2a} - \frac{\delta x^2}{4a} - \delta\gamma^2 x + \frac{\delta x}{4} + \frac{\sigma^2 x^2}{2a} \\ &\le \frac{ra}{2} + \left(\gamma + \frac{1}{2}\right) c + \frac{3\delta\gamma a}{2} + \frac{\delta a}{4} + \frac{\sigma^2 a}{2} \quad \forall x \in [0,a]. \end{aligned}$$

Since $\left(\gamma + \frac{1}{2}\right) c < c$, we have $\sup_{x\in[0,a]} \mathcal{L}_g g(x) \le c$ for sufficiently small a. ∎

We summarize what has so far been established.

(a) For $a > b_\infty$ we have $f_a = f_\infty|_{[0,a]}$ and the maximum $m(a) = \gamma a$ of f_a over $[0,a]$ is attained at the right endpoint a. We have $f_a(x) > \gamma x$ for $x \in (0, b_\infty)$ and $f_a(x) = \gamma x$ for $x \in [b_\infty, a]$.

(b) For $b_0 \le a \le b_\infty$, the maximum $m(a) = \gamma a$ of f_a over $[0,a]$ is attained at the right endpoint a and $f_a(x) > \gamma x$ for all $x \in (0,a)$.

(c) If $b_0 > 0$ (a sufficient condition for this is $0 < \gamma < \frac{1}{2}$), then for $0 < a < b_0$, we have $f_a(x) > \gamma x$ for all $x \in (0,a)$ and the maximum $m(a) > \gamma a$ of f_a over $[0,a]$ is attained at a unique point $x_a \in (0,a)$.

For a fixed call price K we want to define $f^*(x)$ to be $f_a(x)$ for small values of x, where a is the unique parameter such that $m(a) = K$. Denoting

$$K_1 = \gamma b_0, \quad K_2 = \gamma b_\infty,$$

we have the following three situations corresponding to the three cases of Theorem 2.5.

(i) If $K > K_2$, we set $a = \frac{K}{\gamma}$. We define

$$f_*(x) = \begin{cases} f_a(x) = f_\infty(x) & \text{if } 0 \le x \le a, \\ \gamma x & \text{if } x \ge a. \end{cases} \tag{7.10}$$

We see that $f_*(x) = f(x, C_a^*, C_o^*)$ for $C_a^* = \frac{K}{\gamma}$ and $C_o^* = b_\infty < C_a^*$.

(ii) If $K_1 \le K \le K_2$, then again we set $a = \frac{K}{\gamma}$. We define

$$f_*(x) = \begin{cases} f_a(x) & \text{if } 0 \le x \le a, \\ \gamma x & \text{if } x \ge a. \end{cases} \tag{7.11}$$

In this case, $f_*(x) = f(x, C_a^*, C_o^*)$ for $C_a^* = C_o^* = \frac{K}{\gamma}$.

(iii) Assume $K_1 > 0$ and $0 < K < K_1$. Because $m(b_0) = K_1$, there exists a unique $a = m^{-1}(K) < b_0$ such that $m(a) = K$. Since $K < K_1$, Lemma 7.2 implies

that $K = m(a) > \gamma a$ and there exists a unique $x_a \in (0, a)$ such that $f_a(x_a) = m(a)$. Since $f_a' < 1$, we obtain that $K = m(a) = f_a(x_a) < x_a$, so $K < x_a < a < \frac{K}{\gamma}$. We now take $C_a^* = x_a$, $C_o = \frac{K}{\gamma}$ and define

$$f^*(x) = \begin{cases} f_a(x) & \text{for } 0 \le x \le C_a^*, \\ K & \text{for } C_a^* \le x \le C_o^*, \\ \gamma x & \text{for } x \ge C_o^*. \end{cases} \tag{7.12}$$

Again we have $f_*(x) = f(x, C_a^*, C_o^*)$. Since $f_a'(C_a^*) = 0$, f_* is a C^1 function on $(0, \frac{K}{\gamma})$.

It is apparent that the function f_* and the numbers C_o^*, C_a^* just defined have all the properties set forth in Theorem 2.5. The uniqueness of solutions to $\mathcal{N}f = c$ in that theorem follows from Lemma 6.1. We now accept Theorem 2.2, whose proof will be given later in this section, and show that the function f_* defined by (7.10)–(7.12) is indeed the function f_* given by (2.19), and the numbers C_o^* and C_a^* defined above satisfy (2.18). Using f_*, C_o^*, and C_a^* just defined in this way means that the proof of Theorem 2.4 given below also completes the proof of Theorem 2.5.

Proof of Theorem 2.4. We need to prove that

$$f(x, C_a^*, C_o^*) \le f(x, C_a, C_o^*) \text{ for each } C_a \ge K,\ x \in (0, \infty), \tag{7.13}$$

$$f(x, C_a^*, C_o^*) \ge f(x, C_a^*, C_o) \text{ for each } C_o > 0,\ x \in (0, \infty). \tag{7.14}$$

Case (i). $K > K_2 = \gamma b_\infty$.

If $C_a \ge C_a^* = \frac{K}{\gamma}$, then clearly $f(x, C_a^*, C_o^*) = f(x, C_a, C_o^*)$ for $x \in (0, \infty)$.

If $C_o^* < C_a \le \frac{K}{\gamma}$, according to Definition 2.3(i) we have $f(x, C_a, C_o^*) = f(x, C_a^*, C_o^*)$ for $0 \le x < C_a$, and $f(x, C_a, C_o^*) = K \ge f(x, C_a^*, C_a^*)$ for $C_a \le x \le \frac{K}{\gamma}$. For $x \ge \frac{K}{\gamma}$, we have $f(x, C_a^*, C_o^*) = f(x, C_a, C_o^*) = \gamma x$.

Finally, consider the case $K \le C_a \le C_o^* = b_\infty$. Using the Case (i) assumption, we have $K > K_2 = \gamma b_\infty = \gamma C_o^* \ge \gamma C_a$. From Definition 2.3(ii),

$$f(C_a, C_a, C_o^*) = \max\{K, \gamma C_a\} = K \ge \gamma C_a = f_*(C_a).$$

Since $f(\cdot) = f(\cdot, C_a, C_o^*)$ satisfies $\mathcal{L}_f f(x) = c$ on $(0, C_a)$ and $\mathcal{L}_{f_*} f_*(x) = c$ on $(0, C_a)$, an application of Lemma 6.1 yields

$$f(x, C_a^*, C_o^*) = f_*(x) \le f(x, C_a, C_o^*) \text{ for } 0 \le x \le C_a.$$

For $C_a \le x \le \frac{K}{\gamma}$, we have

$$f(x, C_a, C_o^*) = \max\{K, \gamma x\} = K = f_*\Big(\frac{K}{\gamma}\Big) \ge f_*(x) = f(x, C_a^*, C_o^*).$$

For $x > \frac{K}{\gamma}$, we have $f(x, C_a^*, C_o^*) = \gamma x = f(x, C_a, C_o^*)$. This completes the proof of (7.13) in Case (i).

To establish (7.14), we let $C_o > 0$ be given. If $C_o \le C_o^*$, then $f(C_o, C_a^*, C_o) = \gamma C_o \le f_*(C_o)$. Applying Lemma 6.1, we get

$$f(x, C_a^*, C_o) \le f_*(x) = f(x, C_a^*, C_o^*) \text{ for } 0 \le x \le C_o. \tag{7.15}$$

The same inequality is easily verified for $C_o \le x < \infty$.

The case $C_o^* < C_o \leq C_a^*$ is the most interesting. We know that the function $f(\cdot) = f(\cdot, C_a^*, C_o)$ satisfies

$$\begin{cases} \mathcal{L}_f f(x) = c \text{ for } 0 < x < C_o, \\ f(0) = 0,\ f(C_o) = \gamma C_o = f^*(C_o) \text{ (since } C_o > C_o^* = b_\infty). \end{cases}$$

We recall that f_* is a C^1 viscosity supersolution of $\mathcal{L}_{f_*} f_*(x) = c$ on $(0, C_o)$, so Lemma 6.1 can be again used to obtain

$$f(x, C_a^*, C_o) = f(x) \leq f^*(x) = f(x, C_a^*, C_o^*) \text{ for } 0 \leq x \leq C_o.$$

For $C_o \leq x$, we have

$$f(x, C_a^*, C_o) = \gamma x = f^*(x) = f(x, C_a^*, C_o^*).$$

If $C_o \geq C_a^* = \frac{K}{\gamma}$, we just observe that $f(x, C_a^*, C_o) = f(x, C_a^*, \frac{K}{\gamma})$, so we can reduce this case to the case $C_o = C_a^*$ already considered. This completes the proof of (7.14) in Case (i).

Case (ii). $\gamma b_0 = K_1 \leq K \leq K_2 = \gamma b_\infty$.

This is the simplest case, all proofs being based on comparison arguments for C^2 solutions of the equation $\mathcal{L}_f f(x) = c$. The details are left to the reader.

Case (iii). $0 < K < K_1 = \gamma b_o$.

If $C_a^* < C_o < \infty$, there is no change:

$$f(x, C_a^*, C_o) = f(x, C_a^*, C_o^*) \text{ for } 0 \leq x < \infty.$$

If $0 < C_o \leq C_a^*$, then $f(C_o, C_a^*, C_o) = \gamma C_o \leq f_*(C_o)$. The Comparison Lemma 6.1 implies

$$f(x, C_a^*, C_o) \leq f(x, C_a^*, C_o^*) \text{ for } 0 \leq x \leq C_o.$$

For $C_o < x < C_a^*$, we have $f(x, C_a^*, C_o) = \gamma x \leq f(x, C_a^*, C_o^*)$, and for $C_a^* \leq x$ we know that $f(x, C_a^*, C_o) = f(x, C_a^*, C_o^*) = \max\{K, \gamma x\}$. This completes the proof of (7.14) in Case (iii).

We consider (7.13). If $K \leq C_a \leq C_a^*$, then $f(C_a, C_a, C_o^*) = K \geq f(C_a, C_a^*, C_o^*)$. The Comparison Lemma 6.1 implies $f(x, C_a, C_o^*) \geq f(x, C_a^*, C_o^*)$ for $0 \leq x \leq C_a$. For $x \geq C_a$, we have $f(x, C_a, C_o^*) = \max\{K, \gamma x\} \geq f(x, C_a^*, C_o^*)$.

The case $C_a \geq C_o^*$ can be reduced to the case $C_a = C_o^*$ since $f(x, C_a, C_o^*) = f(x, C_a^*, C_o^*)$ for all $x \geq 0$ if $C_a \geq C_o^*$. We do that case now.

Assume $C_a^* < C_a \leq C_o^*$. First we claim that $f_*(\cdot) = f(\cdot, C_a^*, C_o^*)$ is a C^1 viscosity subsolution of

$$\mathcal{L}_{f_*} f_*(x) \leq c \text{ on } \left(0, \frac{K}{\gamma}\right), \tag{7.16}$$

and then we use the Comparison Lemma 6.1 (the difference $f_*(\cdot) - f(\cdot, C_a, C_o^*)$ cannot have a positive maximum in $(0, C_a)$) to conclude that

$$f_*(x) \leq f(x, C_a, C_o^*) \text{ for } 0 \leq x \leq C_a.$$

In the comparison argument we also use the fact that $f(\cdot) = f(\cdot, C_a, C_o^*)$ satisfies $\mathcal{L}_f f(x) = c$ for $0 < x < C_a$, and

$$f_*(0) = 0 = f(0, C_a, C_o^*), \quad f_*(C_a) = K = f(C_a, C_a, C_o^*).$$

For $C_a \le x \le \frac{K}{\gamma}$, we have $f_*(x) = K = f(x, C_a, C_o^*)$, and for $x > \frac{K}{\gamma}$ we know that $f_*(x) = \gamma x = f(x, C_a, C_o^*)$.

This means that the proof of (7.14) is complete, provided we can show that f_* is a viscosity subsolution of (7.16). We know that $\mathcal{L}_{f_*} f_*(x) = c$ for $0 < x < C_a^*$, f_* being a C^2 function on $(0, C_a^*)$. From (7.8) and the Case (iii) assumption, we see that $rK \le c$. Furthermore, $f_*(x) = K$ for $C_o^* \le x \le \frac{K}{\gamma}$. We conclude that $\mathcal{L}_{f_*} f_*(x) \le c$ on $(C_a^*, \frac{K}{\gamma})$.

It remains to show that if $\psi \in C^2(0, C_o^*)$ dominates f^* on $(0, C_o^*)$ and agrees with f^* at C_a^*, then

$$r\psi(C_a^*) - (rC_a^* - c)\psi'(C_a^*) + \delta(C_a^* - \psi(C_a^*))\psi'(C_a^*) - \frac{1}{2}\sigma^2 (C_a^*)^2 \psi''(C_a^*) \le c. \tag{7.17}$$

Since $f_* \in C^1(0, \frac{K}{\gamma})$ and $f_*'(C_a^*) = 0$, we have $\psi'(C_a^*) = 0$. Since $0 < C_a^* < a$, we know $\mathcal{L}_{f_a} f_a(C_a^*) = c$, and since $f_a(C_a^*) = K$ and $f_a'(C_a^*) = 0$, we obtain

$$rK - \frac{1}{2}\sigma^2 (C_a^*)^2 f_a''(C_a^*) = c. \tag{7.18}$$

However, since $\psi(C_a^*) = f_a(C_a^*)$, $\psi'(C_a^*) = f_a'(C_a^*) = 0$, and ψ dominates f_a on $[0, C_a^*]$ (because $f_*(x) = f_a(x)$ on $[0, C_a^*]$), we conclude that

$$f_a''(C_a^*) \le \psi''(C_a^*). \tag{7.19}$$

Substituting this into (7.18), we obtain (7.17). □

Remark 7.5. The proof of the last claim is based on the elementary observation that for a C^2 function, a one-sided maximum is enough to conclude that the second derivative is not positive, provided that the first derivative vanishes. Furthermore, we have proved that f_* is a viscosity solution of the variational inequality $\max\{\mathcal{N} f_*(x) - c, f_*(x) - K\} = 0$ on $(0, \frac{K}{\gamma})$.

Proof of Theorem 2.2. For $y_1 = x_1$, it is easily verified that $f(x) = x$ is a solution of (2.16), and the Comparison Lemma 6.1 establishes uniqueness.

For $0 < y_1 < x_1$, uniqueness again follows from Lemma 6.1 once we have a solution satisfying $f' \le 1$ on $(0, x_1)$. The proof of existence is based on a fixed point argument similar to the proof of Proposition 6.2 with $a < \infty$. In fact, the argument here is simpler, since we deal only with C^2 solutions of the differential equation $\mathcal{L}_g f(x) = c$ rather then viscosity solutions of the variational inequality $\min\{\mathcal{L}_g f(x) - c, f(x) - \gamma x\} = 0$.

For $0 < y_1 < x_1$, we set $A = x_1 - y_1$ and define $\mathcal{G}$ to be the set of all functions $g \in C[0, x_1] \cap C^2(0, x_1)$ such that $g(0) = 0$, $g(x_1) = y_1$, and

$$g(x) \ge \max\{x - A, 0\}, \quad -M(x_1, y_1) \le g'(x) < 1 \quad \forall x \in (0, x_1),$$

where $M(x_1, y_1)$ is a constant to be determined later but depending on only x_1 and y_1. We further define $\overline{\mathcal{G}}$ to be the closure of $\mathcal{G}$ in $C[0, x_1]$ with respect to the supremum norm $\|\cdot\|$. For $g \in \overline{\mathcal{G}}$, we set

$$Tg(x) = \mathbb{E}\left[\int_0^{\tau_0^x \wedge \tau_{x_1}^x} ce^{-ru} du + \mathbb{I}_{\{\tau_{x_1}^x < \tau_0^x\}} e^{-r(\tau_0^x \wedge \tau_{x_1}^x)} y_1\right], \tag{7.20}$$

where $X^x(t)$ is given by (3.4) with $X^x(0) = x$. It is clear from its definition that $Tg \ge 0$ for every $g \in \overline{\mathcal{G}}$. We use the argument in the proof of Proposition 4.6 to

conclude that for $g \in \overline{\mathcal{G}}$ the function Tg is of class C^2 on $(0, x_1)$ and $\mathcal{L}_g Tg(x) = c$ for $0 < x < x_1$. The continuity of Tg at 0 and x_1 follows from Lemma 4.2. The functions $\max\{x - A, 0\}$ and x are respective sub- and supersolutions of $\mathcal{L}_g f = c$ which lie respectively below and above Tg at the endpoints 0 and x_1. Lemma 6.1 implies that for all $g, h \in \overline{\mathcal{G}}$,

$$\max\{x - A, 0\} \le Tg(x) \le x \text{ for } 0 \le x \le x_1, \tag{7.21}$$

$$\|Tg - Th\| \le \sup_{0<x<x_1} |DTg(x)| \|g - h\|. \tag{7.22}$$

We now prove that $T(\mathcal{G}) \subset \mathcal{G}$, the analogue of Proposition 5.6. For $g \in \mathcal{G}$, the first part of the proof of Theorem 5.6 shows that DTg cannot attain a positive local maximum nor a negative local minimum in $(0, x_1)$. This implies that either DTg is nonnegative on $(0, x_1)$ or else $D^- Tg(x_1) \le DTg(x)$ for $0 < x < x_1$. To show that $DTg(x) \ge -M(x_1, y_1)$, it suffices to find a lower bound on $D^- Tg(x_1)$ which may depend on x_1 and y_1 but not on g. For this purpose, we let h be the solution on $[y_1, x_1]$ of the equation $\mathcal{L}_g h = c$ with boundary conditions $h(y_1) = h(x_1) = y_1$. Lemma 6.1 shows that h is nonnegative and dominates Tg on $[y_1, x_1]$, and hence $D^- h(x_1) \le D^- Tg(x_1)$. If h attains a maximum at some point $x_* \in (y_1, x_1)$, the equation $\mathcal{L}_g h(x_*) = c$ implies $h(x_*) \le \frac{c}{r}$. If h does not attain a maximum in (y_1, x_1), then h is dominated by its value y_1 at the endpoints of this interval. In either case, we obtain a bound on $|h|$ which is independent of g. Furthermore, there must be some point $x_0 \in (y_1, x_1)$ where h' vanishes. We solve the equation $\mathcal{L}_g h = c$ for h'' and integrate from x_0 to obtain (5.7). We then use Gronwall's inequality to obtain a bound on $|h'|$ independent of g.

We need also to obtain the upper bound $DTg < 1$. We observe first that since $Tg(x) \ge \max\{x - A, 0\}$ and these two functions agree at $x = x_1$, we must have $D^- Tg(x_1) \le 1$. We use the same arguments used to prove $DT_a g < 1$ if $g \in \mathcal{G}_a$ to conclude that $DTg < 1$ on $(0, x_1)$. This completes the proof that $T(\mathcal{G}) \subset \mathcal{G}$. A relation similar to (6.1) shows that the operator T is continuous on $\overline{\mathcal{G}}$, and hence $T(\overline{\mathcal{G}}) \subset \overline{\mathcal{G}}$. Schauder's fixed point theorem implies the existence of a function $f \in \overline{\mathcal{G}}$ satisfying $Tf = f$. This means, in particular, that $f \in C[0, x_1] \cap C^2(0, x_1)$ and $\mathcal{L}_f f = c$, so f is a solution of (2.16). Since f is differentiable and $f \in \overline{\mathcal{G}}$, we know that $f' \le 1$ on $(0, x_1)$. In fact, $f' < 1$. The proof is identical to the proof of $f'_a < 1$. □

Acknowledgment. The third author thanks John Noddings for introducing him to convertible bonds.

REFERENCES

[1] P. Asquith, *Convertible bonds are not called late*, J. Finance, 50 (1995), pp. 1275–1289.
[2] P. Asquith and D. Mullins, Jr., *Convertible debt: Corporate call policy and voluntary conversion*, J. Finance, 46 (1991), pp. 1273–1289.
[3] A. Bensoussan, M. Crouhy, and D. Galai, *Stochastic equity volatility and the capital structure of the firm*, Philos. Trans. Roy. Soc. London Ser. A, 347 (1994), pp. 531–541.
[4] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Political Economy, 81 (1973), pp. 637–659.
[5] M. Brennan and E. Schwartz, *Convertible bonds: Valuation and optimal strategies for call and conversion*, J. Finance, 32 (1977), pp. 1699–1715.
[6] M. Brennan and E. Schwartz, *Analyzing convertible bonds*, J. Financial Quantitative Analysis, 15 (1980), pp. 907–929.
[7] M. Crandall, H. Ishii, and P. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[8] K. Dunn and K. Eades, *Voluntary conversion of convertible securities and the optimal call strategy*, J. Financial Econom., 23 (1984), pp. 273–301.
[9] W. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[10] M. Harris and A. Raviv, *A sequential model of convertible debt call policy*, J. Finance, 40 (1985), pp. 1263–1282.
[11] J. E. Ingersoll, *A contingent-claims valuation of convertible securities*, J. Financial Econom., 4 (1977), pp. 289–322.
[12] J. E. Ingersoll, *An examination of corporate call policies on convertible securities*, J. Finance, 32 (1977), pp. 463–478.
[13] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1991.
[14] I. Karatzas and S. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.
[15] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, Cambridge, UK, 1990.
[16] R. C. Merton, *Theory of rational option pricing*, Bell J. Econom. Manag. Sci., 4 (1973), pp. 141–183.
[17] M. Miller and F. Modigliani, *The cost of capital, corporation finance, and the theory of investment*, Amer. Econ. Rev., 48 (1958), pp. 261–297.
[18] M. Miller and F. Modigliani, *Dividend policy, growth and the valuation of shares*, J. Business, 34 (1961), pp. 411–433.
[19] B. Øksendal and K. Reikvam, *Viscosity solutions of optimal stopping problems*, Stochastics Stochastics Rep., 62 (1998), pp. 285–301.
[20] A. N. Shiryayev, *Optimal Stopping Rules*, Springer-Verlag, Berlin, 1977.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 86–110

STOCHASTIC GAMES WITH A SINGLE CONTROLLER AND INCOMPLETE INFORMATION*

DINAH ROSENBERG†, EILON SOLAN‡, AND NICOLAS VIEILLE§

Abstract. We study stochastic games with incomplete information on one side, in which the transition is controlled by one of the players.

We prove that if the informed player also controls the transitions, the game has a value, whereas if the uninformed player controls the transitions, the max-min value as well as the min-max value exist, but they may differ.

We discuss the structure of the optimal strategies, and provide extensions to the case of incomplete information on both sides.

Key words. stochastic games, incomplete information, single controller

AMS subject classifications. 91A05, 91A15

DOI. 10.1137/S0363012902407107

1. Introduction. In a seminal work, Aumann and Maschler [1, 2] introduced infinitely repeated two-player zero-sum games with incomplete information on one side. Those are repeated games where the payoff matrix is known to one player, say player 1, but is not known to the other player—all player 2 knows is that the payoff matrix was drawn according to some known probability distribution from a finite set of possible matrices. Aumann and Maschler proved that those games have a value.

The issue faced by player 1 is the optimal use of information. On the one hand, player 1 needs to reveal his information (at least partially) in order to make use of it. On the other hand, any piece of information that is revealed to player 2 can later be exploited against player 1.

In the optimal strategies devised by Aumann and Maschler, player 1 reveals part of his information at the first stage, but no further information is revealed during the game. Player 2, on the other hand, has to play optimally whatever the actual payoff matrix may be. Aumann and Maschler achieved this by using Blackwell's approachability strategies.

When the underlying game is a stochastic game rather than a repeated one, the difficulties the players face are more serious.

Is it optimal for player 1 to reveal information only once in every state, or will he reveal information several times in each state? In repeated games, it does not help to dilute the revelation of information over time, since player 2 would wait until player 1 has revealed all the information he will ever reveal, and since interim payoffs are irrelevant in the long run. In stochastic games, by contrast, the game can move to a different state that can be more or less favorable to the informed player. By

*Received by the editors May 6, 2002; accepted for publication (in revised form) August 9, 2003; published electronically May 25, 2004. The authors acknowledge the financial support of the Arc-en-Ciel/Keshet program for 2001/2002. The research of the second author was supported by the Israel Science Foundation (grant 69/01-1).

http://www.siam.org/journals/sicon/43-1/40710.html

†Laboratoire d'Analyse Géométrie et Applications, Institut Galilée, Université Paris Nord, avenue Jean-Baptiste Clément, 93430 Villetaneuse, France (dinah@zeus.math.univ-paris13.fr).

‡MEDS Department, Kellogg School of Management, Northwestern University and the School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel (eilons@post.tau.ac.il).

§Département Finance et Economie, HEC, 1 rue de la Libération, 78 Jouy-en-Josas, France (vieille@hec.fr).

giving away some information about the true game at the initial stage, player 1 might induce player 2 to adapt in an adverse way, while postponing this disclosure might allow player 1 to escape from specific states. This is a crude explanation for why it may help player 1 to conceal his information for a while.

For player 2 the issue is to devise the analog of Blackwell's approachability strategies for stochastic games.

Sorin [20, 21] and Sorin and Zamir [23] studied classes of stochastic games with incomplete information on one side that have a single nonabsorbing state, and proved that these games have a min-max value, a max-min value, and that the values of the n-stage (resp., λ-discounted) games converge as n goes to infinity (resp., as λ goes to 0) to the max-min value. Rosenberg and Vieille [17] studied recursive games with incomplete information on one side, and proved that the max-min value exists and is equal to the limit of the values of n-stage games (resp., λ-discounted games) as n goes to infinity (resp., as λ goes to 0).

In the present paper we study stochastic games in which one player controls the transitions; that is, the evolution of the stochastic state depends on the actions of one player but is independent of the actions of his opponent.

We show that if player 1 (who is the informed player) controls the transitions, then the game admits a value. We also propose a specific optimal strategy for player 1 and explain the way this strategy uses the additional information he possesses. Roughly speaking, the state space is partitioned into disjoint sets, which are called communicating sets. Whenever the play enters a communicating set, player 1 chooses a stationary nonrevealing strategy, and he plays this strategy until a new communicating set is visited. The random choice of the stationary strategy itself may be revealing, in that the distribution used at stage n to select a stationary strategy depends on the actual payoff function.

If player 2 controls the transitions, then the game admits a min-max value and a max-min value. We use an example to show that the two values may differ.

The techniques and the characterizations we provide extend the ideas of Aumann and Maschler for incomplete information games to our framework.

In the last section of the paper we extend the existence results to the case of stochastic games with a single controller and incomplete information on both sides; that is, to the case when each of the players has some partial private information about the true stochastic game being played.

2. The model and the main results.

2.1. The model. A *two-player zero-sum stochastic game* G is described by (i) a finite set Ω of states, and an initial state $\omega \in \Omega$; (ii) finite action sets I and J for the two players; (iii) a transition rule $q : \Omega \times I \times J \to \Delta(\Omega)$, where $\Delta(\Omega)$ is the simplex of probability distributions over Ω; and (iv) a reward function $g : \Omega \times I \times J \to \mathbf{R}$.

A *two-player zero-sum stochastic game with incomplete information* is described by a finite collection $(G_k)_{k\in K}$ of stochastic games, together with a distribution $p \in \Delta(K)$ over K. We assume that the games G_k differ only through their reward functions g^k, but they all have the same sets of states and actions, and the same transition rule. We denote the common transition rule by q.

The game is played in stages. An element $k \in K$ is chosen according to p. Player 1 is informed of k, while player 2 is not. At every stage $n \in \mathbf{N}$, the two players choose simultaneously actions $i_n \in I$ and $j_n \in J$, and ω_{n+1} is drawn according to $q(\cdot \mid \omega_n, i_n, j_n)$. Both players are informed of (i_n, j_n, ω_{n+1}). We stress that the actual reward $g^k(\omega_n, i_n, j_n)$ is not told to player 2 (but is known to player 1).

We parametrize the game by the initial distribution p and by the initial state ω, and denote it by $\Gamma(p,\omega)$. We write Γ for $(\Gamma(p,\omega))_{(p,\omega)\in\Delta(K)\times\Omega}$.

A few remarks are in order. This model is an extension of the classical model of zero-sum stochastic games. It is also an extension of Aumann and Maschler's model of repeated games with incomplete information, where a zero-sum matrix game is first drawn using p, then played repeatedly over time. Here, nature chooses a stochastic game that is then played over time.

We assume without loss of generality (w.l.o.g.) that $0 \le g^k \le 1$ for every $k \in K$, and we identify each $k \in K$ with the probability measure over K that gives weight 1 to k.

2.2. Strategies and values. Players may base their choices on the stochastic states the play has visited so far, as well as on past choices of actions (of the two players). Player 1 can base his choices also on the state of the world k.

The space of histories of length n is $H_n = (\Omega \times I \times J)^{n-1} \times \Omega$, the space of finite histories is $H = \cup_{n\in\mathbf{N}} H_n$, and the space of plays (infinite histories) is $H_\infty = (\Omega \times I \times J)^\infty$. H_n defines naturally a finite algebra $\mathcal{H}_n$ over H_∞. We equip H_∞ with the σ-algebra $\vee_{n\in\mathbf{N}}\mathcal{H}_n$ spanned by all cylinder sets. A (behavioral) *strategy* of player 1 is a function $\sigma : K \times H \to \Delta(I)$. A strategy for player 2 is a function $\tau : H \to \Delta(J)$. A strategy $\sigma = (\sigma_k)_{k\in K}$ of player 1 is *nonrevealing* if σ_k is independent of $k \in K$.[1]

A strategy σ is stationary if the mixed action played at every stage depends only on the current state. We identify each vector $x = (x_\omega)_{\omega\in\Omega} \in (\Delta(I))^\Omega$ with the stationary strategy that plays the mixed action x_ω whenever the game visits ω. Stationary strategies of player 2 are defined analogously.

Every distribution p, every initial stochastic state ω, and every pair of strategies (σ,τ) induce a probability measure $\mathbf{P}_{p,\omega,\sigma,\tau}$ over $K \times H_\infty$ (equipped with the product σ-algebra). We denote by $\mathbf{E}_{p,\omega,\sigma,\tau}$ the corresponding expectation operator.

We let k, ω_n, i_n and j_n denote, respectively, the actual game being played, the current state at stage n, and the actions played at stage n. These are random variables.

Define the expected average payoff up to stage N by

$$\gamma_N(p,\omega,\sigma,\tau) = \mathbf{E}_{p,\omega,\sigma,\tau}\left[\overline{g}_N\right],$$

where $\overline{g}_N = \frac{1}{N}\sum_{n=1}^N g^k(\omega_n, i_n, j_n)$. For fixed strategies σ,τ, $\gamma_N(p,\omega,\sigma,\tau)$ is linear in p and 1-Lipshitz.

We recall the definitions of the max-min value, the min-max value, and the (uniform) value.

DEFINITION 1. *Player* 1 *can* guarantee $\phi \in \mathbf{R}$ *in the game* $\Gamma(p,\omega)$ *if, for every* $\epsilon > 0$*, there exists a strategy* σ *of player* 1 *and* $N \in \mathbf{N}$ *such that*

$$\forall \tau, \forall n \ge N, \quad \gamma_n(p,\omega,\sigma,\tau) \ge \phi - \epsilon.$$

We then say that the strategy σ guarantees $\phi - \epsilon$ *in* $\Gamma(p,\omega)$.

Player 1 *can guarantee a function* $\phi : \Delta(K) \times \Omega \to \mathbf{R}$ *if player* 1 *can guarantee* $\phi(p,\omega)$ *in the game* $\Gamma(p,\omega)$*, for every* $(p,\omega) \in \Delta(K) \times \Omega$.

Note that, due to the Lipshitz property on payoffs and the compactness of $\Delta(K)$, the integer N in Definition 1 can be chosen to be independent of (p,ω). The definition

[1] The strategy is nonrevealing in the sense that knowledge of the strategy σ and of past play does not enable player 2 to gain information on k. This property relies on the fact that transitions are independent of k.

of a function that is guaranteed by player 2 is similar, with the roles of the two players exchanged.

DEFINITION 2. *Player* 2 *can* defend $\phi \in \mathbf{R}$ *in the game* $\Gamma(p,\omega)$ *if, for every* $\epsilon > 0$ *and every strategy* σ *of player* 1*, there exists a strategy* τ *of player* 2 *and* $N \in \mathbf{N}$ *such that*

$$\forall n \geq N, \quad \gamma_n(p,\omega,\sigma,\tau) \leq \phi + \epsilon. \tag{1}$$

We say that such a strategy τ defends $\phi + \epsilon$ *against* σ *in* $\Gamma(p,\omega)$.

Player 2 *can defend a function* $\phi : \Delta(K) \times \Omega \to \mathbf{R}$ *if player* 2 *can defend* $\phi(p,\omega)$ *in the game* $\Gamma(p,\omega)$*, for every* $(p,\omega) \in \Delta(K) \times \Omega$.

The definition of a function that is defended by player 1 is similar, with the roles of the two players exchanged. The following lemma follows from the definitions.

LEMMA 3. *Player* 1 *can guarantee (resp., defend)* $\max\{\phi,\phi'\}$ *as soon as he can guarantee (resp., defend) both* ϕ *and* ϕ'*. Player* 2 *can guarantee (resp., defend)* $\min\{\phi,\phi'\}$ *as soon as he can guarantee (resp., defend) both* ϕ *and* ϕ'.

DEFINITION 4. *A function* $\phi : \Delta(K) \times \Omega \to \mathbf{R}$ *is*

- *the* (uniform) value *of* Γ *if both players can guarantee* ϕ*;*
- *the* max-min value *of* Γ *if player* 1 *can guarantee* ϕ *and player* 2 *can defend* ϕ*;*
- *the* min-max value *of* Γ *if player* 1 *can defend* ϕ *and player* 2 *can guarantee* ϕ.

Note that the value exists if and only if the max-min value and min-max value exist and coincide.

The value (resp., max-min value, min-max value) is denoted by v (resp., $\underline{v}, \bar{v}$) when it exists. Observe that $\underline{v} \leq \bar{v}$ whenever the two exist. Note that each of the functions $\underline{v}$ and $\bar{v}$ is 1-Lipshitz in p as soon as it exists. When the value v exists, any strategy that guarantees v up to ε is ε-*optimal*. Strategies that are ε-optimal for each $\varepsilon > 0$ are also termed *optimal*.

2.3. Related literature. Most of the literature deals with the polar cases where either Ω or K is a singleton. In the former case, the game is a repeated game with incomplete information. Such games have a value; see Aumann and Maschler [2]. Moreover, an explicit formula for the value exists. Letting $u^*(p)$ be the value of the matrix game with payoff function $\sum_k p_k g^k(\cdot,\cdot)$, the value of the repeated game with incomplete information is the concavification cav u^* of u^* (see section 3.1 for definitions).

When K is a singleton the game is a standard stochastic game. Such games have a value; see Mertens and Neyman [9].

For general stochastic games with incomplete information, little is known, but some classes were studied in the literature. For "Big Match" games Sorin [20, 21] and Sorin and Zamir [23] proved the existence of the max-min value and of the min-max value. These values may differ.

For recursive games, Rosenberg and Vieille [17] proved that the max-min value exists and provided an example where the value does not exist.

Parthasarathy and Raghavan [14] were the first to study the class of stochastic games in which one player controls the transitions. They proved that in this class the value exists, and both players have optimal stationary strategies. They also studied the two-player non-zero-sum game. Filar [5] studied the situation in which states are partitioned into two subsets, and each player controls the transitions from states in his subset of the partition. Several finite-stage algorithms that calculate the value

and optimal stationary strategies were proposed in the literature (see the survey by Raghavan and Filar [15] and the references therein).

Recently Renault [16] studied games where transitions do not depend on the actions chosen by the players and only player 1 observes the current state of the world. All that player 2 observes are the actions of player 1.

2.4. Statements of the results. In the present paper we consider games where a single player controls the transitions.

DEFINITION 5. *Player* 1 controls the transitions *if, for every* $\omega \in \Omega$ *and every* $i \in I$, *the transition* $q(\cdot \mid \omega, i, j)$ *does not depend on* j. *Player* 2 *controls the transitions if the symmetric property holds. We then simply write* $q(\cdot \mid \omega, i)$ *or* $q(\cdot \mid \omega, j)$ *depending on who controls the transitions.*

We prove the following two results.

THEOREM 6. *If player* 1 *controls the transitions, the value exists.*

THEOREM 7. *If player* 2 *controls the transitions, both the min-max value and max-min value exist.*

We provide an example of a game where player 2 controls the transitions and $\overline{v} \neq \underline{v}$. We also provide a characterization of $\overline{v}$ and $\underline{v}$ as a unique solution to a functional equation, and we study the structure of simple optimal strategies of player 1.

We prove no result on the existence of the limit of the values of the finitely repeated games. In the games analyzed so far (see section 2.3), this limit is known to exist and coincides with $\underline{v}$. This property is conjectured to hold in general by Mertens [8].

3. Various tools. This section gathers a few results that we use in subsequent sections. The first three subsections introduce a few extensions of tools used in the analysis of games with incomplete information.

For three vectors $a, b, c \in \mathbf{R}^K$, $c = a + b$ if and only if $c_k = a_k + b_k$ for every $k \in K$, $c = \max\{a, b\}$ if and only if $c_k = \max\{a_k, b_k\}$ for every $k = 1, \ldots, K$, and $a \geq b$ if and only if $a_k \geq b_k$ for every $k = 1, \ldots, K$. For a scalar $r \in \mathbf{R}$, $c = a + r$ if and only if $c_k = a_k + r$ for every $k = 1, \ldots, K$, and $c = ra$ if and only if $c_k = ra_k$ for every $k = 1, \ldots, K$. Finally, the L_1-norm and L_∞-norm will be denoted by $\|\cdot\|_1$ and $\|\cdot\|_\infty$, respectively.

3.1. Concavification. Given a continuous function $u : \Delta(K) \to \mathbf{R}$, we denote by cav u its concavification, namely, the least concave function v defined over $\Delta(K)$, such that $v \geq u$. It is the function whose hypograph is the convex hull of the hypograph of u. Similarly, we denote by vex u its convexification, namely, the largest convex function v such that $v \leq u$. Both cav u and vex u are well defined. Thus, cav and vex are functional operators that act on real-valued functions defined on $\Delta(K)$.

LEMMA 8 (see, e.g., Laraki [7]). *When* $\Delta(K)$ *is endowed with the* L_1*-norm, the two operators* cav *and* vex *map* C*-Lipshitz functions into* C*-Lipshitz functions.*

LEMMA 9. *When the set of functions* $u : \Delta(K) \to \mathbf{R}$ *is endowed with the* L_∞*-norm, the two operators* cav *and* vex *are nonexpansive.*

Proof. For any two real-valued continuous functions over $\Delta(K)$, u, and v, one has

$$\|u^{**} - v^{**}\|_\infty \leq \|u^* - v^*\|_\infty \leq \|u - v\|_\infty,$$

where $u^*(x) = \inf\{\langle y, x\rangle - u(y), y \in \mathbf{R}^K\}$ is the conjugate of u. Since $u^{**} = \text{cav } u$, the result follows.

The argument for the operator vex is analogous. ∎

The following lemma is classical (see, e.g., Mertens, Sorin, and Zamir [10, Corollary V.1.3], or the discussion in Zamir [24, p. 118]).

LEMMA 10. *Assume that player* 1 *can guarantee* u. *Then player* 1 *can guarantee* cav u.

Proof. We briefly recall the main ideas of the proof. Prior to the first stage, player 1 performs a state-dependent lottery, designed as follows. By the Carathéodory theorem there exist $p_e \in \Delta(K), \alpha_e \in [0,1]$, for $e = 1, \ldots, |K|+1$, such that $\sum_e \alpha_e = 1$, $\sum_e \alpha_e p_e = p$, and

$$\text{cav } u(p) \leq \sum_e \alpha_e u(p_e) + \varepsilon. \tag{2}$$

If u is continuous, ε may be set to zero in (2). To guarantee cav $u(p)$ in $\Gamma(p,\omega)$, player 1 chooses a fictitious distribution p_e, and he plays optimally in $\Gamma(p_e,\omega)$. The distributions (p_e) must satisfy that their average is p. We now provide one mechanism player 1 can employ.

For each e set $\mu^k(e) = \alpha_e p_e^k / p^k$ if $p^k > 0$, and we let μ^k be arbitrary if $p^k = 0$. Observe that $\sum_e p^k \mu^k(e) = \sum_e \alpha_e p_e^k = p^k$. The following strategy of player 1 guarantees cav $u(p) - 2\varepsilon$: given k, choose e according to μ^k, and play a strategy σ_e that guarantees $u(p_e) - \varepsilon$. □

The following result will be useful later.

LEMMA 11. *Let* $(\mathcal{A}_i)_{i\in I}$ *be a finite collection of convex closed upwards comprehensive sets, and let* $\mathcal{A}$ *be the set* $\left\{a \in R^K : a = \max_{i\in I} a_i, a_i \in \mathcal{A}_i\right\}$. *Then*

$$f_{\mathcal{A}}(p) = (\text{cav } \max_{i\in I} f_{\mathcal{A}_i})(p),$$

where, for any convex upwards comprehensive set $\mathcal{B}$, $f_{\mathcal{B}}(p) = \inf_{a\in\mathcal{B}} \langle a, p\rangle$.

Proof. Since each $\mathcal{A}_i$ is upwards comprehensive, $\mathcal{A}$ coincides with $\cap_i \mathcal{A}_i$. Therefore $f_{\mathcal{A}} \geq f_{\mathcal{A}_i}$ for each i. In particular $f_{\mathcal{A}} \geq \max_{i\in I} f_{\mathcal{A}_i}$. Since $f_{\mathcal{A}}$ is concave, $f_{\mathcal{A}} \geq \text{cav } \max_{i\in I} f_{\mathcal{A}_i}$.

To prove the opposite inequality, we first observe that if $\mathcal{B}$ is convex, closed, and upwards comprehensive, one has

$$\mathcal{B} = \left\{a \in \mathbf{R}^K : \langle a, p\rangle \geq f_{\mathcal{B}}(p) \text{ for each } p \in \Delta(K)\right\}. \tag{3}$$

Set $g = \text{cav } \max_{i\in I} f_{\mathcal{A}_i}$, and

$$\mathcal{D} = \left\{a \in \mathbf{R}^K : \langle a, p\rangle \geq g(p) \text{ for each } p \in \Delta(K)\right\}.$$

Since $g \geq f_{\mathcal{A}_i}$ for each $i \in I$, and using (3) with $\mathcal{B} = \mathcal{A}_i$, one has $\mathcal{D} \subseteq \mathcal{A}_i$. Therefore, $\mathcal{D} \subseteq \mathcal{A}$, which readily implies $g \geq f_{\mathcal{A}}$. □

3.2. Approachability. We present here the basic approachability result of Blackwell [4], in the framework of stochastic games. Let G be a stochastic game with payoffs in $\mathbf{R}^K$. The description of such a game is the same as that of a two-player zero-sum stochastic game given in section 2.1, except that the reward function g now takes values in $\mathbf{R}^K$. The definition of strategies in this framework is similar to that given in section 2.2.

We denote $\bar{g}_N = \frac{1}{N}\sum_{n=1}^N g(\omega_n, i_n, j_n) \in \mathbf{R}^K$, the average vector payoff in the first N stages.

DEFINITION 12. *A vector $a \in \mathbf{R}^K$ is* approachable *by player 2 at ω if, for every $\varepsilon > 0$, there is a strategy τ of player 2 and $N \in \mathbf{N}$ such that*[2]

$$\forall \sigma, \mathbf{E}_{\omega,\sigma,\tau}\left[\sup_{n \geq N} (\bar{g}_n - a)^+\right] \leq \varepsilon.$$

We say that such a strategy τ approaches $a + \varepsilon$ *at ω.*

In other words, for every ε player 2 has a strategy such that the average payoff vector will eventually not exceed $a + \varepsilon$. Note that a is approachable if and only if $a + \varepsilon$ is approachable for every $\varepsilon > 0$, so that the set of approachable vectors is closed and upwards comprehensive.

Our definition differs slightly from that of Blackwell [4], where the strategy τ is required to be independent of ε (i.e., the original definition of Blackwell reads as $\exists \tau, \forall \varepsilon > 0$, etc.). Any vector a that is approachable in Blackwell's sense is also approachable in our sense. The two definitions are not equivalent. However, it is easily checked that, if a is approachable (in our sense) at each state, it is also approachable in Blackwell's sense.

Every stochastic game with incomplete information $\Gamma(p, \omega)$ induces a stochastic game with vector payoffs $\Gamma^V(\omega)$, in which the payoff coordinates are given by the reward functions of the component games (G_k) of $\Gamma(p, \omega)$. The next lemma relates the two games. Its proof is straightforward.

LEMMA 13. *If $a \in \mathbf{R}^K$ is approachable at ω in the game Γ^V, then player 2 can guarantee $\langle a, p \rangle$ in $\Gamma(p, \omega)$ for every $p \in \Delta(K)$.*

We now state Blackwell's sufficient condition for approachability in this context. Denote by $u_\infty(p, \omega)$ the uniform value of the two-player zero-sum stochastic game with reward function $\sum_{k \in K} p_k g^k(\omega, \cdot, \cdot)$. The existence of u_∞ follows, by Mertens and Neyman [9] or by Parthasarathy and Raghavan [14]. We also denote by $u_n(p, \omega)$ the value of the n-stage version of that game (thus, $\lim_{n \to \infty} u_n = u_\infty$, and the limit is uniform in p).

PROPOSITION 14. *If* cav $u_\infty(p, \omega) \leq \langle a, p \rangle$ *for every $(p, \omega) \in \Delta(K) \times \Omega$, then a is approachable in Γ^V by player 2 at ω, for each $\omega \in \Omega$.*

In this statement (and in later ones), cav u_∞ is the concavification of u_∞ with respect to the first variable, p: cav $u_\infty(p, \omega) = (\text{cav } u_\infty(\cdot, \omega))(p)$.

Sketch of the proof. Let $\varepsilon > 0$, and choose N such that $\|u_N - u_\infty\| \leq \varepsilon$, so that cav $u_N(p, \omega) \leq \langle a + \varepsilon, p \rangle$. We define an auxiliary game with vector payoffs, where each stage corresponds to N stages in the original game. We apply Blackwell's result to the auxiliary game, noting that Blackwell's proof still holds when the stage game changes from stage to stage, with payoffs remaining bounded. □

A more precise result was proved by Milman [13, Theorem 2.1.1]. For results with similar flavor, see Shimkin and Shwartz [19].

3.3. Information revelation. Let σ be a given strategy of player 1. For $n \in \mathbf{N}$, we denote by p_n the conditional distribution over K given $\mathcal{H}_n$: it is the belief held by player 2 about the true game being played.[3] The difference $\|p_n - p_{n+1}\|_1$ may be interpreted as the amount of information that is revealed at stage n.

It is well known (see, e.g., Sorin [22, Lemma 3.4] or Mertens, Sorin, and Zamir

[2] For every real $c \in \mathbf{R}$, $c^+ = \max\{c, 0\}$.

[3] The value of p_n at a specific atom of $\mathcal{H}_n$ depends on σ but not on τ. Since the distribution on $\mathcal{H}_n$ depends on τ, the law of p_n depends on both σ and τ.

[24, Lemma IV.2.1]) that, for each τ,

$$\mathbf{E}_{p,\omega,\sigma,\tau}\left[\sum_{n=1}^{\infty}\|p_n-p_{n+1}\|_2^2\right]\leq|K|\,. \tag{4}$$

Given $p\in\Delta(K)$, we denote by σ^p the average nonrevealing strategy defined by $\sigma^p(h)=\sum_{k\in K}p(k)\sigma(k,h)$ for each finite history h. It is convenient to relate the benefit derived by player 1 from using his information at a given stage to the amount of information revealed at that stage. Let $n\in\mathbf{N}$ be given. The expected payoff at stage n, conditional on past play, is

$$\mathbf{E}_{p,\omega,\sigma,\tau}\left[g_n|\mathcal{H}_n\right]=\sum_{k\in K}p_n(k)g^k(\omega_n,\sigma(k,h_n),\tau(h_n)),$$

where $\sigma(k,h_n)$ and $\tau(h_n)$ are the mixed moves used by the two players at that stage.[4] By Proposition 3.2 and Lemma 3.13 in Sorin [22],

$$\left|\mathbf{E}_{p,\omega,\sigma,\tau}\left[g_n|\mathcal{H}_n\right]-\langle p_n,g(\omega_n,\sigma^{p_n}(h_n),\tau(h_n))\rangle\right|\leq\mathbf{E}\left[\|p_n-p_{n+1}\|_1\,|\mathcal{H}_n\right]. \tag{5}$$

Definition 15. *Let $\widetilde{\mathcal{T}}$ be a set of strategies of player 2. Let $\varepsilon>0$ and σ be given. The strategy $\widetilde{\tau}\in\widetilde{\mathcal{T}}$ is* ε-exhausting information *given (p,ω) and σ if $\widetilde{\tau}$ maximizes $\mathbf{E}_{p,\omega,\sigma,\tau}[\sum_{n=1}^{\infty}\|p_n-p_{n+1}\|_2^2]$ up to ε over $\widetilde{\mathcal{T}}$.*

This notion is relative to the class $\widetilde{\mathcal{T}}$. Which class of strategies is meant will always be clear from the context.

Lemma 16. *Let $\widetilde{\mathcal{T}},\varepsilon,\sigma,(p,\omega)$ as in Definition 15. Let $\widetilde{\tau}\in\widetilde{\mathcal{T}}$ be an ε-exhausting strategy given (p,ω) and σ, and let $N\in\mathbf{N}$ be such that $\mathbf{E}_{p,\omega,\sigma,\widetilde{\tau}}[\sum_{n=N}^{\infty}\|p_n-p_{n+1}\|_2^2]\leq\varepsilon$. Then for each strategy $\tau\in\widetilde{\mathcal{T}}$ that coincides with $\widetilde{\tau}$ until stage N, one has*

$$\mathbf{E}_{p,\omega,\sigma,\tau}\left[\sum_{n=N}^{\infty}\|p_n-p_{n+1}\|_1\right]\leq\sqrt{2\varepsilon},\ \text{and}\ \mathbf{E}_{p,\omega,\sigma,\tau}\left[\|p_l-p_N\|_2\right]\leq\sqrt{2\varepsilon}\ \text{for each}\ l\geq N.$$

Proof. By Jensen's inequality and since (p_n) is a martingale, for every $l\geq N$ one has

$$\left(\mathbf{E}_{p,\omega,\sigma,\tau}\left[\|p_l-p_N\|_2\right]\right)^2\leq\mathbf{E}_{p,\omega,\sigma,\tau}\left[\|p_l-p_N\|_2^2\right]=\mathbf{E}_{p,\omega,\sigma,\tau}\left[\sum_{n=N}^{l-1}\|p_n-p_{n+1}\|_2^2\right]. \tag{6}$$

The equality in (6) is a standard result for martingales; see, e.g., Karatzas and Shreve [6, p. 32]. The second inequality follows. The first inequality follows using Jensen's inequality (applied to each stage independently) and since $\|\cdot\|_1\leq\|\cdot\|_2$. □

The next lemma is specific to stochastic games with incomplete information. In effect, it proves that the amount of information revealed by player 1 up to stage $l\in\mathbf{N}$ is an upper bound on the excess gain from the private information.

Lemma 17. *Let (σ,τ) be given. For every $p\in\Delta(K)$, every $\omega\in\Omega$, and every $l\in\mathbf{N}$, one has*

$$\left|\mathbf{E}_{p,\omega,\sigma,\tau}\left[\overline{g}_l\right]-\mathbf{E}_{p,\omega,\sigma^p,\tau}\left[\overline{g}_l\right]\right|\leq4\mathbf{E}_{p,\omega,\sigma,\tau}\left[\sum_{m=1}^{l}\|p_m-p_{m+1}\|_1\right].$$

[4] There is a small notational inconsistency here, since the right-hand side is the value of the left-hand side on a typical atom of $\mathcal{H}_n$.

Proof. To distinguish between $\mathbf{E}_{p,\omega,\sigma,\tau}$ and $\mathbf{E}_{p,\omega,\sigma^p,\tau}$, we denote the latter by $\widetilde{\mathbf{E}}_{p,\omega,\sigma^p,\tau}$. Let $n \leq l$ be given. Since σ^p is nonrevealing, and by the Lipshitz property,

$$\begin{aligned}
&\left|\langle p_n, g(\omega_n, \sigma^{p_n}(h_n), \tau(h_n))\rangle - \widetilde{\mathbf{E}}_{p,\omega,\sigma^p,\tau}\left[g_n|\mathcal{H}_n\right]\right| \\
&\quad = \left|\langle p_n, g(\omega_n, \sigma^{p_n}(h_n), \tau(h_n))\rangle - \langle p, g(\omega_n, \sigma^{p}(h_n), \tau(h_n))\rangle\right| \\
&\quad \leq 2\left\|p_n - p\right\|_1. \qquad (7)
\end{aligned}$$

By (5), it follows that

$$\left|\mathbf{E}_{p,\omega,\sigma,\tau}\left[g_n|\mathcal{H}_n\right] - \widetilde{\mathbf{E}}_{p,\omega,\sigma^p,\tau}\left[g_n|\mathcal{H}_n\right]\right| \leq 2\left\|p_n - p\right\|_1 + \mathbf{E}_{p,\omega,\sigma,\tau}\left[\left\|p_n - p_{n+1}\right\|_1 |\mathcal{H}_n\right]. \qquad (8)$$

On the other hand, it is easily checked that the probabilities $\mathbf{P}^n_{p,\omega,\sigma,\tau}$ and $\widetilde{\mathbf{P}}^n_{p,\omega,\sigma^p,\tau}$ induced by $\mathbf{P}$ and $\mathbf{P}_{p,\omega,\sigma^p,\tau}$ on $\mathcal{H}_n$ satisfy

$$\left\|\mathbf{P}^n_{p,\omega,\sigma,\tau} - \widetilde{\mathbf{P}}^n_{p,\omega,\sigma^p,\tau}\right\|_1 \leq \mathbf{E}_{p,\omega,\sigma,\tau}\left[\sum_{m=1}^{n}\left\|p_m - p_{m+1}\right\|_1\right]. \qquad (9)$$

By (8) and (9),

$$\left|\mathbf{E}_{p,\omega,\sigma,\tau}\left[g_n\right] - \widetilde{\mathbf{E}}_{p,\omega,\sigma^p,\tau}\left[g_n\right]\right| \leq 4\mathbf{E}_{p,\omega,\sigma,\tau}\left[\sum_{m=1}^{n}\left\|p_m - p_{m+1}\right\|_1\right],$$

which implies the result. □

3.4. A partition of states. In this section we define a partition of the set of states that will be extensively used in what follows. It hinges on the fact that a single player controls the transitions, but it does not matter who is the controller. The partition is similar to the one defined by Ross and Varadarajan [18] for Markov decision processes, who also provide an algorithm to calculate it.

We assume that player 1 controls the transitions. The partition when player 2 controls the transitions is defined analogously. Since transitions are independent of player 2's actions, we here omit player 2's strategy from the notations.

Given $\omega \in \Omega$, we denote by

$$r_\omega \ = \min\{n \in \mathbf{N}, \omega_n = \omega\}$$

the stage of the first visit to ω. By convention, the minimum over an empty set is $+\infty$.

Definition 18. *Let $\omega_1, \omega_2 \in \Omega$. We say that* ω_1 leads to ω_2 *if $\omega_1 = \omega_2$, or if $\mathbf{P}_{\omega_1,\sigma}(r_{\omega_2} < +\infty) = 1$ for some strategy σ of player* 1.

Note that the relation *leads to* is reflexive and transitive.

We define an equivalence relation over Ω by

$$\omega \leftrightarrow \omega' \text{ if and only if } \omega \text{ leads to } \omega' \text{ and } \omega' \text{ leads to } \omega.$$

The equivalence classes of this relation are called *communicating sets.* Given $\omega \in \Omega$, we let C_ω denote the communicating set that contains ω, and we define

$$I_\omega = \{i \in I\colon q(C_\omega \mid \omega, i) = 1\}.$$

Thus, whenever C_ω contains at least two elements, by properly selecting actions in $(I_{\omega'})_{\omega' \in C_\omega}$ player 1 can ensure that the play reaches any state in C_ω infinitely often, provided the play starts in C_ω.

The set I_ω may (but does not have to) be empty only if $|C_\omega| = 1$. Actions in I_ω are called *stay* actions, and any state ω such that $I_\omega = \emptyset$ is a *null* state. The set of nonnull states is denoted by Ω_c. Note that $C_\omega \subseteq \Omega_c$ whenever $\omega \in \Omega_c$.

LEMMA 19. *$\omega \in \Omega_c$ if and only if there is a stationary strategy x_{C_ω} such that C_ω is a recurrent set for x.*

Thus, a state is null if it is visited only finitely many times, whatever player 1 plays: $I_\omega = \emptyset$ if and only if ω is transient for every stationary strategy x.

Proof. We start with the direct implication. Let $\omega \in \Omega_c$. For $\omega' \in C_\omega$, define $x_{\omega'} \in \Delta(A)$ by

$$x_{\omega'}[i] = \begin{cases} 0, & i \notin I_{\omega'}, \\ 1/|I_{\omega'}|, & i \in I_{\omega'}, \end{cases}$$

and let x be any stationary strategy that coincides with $x_{\omega'}$ in each state $\omega' \in C_\omega$. It is easy to show that C_ω is recurrent under x.

The reverse implication is straightforward. □

Some communicating sets are absorbing, in the sense that once entered, the play remains there forever. We now single them out. Let x^* be a fully mixed stationary strategy, i.e., $x^*_\omega[i] > 0$ for every $\omega \in \Omega$ and every $i \in I$. If $R \subseteq \Omega$ is a recurrent set for x^*, then R is a communicating set, and $I_\omega = I$ for every $\omega \in R$.

We denote by Ω_0 the union of these sets:

$$\Omega_0 = \cup\{R : R \text{ recurrent for } x^*\} = \{\omega \in \Omega : I_{\omega'} = I \text{ for every } \omega' \in C_\omega\}.$$

The following lemma implies that the max-min value and the min-max value are constant over C_ω for every $\omega \in \Omega_0$, provided they exist.

LEMMA 20. *Assume player 1 controls transitions. Let $\omega \in \Omega$ and $\omega' \in C_\omega$. If one of the players can guarantee ϕ in $\Gamma(p, \omega)$, he can also guarantee ϕ in $\Gamma(p, \omega')$.*

Proof. Assume first that player 1 can guarantee ϕ in $\Gamma(p, \omega)$. Let σ be a strategy that guarantees $\phi - \varepsilon$ in $\Gamma(p, \omega)$, and let σ^* be the strategy that plays x_{C_ω} until r_ω, then switches to σ. In the game $\Gamma(p, \omega')$, the strategy σ^* guarantees $\phi - \varepsilon'$ for each $\varepsilon' > \varepsilon$.

Assume now that player 2 can guarantee ϕ in $\Gamma(p, \omega)$, but assume to the contrary that he *cannot* guarantee ϕ in $\Gamma(p, \omega')$ for some $\omega' \in C_\omega$. We argue that player 2 cannot guarantee ϕ in $\Gamma(p, \omega)$, a contradiction. Since player 2 cannot guarantee ϕ in $\Gamma(p, \omega')$, there is $\varepsilon > 0$ such that for every strategy τ of player 2 and every $N \in \mathbf{N}$ there is a strategy $\sigma_{\tau,N}$ of player 1 and an integer $n_{\tau,N} \geq N$ such that $\gamma_{n_{\tau,N}}(p, \omega', \sigma_{\tau,N}, \tau) > \phi + \varepsilon$. Let τ and N be given. Let σ^* be the strategy of player 1 defined as follows. Play x_{C_ω} until stage $r_{\omega'}$, then switch to $\sigma_{\tau_\nu,M}$, where τ_ν is the strategy induced by τ after stage ν, and M is sufficiently large so that $\mathbf{P}_{\omega,x_{C_\omega}}(r_{\omega'} < M) > 1 - \frac{\varepsilon}{2}$. One can verify that there is $n' \geq N$ such that $\gamma_{n'}(p, \omega, \sigma^*, \tau) > \phi + \varepsilon/2$, a contradiction. □

When player 2 controls the transitions, we denote by J_ω the set of stay actions at ω:

$$J_\omega = \{j \in J : q(C_\omega \mid \omega, j) = 1\}.$$

3.5. Auxiliary games. As for the analysis of zero-sum repeated games with incomplete information on one side, it is convenient to introduce an average game in which no player is informed of the realization of k.

For notational ease, assume that player 1 is the controller. For every $p \in \Delta(K)$ and every nonnull state $\omega \in \Omega$, we denote by $\widetilde{\Gamma}_R(p,\omega)$ the zero-sum stochastic game with (i) initial state ω, (ii) state space C_ω, (iii) reward function $\sum_k p_k g^k$, (iv) action sets $I_{\omega'}$ and J at each state $\omega' \in C_\omega$, and (v) transition function induced by q.

In the case where player 2 is the controller, the game $\widetilde{\Gamma}_R(p,\omega)$ is defined by restricting player 2's action set to $J_{\omega'}$ in each state $\omega' \in C_\omega$.

Thus, $\widetilde{\Gamma}_R(p,\omega)$ is the stochastic game in which player 1 is not informed of the realization of k (or does not use his information), and the controller is restricted to stay actions. In particular, the game remains in C_ω forever. The letter R is a symbol for *restricted*, while the symbol $\widetilde{\ }$ stands for average.

Note that $\widetilde{\Gamma}_R(p,\omega)$ is a single controller game, so that both players have optimal stationary strategies. Denote by $\widetilde{u}(p,\omega)$ its value. Note that $\widetilde{u}(p,\omega) = u_\infty(p,\omega)$ for each $\omega \in \Omega_0$.

By convention, if ω is a null state, we set $\widetilde{u}(p,\omega) = -\infty$ if player 1 controls the transitions, and $\widetilde{u}(p,\omega) = +\infty$ if player 2 controls the transitions. By Lemma 20, for every communicating set C, $\widetilde{u}(p,\omega)$ is independent of $\omega \in C$.

PROPOSITION 21. *For every $\omega \in \Omega_0$ and every $p \in \Delta(K)$ the value $v(p,\omega)$ of $\Gamma(p,\omega)$ exists and is equal to* cav $\widetilde{u}(p,\omega)(=$ cav $u_\infty(p,\omega))$.

Thus, restricted to Ω_0, the game is similar to a standard repeated game with incomplete information.

Proof. The proof of this lemma is similar to the proof for repeated games with incomplete information on one side. Let $p \in \Delta(K)$ and $\omega \in \Omega_0$ be given. Clearly player 1, by not using his information, can guarantee $\widetilde{u}(p,\omega)$. By Lemma 10, player 1 can guarantee cav $\widetilde{u}(p,\omega)$.

The proof that player 2 can guarantee cav $\widetilde{u}$ is based on approachability results, and closely follows classical lines. Let $a \in \mathbf{R}^K$ be such that

$$\langle a, p\rangle = \text{cav } \widetilde{u}(p,\omega),$$
$$\langle a, q\rangle \geq \text{cav } \widetilde{u}(q,\omega) \text{ for every } q \in \Delta(K).$$

If cav $\widetilde{u}(\cdot,\omega)$ is differentiable at p, then a is defined by the hyperplane tangent to cav $\widetilde{u}(\cdot,\omega)$ at p. By Proposition 14, a is approachable. By Lemma 13, player 2 can guarantee cav $\widetilde{u}(p,\omega)$. □

Let $\Gamma_R(p,\omega)$ be a game similar to $\widetilde{\Gamma}_R(p,\omega)$, but in which player 1 is informed of k. Thus, $\Gamma_R(p,\omega)$ differs from $\Gamma(p,\omega)$ only in that actions of the controller are restricted.

Since in Γ_R, for each nonnull state ω the game cannot leave C_ω, Proposition 21 yields the following.

LEMMA 22. *Let ω be a nonnull state. Then $\Gamma_R(p,\omega)$ has a value, which is* cav $\widetilde{u}(p,\omega)$.

We denote by Γ_R^V the stochastic game with vector payoffs in which the controller is restricted to stay actions.

3.6. Functional equations. Let $\mathcal{B}$ denote the set of functions $\phi\colon \Delta(K) \times \Omega \to [0,1]$ that are 1-Lipshitz with respect to p, when $\Delta(K)$ is endowed with the L_1-norm. We here define three operators on $\mathcal{B}$ that will be used to characterize the solutions of the game.

When transitions are controlled by player 1, we define the operator T_1 by

$$T_1\phi(p,\omega) = \text{cav max}\left\{ \widetilde{u}, \max_{\omega'\in C_\omega, i\notin I_{\omega'}} \mathbf{E}\left[\phi \mid \omega', i\right]\right\}(p,\omega). \tag{10}$$

By convention, a maximum over an empty set is $-\infty$. In this expression, $\mathbf{E}\left[\phi \mid \omega', i\right]$ stands for the expectation of ϕ under $q(\cdot \mid \omega', i)$.

Note that $T_1\phi(p,\omega)$ is equal to $\text{cav max}\left\{\text{cav } \widetilde{u}, \max_{\omega'\in C_\omega, i\notin I_{\omega'}} \mathbf{E}\left[\phi \mid \omega', i\right]\right\}(p,\omega)$ as well.

When transitions are controlled by player 2, we define the operators T_2 and T_3 by

$$T_2\phi(p,\omega) = \text{cav min}\left\{\widetilde{u}, \min_{\omega'\in C_\omega, j\notin J_{\omega'}} \mathbf{E}\left[\phi \mid \omega', j\right]\right\}(p,\omega),$$

$$T_3\phi(p,\omega) = \min\left\{\text{cav } \widetilde{u}, \min_{\omega'\in C_\omega, j\notin J_{\omega'}} \mathbf{E}\left[\phi \mid \omega', j\right]\right\}(p,\omega).$$

Since the maximum (or minimum) of a finite number of elements of $\mathcal{B}$ belongs to $\mathcal{B}$, and since by Lemma 8 concavification preserves Lipshitz properties when $\Delta(K)$ is endowed with the L_1-norm, all three operators T_1, T_2, and T_3 map $\mathcal{B}$ into $\mathcal{B}$. Note that for each $i = 1, 2, 3$ the operator T_i is monotonic: $\phi_1 \leq \phi_2$ implies $T_i\phi_1 \leq T_i\phi_2$. Moreover, for every $\phi \in \mathcal{B}$, $T_i\phi$ is constant over C_ω, for each $\omega \in \Omega$.

We now assume that player 1 controls transitions, and prove a few results on T_1. When transitions are controlled by player 2, identical results hold for both T_2 and T_3. Since the proofs are similar, they are omitted.

PROPOSITION 23.

1. *T_1 has a unique fixed point ϕ.*
2. *The sequences (ϕ_n^0) and (ϕ_n^1) defined by $\phi_0^j = j$, $\phi_{n+1}^j = T_1\phi_n^j$ for $j = 0, 1$, are monotonic and converge uniformly to ϕ.*
3. *ϕ coincides with* cav $\widetilde{u}$ *on Ω_0.*
4. *If $f \in \mathcal{B}$ satisfies $f \leq T_1 f$ (resp., $f \geq T_1 f$), then $f \leq \phi$ (resp., $f \geq \phi$).*

Since $T_1\phi$ and $T_2\phi$ are concave for every $\phi \in \mathcal{B}$, the fixed points of those operators are concave functions. Since 0 is concave, and since T_3 maps concave functions to concave functions, the analog of Proposition 23 for T_3 implies that the fixed point of T_3 is concave as well.

Proof. By monotonicity of T_1, item 2 follows from item 1. Since cav $\widetilde{u}(p,\omega)$ is constant on every communicating set, so is $T_1\phi(p,\omega)$ for every $\phi \in \mathcal{B}$. Since $I_\omega = I$ for every $\omega \in \Omega_0$, $T_1\phi(p,\omega) = \text{cav } \widetilde{u}(p,\omega)$ for every $\phi \in \mathcal{B}$, every $\omega \in \Omega_0$, and every $p \in \Delta(K)$. Thus, item 3 will follow from item 1. We now prove item 1. By Ascoli's characterization, $\mathcal{B}$ is a compact metric space when endowed with the L_∞-norm. By Lemma 9, T_1 is nonexpansive, so that it is continuous on $\mathcal{B}$. Hence T_1 has a fixed point.

We prove uniqueness by contradiction. Let ϕ_1 and ϕ_2 be two distinct fixed points of T_1, and assume w.l.o.g. that $\delta := \max_{(p,\omega)\in\Delta(K)\times\Omega}(\phi_1(p,\omega) - \phi_2(p,\omega)) > 0$. Let

$$D = \{\omega \in \Omega, \phi_1(p,\omega) - \phi_2(p,\omega) = \delta \text{ for some } p \in \Delta(K)\}$$

contain those states where the difference is maximal. Since both $\phi_1(p,\cdot)$ and $\phi_2(p,\cdot)$ are constant on each communicating set, $C_\omega \subseteq D$ whenever $\omega \in D$.

Since $\phi_1 = \phi_2$ on Ω_0, $D \subseteq \Omega \setminus \Omega_0$. Let $\omega \in D$ be given, and let $p_0 \in \Delta(K)$ be an extreme point of the convex hull of the set $\{p \in \Delta(K) : \phi_1(p,\omega) - \phi_2(p,\omega) = \delta\}$.

Thus, $\phi_1(p_0,\omega) - \phi_2(p_0,\omega) = \delta > 0$. Since $\phi_1(\cdot,\omega)$ and $\phi_2(\cdot,\omega)$ are concave, it also follows that $(p_0, \phi_1(p_0,\omega))$ is an extreme point of the hypograph of the concave function $\phi_1(\cdot,\omega)$. This implies

$$\phi_1(p_0,\omega) = \max\left\{\text{cav } \widetilde{u}, \max_{\omega'\in C_\omega, i\notin I_\omega} \mathbf{E}\left[\phi_1|\omega', i\right]\right\}(p_0,\omega).$$

Since $\phi_1(p_0,\omega) > \phi_2(p_0,\omega) \geq \text{cav } \widetilde{u}(p_0,\omega)$, one has $\phi_1(p_0,\omega) = \mathbf{E}\left[\phi_1(p_0,\cdot) \mid \omega', i\right]$ for some $\omega' \in C_\omega$ and $i \notin I_{\omega'}$. Since $T_1\phi_2 = \phi_2$, $\phi_2(p_0,\omega) \geq \mathbf{E}\left[\phi_2(p_0,\cdot) \mid \omega', i\right]$, and therefore

$$\delta = \phi_1(p_0,\omega) - \phi_2(p_0,\omega) \leq \mathbf{E}\left[\phi_1(p_0,\cdot) - \phi_2(p_0,\cdot) \mid \omega', i\right].$$

By the definition of D, this implies that $q(D \mid \omega', i) = 1$.

Thus, for every $\omega \in D$ there exists $\omega' \in C_\omega$ and $i \notin I_{\omega'}$ that satisfy $q(D \mid \omega', i) = 1$. This implies the existence of $\omega_1, \omega_2 \in D$ such that $C_{\omega_1} \neq C_{\omega_2}$ and $\omega_1 \leftrightarrow \omega_2$, a contradiction. This proves 1.

To prove 4, we assume that $\delta = \max_{(p,\omega)\in\Delta(K)\times\Omega}(f(p,\omega) - \phi(p,\omega)) > 0$, and repeat the second part of the proof of 1 to obtain a contradiction. □

4. Incomplete information on one side.

4.1. Preliminaries. We here single out a useful lemma. The lemma concerns a standard two-player zero-sum stochastic game G and its version G_R in which player 1 is restricted to stay actions. Thus, K is a singleton.

LEMMA 24. *Let G be a two-player zero-sum stochastic game with transitions controlled by player* 1*, and let $\omega \in \Omega$. If player* 2 *can guarantee that $\alpha \in \mathbf{R}$ in $G_R(\omega)$ and he can guarantee that $\phi : \Omega \to \mathbf{R}$ in G, then he can also guarantee* $\max\{\alpha, \max_{\omega'\in C_\omega, i\notin I_{\omega'}} \mathbf{E}[\phi|\omega', i]\}$ *in $G(\omega)$.*

Proof. By Lemma 20 player 2 can guarantee α in $G_R(\omega')$ for every $\omega' \in C_\omega$. Let τ_1 be a strategy that guarantees $\alpha+\varepsilon$ in $G_R(\omega')$ for every $\omega' \in C_\omega$, and let τ_2 be a strategy that guarantees $\phi+\varepsilon$ in G. Let $N \in \mathbf{N}$ be such that for every $n \geq N$, every $\omega' \in C_\omega$, and every σ in $G_R(\omega)$, $\gamma_n(\omega', \sigma, \tau_1) \leq \alpha + \varepsilon$, and for every σ in G, $\gamma_n(\omega', \sigma, \tau_2) \leq \phi(\omega') + \varepsilon$.

Define $\nu = 1 + \inf\{n \geq 1, i_n \notin I_{\omega_n}\}$. Define a strategy τ for player 2 as follows.

- Until stage ν, τ plays in blocks of size N (the last block may be shorter). In block $l \geq 0$, where $lN < \nu$, τ forgets past play and follows $\tau_1(\omega_{lN+1})$ for N stages.
- At stage ν, τ forgets past play and starts following τ_2.

Let σ be an arbitrary pure strategy. We will compute an upper bound on $\mathbf{E}_{\omega,\sigma,\tau}[\overline{g}_n]$ for n sufficiently large. Set $L_* = \lceil \frac{\ln\varepsilon}{\ln(1-\varepsilon)} \rceil$[5] and take $n \geq N_1 := \lceil L_* N/\varepsilon^2 \rceil$. Denote by $\overline{g}_{m_1\to m_2}$ the average payoff from stage m_1 to stage m_2. With $\theta^* := N \times \lceil \frac{\nu}{N} \rceil$, since payoffs are nonnegative one has

$$\overline{g}_n \leq \frac{\theta^*}{n}\overline{g}_{\theta^*} + \frac{n+1-\nu}{n}\overline{g}_{\nu\to n}. \tag{11}$$

On the event $\nu \leq n - N$, one has

$$\mathbf{E}_{\omega,\sigma,\tau}\left[\overline{g}_{\nu\to n}|\mathcal{H}_\nu\right] = \mathbf{E}_{\omega_\nu,\sigma_\nu,\tau_2}\left[\overline{g}_{n-\nu+1}\right] \leq \phi(\omega_\nu) + \varepsilon, \tag{12}$$

[5]For every real c, $\lceil c \rceil$ is the smallest integer larger than or equal to c.

where σ_ν is the strategy induced by σ after ν. Since σ is pure, $\nu - 1$ is a stopping time and, using (12),

$$\begin{aligned}\mathbf{E}_{\omega,\sigma,\tau}\left[\overline{g}_{\nu\to n}|\mathcal{H}_{\nu-1}\right] &\leq \mathbf{E}\left[\phi|\omega_{\nu-1}, i_{\nu-1}\right] + \varepsilon \\ &\leq \max_{\omega'\in C_\omega, i\notin I_{\omega'}} \mathbf{E}\left[\phi|\omega', i\right] + \varepsilon.\end{aligned} \tag{13}$$

On the other hand, on the event $\nu > n - N$,

$$\frac{n+1-\nu}{n} \leq \varepsilon. \tag{14}$$

We now proceed to the first term in the decomposition (11) of $\overline{g}_n$. For each l, we let $\pi_l = \mathbf{P}_{\omega,\sigma,\tau}\left(\nu \leq (l+1)N \mid \mathcal{H}_{lN+1}\right)$. By the choice of N,

$$\mathbf{E}_{\omega,\sigma,\tau}\left[\overline{g}_{lN+1\to(l+1)N}|\mathcal{H}_{lN+1}\right] \leq \alpha + \varepsilon \ + \mathbf{P}_{\omega,\sigma,\tau}\left(\nu \leq (l+1)N \mid \mathcal{H}_{lN+1}\right)$$

on the event $lN + 1 < \nu$. By taking expectations, this yields

$$\mathbf{E}_{\omega,\sigma,\tau}\left[\overline{g}_{lN+1\to(l+1)N}\mathbf{1}_{lN+1<\nu}\right] \leq (\alpha+\varepsilon)\mathbf{P}_{\omega,\sigma,\tau}(lN+1<\nu) + \mathbf{P}_{\omega,\sigma,\tau}((l+1)N \geq \nu).$$

By summation over l, and using the definition of θ^*, this yields

$$\mathbf{E}_{\omega,\sigma,\tau}\left[\sum_{l=0}^{\theta^*-1} \overline{g}_{lN+1\to(l+1)N}\right] \leq (\alpha+\varepsilon)\mathbf{E}_{\omega,\sigma,\tau}\left[\theta^*\right] + 1,$$

hence

$$\mathbf{E}_{\omega,\sigma,\tau}\left[\frac{N\theta^*}{n}\overline{g}_{N\theta^*}\right] \leq (\alpha+\varepsilon)\mathbf{E}_{\omega,\sigma,\tau}\left[\frac{N\theta^*}{n}\right] + \frac{N}{n}. \tag{15}$$

The result follows by (11), (13), (14), and (15). □

We shall need a variant of the previous result whose proof is identical to the previous proof. Consider the stochastic game with incomplete information $\Gamma(p,\omega)$ where ω is a nonnull state and assume that transitions are controlled by player 2. Assume that player 1 can guarantee a function ϕ. Then player 1 can also guarantee $\min\left\{\widetilde{u}, \min_{\omega'\in C_\omega, j\notin J_{\omega'}} \mathbf{E}\left[\phi|\omega', j\right]\right\}(p,\omega)$ in $\Gamma(p,\omega)$.

4.2. Transitions controlled by player 1. In this section we assume that transitions are controlled by player 1.

4.2.1. Existence of the value.

PROPOSITION 25. *The unique fixed point of T_1 is the value of Γ.*

Proof. Let ϕ be the unique fixed point of T_1, and fix $\epsilon > 0$ once and for all.

Step 1. Player 1 can guarantee ϕ in Γ. By Lemma 22 player 1 can guarantee cav $\widetilde{u}$. Set $\phi_0^0 = 0$, and, for $n \geq 0$, define $\phi_{n+1}^0 = T_1\phi_n^0$. Assume that player 1 can guarantee ϕ_n^0 for some $n \in \mathbf{N}$. Let $p \in \Delta(K)$ and $\omega \in \Omega$ be given. Plainly, for every $\omega' \in C_\omega$ and every $i \notin I_\omega$, player 1 can guarantee $\mathbf{E}\left[\phi_n^0 \mid \omega', i\right](p,\omega)$ in $\Gamma(p,\omega')$; first he plays the action i at ω', and then a strategy that guarantees $\phi_n^0(p,\cdot)$ (up to ε). By Lemma 20, he can guarantee $\mathbf{E}\left[\phi_n^0 \mid \omega', i\right]$ (p,ω) in $\Gamma(p,\omega)$. By Lemmas 3 and 10 he can guarantee $T_1\phi_n^0 = \phi_{n+1}^0$ in Γ. Since player 1 can guarantee $\phi_0^0 = 0$, and since $\lim_{n\to\infty}\phi_n^0 = \phi$, the result follows.

We now prove that player 2 can guarantee ϕ.

Step 2. Definition of approachable sets. For $\omega \in \Omega$, let $\mathcal{B}_\omega$ be the set of vectors approachable in Γ^V by player 2 at ω. We also define

$$\mathcal{A}_\omega = \left\{a \in \mathbf{R}^K \colon \langle a, p\rangle \geq \operatorname{cav} \widetilde{u}(p,\omega) \text{ for every } p\right\}.$$

By Proposition 14 and Lemma 22, $\mathcal{A}_\omega$ is the set of vectors approachable by player 2 at ω in the stochastic game with vector payoffs Γ_R^V. Both sets $\mathcal{A}_\omega$ and $\mathcal{B}_\omega$ are nonempty, closed, convex, and upwards comprehensive.

For every $\omega \in \Omega$ define

$$\mathcal{D}_\omega = \left\{d = \max\left\{a, \max_{\omega' \in C_\omega, i \notin I_{\omega'}} \mathbf{E}\left[b(\cdot) \mid \omega', i\right]\right\} : a \in \mathcal{A}_\omega, b(\omega'') \in \mathcal{B}_{\omega''} \text{ for every } \omega'' \in \Omega\right\}.$$

Step 3. $\mathcal{D}_\omega \subseteq \mathcal{B}_\omega$. Fix $d \in \mathcal{D}_\omega$. Let τ_1 be a strategy that approaches $a + \varepsilon$ at ω, and let τ_2 be a strategy that approaches $b(\omega'') + \varepsilon$ at each state ω''. For each k the strategy τ_1 guarantees $a^k + \varepsilon$ in the game $\Gamma(k,\omega)$, and τ_2 has a similar property. By Lemma 24, applied independently to each G^k, the strategy obtained by concatenation of τ_1 and τ_2 guarantees $\max\left\{a^k, \max_{\omega' \in C_\omega, i \notin I_{\omega'}} \mathbf{E}\left[b^k(\cdot) \mid \omega', i\right]\right\} + 3\varepsilon = d^k + 3\varepsilon$ in G^k. Lemma 13 implies that $d \in \mathcal{B}_\omega$.

Step 4. Player 2 can guarantee ϕ. Let $f(p,\omega) = \inf_{a \in \mathcal{B}_\omega} \langle a, p\rangle$ and $h(p,\omega) = \inf_{a \in \mathcal{D}_\omega} \langle a, p\rangle$, so that by Step 3 $f \leq h$. By Lemma 13 player 2 can guarantee $\langle a, p\rangle$ in $\Gamma(p,\omega)$ for every $a \in \mathcal{B}_\omega$. Therefore he can guarantee $f(p,\omega)$ as well. By Lemma 11, the definition of $\mathcal{D}_\omega$ may be rephrased as

$$h = \operatorname{cav}\ \max\left\{\operatorname{cav}\ \widetilde{u}, \max_{\omega' \in C_\omega, i \notin I_{\omega'}} \mathbf{E}\left[f \mid \omega', i\right]\right\} = T_1 f.$$

Thus, $f \leq T_1 f$. By item 4 in Proposition 23, $f \leq \phi$. Therefore, player 2 can guarantee ϕ. □

4.2.2. Optimal strategies. The proof of Proposition 25 yields no information on ε-optimal strategies for player 1. We argue here that player 1 has an optimal strategy σ of a simple type. In effect, σ has the following structure. Whenever the play enters a communicating set, say at stage $n \geq 0$, σ randomly selects a nonrevealing stationary strategy that is used until the play moves to a new communicating set, if ever. The random choice of the stationary strategy may itself be revealing, in that the distribution used at stage n to select a stationary strategy depends both on p_n and on k. We describe below such a strategy in more detail.

Let v be the value of the game. Let $(p,\omega) \in \Delta(K) \times \Omega$ be given. Upon entering a communicating set at stage $n \geq 0$, player 1 computes $v(p_n) = \operatorname{cav} \max\{\widetilde{u}, \max_{\omega' \in C_{\omega_n}, i \notin I_{\omega'}} \mathbf{E}[v \mid \omega', i]\}(p_n, \omega_n)$, and performs a state-dependent lottery, as described in the proof of Lemma 10. To be specific, one determines $\widetilde{p}_e \in \Delta(K), \alpha_e \in [0,1]$, for $e = 1, \ldots, |K| + 1$, such that $\sum_e \alpha_e = 1$, $\sum_e \alpha_e \widetilde{p}_e = p_n$, and

$$v(p_n) = \sum_e \alpha_e \max\left\{\widetilde{u}, \max_{\omega' \in C_{\omega_n}, i \notin I_{\omega'}} \mathbf{E}\left[v \mid \omega', i\right]\right\}(\widetilde{p}_e, \omega_n).$$

If G^k is the actual game that is played, player 1 chooses e according to a state-dependent lottery μ^k, where $\mu^k(e) = \alpha_e \widetilde{p}_e^k / p_n^k$.

If $\max\left\{\widetilde{u}, \max_{\omega' \in C_{\omega_n}, i \notin I_{\omega'}} \mathbf{E}\left[v \mid \omega', i\right]\right\}(p^e, \omega_n) = \widetilde{u}(p^e, \omega_n)$, player 1 plays a stationary (nonrevealing) strategy which guarantees $\widetilde{u}(p^e, \omega_n)$ in the restricted game $\widetilde{\Gamma}_R(p^e, \omega_n)$. Recall that there are finite-stage algorithms that compute this strategy.

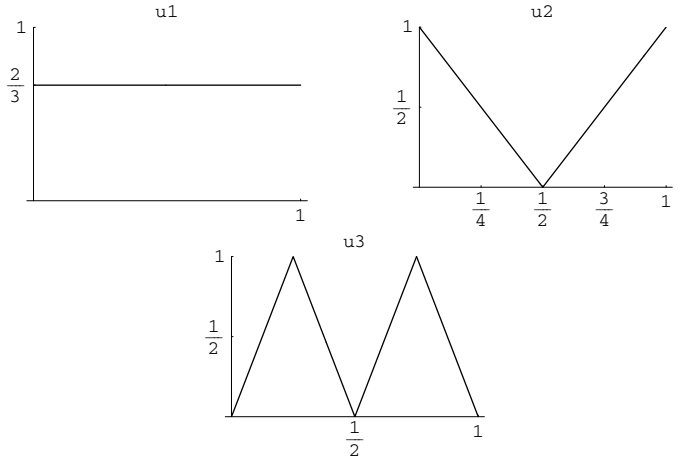


FIG. 1. *The value of the restricted games* $\widetilde{\Gamma}_R(p, \omega_i)$.

If, on the other hand, $\max\left\{\widetilde{u}, \max_{\omega' \in C_{\omega_n}, i \notin I_{\omega'}} \mathbf{E}\left[v \mid \omega', i\right]\right\}(p^e, \omega_n) = \mathbf{E}[v(p^e, \cdot) \mid \omega', i]$ for some $\omega' \in C_{\omega_n}$ and $i \notin I_{\omega'}$, player 1 plays the stationary strategy $x_{C_{\omega_n}}$ until the play reaches ω', and at ω' he plays the action i. He then recursively switches to a strategy that guarantees $v(p^e, \cdot)$.

Under σ, player 1 will end up in finite time playing an optimal stationary strategy in some restricted game $\widetilde{\Gamma}_R(p', \omega')$, with $p' \in \Delta(K)$ and $\omega' \in \Omega$. It can be checked that σ guarantees $v(p, \omega) - \varepsilon$ for every $\varepsilon > 0$. In that sense, σ is optimal. The proof is standard and therefore omitted.

4.2.3. An example. We here provide a simple example that illustrates the basic issues of splitting and information revelation. In particular, in this example the informed player will perform two state-dependent lotteries and therefore reveal information in two different stages of the game, unlike what happens in standard repeated games with incomplete information. The game has three states $\Omega = \{\omega_1, \omega_2, \omega_3\}$, where ω_2 and ω_3 are absorbing. There are two possible payoff functions, so that $K = \{1, 2\}$. A distribution over K is identified with the probability $p \in [0, 1]$ assigned to $k = 2$.

We first describe the main features of the example before providing the payoff and transition matrices of the game.

All actions of player 1 at state ω_1 are stay actions, except one, which leads to either state ω_2 and ω_3 with equal probability.

The value $u_i(p)$ of the restricted game $\widetilde{\Gamma}_R(p, \omega_i)$ is given by (see Figure 1)

$$u_1(p) = 2/3,$$

$$u_2(p) = \max\{1 - 2p, 2p - 1\},$$

$$u_3(p) = \begin{cases} 4p & 0 \le p \le 1/4, \\ 2 - 4p & 1/4 \le p \le 1/2, \\ 4p - 2 & 1/2 \le p \le 3/4, \\ 4 - 4p & 3/4 \le p \le 1. \end{cases}$$

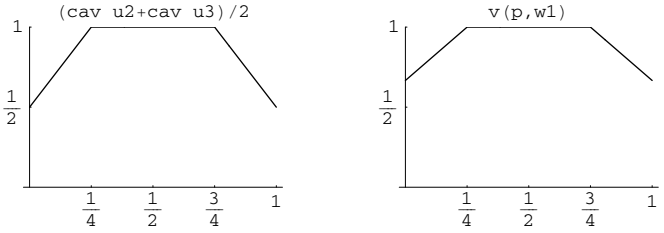


FIG. 2.

Note that $1/2(\text{cav } u_2 + \text{cav } u_3)$ is given by (see Figure 2)

$$1/2(\text{cav } u_2 + \text{cav } u_3)(p) = \begin{cases} 2p + 1/2 & 0 \le p \le 1/4, \\ 1 & 1/4 \le p \le 3/4, \\ 5/2 - 2p & 3/4 \le p \le 1. \end{cases}$$

It is the payoff that is guaranteed by player 1, when starting at (p, ω_1), and exiting from ω_1 without revealing any information. Indeed, once in ω_2 or in ω_3, the game will stay there and therefore the value is given by cav u_2 and cav u_3, respectively.

By section 4.2, the value $v(p, \omega_1) = \text{cav } \max\left\{\frac{2}{3}, \frac{1}{2}(\text{cav } u_2 + \text{cav } u_3)\right\}$ is given by (see Figure 2)

$$v(p, \omega_1) = \begin{cases} (2 + 4p)/3 & 0 \le p \le 1/4, \\ 1 & 1/4 \le p \le 3/4, \\ 2 - 4p/3 & 3/4 \le p \le 1. \end{cases}$$

Assume that the game starts in state ω_1, with $p = 1/8$. Note that $v(1/8, \omega_1) = \frac{1}{2}u_1(0) + \frac{1}{2}(\text{cav } u_2(1/4) + \text{cav } u_3(1/4))$, and that cav $u_3(1/4) = \frac{3}{4}u_3(0) + \frac{1}{4}u_3(1)$, while cav $u_2(1/4) = u_2(1/4)$.

The optimal strategy described in section 4.2.2 is as follows. Player 1 starts by tossing a state-dependent coin. If the coin comes up heads, player 1 plays forever an optimal stationary strategy in the game $k = 1$. If the coin comes up tails, player 1 first plays the nonstay action. The game then moves with equal probability to states ω_2 and ω_3. In the former case, player 1 continues with an optimal nonrevealing stationary strategy in the average game $\tilde{\Gamma}_R(1/4, \omega_2)$. In the latter case, player 2 again tosses a (degenerate) state-dependent coin. If $k = 1$ (resp., $k = 2$), player 1 continues with an optimal strategy in the game $\tilde{\Gamma}_R(0, \omega_3)$ (resp., $\tilde{\Gamma}_R(1, \omega_3)$).

Note that the amount of information revealed by player 1 depends on the actual play.

To complete the example, we provide in Figure 3 payoff matrices that satisfy the required specifications. The vertical arrows that appear in the bottom row of the two top matrices stand for the random transition to either ω_2 or ω_3.

In both states ω_1 and ω_2, player 2 is a dummy. The incomplete information game that corresponds to the game of state ω_3 coincides with example 1.3 in Zamir [24].

4.3. Transitions controlled by player 2. In this section we assume that transitions are controlled by player 2. We prove that both the min-max value and the max-min value exist, but that they may differ.

4.3.1. The max-min value.

LEMMA 26. *The unique fixed point of T_2 is the max-min value of Γ .*

$k=1$

	j_1	j_2	j_3
T	2/3	2/3	2/3
B	↓	↓	↓

$k=2$

	j_1	j_2	j_3	
T	2/3	2/3	2/3	ω_1
B	↓	↓	↓	

	j_1	j_2	j_3
T	1	1	1
B	-1	-1	-1

	j_1	j_2	j_3	
T	-1	-1	-1	ω_2
B	1	1	1	

	j_1	j_2	j_3
T	4	0	2
B	4	0	-2

	j_1	j_2	j_3	
T	0	4	-2	ω_3
B	0	4	2	

FIG. 3. *The payoff matrices.*

Proof. Let ϕ be the unique fixed point of T_2, and fix $\varepsilon > 0$.

Step 1. Player 1 can guarantee ϕ. Set ϕ_0^0 and, for $m \geq 0$, set $\phi_{m+1}^0 = T_2\phi_m^0$. Assume that player 1 can guarantee ϕ_m^0 for some $m \in \mathbf{N}$. By the remark following Lemma 24, player 1 can guarantee $\min\{\widetilde{u}, \min_{\omega' \in C_\omega, j \notin J_{\omega'}} \mathbf{E}[\phi_m^0 \mid \omega', j]\}$. Hence player 1 can also guarantee $\operatorname{cav}\ \min\{\widetilde{u}, \min_{\omega' \in C_\omega, j \notin J_{\omega'}} \mathbf{E}[\phi_m^0 \mid \omega', j]\} = \phi_{m+1}^0$. Since player 1 can guarantee $\phi_0^0 \equiv 0$, and since $\phi = \lim_{m\to\infty} \phi_m^0$, the result follows.

We now prove that player 2 can defend ϕ. Assume that player 2 can defend ϕ_m^1 for some $m \in \mathbf{N}$, and let σ be an arbitrary strategy of player 1. We prove in Steps 2 and 3 below that in this case player 2 can defend ϕ_{m+1}^1. Since $\phi = \lim_{m\to\infty} \phi_m^1$, and since player 2 can defend $\phi_0^1 \equiv 1$, he can defend ϕ as well.

Step 2. Definition of a reply. Given (p, ω), we let $\tau_1(p, \omega)$ be a (stationary) strategy that guarantees $\widetilde{u}(p, \omega) + \varepsilon$ in $\widetilde{\Gamma}_R(p, \omega)$. Choose $N_1 \in \mathbf{N}$ such that $\gamma_n(p, \omega, \widetilde{\sigma}, \tau_1(p, \omega)) \leq \widetilde{u}(p, \omega) + 2\varepsilon$ for every $n \geq N_1$ and every nonrevealing strategy $\widetilde{\sigma}$ of player 1.

By the remark that follows Definition 1, N_1 can be chosen independently of (p, ω). Let $\widetilde{\mathcal{T}}$ be the set of strategies of player 2 in $\widetilde{\Gamma}_R(p, \omega)$, and let $\widetilde{\tau} \in \widetilde{\mathcal{T}}$ be an $\varepsilon^2/32N_1^2$-exhausting information strategy given σ and (p, ω). Choose $N \in \mathbf{N}$ such that

$$\mathbf{E}_{p,\omega,\sigma,\widetilde{\tau}}\left[\sum_{n=N}^{+\infty} \|p_n - p_{n+1}\|_2^2\right] \leq \frac{\varepsilon^2}{32N_1^2}.$$

By Lemma 16,

$$\mathbf{E}_{p,\omega,\sigma,\widetilde{\tau}}\left[\sum_{n=N}^{+\infty} \|p_n - p_{n+1}\|_1\right] \leq \frac{\varepsilon}{\sqrt{32}N_1} \leq \frac{\varepsilon}{4}. \tag{16}$$

We define τ as follows.

- Play $\widetilde{\tau}$ up to stage N.

- At stage N compute $\beta_N := \min\left\{\widetilde{u}, \min_{\omega' \in C_\omega, j \notin J_{\omega'}} \mathbf{E}\left[\phi^1_m \mid \omega', j\right]\right\}(p_N, \omega_N)$.
  - If $\beta_N = \widetilde{u}(p_N, \omega_N)$, play by successive blocks of length N_1: in the $b+1$th block play the strategy $\tau_1(p_{N+bN_1}, \omega_{N+bN_1})$.
  - Otherwise, switch to a strategy that defends the quantity $\min_{\omega' \in C_\omega, j \notin J_{\omega'}} \mathbf{E}\left[\phi^1_m \mid \omega', j\right](p_N, \omega_N) + \varepsilon$ against σ_N, where σ_N is the strategy induced by σ after stage N.

Step 3. The computation. We here prove that τ defends $\phi^1_{m+1}(p,\omega) + 8\sqrt{\varepsilon}$ in $\Gamma(p,\omega)$. We abbreviate $\mathbf{E}_{p,\omega,\sigma,\tau}$ to $\mathbf{E}$. First, we provide an upper bound on the average payoff $\mathbf{E}\left[\overline{g}_{N\to N+n-1}|\mathcal{H}_N\right]$ between stages N and $N+n$ on the event

$$A := \{\beta_N = \widetilde{u}(p_N, \omega_N)\}. \tag{17}$$

First take $n = N_1$. By definition,

$$\mathbf{E}\left[\overline{g}_{N\to N+N_1-1}|\mathcal{H}_N\right] = \mathbf{E}_{p_N,\omega_N,\sigma_N,\tau_1(p_N,\omega_N)}\left[\overline{g}_{N_1}\right].$$

By the choice of N_1,

$$\mathbf{E}_{p_N,\omega_N,\sigma_N^{p_N},\tau_1(p_N,\omega_N)}\left[\overline{g}_{N_1}\right] \le \widetilde{u}(p_N,\omega_N) + 2\varepsilon. \tag{18}$$

On the other hand, by Lemma 17,

$$\begin{aligned}\left|\mathbf{E}_{p_N,\omega_N,\sigma_N^{p_N},\tau_1(p_N,\omega_N)}\left[\overline{g}_{N_1}\right] - \mathbf{E}_{p_N,\omega_N,\sigma_N,\tau_1(p_N,\omega_N)}\left[\overline{g}_{N_1}\right]\right| \\ \le 4\mathbf{E}_{p_N,\omega_N,\sigma_N,\tau_1(p_N,\omega_N)}\left[\sum_{m=1}^{N_1}\|p_m - p_{m+1}\|_1\right].\end{aligned}$$

Thus, using (18),

$$\mathbf{E}\left[\overline{g}_{N\to N+N_1-1}|\mathcal{H}_N\right] \le \widetilde{u}(p_N,\omega_N) + 2\varepsilon + 4\mathbf{E}\left[\sum_{m=N}^{N+N_1-1}\|p_m - p_{m+1}\|_1 |\mathcal{H}_N\right].$$

The same computation applies to any block of N_1 stages. Specifically, for each $b \ge 0$,

$$\begin{aligned}\mathbf{E}\left[\overline{g}_{N+bN_1\to N+(b+1)N_1-1}|\mathcal{H}_{N+bN_1}\right] &\le \widetilde{u}(p_{N+bN_1},\omega_{N+bN_1}) + 2\varepsilon \\ &\quad + 4\mathbf{E}\left[\sum_{m=N+bN_1}^{N+(b+1)N_1-1}\|p_m - p_{m+1}\|_1 |\mathcal{H}_{N+bN_1}\right].\end{aligned}$$

Since $\widetilde{u}(p,\cdot)$ is constant on every communicating set, and since $\widetilde{u}(\cdot,\omega)$ is 1-Lipshitz, $\widetilde{u}(p_{N+bN_1},\omega_{N+bN_1}) \le \widetilde{u}(p_N,\omega_N) + \|p_{N+bN_1} - p_N\|_1$. By taking expectations on the event A (defined by (17)), one gets, by Lemma 16, (16), and since $\|\cdot\|_1 \le \|\cdot\|_2$,

$$\begin{aligned}\mathbf{E}\left[\mathbf{1}_A\overline{g}_{N+bN_1\to N+(b+1)N_1-1}\right] &\le \mathbf{E}\left[\mathbf{1}_A\widetilde{u}(p_N,\omega_N)\right] + 2\varepsilon + \mathbf{E}\left[\mathbf{1}_A\|p_{N+bN_1} - p_N\|_2\right] \\ &\quad + 4\mathbf{E}\left[\mathbf{1}_A\sum_{m=N+bN_1}^{N+(b+1)N_1-1}\|p_m - p_{m+1}\|_2\right] \\ &\le \mathbf{E}\left[\mathbf{1}_A\widetilde{u}(p_N,\omega_N)\right] + 5\sqrt{\varepsilon}.\end{aligned}$$

By averaging over blocks, one obtains for every $n \geq \frac{2}{\varepsilon}(N+N_1)$

$$\mathbf{E}\left[\mathbf{1}_A \overline{g}_n\right] \leq \mathbf{E}\left[\mathbf{1}_A \widetilde{u}(p_N,\omega_N)\right] + 6\sqrt{\varepsilon}. \tag{19}$$

On the other hand, there is $N_2 \in \mathbf{N}$ such that for every $n \geq N_2$,

$$\mathbf{E}\left[\overline{g}_n|\mathcal{H}_N\right] \leq \min_{\omega' \in C_\omega, j \notin J_{\omega'}} \mathbf{E}\left[\phi_m^1 \mid \omega', j\right](p_N,\omega_N) + 2\varepsilon \text{ on the event } \overline{A}. \tag{20}$$

By taking expectations, (19) and (20) yield

$$\begin{aligned}\mathbf{E}\left[\overline{g}_n\right] &\leq \mathbf{E}\left[\min\left\{\widetilde{u}, \min_{\omega' \in C_\omega, j \notin J_{\omega'}} \mathbf{E}\left[\phi_m^1 \mid \omega', j\right]\right\}(p_N,\omega_N)\right] + 8\sqrt{\varepsilon} \\ &\leq \operatorname{cav}\ \min\left\{\widetilde{u}, \min_{\omega' \in C_\omega, j \notin J_{\omega'}} \mathbf{E}\left[\phi_m^1 \mid \omega', j\right]\right\}(p,\omega) + 8\sqrt{\varepsilon}\end{aligned}$$

for every $n \geq \max\{N_2, \frac{2}{\varepsilon}(N+N_1)\}$. □

Let $\underline{v}$ denote the max-min of the game, and let (p,ω) be given. Similar to the discussion in section 4.2.2, it can be checked that there is a simple strategy for player 1 that guarantees $\underline{v}(p,\omega) - \varepsilon$ for each $\varepsilon > 0$. Under this strategy, player 1 chooses at random a nonrevealing stationary strategy whenever the play enters a communicating set, and uses it until the play moves to a new communicating set.

4.4. The min-max value.

LEMMA 27. *The unique fixed point of T_3 is the min-max value of* Γ.

Proof. Let ϕ be the unique fixed point of T_3, and fix $\varepsilon > 0$.

We first prove by induction that player 2 can guarantee ϕ. Set $\phi_0^1 \equiv 1$ and, for $m \geq 0$, set $\phi_{m+1}^1 = T_3\phi_m^1$. Assume that player 2 can guarantee ϕ_m^1 for some $m \in \mathbf{N}$, and let (p,ω) be given. Plainly, for each $\omega' \in C_\omega, j \notin J_{\omega'}$, player 2 can guarantee $\mathbf{E}\left[\phi_m^1 \mid \omega', j\right]$ in $\Gamma(p,\omega')$ by first playing j at ω', and then a strategy that guarantees ϕ_m^1 (up to ε). By Lemma 20, he can guarantee $\mathbf{E}\left[\phi_m^1 \mid \omega', j\right]$ in $\Gamma(p,\omega)$ as well. By Lemma 22, player 2 can guarantee cav $\widetilde{u}$. Thus, he can guarantee $T_3\phi_m^1 = \phi_{m+1}^1$. Since he can guarantee ϕ_0^1, and since $\phi = \lim_{m\to\infty} \phi_m^1$, the result follows.

We now prove that player 1 can defend ϕ_m^0 for each $m \in \mathbf{N}$. Clearly, player 1 can defend $\phi_0^0 \equiv 0$. Assume that player 1 can defend ϕ_m^0 for some $m \in \mathbf{N}$. Let a strategy τ of player 2 and $(p,\omega) \in \Delta(K) \times \Omega$ be given. Set $\nu = 1 + \inf\{n \geq 1, j_n \notin J_{\omega_n}\}$. The supremum of $\mathbf{P}_{p,\omega,\sigma,\tau}(\nu < \infty)$ over all strategies σ coincides with the supremum over all nonrevealing strategies σ.[6] Denote by σ^* a *nonrevealing* strategy that achieves the supremum up to ε. We choose N such that $\mathbf{P}_{p,\omega,\sigma^*,\tau}(\nu > N) \leq \varepsilon$. The strategy σ^* thus exhausts the probability of leaving the initial communicating set. Denote by $\tau_{\min\{\nu,N\}}$ the strategy induced by τ after stage $\min\{\nu,N\}$.

On the event $\nu > N$, there is a strategy $\widetilde{\tau}$ in $\Gamma_R(p,\omega)$ such that $\|\mathbf{P}_{p,\omega_N,\sigma,\widetilde{\tau}} - \mathbf{P}_{p,\omega_N,\sigma,\tau_N}\| \leq \mathbf{P}_{p,\omega_N,\sigma,\tau_N}(\nu < +\infty)$ for every nonrevealing strategy σ in $\Gamma_R(p,\omega)$. This strategy depends on the history up to stage N.

We now define the reply σ of player 1 to τ as follows: play σ^* up to stage $\min\{\nu,N\}$.

- If $\nu > N$, switch to a strategy that defends cav $\widetilde{u}(p,\omega) + \varepsilon$ in $\Gamma_R(p,\omega_N)$ against $\widetilde{\tau}$.

[6] Indeed, for every strategy $\sigma = (\sigma^k)_k$, one has $\mathbf{P}_{p,\omega,(\sigma^k)_k,\tau}(\nu < +\infty) = \sum_k p_k \mathbf{P}_{k,\omega,\sigma^k,\tau}(\nu < +\infty) \leq \max_k \mathbf{P}_{k,\omega,\sigma^k,\tau}(\nu < +\infty)$. Let k_0 achieve the maximum, and let σ' be the nonrevealing strategy that plays σ_{k_0} irrespective of k. Since transitions are independent of k, one has $\mathbf{P}_{p,\omega,\sigma,\tau}(\nu < +\infty) \leq \mathbf{P}_{p,\omega,\sigma',\tau}(\nu < +\infty)$.

$k=1$

	j_1	j_2	j_3	j_4	j_5
T	4	0	0	0	0
B	0	0	0	0	0

$k=2$

	j_1	j_2	j_3	j_4	j_5	
T	0	0	0	0	0	ω_2
B	0	4	4	4	4	

	j_1	j_2	j_3	j_4	j_5
T	0	1	1	3	$\uparrow$
B	0	1	0	3	$\uparrow$

	j_1	j_2	j_3	j_4	j_5	
T	3	0	1	0	$\uparrow$	ω_1
B	3	1	1	0	$\uparrow$	

FIG. 4. *The payoff matrices.*

- If $\nu \leq N$, switch to a strategy that defends $\phi^0_m(p,\omega_\nu)+\varepsilon$ against τ_ν.

Since there are finitely many histories of length N, the set of strategies $(\tau_{\min\{\nu,N\}})$ is finite. It is straightforward to check that σ defends

$$\min\left\{\operatorname{cav}\,\widetilde{u}, \min_{\omega'\in C_\omega, j\notin J_{\omega'}} \mathbf{E}\left[\phi^0_m \mid \omega', i\right]\right\}(p,\omega)+2\varepsilon = \phi^0_{m+1}(p,\omega)+2\varepsilon$$

against τ. □

4.5. An example. Here we provide an example where $\min\{\operatorname{cav} f, g\}$ is strictly larger than $\operatorname{cav}\,\min\{f,g\}$, so that the max-min value and the min-max value differ.

Consider the game depicted in Figure 4, where player 2 controls the transitions, and $|\Omega|=|K|=2$, $|I|=2$, and $|J|=5$. The initial state is ω_1 (bottom two matrices). If in ω_1 player 2 chooses j_5, the game moves to ω_2, which is absorbing. If player 2 chooses another action in ω_1, the game remains in ω_1. Payoffs are as depicted in Figure 4 (the definition of $g^k(\omega_1,\cdot,j_5)$ is irrelevant).

Note that $I_{\omega_1}=\{j_1,j_2,j_3,j_4\}$, $\Omega_0=\{\omega_2\}$, and $C_{\omega_1}=\{\omega_1\}$.

The game $\Gamma_R(p,\omega_1)$ is similar to Example 3.3 in Aumann and Maschler [2]. As calculated in Aumann and Maschler,

$$f(p)=\widetilde{u}(p,\omega_1)=\begin{cases} 3p & 0\leq p\leq 2-\sqrt{3},\\ 1-p(1-p) & 2-\sqrt{3}\leq p\leq \sqrt{3}-1,\\ 3(1-p) & \sqrt{3}-1\leq p\leq 1\end{cases}$$

(see Figure 5). Note that $\operatorname{cav} f \neq f$.

The game $\Gamma_R(p,\omega_2)$ is similar to the game presented in Aumann and Maschler [2, I.2], with all payoffs multiplied by 4.[7] As calculated in Aumann and Maschler,

$$g(p)=\widetilde{u}(p,\omega_2)=4p(1-p).$$

As proved above, the max-min value when the initial state is ω_1 is $(\operatorname{cav}\,\min\{f,g\})(p)$, while the min-max value is $\min\{\operatorname{cav} f, g\}(p)$. A straightforward calculation shows that $\min\{\operatorname{cav} f,g\}(1/2)=3(2-\sqrt{3})$ while $\operatorname{cav}\,\min\{f,g\}(1/2)=4/5$, so the two functions differ. The graphs of the two functions appear in Figure 6.

[7] We added the actions j_3,j_4,j_5, which do not change the calculation of the value. For our purposes, we could have multiplied all payoffs by any α, $3<\alpha<3/(\sqrt{3}-1)$.

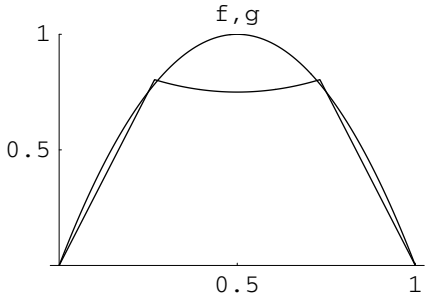


FIG. 5. *The value functions of the restricted games* $\widetilde{\Gamma}_R(p, \omega_i)$.

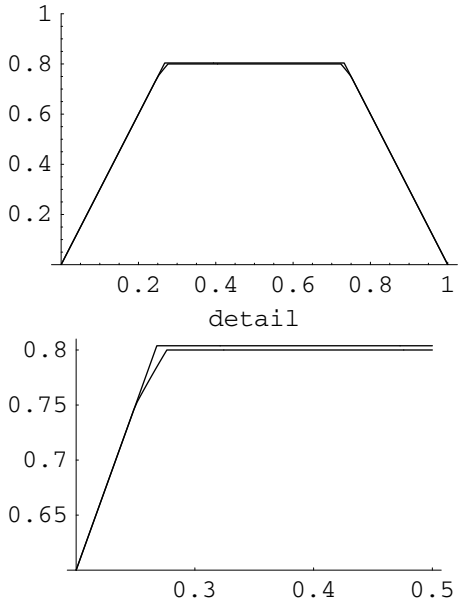


FIG. 6. *The functions* $\min\{\text{cav } f, g\}$ *and* $\text{cav } \min\{f, g\}$.

5. Incomplete information on both sides.

5.1. The model. We now extend our model to the case of incomplete information on both sides; that is, each player has some private information on the game that is to be played. Formally the model is extended as follows. For more details we refer to Mertens, Sorin, and Zamir [10] or Sorin [22].

A *two-player zero-sum stochastic game with incomplete information on both sides* is described by a finite collection $(G_{k,l})_{k\in K, l\in L}$ of stochastic games, together with a distribution $p \in \Delta(K)$ and a distribution $s \in \Delta(L)$. We assume that the games $G_{k,l}$ differ only through their reward functions $g^{k,l}$, but they all have the same sets of states Ω and actions I and J, and the same transition rule q.

The game is played in stages. At the outset of the game a pair $(k, l) \in K \times L$ is chosen according to $p \otimes s$. Player 1 is informed of k, and player 2 of l. At every stage n, the two players choose simultaneously actions $i_n \in I$ and $j_n \in J$, and ω_{n+1} is drawn according to $q(\cdot \mid \omega_n, i_n, j_n)$. Both players are informed of (i_n, j_n, ω_{n+1}).

W.l.o.g. we assume throughout this section that transitions are controlled by player 1. We will only sketch the proofs, since none of them involves any new idea.

5.2. Related literature. The main results in this framework are related to the case $|\Omega| = 1$ (repeated games with incomplete information) and are due to Aumann, Maschler, and Stearns [3] (see also Aumann and Maschler [2]) and Mertens and Zamir [11, 12]. As in the case of incomplete information on one side, we denote by $u(p,s)$ the value of the matrix game with action sets I and J and matrix payoff $(\sum_{k\in K,l\in L} p^k s^l g^{k,l}(i,j))_{i,j}$. Given $f : \Delta(K)\times\Delta(L) \to \mathbf{R}$, we let $\mathrm{cav}_p f$ denote the smallest function that is above f and concave in p, and $\mathrm{vex}_s f$ denotes the largest function that is below f and convex in s.

The min-max value of a repeated game with incomplete information exists and is equal to $\mathrm{vex}_s\mathrm{cav}_p u(p,s)$. The max-min value exists and is equal to $\mathrm{cav}_p\mathrm{vex}_s u(p,s)$.

5.3. Partitioning the states and the average restricted game. Since player 1 controls transitions, the partition defined in section 4 extends to this case, as well as the definition of the average restricted game $\widetilde{\Gamma}_R(p,s,\omega)$ in which none of the players has any information. Denote by $\widetilde{u}(p,s,\omega)$ the value of $\widetilde{\Gamma}_R(p,s,\omega)$. In addition, we define the average restricted game $\widetilde{\Gamma}^1_R(p,s,\omega)$ (resp., $\widetilde{\Gamma}^2_R(p,s,\omega)$) in which player 1 (resp., player 2) is informed of k (resp., l) while his opponent gets no information. Our first goal is to extend Proposition 21.

PROPOSITION 28. *For every $(\omega,p,s)\in\Omega_0\times\Delta(K)\times\Delta(L)$, the min-max value of $\Gamma(p,s,\omega)$ exists and is equal to $\mathrm{vex}_s\mathrm{cav}_p\widetilde{u}(p,s,\omega)$. Similarly the max-min value exists and is equal to $\mathrm{cav}_p\mathrm{vex}_s\widetilde{u}(p,s,\omega)$.*

Proof. The proof follows the proof for repeated games with incomplete information, using the tools developed in the previous sections. We shall only sketch the arguments for the min-max value, and refer for details to Zamir [24].

First, we explain how player 2 can guarantee $\mathrm{vex}_s\mathrm{cav}_p\widetilde{u}(p,s,\omega)$. When player 2 ignores his information, he faces a game with incomplete information on one side with parameter set K and payoffs $\sum_{l\in L} s^l g^{k,l}$. By Proposition 21, player 2 can guarantee $\mathrm{cav}_p\widetilde{u}(p,s,\omega)$ in this game. Therefore by Lemma 10 (with the roles of the two players exchanged) he can also guarantee $\mathrm{vex}_s\mathrm{cav}_p\widetilde{u}(p,s,\omega)$.

To prove that player 1 can defend $\mathrm{vex}_s\mathrm{cav}_p\widetilde{u}(p,s,\omega)$, we adapt Zamir [24, Theorem 4.1]. Let τ be a given strategy of player 2. As in Step 2 of the proof of Lemma 26, we let player 1 play first an ε-exhausting strategy $\widetilde{\sigma}$ given τ. This strategy may be chosen to be nonrevealing (see, e.g., Sorin [22, Lemma IV.4.1]). Player 1 switches at some stage N to a strategy that defends $\mathrm{cav}_p\widetilde{u}(p,s_N,\omega_N)$ (up to ε) in $\Gamma(p,s_N,\omega_N)$ against the continuation strategy τ_N (see Step 3 of Lemma 26). Since $\widetilde{u}(\cdot,\cdot,\omega) = \widetilde{u}(\cdot,\cdot,\omega_N)$, $\mathrm{cav}_p\widetilde{u}(p,s_N,\omega_N) = \mathrm{cav}_p\widetilde{u}(p,s_N,\omega)$. Therefore player 1 defends $\mathbf{E}_{p,s,\omega,\widetilde{\sigma},\tau}\left[\mathrm{cav}_p\widetilde{u}(p,s_N,\omega)\right] \geq \mathrm{vex}_s\mathrm{cav}_p\widetilde{u}(p,s,\omega)$. □

5.4. The max-min value and the min-max value. Let $\mathcal{B}$ denote the set of all functions $\phi : \Delta(K)\times\Delta(L)\times\Omega \to [0,1]$ that are 1-Lipshitz with respect to p and s. Denote by T_4 and T_5 the operators on $\mathcal{B}$ defined by

$$T_4\phi(p,s,\omega) = \mathrm{cav}_p \max\left\{\mathrm{cav}_p\mathrm{vex}_s\widetilde{u}, \max_{\omega'\in C_\omega, i\notin I_{\omega'}} \mathbf{E}[\phi \mid \omega', i]\right\}(p,s,\omega) \tag{21}$$

and

$$T_5\phi(p,s,\omega) = \mathrm{vex}_s\mathrm{cav}_p \max\left\{\mathrm{cav}_p\widetilde{u}, \max_{\omega'\in C_\omega, i\notin I_{\omega'}} \mathbf{E}[\phi \mid \omega', i]\right\}(p,s,\omega). \tag{22}$$

Our main result is the following.

THEOREM 29.

1. *The mappings T_4 and T_5 have unique fixed points, denoted respectively by $\underline{v}$ and $\overline{v}$.*
2. *The function $\underline{v}$ is the max-min value of the game.*
3. *The function $\overline{v}$ is the min-max value of the game.*

Note that if player 2 has no information, there is no vex operator in (21) and (22), and both T_4 and T_5 reduce to T_1. If player 1 has no information, there is no cav operator in (21) and (22), and T_4 and T_5 reduce respectively to T_3 and T_2 with the roles of the players reversed.

Proof. The first assertion follows the same lines as the proof of Proposition 23.

We now prove the second assertion. For $j = 0, 1$, we define the sequence $(\phi_n^j)_{n\geq 0}$ by $\phi_0^j = j$ and $\phi_{n+1}^j = T_4\phi_n^j$. We follow the inductive proof of Proposition 25, Step 1, or the first part of Lemma 27.

The sequence (ϕ_n^0) is increasing and converges uniformly to $\underline{v}$. It is clear that player 1 can guarantee ϕ_0^0. Assuming player 1 can guarantee ϕ_n^0, we prove that he can guarantee ϕ_{n+1}^0. By Lemma 10 it is sufficient to show that he can guarantee both $\mathrm{cav}_p\mathrm{vex}_s\widetilde{u}(p,s,\omega)$ and $\max_{\omega'\in C_\omega, i\notin I_{\omega'}} \mathbf{E}[\phi_n^0 \mid \omega', i]$, which is true by Proposition 28 and by Step 1 of Proposition 25.

To prove that player 2 can defend $\underline{v}$, we combine several ideas from the preceding sections. Let σ be given, and let $\widetilde{T}$ be the set of nonrevealing strategies of player 2. We let τ_σ be a nonrevealing strategy that ε-exhausts the information contained in σ, and choose N as in Step 2 of Lemma 26. Denote by $\nu = 1 + \min\{n \geq 1, i_n \notin I_{\omega_n}\}$. Player 2 plays according to τ_σ up to stage $\min\{\nu, N\}$.

- If $\nu \leq N$, from stage ν on he defends $\phi_n^1(p_\nu, s, \omega_\nu)$.
- If $\nu > N$, we first use the idea of Lemma 27, with the roles of the two players exchanged. Specifically, we define a nonrevealing strategy τ_N^σ that exhausts the probability of leaving the initial communicating set, given the strategy σ_N induced by σ after stage N. Choose N' such that $\mathbf{P}_{p_N,s,\omega_N,\sigma_N,\tau_N^\sigma}(\nu > N') \leq \varepsilon$. Player 2 plays τ_N^σ up to stage $\min\{\nu, N + N'\}$.
  - If $\nu \leq N+N'$, player 2 switches to a strategy that defends $\phi_n(p_\nu, s, \omega_\nu)+\varepsilon$.
  - If $\nu > N + N'$, following Steps 2 and 3 of Lemma 26, player 2 starts to play in blocks of length N_1. In the bth block he forgets past play and follows a strategy that defends $\mathrm{vex}_s\widetilde{u}(p_{N+N'+bN_1}, s, \omega_{N+N'+bN_1})$ in the restricted game $\widetilde{\Gamma}_R^2(p_{N+N'+bN_1}, s, \omega_{N+N'+bN_1})$ against the average continuation strategy $\sigma_{N+N'+bN_1}^{p_{N+N'+bN_1}}$ of player 1.

We now turn to the third assertion. We first prove that player 2 can guarantee $\overline{v}$. By following Steps 2, 3, and 4 of Lemma 25, one proves that player 2 guarantees $\mathrm{cav}_p \max\{\mathrm{cav}_p\widetilde{u}, \max_{\omega'\in C_\omega, i\notin I_{\omega'}} \mathbf{E}[\overline{v} \mid \omega', i]\}(p,s,\omega)$. Hence, by Lemma 10 (with the roles of the two players exchanged), he can guarantee $\mathrm{vex}_s\mathrm{cav}_p \max\{\mathrm{cav}_p\widetilde{u}, \max_{\omega'\in C_\omega, i\notin I_{\omega'}} \mathbf{E}[\overline{v} \mid \omega', i]\} = \overline{v}$.

We now prove that player 1 can defend $\overline{v}$. We first follow Step 2 of Lemma 26. Given τ, we let σ^τ be a strategy in $\widetilde{\Gamma}_R^2(p,s,\omega)$ that exhausts the information contained in τ, and we choose N such that $\mathbf{E}_{p,s,\omega,\sigma^\tau,\tau}[\sum_{n=N}^\infty \|p_n - p_{n+1}\|_1^2] \leq \varepsilon$. Player 1 plays σ^τ up to stage N. He then switches to a strategy that guarantees $\mathrm{cav}_p \max\left\{\mathrm{cav}_p\widetilde{u}(\cdot, s_N, \omega), \max_{\omega'\in C_\omega, i\notin I_{\omega'}} \mathbf{E}[\overline{v} \mid \omega', i]\right\}(p, s_N)$ in $\widetilde{\Gamma}_R^1(p, s_N, \omega_N)$, as in the proof of Proposition 25. The result follows. ∎

REFERENCES

[1] R. J. Aumann and M. B. Maschler, *Repeated games of incomplete information: The zero-sum extensive case*, in Report of the U.S. Arms Control and Disarmament Agency ST-143, Washington, D.C., 1968, Chapter III, pp. 37–116.

[2] R. J. Aumann and M. B. Maschler, *Repeated Games with Incomplete Information*, MIT Press, Cambridge, MA, 1995.

[3] R. J. Aumann, M. B. Maschler, and R. E. Stearns, *Repeated games of incomplete information: An approach to the non-zero sum case*, in Report of the U.S. Arms Control and Disarmament Agency ST-143, Washington, D.C., 1968, Chapter IV, pp. 117–216.

[4] D. Blackwell, *An analog of the minimax theorem for vector payoffs*, Pacific J. Math., 6 (1956), pp. 1–8.

[5] J. A. Filar, *Ordered field property for stochastic games when the player who controls transitions changes from state to state*, J. Optim. Theory Appl., 34 (1981), pp. 503–515.

[6] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[7] R. Laraki, *On the regularity of the convexification operator on a compact set*, J. Convex Anal, to appear.

[8] J.-F. Mertens, *Repeated games*, Proceedings of the International Congress of Mathematicians, Berkeley, CA, 1986, A. M. Gleason, ed., American Mathematical Society, Providence, RI, 1987, pp. 1528–1577.

[9] J.-F. Mertens and A. Neyman, *Stochastic games*, Internat. J. Game Theory, 10 (1981), pp. 53–66.

[10] J.-F. Mertens, S. Sorin, and S. Zamir, *Repeated Games*, CORE Discussion Paper 9420-2, 1994.

[11] J.-F. Mertens and S. Zamir, *The value of two-person zero-sum repeated games with lack of information on both sides*, Internat. J. Game Theory, 1 (1971), pp. 39–64.

[12] J.-F. Mertens and S. Zamir, *Minmax and maxmin of repeated games with incomplete information*, Internat. J. Game Theory, 9 (1980), pp. 201–215.

[13] E. Milman, *Uniform Properties of Stochastic Games and Approachability*, Master Thesis, Tel Aviv University, 2000.

[14] T. Parthasarathy and T. E. S. Raghavan, *An orderfield property for stochastic games when one player controls transition probabilities*, J. Optim. Theory Appl., 33 (1981), pp. 375–392.

[15] T. E. S. Raghavan and J. A. Filar, *Algorithms for stochastic games—a survey*, Z. Oper. Res., 35 (1991), pp. 437–472.

[16] J. Renault, *The Value of Markov Chain Games with Lack of Information on One Side*, preprint.

[17] D. Rosenberg and N. Vieille, *The maxmin of recursive games with incomplete information on one side*, Math. Oper. Res., 25 (2000), pp. 23–35.

[18] K. W. Ross and R. Varadarajan, *Multichain Markov decision processes with a sample path constraint: A decomposition approach*, Math. Oper. Res., 16 (1991), pp. 195–207.

[19] N. Shimkin and A. Shwartz, *Guaranteed performance regions in Markovian systems with competing decision makers*, IEEE Trans. Automat. Control, 38 (1993), pp. 84–95.

[20] S. Sorin, *Big match with lack of information on one side* I, Internat. J. Game Theory, 13 (1984), pp. 201–255.

[21] S. Sorin, *Big match with lack of information on one side* II, Internat. J. Game Theory, 14 (1985), pp. 173–204.

[22] S. Sorin, *A First Course on Zero-Sum Repeated Games*, Mathematiques et Applications 37, Springer-Verlag, Berlin, 2002.

[23] S. Sorin and S. Zamir, *Big match with lack of information on one side* II, in Stochastic Games and Related Topics, T. E. S. Raghavan et al., eds., Kluwer, Dordrecht, The Netherlands, 1991, pp. 101–112.

[24] S. Zamir, *Repeated games of incomplete information: zero-sum*, in Handbook of Game Theory with Economic Applications, Vol. 1, R. J. Aumann and S. Hart, eds., Elsevier Science Publishers B.V., North-Holland, Amsterdam, 1992, pp. 109–154.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 111–136

OPTIMAL CONTROL OF NEUTRAL FUNCTIONAL-DIFFERENTIAL INCLUSIONS*

BORIS S. MORDUKHOVICH† AND LIANWEN WANG‡

Dedicated to Jack Warga in honor of his 80th birthday

Abstract. This paper deals with optimal control problems for dynamical systems governed by constrained functional-differential inclusions of neutral type. Such control systems contain time delays not only in state variables but also in velocity variables, which make them essentially more complicated than delay-differential (or differential-difference) inclusions. Our main goal is to derive necessary optimality conditions for general optimal control problems governed by neutral functional-differential inclusions with endpoint constraints. While some results are available for smooth control systems governed by neutral functional-differential equations, we are not familiar with any results for neutral functional-differential inclusions, even with smooth cost functionals in the absence of endpoint constraints. Developing the method of discrete approximations (which is certainly of independent interest) and employing advanced tools of generalized differentiation, we conduct a variational analysis of neutral functional-differential inclusions and obtain new necessary optimality conditions of both Euler–Lagrange and Hamiltonian types.

Key words. optimal control, functional-differential inclusions of neutral type, variational analysis, discrete approximations, generalized differentiation, necessary optimality conditions

AMS subject classifications. 49K24, 49K25, 49J52, 49M25, 90C31

DOI. 10.1137/S0363012902420376

1. Introduction.

This paper concerns the study of optimal control problems for the so-called *neutral functional-differential inclusions*, which contain time delays in both state and velocity variables. Such inclusions belong to the broad class of *hereditary* systems known also as systems with *memory* or *aftereffect*. They have been investigated in the form of controlled functional-differential *equations* being important for various practical applications, particularly to problems of automatic control, economic dynamics, modeling of ecological, biological, and chemical processes, etc.; see examples and discussions in [1, 4, 10, 12, 14, 15, 19] and their references. Note that some classes of *hyperbolic PDEs* (e.g., the so-called telegraph equations) can be reduced to neutral functional-differential equations, as shown in the above references.

To our knowledge, control problems for neutral functional-differential *inclusions* have not been sufficiently studied in the literature. We are only familiar with results concerning the existence of optimal solutions, local controllability, and relaxation procedures mostly collected in [14].

In this paper we consider the following dynamic optimization (generalized optimal control) problem (P):

$$\text{minimize} \quad J[x] := \varphi(x(a), x(b)) + \int_a^b f(x(t), x(t-\Delta), t)\, dt \tag{1.1}$$

*Received by the editors December 24, 2002; accepted for publication (in revised form) September 30, 2003; published electronically June 4, 2004. This research was partly supported by the National Science Foundation under grants DMS-0072179 and DMS-0304989.
http://www.siam.org/journals/sicon/43-1/42037.html
†Department of Mathematics, Wayne State University, Detroit, MI 48202 (boris@math.wayne.edu).
‡Department of Mathematics and Computer Science, Central Missouri State University, Warrensburg, MO 64093 (lwang@cmsu1.cmsu.edu).

over feasible arcs $x : [a-\Delta, b] \to \mathbb{R}^n$, which are continuous on $[a-\Delta, a)$ and $[a, b]$ (with a possible *jump* at $t = a$) and are such that the combination $x(\cdot) - Ax(\cdot - \Delta)$ is absolutely continuous on $[a, b]$, satisfying the *neutral functional-differential inclusion*

$$\frac{d}{dt}[x(t) - Ax(t-\Delta)] \in F(x(t), x(t-\Delta), t) \quad \text{for a.e. } t \in [a, b], \tag{1.2}$$

$$x(t) = c(t), \quad t \in [a-\Delta, a), \tag{1.3}$$

with the endpoint constraints

$$(x(a), x(b)) \in \Omega \subset \mathbb{R}^{2n}. \tag{1.4}$$

We always assume that $F : \mathbb{R}^n \times \mathbb{R}^n \times [a, b] \rightrightarrows \mathbb{R}^n$ is a set-valued mapping of closed graph, Ω is a closed set, $\Delta > 0$ is a constant delay, and A is a constant $n \times n$ matrix.

Note that the neutral-type operator on the left-hand side of (1.2) is given in the *Hale form* [10], which seems to be essential for our consideration. Observe to this end that trajectories $x(t)$ of (1.2) are generally *discontinuous* on $[a-\Delta, b]$ while their linear combinations $x(t) - Ax(t-\Delta)$ behave nicely on $[a, b]$. Note also that trajectories of the neutral inclusion may be assumed to be discontinuous not only at $t = a$ but also at the points $t = a + j\Delta \in [a, b]$, $j = 1, 2, \ldots$. Moreover, the results obtained in this paper can be easily extended to problems with the cost function φ depending on $x(a + j\Delta)$, the constraints (1.4) given at these intermediate points, and the integrand f in (1.1) depending on the velocity $\dot{z}(t)$ for $z(t) := x(t) - Ax(t-\Delta)$, $t \in [a, b]$. We can also consider the cases of *multiple delays* $\Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_m > 0$ as well as *variable delays* $\Delta(t)$, where $\Delta(\cdot)$ is a Lipschitz continuous (hence almost everywhere differentiable) function satisfying the assumption

$$|\dot{\Delta}(t)| < \alpha \in (0, 1) \quad \text{for a.e. } t \in [a, b],$$

which ensures that the function $t - \Delta(t)$ is invertible on $[a, b]$. For simplicity we focus in what follows on problem (P) formulated in (1.1)–(1.4).

Our primary goal is to derive *necessary optimality conditions* for problem (P) under general assumptions on the initial data. For *nondelayed* systems governed by differential inclusions ($\Delta = 0, A = 0$) necessary optimality conditions have been studied intensively during recent years; see [5, 11, 16, 21, 27, 28, 29, 31] and the references therein. Some results are known for *delay-differential* (or differential-difference) inclusions corresponding to $A = 0$ in (1.2); see [6, 7, 17, 23, 24]. We are not familiar with any necessary optimality conditions obtained for problem (P) governed by neutral functional-differential inclusions with $A \neq 0$ in (1.2) besides the case of *smooth* control systems corresponding to

$$F(x, y, t) = \{v \in \mathbb{R}^n \mid v = g(x, y, u, t), u \in U\} \tag{1.5}$$

with continuously differentiable functions φ, f, g in (1.1) and (1.2) as well as those describing endpoint constraints; see [3, 9, 13, 19] and their references.

Observe that neutral-type systems are essentially different from their counterparts with $A = 0$. In particular, it is well known that an analog of the Pontryagin maximum principle does *not* generally hold for neutral systems, even in the classical smooth framework of (1.5), with *no convexity* assumptions; see, e.g., [9, 15]. In a sense, neutral-type systems combine properties of continuous-time and discrete-time

control systems; indeed, they can be treated as discrete-time systems regarding velocity variables. On the other hand, neutral systems have some similarities with the so-called *hybrid* and *differential-algebraic* equations that are important in engineering control applications.

In this paper we derive necessary optimality conditions for the neutral-type control problem (P) under general assumptions on its initial data involving nonsmooth functions and nonconvex sets. These conditions are obtained in extended *Euler–Lagrange* and *Hamiltonian* forms involving advanced generalized differential constructions of variational analysis; see section 4. Note that the results obtained seem to be new even in the case of nondelayed problems with $A \neq 0$ corresponding to *implicit* differential inclusions.

Our approach is based on the *method of discrete approximations*, in the manner developed in [19, 21] for nondelayed differential inclusions and in [23, 24] for delay-differential systems with $A = 0$. This method, which is certainly of independent interest from both qualitative and numerical viewpoints, allows us to construct a *well-posed* parametric family of optimal control problems for approximating systems governed by *discrete-time* analogs of neutral functional-differential inclusions. A crucial issue is to establish *stability* of such approximations that ensures an appropriate *strong convergence* of optimal solutions. Convergence analysis of this method and its application to necessary optimality conditions for neutral systems are essentially more involved in comparison with the cases of differential and delay-differential inclusions.

The approximating discrete-time control problems can be reduced to special *finite-dimensional* problems of *nonsmooth programming* with an increasing number of *geometric constraints* that may have empty interiors. To handle such problems, we use suitable generalized differential tools of variational analysis satisfying a comprehensive calculus that allows us to derive general necessary optimality conditions for finite-difference analogs of neutral functional-differential inclusions. Then passing to the limit from well-posed discrete approximations with the strong convergence of optimal solutions and employing generalized differential calculus, we obtain necessary optimality conditions for the original problem (P).

The rest of the paper is organized as follows. In section 2 we show that some combination built upon a given admissible trajectory of the neutral inclusion (1.2) can be *strongly approximated* by the corresponding combination built upon admissible trajectories of discrete-time systems. This result, important for its own sake, plays a crucial role in the construction of well-posed discrete approximations to the original problem (P) and in the subsequent justification of a strong convergence of their optimal solutions to the given optimal trajectory for (P).

Such a convergence analysis is conducted in section 3 for a sequence of well-posed discrete approximations to (P) involving an appropriate perturbation of the endpoint constraints (1.4) that is *consistent* with the step of discretization. The required strong convergence of optimal solutions is established under an intrinsic property of the original problem (P) called *relaxation stability*. This property, imposing the equality between the optimal values in (P) and its relaxation (convexification), goes far beyond the convexity assumption on the velocity sets $F(x, y, t)$.

Section 4 contains the basic constructions and required material on generalized differentiation needed for performing a variational analysis of discrete-time and continuous-time optimal control problems in the subsequent sections. These constructions and calculus rules are used in section 5 for deriving general necessary optimality conditions for nonconvex discrete-time inclusions arising in discrete approximations

of the original problem (P). The main results on the extended Euler–Lagrange and Hamiltonian conditions for neutral functional-differential inclusions are derived in section 6 via passing to the limit from discrete approximations.

Our notation is basically standard; cf. [21] and [26]. Recall that, given a set-valued mapping (multifunction) $F : X \rightrightarrows Y$ between finite-dimensional spaces, the *Painlevé–Kuratowski upper/outer limit* of $F(x)$ as $x \to \bar{x}$ is defined by

$$\operatorname*{Lim\,sup}_{x\to\bar{x}} F(x) := \{y \in Y \mid \exists\, x_k \to \bar{x},\ \exists\, y_k \to y \text{ with } y_k \in F(x_k) \text{ for all } k \in \mathbb{N}\},$$

where $\mathbb{N}$ stands for the collection of all natural numbers.

2. Discrete approximations of neutral inclusions. This section concerns the study of discrete approximations of an *arbitrary* admissible trajectory to the neutral functional-differential inclusion (1.2) with the initial condition (1.3). We show that, under fairly general assumptions, any admissible trajectory to (1.2) and (1.3) can be *strongly approximated*, in the sense indicated below, by the corresponding trajectories to finite-difference inclusions obtained from (1.2) by the Euler scheme. This result is constructive, providing efficient estimates for the approximation rate, and hence it is certainly of independent interest for numerical analysis and applications.

Let $\bar{x}(\cdot)$ be an *admissible trajectory* in (P), i.e., it is continuous on $[a-\Delta, a)$ and $[a, b]$ (with a possible jump at $t = a$), the combination $x(\cdot) - Ax(\cdot - \Delta)$ is absolutely continuous on $[a, b]$, and relations (1.2) and (1.3) are satisfied. Note that the endpoint constraints (1.4) may not hold for $\bar{x}(\cdot)$; if they do hold, $\bar{x}(\cdot)$ is *feasible* to (P). The following *standing assumptions* are imposed throughout the paper.

(H1) There are an open set $U \subset \mathbb{R}^n$ and two positive numbers l_F and m_F such that $\bar{x}(t) \in U$ for all $t \in [a-\Delta, b]$, the sets $F(x, y, t)$ are closed, and one has

$$\begin{aligned} F(x, y, t) &\subset m_F\mathbb{B} \quad \text{for all } (x, y, t) \in U \times U \times [a, b],\\ F(x_1, y_1, t) &\subset F(x_2, y_2, t) + \ell_F(|x_1 - x_2| + |y_1 - y_2|)\mathbb{B} \end{aligned}$$

if $(x_1, y_1), (x_2, y_2) \in U \times U$ and $t \in [a, b]$, where $\mathbb{B}$ stands for the closed unit ball in $\mathbb{R}^n$.

(H2) $F(x, y, t)$ is Hausdorff continuous for a.e. $t \in [a, b]$ uniformly in $(x, y) \in U \times U$.

(H3) The function $c(\cdot)$ is continuous on $[a-\Delta, a]$.

Following [8], we consider the so-called *averaged modulus of continuity* for the multifunction $F(x, y, t)$ with $(x, y) \in U \times U$ and $t \in [a, b]$ that is defined by

$$\tau(F; h) := \int_a^b \sigma(F; t, h)\, dt,$$

where $\sigma(F; t, h) := \sup\{\vartheta(F; x, y, t, h) \mid (x, y) \in U \times U\}$ with

$$\vartheta(F; x, y, t, h) := \sup\left\{\operatorname{haus}(F(x, y, t_1), F(x, y, t_2)) \mid (t_1, t_2) \in \left[t - \frac{h}{2}, t + \frac{h}{2}\right] \cap [a, b]\right\},$$

and where $\operatorname{haus}(\cdot, \cdot)$ stands for the Hausdorff distance between two compact sets. It is proved in [8] that if $F(x, y, t)$ is Hausdorff continuous for a.e. $t \in [a, b]$ uniformly in $(x, y) \in U \times U$, then $\tau(F; h) \to 0$ as $h \to 0$. This fact is essentially used in what follows.

Let us construct a sequence of discrete approximations of the given neutral-differential inclusion replacing the derivative in (1.2) by the *Euler finite difference*

$$\frac{d}{dt}[x(t)-Ax(t-\Delta)]\approx\frac{x(t+h)-Ax(t+h-\Delta)-x(t)+Ax(t-\Delta)}{h}.$$

For any $N\in\mathbb{N}:=\{1,2,\dots\}$ we consider the *step of discretization* $h_N:=\frac{\Delta}{N}$ and define the *discrete grid/partition* $t_j:=a+jh_N$ as $j=-N,\dots,k$ and $t_{k+1}:=b$, where k is a natural number determined from $a+kh_N\le b<a+(k+1)h_N$. One clearly has $t_{-N}=a-\Delta$, $t_0=a$, and $h_N\to 0$ as $N\to\infty$. Then the corresponding *neutral functional-difference inclusions* associated with (1.2) and (1.3) are given by

$$\begin{cases}x_N(t_{j+1})-Ax_N(t_{j+1}-\Delta)\in x_N(t_j)-Ax_N(t_j-\Delta),\\ +h_NF(x_N(t_j),x_N(t_j-\Delta),t_j)\quad\text{for } j=0,\dots,k,\\ x_N(t_j)=c(t_j)\quad\text{for } j=-N,\dots,-1.\end{cases}\tag{2.1}$$

A collection of vectors $\{x_N(t_j)\mid j=-N,\dots,k+1\}$ satisfying (2.1) is a *discrete trajectory* and the corresponding collection

$$\left\{\frac{x_N(t_{j+1})-Ax_N(t_{j+1}-\Delta)-x_N(t_j)+Ax_N(t_j-\Delta)}{h_N}\mid j=0,\dots,k\right\}$$

is a *combined discrete velocity* for (2.1). We consider *extensions* $x_N(t)$ *of discrete trajectories* to the continuous-time interval $[a-\Delta,b]$ defined piecewise linearly on $[a,b]$ and piecewise constantly, continuously from the right on $[a-\Delta,a)$. We also define piecewise constant *extensions of combined discrete velocities* on $[a,b]$ by

$$v_N(t):=\frac{x_N(t_{j+1})-Ax_N(t_{j+1}-\Delta)-x_N(t_j)+Ax_N(t_j-\Delta)}{h_N}$$

for $t\in[t_j,t_{j+1})$, $j=0,\dots,k$. It is easy to verify that

$$x_N(t)-Ax_N(t-\Delta)=x_N(a)-Ax_N(a-\Delta)+\int_a^t v_N(s)\,ds\quad\text{for } t\in[a,b]\quad\text{and}$$

$$\frac{d}{dt}[x_N(t)-Ax_N(t-\Delta)]=v_N(t)\quad\text{for a.e. } t\in[a,b].$$

Let $W^{1,2}[a,b]$ be a standard Sobolev space of absolutely continuous functions $x:[a,b]\to\mathbb{R}^n$ with the norm

$$\|x(\cdot)\|_{W^{1,2}}:=\max_{t\in[a,b]}|x(t)|+\left(\int_a^b|\dot{x}(t)|^2dt\right)^{1/2}.$$

The next theorem, important in what follows, establishes a *strong approximation* of any admissible trajectory for the given neutral functional-differential inclusion by solutions corresponding to discrete approximations (2.1).

THEOREM 2.1. *Let* $\bar{x}(\cdot)$ *be an admissible trajectory for* (1.2) *and* (1.3) *under hypotheses* (H1)–(H3). *Then there is a sequence* $\{z_N(t_j)\mid j=-N,\dots,k+1\}$, $N\in\mathbb{N}$, *of solutions to discrete inclusions* (2.1) *such that* $z_N(t_0)=\bar{x}(a)$ *for all* $N\in\mathbb{N}$, *the extended discrete trajectories* $z_N(t)$, $a-\Delta\le t\le b$, *converge uniformly to* $\bar{x}(\cdot)$ *on*

$[a-\Delta, b]$, and their extended combinations $z_N(t) - Az_N(t-\Delta)$ converge to $\bar{x}(t) - A\bar{x}(t-\Delta)$ in the $W^{1,2}$-norm on $[a,b]$ as $N \to \infty$. In particular, some subsequence of $\{\frac{d}{dt}[z_N(t) - Az_N(t-\Delta)]\}$ converges pointwisely to $\frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)]$ for a.e. $t \in [a,b]$.

Proof. Using the density of step functions in $L^1[a,b]$, we first select a sequence $\{\omega_N(\cdot)\}$, $N \in \mathbb{N}$, such that each $\omega_N(t)$ is constant on the interval $[t_j, t_{j+1})$ for $j = 0, \ldots, k$ and that $\omega_N(\cdot)$ converge to $\frac{d}{dt}[\bar{x}(\cdot) - A\bar{x}(\cdot - \Delta)]$ as $N \to \infty$ in the norm topology of $L^1[a,b]$. It follows from (H1) that

$$|\omega_N(t)| \le \left|\omega_N(t) - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)]\right| + \left|\frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)]\right| \le 1 + m_F$$

for all $t \in [a,b]$ and $N \in \mathbb{N}$. In the estimates below we use the sequence

$$\xi_N := \int_a^b \left|\frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] - \omega_N(t)\right| dt \to 0 \quad \text{as } N \to \infty.$$

Denote $\omega_{N_j} := \omega_N(t_j)$ and define discrete functions $\{u_N(t_j) \mid j = -N, \ldots, k+1\}$ recurrently by

$$\begin{cases} u_N(t_j) := \bar{x}(t_j) \quad \text{for } j = -N, \ldots, 0, \\ u_N(t_{j+1}) := Au_N(t_{j+1} - \Delta) + u_N(t_j) - Au_N(t_j - \Delta) \\ \qquad\qquad + h_N\omega_{N_j} \quad \text{for } j = 0, \ldots, k. \end{cases}$$

Then the extended (in the above way) discrete functions satisfy

$$\begin{cases} u_N(t) = \bar{x}(t_j) \quad \text{for } t \in [t_j, t_{j+1}),\ j = -N, \ldots, -1, \\ u_N(t) - Au_N(t-\Delta) = \bar{x}(a) - A\bar{x}(a-\Delta) + \displaystyle\int_a^t \omega_N(s)\, ds \quad \text{for } t \in [a,b]. \end{cases}$$

Next we denote $r_N(t) := u_N(t) - \bar{x}(t)$, $y_N(t) := |r_N(t) - Ar_N(t-\Delta)|$ and prove that $|r_N(t)| \to 0$ uniformly in $[a,b]$ as $N \to \infty$. Indeed, for any $t \in [a,b]$ one has

$$\begin{aligned} y_N(t) :&= |u_N(t) - Au_N(t-\Delta) - [\bar{x}(t) - A\bar{x}(t-\Delta)]| \\ &\le \int_a^t \left|\omega_N(s) - \frac{d}{dt}[\bar{x}(s) - A\bar{x}(s-\Delta)]\right| ds \le \xi_N, \end{aligned}$$

which implies the estimates

$$\begin{aligned} |r_N(t)| &\le y_N(t) + |A| \cdot |r_N(t-\Delta)| \\ &\le y_N(t) + |A| y_N(t-\Delta) + |A|^2 |r_N(t-2\Delta)| \le \cdots \\ &\le y_N(t) + |A| y_N(t-\Delta) + \cdots + |A|^m y_N(t-m\Delta) \\ &\quad + |A|^{m+1} |r_N(t-(m+1)\Delta)|. \end{aligned}$$

Observe that $c(\cdot)$ is *uniformly continuous* on $[a-\Delta, a]$ due to assumption (H3). Picking an arbitrary sequence $\beta_N \downarrow 0$ as $N \to \infty$, we therefore have

$$|c(t') - c(t'')| \le \beta_N \quad \text{whenever } t', t'' \in [t_j, t_{j+1}],\ j = -N, \ldots, -1.$$

Choose an integer number m such that $a - \Delta \le b - (m+1)\Delta < a$. Then $t - (m+1)\Delta \in [t_j, t_{j+1})$ for some $j \in \{-N, \ldots, -1\}$, which implies that

$$|r_N(t-(m+1)\Delta)| \le |c(t_j) - c(t-(m+1)\Delta)| \le \beta_N.$$

Since $m \in \mathbb{N}$ does not depend on N, this gives

$$|r_N(t)| \le \xi_N(1 + |A| + \cdots + |A|^m) + |A|^{m+1}\beta_N \to 0 \quad \text{as } N \to \infty. \tag{2.2}$$

Now consider a sequence $\{\zeta_N\}$ defined by

$$\zeta_N := h_N \sum_{j=0}^{k} \operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t_j))$$

and show that $\zeta_N \downarrow 0$ as $N \to \infty$. By construction of ζ_N and the averaged modulus of continuity $\tau(F; h)$ we get the following estimates:

$$\begin{aligned}
\zeta_N &= \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t_j))\, dt \\
&= \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t))\, dt \\
&\quad + \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} [\operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t_j)) \\
&\qquad\qquad - \operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t))]\, dt \\
&\le \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t))\, dt + \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \sigma(F; t, h_N)\, dt \\
&\le \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t))\, dt + \tau(F; h_N).
\end{aligned}$$

Further, assumption (H1) implies that for any $t \in [t_j, t_{j+1})$ with $j = 0, \dots, k$ one has

$$\begin{aligned}
&\operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t)) - \operatorname{dist}(\omega_{N_j}; F(u_N(t), u_N(t - \Delta), t)) \\
&\quad \le \operatorname{dist}(F(u_N(t_j), u_N(t_j - \Delta), t), F(u_N(t), u_N(t - \Delta), t)) \\
&\quad \le \ell_F(|u_N(t_j) - u_N(t)| + |u_N(t_j - \Delta) - u_N(t - \Delta)|).
\end{aligned}$$

Taking into account that

$$\begin{aligned}
&|u_N(t_j) - Au_N(t_j - \Delta) - [u_N(t) - Au_N(t - \Delta)]| \\
&\quad = \left| \int_{t_j}^{t} \omega_N(s)\, ds \right| \le (1 + m_F)(t_{j+1} - t_j) = (1 + m_F)h_N := a_N \downarrow 0,
\end{aligned}$$

we arrive at

$$\begin{aligned}
|u_N(t) - u_N(t_j)| &\le a_N + |A| \cdot |u_N(t - \Delta) - u_N(t_j - \Delta)| \\
&\le a_N(1 + |A| + \cdots + |A|^m) + |A|^{m+1}|u_N(t - (m+1)\Delta) - u_N(t_j - (m+1)\Delta)| \\
&\le a_N(1 + |A| + \cdots + |A|^m) + |A|^{m+1}\beta_N := b_N \downarrow 0 \quad \text{as } N \to \infty
\end{aligned}$$

and hence ensure that

$$\operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t)) - \operatorname{dist}(\omega_{N_j}; F(u_N(t), u_N(t - \Delta), t)) \le 2\ell_F b_N.$$

It follows from (H1) and (2.2) that for any $t \in [t_j, t_{j+1})$ and $j = 0, \ldots, k$ one has

$$\begin{aligned}
&\operatorname{dist}(\omega_{N_j}; F(u_N(t), u_N(t-\Delta), t)) - \operatorname{dist}(\omega_N(t); F(\bar{x}(t), \bar{x}(t-\Delta), t)) \\
&\quad \le \operatorname{dist}(F(u_N(t), u_N(t-\Delta), t),\ F(\bar{x}(t), \bar{x}(t-\Delta), t)) \\
&\quad \le \ell_F(|u_N(t) - \bar{x}(t)| + |u_N(t-\Delta) - \bar{x}(t-\Delta)|) \\
&\quad \le 2\ell_F \xi_N (1 + |A| + \cdots + |A|^m) + 2\ell_F |A|^{m+1} \beta_N.
\end{aligned}$$

Combining the above estimates and denoting

$$\mu_N := 2b_N + 2\xi_N(1 + |A| + \cdots + |A|^m) + 2|A|^{m+1}\beta_N,$$

we arrive at

$$\begin{aligned}
\operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j-\Delta), t)) &\le\ 2\ell_F b_N + \operatorname{dist}(\omega_{N_j}; F(u_N(t), u_N(t-\Delta), t)) \\
&\le 2\ell_F b_N + 2\ell_F \xi_N (1 + |A| + \cdots + |A|^m) + 2\ell_F |A|^{m+1}\beta_N \\
&\quad + \operatorname{dist}(\omega_{N_j}; F(\bar{x}(t), \bar{x}(t-\Delta), t)) \le \ell_F \mu_N + \left| \omega_{N_j} - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|
\end{aligned}$$

and finally conclude that

$$\begin{aligned}
\zeta_N &\le \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \left(\left| \omega_{N_j} - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right| + \ell_F \mu_N \right) dt + \tau(F; h_N) \\
&= \xi_N + \ell_F \mu_N (b-a) + \tau(F; h_N) := \gamma_N \downarrow 0 \quad \text{as } N \to \infty.
\end{aligned} \tag{2.3}$$

Note that the discrete functions $\{u_N(t_j) \mid j = -N, \ldots, k+1\}$ may *not* be trajectories for (2.1), since one does not generally have $\omega_{N_i} \in F(u_N(t_j), u_N(t_j - \Delta), t_j)$ for $j = 0, \ldots, k$. Let us construct the desired trajectories $\{z_N(t_j) \mid j = -N, \ldots, k+1\}$ by the following *proximal algorithm*:

$$\begin{cases}
z_N(t_j) = c(t_j) \quad \text{for } j = -N, \ldots, -1,\ z_N(t_0) = \bar{x}(a), \\
z_N(t_{j+1}) - A z_N(t_{j+1} - \Delta) \\
\qquad = z_N(t_j) - A z_N(t_j - \Delta) + h_N v_{N_j} \quad \text{for } j = 0, \ldots, k, \\
v_{N_j} \in F(z_N(t_j), z_N(t_j - \Delta), t_j) \quad \text{with} \\
|v_{N_j} - \omega_{N_j}| = \operatorname{dist}(\omega_{N_j}; F(z_N(t_j), z_N(t_j - \Delta), t_j)) \quad \text{for } j = 0, \ldots, k.
\end{cases} \tag{2.4}$$

It follows from construction (2.4) that $z_N(t_j)$ is a feasible trajectory to the discrete inclusion (2.1) for each $N \in \mathbb{N}$. Note that

$$|z_N(t) - \bar{x}(t)| = |z_N(t_j) - \bar{x}(t)| = |c(t_j) - c(t)| < \beta_N$$

for $t \in [t_j, t_{j+1})$, $j = -N, \dots, -1$, which implies that the extensions of $z_N(\cdot)$ converge to $\bar{x}(t)$ uniformly on $[a - \Delta, a)$. Let us analyze the situation on $[a, b]$.

First we claim that $z_N(t_j) \in U$ for $j = 0, \dots, k+1$, where $U \subset \mathbb{R}^n$ is a neighborhood of $\bar{x}(\cdot)$ given in (H1). Arguing by induction, we obviously have $z_N(t_0) \in U$ and assume that $z_N(t_j) \in U$ for all $j = 1, \dots, m$ with some fixed $m \in \{1, \dots, k\}$. Then

$$\begin{aligned}
&|z_N(t_{m+1}) - u_N(t_{m+1})| \\
&\quad = |Az_N(t_{m+1} - \Delta) + z_N(t_m) - Az_N(t_m - \Delta) + h_N v_{N_m} \\
&\qquad -(Au_N(t_{m+1} - \Delta) + u_N(t_m) - Au_N(t_m - \Delta) + h_N \omega_{N_m})| \\
&\quad \le |A| \cdot |z_N(t_{m+1} - \Delta) - u_N(t_{m+1} - \Delta)| + |A| \cdot |z_N(t_m - \Delta) - u_N(t_m - \Delta)| \\
&\qquad + |z_N(t_m) - u_N(t_m)| + h_N \operatorname{dist}(\omega_{N_m}; F(z_N(t_m), z_N(t_m - \Delta), t_m)).
\end{aligned}$$

Taking into account that

$$\begin{aligned}
|z_N(t_m) - u_N(t_m)| &\le |z_N(t_{m-1}) - u_N(t_{m-1})| \\
&\quad + |A||z_N(t_{m-1-N}) - u_N(t_{m-1-N})| + |A||z_N(t_{m-N}) - u_N(t_{m-N})| \\
&\quad + h_N \operatorname{dist}(\omega_{N_{m-1}}; F(z_N(t_{m-1}), z_N(t_{m-1} - \Delta), t_{m-1}))
\end{aligned}$$

and that $|z_N(t_j) - u_N(t_j)| = 0$ for $j \le 0$, one has

$$\begin{aligned}
&|z_N(t_{m+1}) - u_N(t_{m+1})| \\
&\qquad \le M h_N \sum_{j=0}^{m} \operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t_j)) \le M\gamma_N
\end{aligned} \tag{2.5}$$

with some constant $M > 0$. Now invoking (2.2) and increasing M if necessary, we arrive at

$$|z_N(t_{m+1}) - \bar{x}(t_{m+1})| \le \xi_N + M\gamma_N \to 0 \quad \text{as } N \to \infty.$$

It remains to prove that the extended combinations $z_N(t) - Az_N(t - \Delta)$ converge to $\bar{x}(t) - A\bar{x}(t - \Delta)$ in the $W^{1,2}$-norm on $[a, b]$, which means that

$$\begin{aligned}
&\max_{t \in [a,b]} |z_N(t) - Az_N(t - \Delta) - [\bar{x}(t) - A\bar{x}(t - \Delta)]| \\
&\qquad + \int_a^b \left| \frac{d}{dt}[z_N(t) - Az_N(t - \Delta)] - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t - \Delta)] \right|^2 dt \to 0
\end{aligned} \tag{2.6}$$

as $N \to \infty$. To furnish this, we use (2.5) and get the estimate

$$\begin{aligned}
\sum_{j=0}^{k+1} |z_N(t_j) - u_N(t_j)| &\le \sum_{j=0}^{k+1} M \sum_{m=0}^{j-1} h_N \operatorname{dist}(\omega_{N_m}; F(u_N(t_m), u_N(t_m - \Delta), t_m)) \\
&\le M(b - a) \sum_{j=0}^{k} \operatorname{dist}(\omega_{N_j}; F(u_N(t_j), u_N(t_j - \Delta), t_j)),
\end{aligned}$$

which by (H1) implies that

$$
\begin{aligned}
&\int_a^b \left|\frac{d}{dt}[z_N(t)-Az_N(t-\Delta)]-\omega_N(t)\right| dt\\
&\quad=\sum_{j=0}^{k}\int_{t_j}^{t_{j+1}}\left|\frac{d}{dt}[z_N(t)-Az_N(t-\Delta)]-\omega_N(t)\right| dt\\
&\quad=\sum_{j=0}^{k}\int_{t_j}^{t_{j+1}}|v_{N_j}-\omega_{N_j}|\,dt\\
&\quad=\sum_{j=0}^{k}h_N\,\mathrm{dist}(\omega_{N_j};F(z_N(t_j),z_N(t_j-\Delta),t_j))\\
&\quad=\sum_{j=0}^{k}h_N\,\mathrm{dist}(\omega_{N_j};F(u_N(t_j),u_N(t_j-\Delta),t_j))\\
&\qquad+\sum_{j=0}^{k}h_N[\mathrm{dist}(\omega_{N_j};F(z_N(t_j),z_N(t_j-\Delta),t_j))\\
&\qquad\qquad-\mathrm{dist}(\omega_{N_j};F(u_N(t_j),u_N(t_j-\Delta),t_j))]\\
&\quad\le\sum_{j=0}^{k}h_N\,\mathrm{dist}(\omega_{N_j};F(u_N(t_j),u_N(t_j-\Delta),t_j))\\
&\qquad+\sum_{j=0}^{k}\ell_F h_N[|z_N(t_j)-u_N(t_j)|+|z_N(t_j-\Delta)-u_N(t_j-\Delta)|]\\
&\quad\le\gamma_N+\sum_{j=0}^{k}\ell_F h_N[|z_N(t_j)-u_N(t_j)|+|z_N(t_j-\Delta)-u_N(t_j-\Delta)|]\\
&\quad\le\gamma_N+2M(b-a)\ell_F\sum_{j=0}^{k}h_N\,\mathrm{dist}(\omega_{N_j};F(u_N(t_j),u_N(t_j-\Delta),t_j))\\
&\quad\le\gamma_N+2M\ell_F(b-a)\gamma_N.
\end{aligned}
$$

The latter ensures the estimate

$$
\begin{aligned}
&\int_a^b\left|\frac{d}{dt}[z_N(t)-Az_N(t-\Delta)]-\frac{d}{dt}[\bar{x}(t)-A\bar{x}(t-\Delta)]\right| dt\\
&\qquad\le\int_a^b\left|\frac{d}{dt}[z_N(t)-Az_N(t-\Delta)]-\omega_N(t)\right| dt\\
&\qquad\quad+\int_a^b\left|\omega_N(t)-\frac{d}{dt}[\bar{x}(t)-A\bar{x}(t-\Delta)]\right| dt\\
&\qquad\le\gamma_N(1+2M(b-a)\ell_F)+\xi_N.
\end{aligned}
$$

Since $z_N(t)\in U$, it follows from (H1) by (1.2) and (2.4) that $|\frac{d}{dt}[z_N(t)-Az_N(t-\Delta)]|\le$

m_F, $|\frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)]| \le m_F$, and hence

$$\begin{aligned}
&\int_a^b \left| \frac{d}{dt}[z_N(t) - Az_N(t-\Delta)] - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt \\
&\quad = \int_a^b \left| \frac{d}{dt}[z_N(t) - Az_N(t-\Delta)] - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right| \\
&\qquad \times \left| \frac{d}{dt}[z_N(t) - Az_N(t-\Delta)] + \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right| dt \\
&\quad \le 2m_F[\gamma_N(1 + 2M(b-a)\ell_F) + \xi_N] \downarrow 0 \quad \text{as } N \to \infty.
\end{aligned}$$

Observing that

$$\begin{aligned}
&\max_{t\in[a,b]} |z_N(t) - Az_N(t-\Delta) - [\bar{x}(t) - A\bar{x}(t-\Delta)]| \\
&\quad \le \int_a^b \left| \frac{d}{dt}[z_N(t) - Az_N(t-\Delta)] - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt,
\end{aligned}$$

we arrive at (2.6) and complete the proof of the theorem. □

3. Strong convergence of discrete optimal solutions. Our next goal is to construct a sequence of *well-posed* discrete approximations of the whole dynamic optimization problem (P) given in (1.1)–(1.4) such that optimal solutions to discrete approximation problems *strongly converge*, in the sense described below, to a *given* optimal solution $\bar{x}(\cdot)$ to the original optimization problem governed by neutral functional-differential inclusions. The following construction explicitly involves the optimal solution $\bar{x}(\cdot)$ to the problem (P) under consideration, for which we aim to derive necessary optimality conditions in the subsequent sections.

Given $\bar{x}(t)$, $a - \Delta \le t \le b$, take its approximation $z_N(t)$ from Theorem 2.1 and denote $\eta_N := |z_N(t_{k+1}) - \bar{x}(b)|$. For any natural number N we consider the following *discrete-time* dynamic optimization problem (P_N):

(3.1)

$$\begin{aligned}
\text{minimize} \quad J_N[x_N] &:= \varphi(x_N(t_0), x_N(t_{k+1})) + |x_N(t_0) - \bar{x}(a)|^2 \\
&\quad + h_N \sum_{j=0}^{k} f(x_N(t_j), x_N(t_j - \Delta), t_j) \\
&\quad + \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \left| \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right. \\
&\quad \left. - \frac{x_N(t_{j+1}) - Ax_N(t_{j+1} - \Delta) - x_N(t_j) + Ax_N(t_j - \Delta)}{h_N} \right|^2 dt
\end{aligned}$$

subject to the *dynamic constraints* governed by neutral functional-difference inclusions (2.1), the *endpoint constraints*

$$(x_N(t_0), x_N(t_{k+1})) \in \Omega_N := \Omega + \eta_N \mathbb{B}, \tag{3.2}$$

which are η_N-*perturbations* of the original endpoint constraints (1.4), and the auxiliary constraints

$$|x_N(t_j) - \bar{x}(t_j)| \le \varepsilon, \quad j = 1, \dots, k+1, \tag{3.3}$$

with some $\varepsilon > 0$. The latter auxiliary constraints are needed to guarantee the existence of optimal solutions in (P_N) and can be ignored in the derivation of necessary optimality conditions; see below.

In what follows we select $\varepsilon > 0$ in (3.3) such that $\bar{x}(t) + \varepsilon\mathbb{B} \subset U$ for all $t \in [a - \Delta, b]$ and take sufficiently large N ensuring that $\eta_N < \varepsilon$. Note that problems (P_N) have *feasible* solutions, since the trajectories z_N from Theorem 2.1 satisfy all the constraints (2.1), (3.2), and (3.3). Therefore, by the classical Weierstrass theorem in finite dimensions, each (P_N) admits an *optimal* solution $\bar{x}_N(\cdot)$ under the following assumption imposed in addition to (H1)–(H3), where U stands for a neighborhood of the optimal trajectory $\bar{x}(\cdot)$ to (P).

(H4) φ is continuous on $U \times U$, $f(x, y, \cdot)$ is continuous for a.e. $t \in [a, b]$ uniformly in $(x, y) \in U \times U$, $f(\cdot, \cdot, t)$ is continuous on $U \times U$ uniformly in $t \in [a, b]$, and Ω is locally closed around $(\bar{x}(a), \bar{x}(b))$.

We are going to justify the strong convergence of $\bar{x}_N(\cdot) \to \bar{x}(\cdot)$ in the sense of Theorem 2.1. To proceed, we need to involve an important intrinsic property of the original problem (P) called *relaxation stability*. Following the line originated by Jack Warga in optimal control theory (see the book [30] and its references), we consider the *relaxed* problem (R) of minimizing the cost functional (1.1) on admissible trajectories of the *convexified* functional-differential inclusion of the neutral type

$$\frac{d}{dt}[x(t) - Ax(t - \Delta)] \in \operatorname{co} F(x(t), x(t - \Delta), t) \quad \text{for a.e. } t \in [a, b] \tag{3.4}$$

with the initial "tail" condition (1.3) and the endpoint constraints (1.4). Any admissible trajectory for (3.4) satisfying (1.3) is called a *relaxed trajectory* for (1.2).

One clearly has $\inf(\mathrm{R}) \le \inf(\mathrm{P})$ for the optimal values of the cost functionals in the relaxed and original problems. We say that the original problem (P) is *stable with respect to relaxation* if

$$\inf(\mathrm{P}) = \inf(\mathrm{R}).$$

This property, which obviously holds under the convexity assumption on the sets $F(x, y, t)$, goes far beyond the convexity. General sufficient conditions for the relaxation stability of the neutral-type problem (P) follow from [14]. We also refer the reader to [2, 21, 23, 30] for more detailed discussions on the validity of the relaxation stability property for various classes of differential, functional-differential, and functional-integral control systems.

Now we are ready to establish the following strong convergence theorem for optimal solutions to discrete approximations, which makes a *bridge* between optimal control problems governed by neutral functional-differential and functional-difference inclusions.

THEOREM 3.1. *Let $\bar{x}(\cdot)$ be an optimal solution to problem* (P)*, which is assumed to be stable with respect to relaxation. Suppose also that hypotheses* (H1)–(H4) *hold. Then any sequence $\{\bar{x}_N(\cdot)\}$, $N \in \mathbb{N}$, of optimal solutions to (P_N) extended to the continuous interval $[a-\Delta, b]$ converges uniformly to $\bar{x}(\cdot)$ on $[a-\Delta, b]$, and the sequence of their combinations $\bar{x}_N(\cdot) - A\bar{x}_N(\cdot - \Delta)$ converges to $\bar{x}(\cdot) - A\bar{x}(\cdot - \Delta)$ in the $W^{1,2}$-norm on $[a, b]$ as $N \to \infty$.*

Proof. We know from the above discussion that (P_N) has an optimal solution $\bar{x}_N(\cdot)$ for all N sufficiently large; suppose that it happens for all $N \in \mathbb{N}$ without loss of generality. Given $\bar{x}(\cdot)$, we consider the sequence $\{z_N(\cdot)\}$ strongly approximating

$\bar{x}(\cdot)$ by Theorem 2.1. Since each z_N is feasible to (P_N), one has

$$J_N[\bar{x}_N] \le J_N[z_N] \quad \text{for all } N \in \mathbb{N}.$$

For convenience we represent $J_N[z_N]$ as the sum of three terms:

$$\begin{aligned} J_N[z_N] &= \varphi(z_N(t_0), z_N(t_{k+1})) + h_N \sum_{j=0}^{k} f(z_N(t_j), z_N(t_j-\Delta), t_j) \\ &\quad + \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \left| \frac{z_N(t_{j+1}) - Az_N(t_{j+1}-\Delta) - z_N(t_j) + Az_N(t_j-\Delta)}{h_N} \right. \\ &\qquad \left. - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt := I_{1N} + I_{2N} + I_{3N}. \end{aligned}$$

It follows from Theorem 2.1 and the assumption on φ in (H4) that

$$I_{1N} \to \varphi(\bar{x}(a), \bar{x}(b)) \quad \text{as } N \to \infty$$

and that, using the sign $\approx$ for expressions that are equivalent as $N \to \infty$,

$$\begin{aligned} I_{2N} &= h_N \sum_{j=0}^{k} f(z_N(t_j), z_N(t_j-\Delta), t_j) \\ &= \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} f(z_N(t_j), z_N(t_j-\Delta), t)\, dt \\ &\quad + \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} [f(z_N(t_j), z_N(t_j-\Delta), t_j) - f(z_N(t_j), z_N(t_j-\Delta), t)]\, dt \\ &= \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} f(z_N(t_j), z_N(t_j-\Delta), t)\, dt + \tau(f; h_N) \\ &\approx \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} f(z_N(t_j), z_N(t_j-\Delta), t)\, dt \\ &\approx \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} f(\bar{x}(t_j), \bar{x}(t_j-\Delta), t)\, dt \\ &\to \int_a^b f(\bar{x}(t), \bar{x}(t-\Delta), t)\, dt \quad \text{as } N \to \infty, \\ I_{3N} &= \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \left| v_N(t) - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt \\ &= \int_a^b \left| v_N(t) - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt \\ &= \int_a^b \left| \frac{d}{dt}[z_N(t) - Az_N(t-\Delta)] - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt \to 0 \quad \text{as } N \to \infty, \end{aligned}$$

where $v_N(t)$ stands for the extensions of combined discrete velocities for $z_N(\cdot)$; see

section 2. This implies that $J_N[z_N] \to J[\bar{x}]$ as $N \to \infty$, and therefore

$$\limsup_{N\to\infty} J_N[\bar{x}_N] \le J[\bar{x}]. \tag{3.5}$$

It is easy to observe that the strong convergence claimed in the theorem follows from

$$\rho_N := |\bar{x}_N(a) - \bar{x}(a)|^2 + \int_a^b \left| \frac{d}{dt}[\bar{x}_N(t) - A\bar{x}_N(t-\Delta)] - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt \to 0 \quad \text{as } N \to \infty.$$

On the contrary, suppose that the latter does not hold. Then there are a constant $\alpha > 0$ and a subsequence $\{N_m\} \subset \mathbb{N}$ for which $\rho_{N_m} \to \alpha$ as $m \to \infty$. Employing the standard compactness arguments based on (2.1) and the boundedness assumption in (H1), we find an absolutely continuous function $\widetilde{z} : [a,b] \to \mathbb{R}^n$ and a function $\widetilde{x} : [a-\Delta, b]$ continuous on $[a-\Delta, a)$ and $[a,b]$ such that

$$\frac{d}{dt}[\bar{x}_N(t) - A\bar{x}_N(t-\Delta)] \to \dot{\widetilde{z}}(t) \quad \text{weakly in } L^2[a,b],$$

that $\bar{x}_N(t) \to \widetilde{x}(t)$ uniformly on $[a-\Delta, b]$ as $N \to \infty$ (without loss of generality), and that $\widetilde{z}(t) = \widetilde{x}(t) - A\widetilde{x}(t-\Delta)$ for $t \in [a,b]$. By the classical Mazur theorem, there is a sequence of *convex combinations* of $\frac{d}{dt}[\bar{x}_N(t) - A\bar{x}_N(t-\Delta)]$ that converges to $\frac{d}{dt}[\widetilde{x}(t) - A\widetilde{x}(t-\Delta)]$ in the norm topology of $L^2[a,b]$ and hence *pointwisely* for a.e. $t \in [a,b]$ along some subsequence. Therefore

$$\frac{d}{dt}[\widetilde{x}(t) - A\widetilde{x}(t-\Delta)] \in \operatorname{co} F(\widetilde{x}(t), \widetilde{x}(t-\Delta), t) \quad \text{for a.e. } t \in [a,b].$$

Since $\widetilde{x}(\cdot)$ obviously satisfies the tail condition (1.3) and the endpoint constraints (1.4), it is a feasible solution to the relaxed problem (R). Note that

$$\begin{aligned} h_N \sum_{j=0}^{k} f(\bar{x}_N(t_j), \bar{x}_N(t_j-\Delta), t_j) &= \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} f(\bar{x}_N(t_j), \bar{x}_N(t_j-\Delta), t_j)dt \\ &\to \int_a^b f(\widetilde{x}(t), \widetilde{x}(t-\Delta), t)dt \end{aligned}$$

as $N \to \infty$ due to the assumptions made. Observe also that the integral functional

$$I[v] := \int_a^b \left| v(t) - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt$$

is lower semicontinuous in the weak topology of $L^2[a,b]$ by the *convexity* of the integrand in v. Since one has

$$\begin{aligned} \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \Bigg| &\frac{\bar{x}_N(t_{j+1}) - A\bar{x}_N(t_{j+1}-\Delta) - [\bar{x}_N(t_j) - A\bar{x}_N(t_j-\Delta)]}{h_N} \\ &- \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \Bigg|^2 dt \\ = \int_a^b &\left| \frac{d}{dt}[\bar{x}_N(t) - A\bar{x}_N(t-\Delta)] - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt, \end{aligned}$$

the latter implies that

$$\int_a^b \left| \frac{d}{dt}[\widetilde{x}(t) - A\widetilde{x}(t-\Delta)] - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt$$
$$\le \liminf_{N\to\infty} \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \left| \frac{\bar{x}_N(t_{j+1}) - A\bar{x}_N(t_{j+1}-\Delta) - [\bar{x}_N(t_j) - A\bar{x}_N(t_j-\Delta)]}{h_N} \right.$$
$$\left. - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt.$$

Using the above relationships and passing to the limit in the expression (3.1) for $J_N[\bar{x}_N]$, we get

$$J[\widetilde{x}] + \alpha \le \lim_{N\to\infty} J_N[\bar{x}_N].$$

By (3.5) one therefore has

$$J[\widetilde{x}] \le J[\bar{x}] - \alpha < J[\bar{x}] \quad \text{if } \alpha > 0.$$

This clearly contradicts the optimality of $\bar{x}(\cdot)$ in the relaxed problem (R) due to the assumption on relaxation stability. Thus $\alpha = 0$, which completes the proof of the theorem. □

4. Tools of variational analysis. Convegence/stability results of the previous section allow us to make a bridge between the original infinite-dimensional optimization problem (P) for neutral functional-differential inclusions and the sequence of finite-dimensional dynamic optimization problems (P_N) for neutral functional-difference inclusions. Our strategy is first to obtain necessary optimality conditions for the latter finite-dimensional problems and then to derive necessary optimality conditions for the original problem (P) by passing to the limit from the ones for (P_N) as $N \to \infty$.

Observe that problems (P_N) are essentially *nonsmooth*, even in the case of smooth functions φ and f in the cost functional and the absence of endpoint constraints. The main source of nonsmoothness comes from the (increasing number of) geometric constraints in (2.1), which reflect the discrete dynamics and may have empty interiors. To conduct a variational analysis of such problems, we use appropriate tools of generalized differentiation introduced in [18] and then developed and applied in many publications; see, in particular, the books [19, 26, 29] for detailed treatments and further references.

Recall that the *basic/limiting normal cone* to the set $\Omega \subset \mathbb{R}^n$ at the point $\bar{x} \in \Omega$ is

$$N(\bar{x};\Omega) := \operatorname*{Limsup}_{x \overset{\Omega}{\to} \bar{x}} \widehat{N}(x;\Omega), \tag{4.1}$$

where $x \overset{\Omega}{\to} \bar{x}$ means that $x \to \bar{x}$ with $x \in \Omega$, and where

$$\widehat{N}(\bar{x};\Omega) := \left\{ x^* \in \mathbb{R}^n \mid \limsup_{x \overset{\Omega}{\to} \bar{x}} \frac{\langle x^*, x - \bar{x}\rangle}{|x - \bar{x}|} \le 0 \right\} \tag{4.2}$$

is the cone of Fréchet (or regular) normals to Ω at $\bar{x}$. For convex sets Ω both cones

$N(\bar{x};\Omega)$ and $\widehat{N}(\bar{x};\Omega)$ reduce to the normal cone of convex analysis. Note that the basic normal cone (4.1) is often *nonconvex* while satisfying a comprehensive calculus, in contrast to (4.2).

Given an extended real-valued function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}} := [-\infty, \infty]$ finite at $\bar{x}$, the *subdifferential* of φ at $\bar{x}$ is defined geometrically as

$$\partial\varphi(\bar{x}) := \{x^* \in \mathbb{R}^n \mid (x^*, -1) \in N((\bar{x}, \varphi(\bar{x})); \operatorname{epi}\varphi)\} \tag{4.3}$$

via basic normals to the epigraph $\operatorname{epi}\varphi := \{(x,\mu) \in \mathbb{R}^{n+1} \mid \mu \geq \varphi(x)\}$; equivalent analytic representations of (4.3) can be found in [19, 26, 29].

Given a set-valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ with the graph $\operatorname{gph} F := \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^m \mid y \in F(x)\}$, the *coderivative* $D^*F(\bar{x},\bar{y}) : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ of F at $(\bar{x},\bar{y}) \in \operatorname{gph} F$ is defined by

$$D^*F(\bar{x},\bar{y})(y^*) := \{x^* \in \mathbb{R}^n \mid (x^*, -y^*) \in N((\bar{x},\bar{y}); \operatorname{gph} F)\}. \tag{4.4}$$

Note the useful relationships

$$\partial\varphi(\bar{x}) = D^*E_\varphi(\bar{x}, \varphi(\bar{x}))(1) \qquad \text{and} \qquad D^*g(\bar{x})(y^*) = \partial\langle y^*, g\rangle(\bar{x}), \quad y^* \in \mathbb{R}^m,$$

between the subdifferential and coderivative introduced, where $E_\varphi(x) := \{\mu \in \mathbb{R} \mid \mu \geq \varphi(x)\}$ is the epigraphical multifunctions associated with $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ and where $\langle y^*, g\rangle(x) := \langle y^*, g(x)\rangle$ is the scalarized function associated with a locally Lipschitzian mapping $g : \mathbb{R}^n \to \mathbb{R}^m$.

The subdifferential/coderivative constructions (4.3) and (4.4) enjoy a variety of useful calculus rules that can be found in the books mentioned above and their references. Let us formulate two results crucial in what follows. The first one gives a complete coderivative characterization of the classical local Lipschitzian property of multifunctions imposed in our standing assumption (H1); cf. [20, Theorem 5.11] and [26, Theorem 9.40].

THEOREM 4.1. *Let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be a closed-graph multifunction locally bounded around $\bar{x}$. Then the following conditions are equivalent.*

(i) *F is locally Lipschitzian around $\bar{x}$.*

(ii) *There exist a neighborhood U of $\bar{x}$ and a number $\ell > 0$ such that*

$$\sup\{|x^*| \mid x^* \in D^*F(x,y)(y^*)\} \leq \ell|y^*| \quad \text{for all } x \in U,\ y \in F(x),\ y^* \in \mathbb{R}^m.$$

The next result taken from [19, Corollary 7.5] provides necessary optimality conditions for a general problem (MP) of *nonsmooth mathematical programming* with many geometric constraints:

$$\begin{cases} \text{minimize } \phi_0(z) & \text{subject to} \\ \phi_j(z) \leq 0, & j = 1, \ldots, r, \\ g_j(z) = 0, & j = 0, \ldots, m, \\ z \in \Lambda_j, & j = 0, \ldots, l, \end{cases}$$

where $\phi_j : \mathbb{R}^d \to \mathbb{R}$, $g_j : \mathbb{R}^d \to \mathbb{R}^n$, and $\Lambda_j \subset \mathbb{R}^d$.

THEOREM 4.2. *Let $\bar{z}$ be an optimal solution to* (MP). *Assume that all ϕ_i are Lipschitz continuous, that g_j are continuously differentiable, and that Λ_j are locally*

closed near $\bar{z}$. *Then there exist real numbers* $\{\mu_j \mid j = 0, \dots, r\}$ *as well as vectors* $\{\psi_j \in \mathbb{R}^n \mid j = 0, \dots, m\}$ *and* $\{z_j^* \in \mathbb{R}^d \mid j = 0, \dots, l\}$, *not all zero, such that*

$$\mu_j \geq 0 \quad \text{for } j = 0, \dots, r, \tag{4.5}$$

$$\mu_j \phi_j(\bar{z}) = 0 \quad \text{for } j = 1, \dots, r, \tag{4.6}$$

$$z_j^* \in N(\bar{z}; \Lambda_j) \quad \text{for } j = 0, \dots, l, \tag{4.7}$$

$$-\sum_{j=0}^{l} z_j^* \in \partial\left(\sum_{j=0}^{r} \mu_j \phi_j\right)(\bar{z}) + \sum_{j=0}^{m} \nabla g_j(\bar{z})^* \psi_j. \tag{4.8}$$

For applications in this paper in the case of nonautonomous continuous-time systems we need the following modifications of the basic constructions (4.1), (4.3), and (4.4) for sets, functions, and set-valued mappings *depending on a parameter* t from a topological space T (in our case $T = [a, b]$).

Given $\Omega : T \rightrightarrows \mathbb{R}^n$ and $\bar{x} \in \Omega(\bar{t})$, we define the *extended normal cone* to $\Omega(\bar{t})$ at $\bar{x}$ by

$$\widetilde{N}(\bar{x}; \Omega(\bar{t})) := \operatorname*{Limsup}_{(t,x) \overset{\text{gph}\Omega}{\to} (\bar{t},\bar{x})} \widehat{N}(x; \Omega(t)). \tag{4.9}$$

For $\varphi : \mathbb{R}^n \times T \to \overline{\mathbb{R}}$ finite at $(\bar{x}, \bar{t})$ and for $F : \mathbb{R}^n \times T \rightrightarrows \mathbb{R}^m$ with $\bar{y} \in F(\bar{x}, \bar{t})$, the *extended subdifferential* of φ at $(\bar{x}, \bar{t})$ and the *extended coderivative* of F at $(\bar{x}, \bar{y}, \bar{t})$ with respect to x are given, respectively, by

$$\widetilde{\partial}_x \varphi(\bar{x}, \bar{t}) := \{x^* \in \mathbb{R}^n \mid (x^*, -1) \in \widetilde{N}((\bar{x}, \varphi(\bar{x}, \bar{t})); \operatorname{epi} \varphi(\cdot, \bar{t}))\} \tag{4.10}$$

and, whenever $y^* \in \mathbb{R}^m$, by

$$\widetilde{D}_x^* F(\bar{x}, \bar{y}, \bar{t})(y^*) := \{x^* \in \mathbb{R}^n \mid (x^*, -y^*) \in \widetilde{N}((\bar{x}, \bar{y}); \operatorname{gph} F(\cdot, \bar{t}))\}. \tag{4.11}$$

Note that the sets (4.9)–(4.11) may be bigger in some situations than the corresponding sets $N(\bar{x}; \Omega(\bar{t}))$, $\partial_x \varphi(\bar{x}, \bar{t})$, and $D_x^* F(\bar{x}, \bar{y}, \bar{t})(y^*)$, where the latter two sets stand for the subdifferential (4.3) of $\varphi(\cdot, \bar{t})$ at $\bar{x}$ and the coderivative (4.4) of $F(\cdot, \bar{t})$ at $(\bar{x}, \bar{y}, \bar{t})$, respectively. Efficient conditions ensuring equalities for these sets are discussed in [21, 22, 24].

It is not hard to check that the extended constructions (4.9)–(4.11) are *robust* with respect to their variables, which is important for performing limiting procedures in what follows. In particular,

$$\widetilde{N}(\bar{x}; \Omega(\bar{t})) = \operatorname*{Lim\,sup}_{(t,x) \overset{\text{gph}\Omega}{\to} (\bar{t},\bar{x})} \widetilde{N}(x; \Omega(t)). \tag{4.12}$$

Note also that the constructions (4.9)–(4.11) enjoy a full generalized differential calculus similar to (4.1), (4.3), and (4.4). We are not going to use this calculus in the present paper.

5. Necessary optimality conditions for discrete approximations. This section concerns necessary optimality conditions for discrete approximation problems (P_N) governed by neutral functional-difference inclusions. We derive such conditions in the extended Euler–Lagrange form by reducing (P_N) to nonsmooth mathematical

programs with many geometric constraints and employing generalized differential calculus for the basic constructions (4.1), (4.3), and (4.4).

Let us reduce the dynamic optimization problem (P_N) for each $N \in \mathbb{N}$ to the mathematical programming problem (MP) considered in section 4 with the decision vector

$$z := (x_0^N, x_1^N, \dots, x_{k+1}^N, v_0^N, v_1^N, \dots, v_k^N) \in \mathbb{R}^{n(2k+3)}$$

and the following data:

$$\begin{aligned}
\phi_0(z) :&= \varphi(x_0^N, x_{k+1}^N) + |x_0^N - \bar{x}(a)|^2 + h_N \sum_{j=0}^{k} f(x_j^N, x_{j-N}^N, t_j) \\
&\quad + \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \left| v_j^N - \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \right|^2 dt, \\
\phi_j(z) :&= |x_j^N - \bar{x}(t_j)| - \varepsilon, \quad j = 1, \dots, k+1, \\
\Lambda_j :&= \{(x_0^N, \dots, v_k^N) \mid v_j^N \in F(x_j^N, x_{j-N}^N, t_j)\}, \quad j = 0, \dots, k, \\
\Lambda_{k+1} :&= \{(x_0^N, \dots, v_k^N) \mid (x_0^N, x_{k+1}^N) \in \Omega_N\}, \\
g_j(z) :&= x_{j+1}^N - Ax_{j+1-N}^N - x_j^N + Ax_{j-N}^N - h_N v_j^N, \quad j = 0, \dots, k,
\end{aligned} \tag{5.1}$$

where $x_j^N := c(t_j)$ for $j < 0$. Let $\bar{z}^N = (\bar{x}_0^N, \dots, \bar{x}_{k+1}^N, \bar{v}_0^N, \dots, \bar{v}_k^N)$ be an optimal solution to (MP). Applying Theorem 4.2, we find real numbers μ_j^N and vectors $z_j^* \in \mathbb{R}^{n(2k+3)}$ for $j = 0, \dots, k+1$ as well as vectors $\psi_j^N \in \mathbb{R}^n$ for $j = 0, \dots, k$, not all zero, such that conditions (4.5)–(4.8) are satisfied.

Taking $z_j^* = (x_{0,j}^*, \dots, x_{k+1,j}^*, v_{0,j}^*, \dots, v_{k,j}^*) \in N(\bar{z}^N; \Lambda_j)$ for $j = 0, \dots, k$, we observe that all components of z_j^* are zero except one, and the remaining one satisfies

$$(x_{j,j}^*, x_{j-N,j}^*, v_{j,j}^*) \in N((\bar{x}_j^N, \bar{x}_{j-N}^N, \bar{v}_j^N); \operatorname{gph} F(\cdot, \cdot, t_j)), \quad j = 0, \dots, k.$$

Similarly, the condition $z_{k+1}^* \in N(\bar{z}^N; \Lambda_{k+1})$ is equivalent to

$$(x_{0,k+1}^*, x_{k+1,k+1}^*) \in N((\bar{x}_0^N, \bar{x}_{k+1}^N); \Omega_N),$$

with all the other components of z_{k+1}^* equal to zero. Applying Theorem 3.1 to the convergence of discrete approximations, we have $\phi_j(\bar{z}^N) < 0$ for $j = 1, \dots, k+1$ whenever N is sufficiently large. Thus $\mu_j^N = 0$ for these indexes due to the complementary slackness conditions (4.6). Let $\lambda^N := \mu_0^N \geq 0$. Observe further that

$$\begin{aligned}
&\sum_{j=0}^{k} (\nabla g_j(\bar{z}^N))^* \psi_j^N \\
&\quad = \big(-\psi_0 + A^*(\psi_N^N - \psi_{N-1}^N), \psi_0 - \psi_1 + A^*(\psi_{N+1}^N - \psi_N^N), \dots, \\
&\qquad \psi_{k-N-1} - \psi_{k-N} + A^*(\psi_k^N - \psi_{k-1}^N), \psi_{k-N} - \psi_{k-N+1} + A^*\psi_k^N, \dots, \\
&\qquad \psi_{k-1}^N - \psi_k^N, \psi_k^N, -h_N\psi_0^N, \dots, -h_N\psi_k^N\big).
\end{aligned}$$

From the subdifferential sum rule for ϕ_0 in (5.1) one has

$$\partial\phi(\bar z^N) \subset \partial\varphi(\bar x_0^N, \bar x_{k+1}^N) + 2(\bar x_0^N - \bar x(a)) + h_N \sum_{j=0}^{k} \partial f(\bar x_j^N, \bar x_{j-N}^N, t_j) + \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} 2\left(\bar v_j^N - \frac{d}{dt}[\bar x(t) - A\bar x(t-\Delta)]\right) dt,$$

with ∂f standing here and in what follows for the basic subdifferential of f with respect to the first two variables. Thus the inclusion (4.8) in Theorem 4.2 is equivalent to the relationships

$$\begin{aligned}
-x_{0,0}^* - x_{0,N}^* - x_{0,k+1}^* &= \lambda^N u_0^N + \lambda^N h_N \vartheta_0^N + \lambda^N h_N \kappa_0^N + 2\lambda^N(\bar x_0^N - \bar x(a)) \\
&\quad -\psi_0^N - A^*(\psi_{N-1}^N - \psi_N^N), \\
-x_{j,j}^* - x_{j,j+N}^* &= \lambda^N h_N \kappa_j^N + \lambda^N h_N \vartheta_j^N + \psi_{j-1}^N - \psi_j^N \\
&\quad -A^*(\psi_{j+N-1}^N - \psi_{j+N}^N), \quad j = 1, \dots, k-N, \\
-x_{k-N+1,k-N+1}^* &= \lambda^N h_N v_{k-N+1}^N + \psi_{k-N}^N - \psi_{k-N+1}^N + A^*\psi_k^N, \\
-x_{j,j}^* &= \lambda^N h_N \vartheta_j^N + \psi_{j-1}^N - \psi_j^N, \quad j = k-N+2, \dots, k, \\
-x_{k+1,k+1}^* &= \lambda^N u_{k+1}^N + \psi_k^N, \\
-v_{j,j}^* &= \lambda^N \theta_j^N - h_N \psi_j^N, \quad j = 0, \dots, k,
\end{aligned}$$

with the notation

$$(u_0^N, u_{k+1}^N) \in \partial\varphi(\bar x_0^N, \bar x_{k+1}^N), \qquad (\vartheta_j^N, \kappa_{j-N}^N) \in \partial f(\bar x_j^N, \bar x_{j-N}^N, t_j),$$
$$\theta_j^N := -2\int_{t_j}^{t_{j+1}} \left(\frac{d}{dt}[\bar x(t) - A\bar x(t-\Delta)] - \bar v_j^N\right) dt.$$

Based on the above relationships, we arrive at the following necessary optimality conditions for discrete-time problems (P_N), where $f_j(\cdot,\cdot) := f(\cdot,\cdot,t_j)$ and $F_j(\cdot,\cdot) := F(\cdot,\cdot,t_j)$.

THEOREM 5.1. *Let $\bar z^N$ be an optimal solution to problem (P_N). Assume that the sets Ω and $\operatorname{gph} F_j$ are closed and that the functions φ and f_j are Lipschitz continuous around the points $(\bar x_0^N, \bar x_{k+1}^N)$ and $(\bar x_j^N, \bar x_{j-N}^N)$, respectively, for all $j = 0, \dots, k$. Then there exist $\lambda^N \ge 0$, p_j^N $(j = 0, \dots, k+N+1)$, and q_j^N $(j = -N, \dots, k+1)$, not all zero, such that*

$$p_j^N = 0, \quad j = k+2, \dots, k+N+1, \tag{5.2}$$

$$q_j^N = 0, \quad j = k-N+1, \dots, k+1, \tag{5.3}$$

$$(p_0^N + q_0^N, -p_{k+1}^N) \in \lambda^N \partial\varphi(\bar x_0^N, \bar x_{k+1}^N) + N((\bar x_0^N, \bar x_{k+1}^N); \Omega_N), \tag{5.4}$$

$$\begin{aligned}
&\left(\frac{P_{j+1}^N - P_j^N}{h_N}, \frac{Q_{j-N+1}^N - Q_{j-N}^N}{h_N}, -\frac{\lambda^N \theta_j^N}{h_N} + p_{j+1}^N + q_{j+1}^N\right) \\
&\quad \in \lambda^N(\partial f_j(\bar x_j^N, \bar x_{j-N}^N), 0) + N((\bar x_j^N, \bar x_{j-N}^N, \bar v_j^N); \operatorname{gph} F_j), \quad j = 0, \dots, k,
\end{aligned} \tag{5.5}$$

with the notation

$$P_j^N := p_j^N - A^* p_{j+N}^N, \qquad Q_j^N := q_j^N - A^* q_{j+N}^N,$$
$$\bar{v}_j^N := \frac{(\bar{x}_{j+1}^N - \bar{x}_j^N) + A(x_{j-N}^N - x_{j-N+1}^N)}{h_N}.$$

Proof. Actually, most of the proof has been done above—we just need to change notation in the relationships formulated right before the theorem. First let

$$\widetilde{p}_j^N := \begin{cases} \psi_{j-1}^N & \text{for } j = 1, \dots, k+1, \\ 0 & \text{for } j = k+2, \dots, k+N+1, \end{cases}$$
$$\widetilde{q}_j^N := \begin{cases} x_{j,j+N}^*/h_N & \text{for } j = -N, \dots, -1, \\ \lambda^N \kappa_j^N + x_{j,j+N}^*/h_N & \text{for } j = 0, \dots, k-N, \\ 0 & \text{for } j = k-N+1, \dots, k+1, \end{cases}$$

and then define q_j^N for $j = -N, \dots, k+1$ by the recurrent formula

$$q_j^N := q_{j+1}^N - A^*(q_{j+N+1}^N + q_{j+N}^N) - h_N \widetilde{q}_j^N,$$

where we put $q_j^N := 0$ for $j > k+1$. Observe that

$$\left(\frac{(\widetilde{p}_{j+1}^N - q_{j+1}^N) - A^*(\widetilde{p}_{j+N+1}^N - q_{j+N+1}^N) - (\widetilde{p}_j^N - q_j^N) + A^*(\widetilde{p}_{j+N}^N - q_{j+N}^N)}{h_N}, \right.$$
$$\left. \frac{(q_{j-N+1}^N - A^* q_{j+1}^N) - (q_{j-N}^N - A^* q_j^N)}{h_N}, -\frac{\lambda^N \theta_j^N}{h_N} + \widetilde{p}_{j+1}^N \right)$$
$$\in \lambda^N(\partial f_j(\bar{x}_j^N, \bar{x}_{j-N}^N), 0) + N((\bar{x}_j^N, \bar{x}_{j-N}^N, \bar{v}_j^N); \operatorname{gph} F_j), \quad j = 0, \dots, k.$$

Finally letting

$$p_0^N := \lambda^N u_0^N + x_{0,k+1}^* - q_0^N, \qquad p_j^N := \widetilde{p}_j^N - q_j^N \quad \text{for } j = 1, \dots, k+N+1,$$

we can easily check that all the relationships (5.2)–(5.5) hold. □

Corollary 5.2. *In addition to the assumptions of Theorem* 5.1*, suppose that the mapping F_j is bounded and Lipschitz continuous around $(\bar{x}_j^N, \bar{x}_{j-N}^N)$ for each $j = 0, \dots, k$. Then conditions $\lambda^N \ge 0$ and* (5.2)–(5.5) *hold with $(\lambda^N, p_{k+1}^N) \ne 0$; i.e., one can let*

$$(\lambda^N)^2 + |p_{k+1}^N|^2 = 1. \tag{5.6}$$

Proof. If $\lambda^N = 0$, then (5.5) together with (5.2) and (5.3) imply that

$$\left(\frac{p_{k+1}^N - p_k^N}{h_N}, \frac{-q_{k-N}^N}{h_N} \right) \in D^* F_k(\bar{x}_k^N, \bar{x}_{k-N}^N, \bar{v}_k^N)(-p_{k+1}^N).$$

Assuming now that $p_{k+1}^N = 0$, we get

$$\left(\frac{-p_k^N}{h_N}, \frac{-q_{k-N}^N}{h_N} \right) \in D^* F_k(\bar{x}_k^N, \bar{x}_{k-N}^N, \bar{v}_k^N)(0),$$

which yield $p_k^N = q_{k-N}^N = 0$ by Theorem 4.1. Repeating the above procedure, we arrive at a contradiction with the nontriviality assertion in Theorem 5.1. □

6. Optimality conditions for functional-differential inclusions. In this section we obtain the main results of the paper, providing necessary optimality conditions for the original dynamic optimization problem (P) in both extended Euler–Lagrange and Hamiltonian forms involving generalized differential constructions of section 4. Our major theorem establishes the following conditions of the Euler–Lagrange type derived by the limiting procedure from discrete approximations. Note that here $\Delta > 0$ as was assumed in section 2.

THEOREM 6.1. *Let $\bar{x}(\cdot)$ be an optimal solution to problem* (P) *under hypotheses* (H1)–(H4)*, where φ and $f(\cdot,\cdot,t)$ are assumed to be Lipschitz continuous instead of the plain continuity. Suppose also that* (P) *is stable with respect to relaxation. Then there exist a number $\lambda \ge 0$ and piecewise continuous functions $p : [a, b+\Delta] \to \mathbb{R}^n$ and $q : [a-\Delta, b] \to \mathbb{R}^n$ (whose points of discontinuity are confined to multiples of the delay time Δ) such that $p(t) - A^*p(t+\Delta)$ and $q(t-\Delta) - A^*q(t)$ are absolutely continuous on $[a,b]$ and the following conditions hold:*

$$\lambda + |p(b)| = 1, \tag{6.1}$$

$$p(t) = 0 \quad \text{for } t \in (b, b+\Delta], \qquad q(t) = 0 \quad \text{for } t \in (b-\Delta, b], \tag{6.2}$$

$$(p(a) + q(a), -p(b)) \in \lambda \partial \varphi(\bar{x}(a), \bar{x}(b)) + N((\bar{x}(a), \bar{x}(b)); \Omega), \tag{6.3}$$

$$\begin{aligned} &\left(\frac{d}{dt}[p(t) - A^*p(t+\Delta)], \frac{d}{dt}[q(t-\Delta) - A^*q(t)] \right) \\ &\quad \in \operatorname{co} \Big\{ (u, w, p(t) + q(t)) \in \lambda(\widetilde{\partial} f(\bar{x}(t), \bar{x}(t-\Delta), t), 0) \\ &\qquad + \widetilde{N}\Big(\Big(\bar{x}(t), \bar{x}(t-\Delta), \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \Big); \\ &\qquad\qquad \operatorname{gph} F(\cdot,\cdot,t) \Big) \Big\} \quad \text{for a.e. } t \in [a,b]. \end{aligned} \tag{6.4}$$

Proof. To prove this theorem by the method of discrete approximations, we first construct a sequence of discrete-time problems (P_N) whose optimal solutions $\bar{x}_N$ strongly approximate $\bar{x}(\cdot)$ in the sense of Theorem 2.1. By necessary optimality conditions for $\bar{x}_N$ from Corollary 5.2 we find $\lambda_N \ge 0$, p_j^N, and q_j^N satisfying relationships (5.2)–(5.6) for all $N \in \mathbb{N}$.

Without loss of generality we suppose that $\lambda^N \to \lambda$ as $N \to \infty$ for some $\lambda \ge 0$. As usual, the symbols $\bar{x}^N(t)$, $p^N(t)$, $q^N(t-\Delta)$, $P^N(t)$, and $Q^N(t)$ stand for the piecewise linear extensions of the corresponding discrete functions from Theorem 5.1 with their piecewise constant derivatives on the continuous-time interval $[a,b]$.

Considering θ_j from Theorem 5.1, we define $\theta^N(t) := \theta_j^N / h_N$ for $t \in [t_j, t_{j+1})$ as $j = 0, \dots, k$ and conclude by Theorem 2.1 that

$$\begin{aligned} \int_a^b |\theta^N(t)|\, dt &= \sum_{j=0}^k |\theta_j^N| \le 2 \sum_{j=0}^k \int_{t_j}^{t_{j+1}} \left| \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] - \bar{v}_j^N \right| dt \\ &= 2 \int_a^b \left| \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] - \frac{d}{dt}[\bar{x}^N(t) - A\bar{x}^N(t-\Delta)] \right| dt := \nu_N \to 0 \end{aligned}$$

as $N \to \infty$. We may assume without loss of generality that

$$\bar{v}^N(t) := \frac{d}{dt}[\bar{x}^N(t) - A\bar{x}^N(t-\Delta)] \to \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)] \qquad \text{and} \qquad \theta^N(t) \to 0$$

for a.e. $t \in [a,b]$ as $N \to \infty$; such a *pointwise* convergence plays a significant role in what follows.

Let us estimate $(p^N(t), q^N(t-\Delta))$ for large N. Using (5.2) and (5.3), we derive from (5.5) that

$$\left(\frac{p^N_{j+1}-p^N_j}{h_N} - \lambda^N\vartheta^N_j, \frac{q^N_{j-N+1}-q^N_{j-N}}{h_N} - \lambda^N\kappa^N_{j-N}, -\frac{\lambda^N\theta^N_j}{h_N} + p^N_{j+1}\right)$$
$$\in N((\bar{x}^N_j, \bar{x}^N_{j-N}, \bar{v}^N_j); \operatorname{gph} F_j) \quad \text{with some } (\vartheta^N_j, \kappa^N_{j-N}) \in \partial f_j(\bar{x}^N_j, \bar{x}^N_{j-N})$$

for $j = k-N+2, \ldots, k+N+1$. This means, by definition of the coderivative (4.4), that

$$\left(\frac{p^N_{j+1}-p^N_j}{h_N} - \lambda^N\vartheta^N_j, \frac{q^N_{j-N+1}-q^N_{j-N}}{h_N} - \lambda^N\kappa^N_{j-N}\right)$$
$$\in D^*F_j(\bar{x}^N_j, \bar{x}^N_{j-N}, \bar{v}^N_j)\left(\frac{\lambda^N\theta^N_j}{h_N} - p^N_{j+1}\right)$$

for such j. Thus it follows from Theorem 4.1 that

$$\left|\left(\frac{p^N_{j+1}-p^N_j}{h_N} - \lambda^N\vartheta^N_j, \frac{q^N_{j-N+1}-q^N_{j-N}}{h_N} - \lambda^N\kappa^N_{j-N}\right)\right| \le \ell_F\left|\frac{\lambda^N\theta^N_j}{h_N} - p^N_{j+1}\right|$$

for $j = k-N+2, \ldots, k+N+1$. Since $|(\vartheta^N_j, \kappa^N_{j-N})| \le \ell_f$ due to the Lipschitz continuity of f with modulus ℓ_f, we derive from the above that

$$\begin{aligned}
&|(p^N_j, q^N_{j-N})| \\
&\quad \le \ell_F|\theta^N_j| + (\ell_F+1)h_N\ell_f + (\ell_F h_N+1)|(p^N_{j+1}, q^N_{j-N+1})| \\
&\quad \le \ell_F|\theta^N_j| + (\ell_F h_N+1)\ell_F|\theta^N_{j+1}| + (\ell_F+1)h_N\ell_f \\
&\qquad + (\ell_F h_N+1)(\ell_F+1)h_N\ell_f + (\ell_F h_N+1)^2|(p^N_{j+2}, q^N_{j-N+2})| \le \cdots \\
&\quad \le \exp[\ell_F(b-a)](1+\ell_f(\ell_F+1)/\ell_F + \ell_F\nu_N), \quad j = k-N+2, \ldots, k+N+1,
\end{aligned}$$

which implies the uniform boundedness of $\{(p^N_j, q^N_{j-N}) \mid j = k-N+2, \ldots, k+N+1\}$ and hence of $(p^N(t), q^N(t-\Delta))$ on $[b-\Delta, b]$.

Next we consider $j = k-2N+2, \ldots, k+1$ and derive from (5.5) that

$$\left|\left(\frac{p^N_{j+1}-p^N_j}{h_N} - \lambda^N\vartheta^N_j, \frac{q^N_{j-N+1}-q^N_{j-N}}{h_N} - \lambda^N\kappa^N_{j-N}\right)\right|$$
$$\le \ell_F\left|\frac{\lambda^N\theta^N_j}{h_N} - p^N_{j+1} - q^N_{j+1}\right| + \left|\left(\frac{A^*p^N_{j+N+1} - A^*p^N_{j+N}}{h_N}, \frac{A^*q^N_{j+1} - A^*q^N_j}{h_N}\right)\right|.$$

This implies, due to Theorem 4.1 and the uniform boundedness of p^N_{j+N} and q^N_j by some constant $\alpha > 0$ for such j, that

$$\left|\left(\frac{p^N_{j+1}-p^N_j}{h_N} - \lambda^N\vartheta^N_j, \frac{q^N_{j-N+1}-q^N_{j-N}}{h_N} - \lambda^N\kappa^N_{j-N}\right)\right|$$
$$\le \ell_F\left|\frac{\lambda^N\theta^N_j}{h_N} - p^N_{j+1} - q^N_{j+1}\right| + \frac{\alpha}{h_N}$$

for $j = k - 2N + 2, \dots, k + 1$. Therefore

$$\begin{aligned}
&|(p_j^N, q_{j-N}^N)| \\
&\quad \le \ell_F|\theta_j^N| + (\ell_F + 1)h_N\ell_f + (\ell_F h_N + 1)|(p_{j+1}^N, q_{j-N+1}^N)| + (\ell_F h_N + 1)\alpha \\
&\quad \le \ell_F|\theta_j^N| + (\ell_F h_N + 1)\ell_F|\theta_{j+1}^N| + (\ell_F + 1)h_N\ell_f + (\ell_F h_N + 1)(\ell_F + 1)h_N\ell_f \\
&\qquad + (\ell_F h_N + 1)(\ell_F + 1)\alpha + (\ell_F h_N + 1)^2|(p_{j+2}^N, q_{j-N+2}^N)| \le \cdots \\
&\quad \le \exp[\ell_F(b-a)](1 + (\ell_f + \alpha)(\ell_F + 1)/\ell_F + \ell_F\nu_N), \quad j = k - 2N + 2, \dots, k + 1.
\end{aligned}$$

This shows that p_j^N and q_{j-N}^N are uniformly bounded for $j = k - 2N + 2, \dots, k + 1$, and hence the sequence $\{p^N(t), q^N(t - \Delta)\}$ is uniformly bounded on $[b - 2\Delta, b - \Delta]$. Repeating the above procedure, we conclude that both sequences $\{p^N(t), q^N(t - \Delta)\}$ and $\{P^N(t), Q^N(t - \Delta)\}$ are uniformly bounded on the whole interval $[a, b]$.

Next we estimate $(\dot{P}^N(t), \dot{Q}^N(t - \Delta))$ on $[a, b]$ using (5.5) and Theorem 4.1. This yields, for $t_j \le t < t_{j+1}$ with $j = 0, \dots, k$, that

$$\begin{aligned}
&|(\dot{P}^N(t), \dot{Q}^N(t - \Delta))| \\
&\qquad = \left| \left(\frac{P_{j+1}^N - P_j^N}{h_N}, \frac{Q_{j-N+1}^N - Q_{j-N}^N}{h_N} \right) \right| \\
&\qquad \le \ell_F \left| \frac{\lambda^N\theta_j^N}{h_N} - p_{j+1}^N - q_{j+1}^N \right| + \ell_f \le \ell_F|\theta_j^N| + \ell_F|p_{j+1}^N| + \ell_F|q_{j+1}^N| + \ell_f.
\end{aligned}$$

Thus the sequence $\{\dot{P}^N(t), \dot{Q}^N(t - \Delta)\}$ is weakly compact in $L^1[a, b]$. Taking the whole sequence of $N \in \mathbb{N}$ without loss of generality, we find two absolutely continuous functions $P(\cdot)$ and $Q(\cdot - \Delta)$ on $[a, b]$ such that

$$\dot{P}^N(t) \to \dot{P}(t), \quad \dot{Q}^N(t - \Delta) \to \dot{Q}(t - \Delta) \quad \text{weakly in } L^1[a, b]$$

and $P^N(t) \to P(t)$, $Q^N(t - \Delta) \to Q(t - \Delta)$ uniformly on $[a, b]$ as $N \to \infty$. Since $p^N(t)$ and $q^N(t - \Delta)$ are uniformly bounded on $[a, b + \Delta]$, they surely converge to some functions $p(t)$ and $q(t - \Delta)$ weakly in $L^1[a, b + \Delta]$. Taking into account the above convergence of $P^N(t)$ and $Q^N(t - \Delta)$, we get that $p(\cdot)$ and $q(\cdot)$ satisfy (6.2), that

$$P(t) = p(t) - A^*p(t + \Delta), \quad Q(t - \Delta) = q(t - \Delta) - A^*q(t), \quad t \in [a, b],$$

and that $p(t)$ and $q(t)$ are piecewise continuous on $[a, b+\Delta]$ and $[a-\Delta, b]$, respectively, with possible discontinuity (from the right) at the points $b - i\Delta$ at $i = 0, 1, \dots$. Conditions (6.1) and (6.3) follow by passing to the limit from (5.6) and (5.4), respectively, taking into account the robustness of the basic subdifferential (4.3) and the normal cone (4.1).

It remains to justify the Euler–Lagrange inclusion (6.4). To furnish this, we rewrite the discrete Euler–Lagrange inclusion (5.5) in the form

$$\begin{aligned}
&(\dot{P}^N(t), \dot{Q}^N(t - \Delta)) \\
&\qquad \in \left\{ (u, w) \mid \left(u, w, p^N(t_{j+1}) + q^N(t_{j+1}) - \frac{\lambda^N\theta_j^N}{h_N} \right) \right. \\
&\qquad\qquad \left. \in \lambda^N(\partial f(\bar{x}(t_j), \bar{x}(t_j - \Delta), t_j), 0) + (N(\bar{x}_j^N, \bar{x}_{j-N}^N, \bar{v}_j^N); \operatorname{gph} F_j) \right\}
\end{aligned} \tag{6.5}$$

for $t \in [t_j, t_{j+1}]$ with $j = 0, \ldots, k$. By the classical Mazur theorem there is a sequence of convex combinations of the functions $(\dot{P}^N(t), \dot{Q}^N(t-\Delta))$ that converges to $(\dot{P}(t), \dot{Q}(t-\Delta))$ for a.e. $t \in [a,b]$. Passing the limit in (6.5), taking into account the pointwise convergence of $\theta^N(t)$ and $\bar{v}^N(t)$ established above, as well as the constructions of the extended normal cone (4.9) and the extended subdifferential (4.10) and their robustness property (4.12) with respect to all variables and parameters, we arrive at (6.4) and complete the proof of the theorem. □

Observe that for the *Mayer problem* (P_M), which is (1.1)–(1.4) with $f = 0$, the generalized Euler–Lagrange inclusion (6.4) is equivalently expressed in terms of the extended coderivative (4.11) with respect to the first two variables of $F = F(x,y,t)$, i.e., in the form

$$\begin{aligned} &\left(\frac{d}{dt}[p(t) - A^*p(t+\Delta)], \frac{d}{dt}[q(t-\Delta) - A^*q(t)]\right) \\ &\qquad \in \operatorname{co} \widetilde{D}^*_{x,y} F\left(\bar{x}(t), \bar{x}(t-\Delta), \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)], t\right)(-p(t) - q(t)) \end{aligned} \tag{6.6}$$

for almost all $t \in [a,b]$. It turns out that the extended Euler–Lagrange inclusion obtained above implies, under the *relaxation stability* of the original problems, two other principal optimality conditions expressed in terms of the Hamiltonian function built on the mapping F in (1.2). The first condition called the extended *Hamiltonian inclusion* is given below in terms of a *partial convexification* of the basic subdifferential (4.3) for the Hamiltonian function. The second is an analog of the classical *Weierstrass–Pontryagin maximum condition* (maximum principle) for neutral functional-differential inclusions. Recall that an analog of the maximum principle does *not generally hold* even in the case of optimal control problems governed by smooth functional-differential equations of neutral type.

The following relationships between the extended Euler–Lagrange and Hamiltonian inclusions are based on Rockafellar's dualization theorem [25] (see also [29, section 7.6]) that concerns subgradients of abstract Lagrangians and Hamiltonians associated with set-valued mappings regardless of the dynamics in (1.2). For simplicity we consider the case of the Mayer problem (P_M) for autonomous functional-differential inclusions of neutral type. Then the *Hamiltonian* function for F in (1.2) is defined by

$$H(x,y,p) := \sup\{\langle p, v\rangle \mid v \in F(x,y)\}. \tag{6.7}$$

Corollary 6.2. *Let $\bar{x}(\cdot)$ be an optimal solution to the Mayer problem (P_M) for the autonomous neutral functional-differential inclusion (1.2) under the assumptions of Theorem 6.1. Then there exist a number $\lambda \ge 0$ and piecewise continuous functions $p : [a, b+\Delta] \to \mathbb{R}^n$ and $q : [a-\Delta, b] \to \mathbb{R}^n$ such that $p(t) - A^*p(t+\Delta)$ and $q(t-\Delta) - A^*q(t)$ are absolutely continuous on $[a,b]$ and, besides (6.1)–(6.4), one has the extended Hamiltonian inclusion*

$$\begin{aligned} &\left(\frac{d}{dt}[p(t) - A^*p(t+\Delta)], \frac{d}{dt}[q(t-\Delta) - A^*q(t)]\right) \\ &\qquad \in \operatorname{co}\left\{(u,w) \mid \left(-u, -w, \frac{d}{dt}[\bar{x}(t) - A\bar{x}(t-\Delta)]\right)\right. \\ &\qquad\qquad \left. \in \partial H\left(\bar{x}(t), \bar{x}(t-\Delta), p(t) + q(t)\right)\right\} \end{aligned} \tag{6.8}$$

and the maximum condition

$$\left\langle p(t)+q(t),\frac{d}{dt}[\bar{x}(t)-A\bar{x}(t-\Delta)]\right\rangle = H(\bar{x}(t),\bar{x}(t-\Delta),p(t)+q(t)) \tag{6.9}$$

for almost all $t\in[a,b]$. *If, moreover,* F *is convex-valued around* $(\bar{x}(t),\bar{x}(t-\Delta))$, *then* (6.8) *is equivalent to the Euler–Lagrange inclusion*

$$\begin{aligned}&\left(\frac{d}{dt}[p(t)-A^*p(t+\Delta)],\frac{d}{dt}[q(t-\Delta)-A^*q(t)]\right)\\ &\quad\in \operatorname{co} D^*F\left(\bar{x}(t),\bar{x}(t-\Delta),\frac{d}{dt}[\bar{x}(t)-A\bar{x}(t-\Delta)]\right)(-p(t)-q(t)) \quad \textit{for a.e. } t\in[a,b],\end{aligned} \tag{6.10}$$

which in this case automatically implies the maximum condition (6.9).

Proof. Since (P_M) is stable with respect to relaxation, $\bar{x}(\cdot)$ is an optimal solution to the relaxed problem (R_M), whose only difference from (P_M) is that the neutral functional-differential inclusion (1.2) is replaced by its convexification (3.4). By Theorem 6.1 the optimal solution $\bar{x}(\cdot)$ satisfies conditions (6.1)–(6.4) and the relaxed counterpart of (6.6), which is the same as (6.10) in this case with F replaced by $\operatorname{co} F$. According to [25, Theorem 3.3] one has

$$\begin{aligned}&\operatorname{co}\{(u,v)\mid (u,w,p)\in N((x,y,v);\operatorname{gph}(\operatorname{co}F)\}\\ &\qquad=\operatorname{co}\{(u,w)\mid(-u,-w,v)\in\partial H_R(x,y,p)\},\end{aligned}$$

where H_R stands for the Hamiltonian (6.7) of the relaxed system, i.e., with F replaced by $\operatorname{co} F$. It is easy to check that $H_R=H$. Thus the extended Euler–Lagrange inclusion for the relaxed system implies the extended Hamiltonian inclusion (6.8), which surely yields the maximum condition (6.9). When F is *convex-valued*, (6.8) and (6.10) are *equivalent* due to the mentioned result of [25]. This completes the proof of the corollary. □

REFERENCES

[1] T. S. Angell, *On conrollability for nonlinear hereditary systems: A fixed point approach*, Nonlinear Anal., 4 (1980), pp. 93–107.

[2] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[3] H. T. Banks and A. Manitius, *Applications of abstract variational theory to hereditary systems—a survey*, IEEE Trans. Automat. Control, 19 (1974), pp. 524–533.

[4] R. Bellman and K. L. Cooke, *Differential-Difference Equations*, Academic Press, New York, 1963.

[5] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[6] F. H. Clarke and G. G. Watkins, *Necessary conditions, controllability and the value function for differential-difference inclusions*, Nonlinear Anal., 10 (1986), pp. 1155–1179.

[7] F. H. Clarke and P. R. Wolenski, *Necessary conditions for functional differential inclusions*, Appl. Math. Optim., 34 (1996), pp. 34–51.

[8] A. L. Dontchev and E. M. Farhi, *Error estimates for discretized differential inclusion*, Computing, 41 (1989), pp. 349–358.

[9] R. Gabasov and F. M. Kirillova, *Maximum Principle in Theory of Optimal Control*, Nauka i Technika, Minsk, Russia, 1974.

[10] J. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[11] A. D. Ioffe, *Euler-Lagrange and Hamiltonian formalisms in dynamic optimization*, Trans. Amer. Math. Soc., 349 (1997), pp. 2871–2900.

[12] M. Q. Jacobs and C. E. Langenhop, *Criteria for function space controllability of linear neutral systems*, SIAM J. Control Optim., 14 (1976), pp. 1009–1948.
[13] G. A. Kent, *A maximum principle for optimal control problems with neutral functional differential systems*, Bull. Amer. Math. Soc., 77 (1971), pp. 565–570.
[14] M. Kisielewicz, *Differential Inclusions and Optimal Control*, Kluwer, Dordrecht, The Netherlands, 1991.
[15] V. B. Kolmanovskii and L. E. Shaikhet, *Control of Systems with Aftereffect*, Academic Press, New York, 1996.
[16] P. D. Loewen and R. T. Rockafellar, *Bolza problems with general time constraints*, SIAM J. Control Optim., 35 (1997), pp. 2050–2069.
[17] L. I. Minchenko, *Necessary optimality conditions for differential-difference inclusions*, Nonlinear Anal., 35 (1999), pp. 307–322.
[18] B. S. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[19] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988.
[20] B. S. Mordukhovich, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.
[21] B. S. Mordukhovich, *Discrete approximations and refined Euler–Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.
[22] B. S. Mordukhovich, J. S. Treiman, and Q. J. Zhu, *An extended extremal principle with applications to multiobjective optimization*, SIAM J. Optim., 14 (2003), pp. 359–379.
[23] B. S. Mordukhovich and R. Trubnik, *Stability of discrete approximation and necessary optimality conditions for delay-differential inclusions*, Ann. Oper. Res., 101 (2001), pp. 149–170.
[24] B. S. Mordukhovich and L. Wang, *Optimal control of constrained delay-differential inclusions with multivalued initial conditions*, Control Cybernet., 28 (2003), No. 3.
[25] R. T. Rockafellar, *Equivalent subgradient versions of Hamiltonian and Euler–Lagrange equations in variational analysis*, SIAM J. Control Optim., 34 (1996), pp. 1300–1314.
[26] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[27] G. V. Smirnov, *Introduction to the Theory of Differential Inclusions*, American Mathematical Society, Providence, RI, 2002.
[28] H. J. Sussmann, *New theories of set-valued differentials and new versions of the maximum principle in optimal control theory*, in Nonlinear Control in the Year 2000, A. Isidori et al., eds., Springer-Verlag, Berlin, 2000, pp. 487–472.
[29] R. B. Vinter, *Optimal Control*, Birkhäuser, Boston, 2000.
[30] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[31] Q. J. Zhu, *Hamiltonian necessary conditions for a multiobjective optimal control problem with endpoint constraints*, SIAM J. Control Optim., 39 (2000), pp. 97–112.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 137–153

COUNTEREXAMPLES CONCERNING OBSERVATION OPERATORS FOR C_0-SEMIGROUPS*

BIRGIT JACOB† AND HANS ZWART‡

Abstract. This paper concerns systems of the form $\dot{x}(t) = Ax(t)$, $y(t) = Cx(t)$, where A generates a C_0-semigroup. Two conjectures which were posed in 1991 and 1994 are shown not to hold. The first conjecture (by G. Weiss) states that if the range of C is one-dimensional, then C is admissible if and only if a certain resolvent estimate holds. The second conjecture (by D. Russell and G. Weiss) states that a system is exactly observable if and only if a test similar to the Hautus test for finite-dimensional systems holds. The C_0-semigroup in both counterexamples is analytic and possesses a basis of eigenfunctions. Using the (A, C)-pair from the second counterexample, we construct a generator A_e on a Hilbert space such that $(sI - A_e)$ is uniformly left-invertible, but its semigroup does not have this property.

Key words. infinite-dimensional system, admissible observation operator, exact observability, conditional basis, C_0-semigroup, left-invertibility

AMS subject classifications. 93C25, 93A05, 93B07, 47D60

DOI. 10.1137/S0363012903423235

1. Introduction.

Consider the abstract system

$$\dot{x}(t) = Ax(t), \qquad y(t) = Cx(t), \qquad x(0) = x_0 \tag{1.1}$$

with $x(t) \in H$ and $y(t) \in Y$, where H and Y are Hilbert spaces. For this abstract differential equation one would like to obtain conditions in terms of A and C such that it has a solution with certain properties. If one only considers the differential equation $\dot{x}(t) = Ax(t)$, then it is well known that it has a unique (weak) solution which is strongly continuous and depends continuously on the initial state $x(0) = x_0 \in H$ if and only if A satisfies the estimates of the Hille–Yosida theorem (see, e.g., [4, Theorem 2.1.12]). Since $\dot{x}(t) = Ax(t)$ is a part of (1.1) we have to assume that A satisfies the estimates of the Hille–Yosida theorem, or equivalently, that A generates a C_0-semigroup. If in addition C is a bounded linear operator from H to Y, then it is straightforward to see that $y(\cdot)$ in (1.1) is well defined and continuous. However, many PDEs rewritten in the form (1.1) do not have a bounded operator C, although the output is a well-defined square integrable function. We assume that C is a bounded operator from $D(A)$ (with the graph norm) to a Hilbert space Y. If the output is locally square integrable, then C is called an *admissible observation operator* (see Weiss [20] and the survey article by Jacob and Partington [7]). In other words, C is an admissible observation operator if and only if for some $t_0 > 0$ (and hence any $t_0 > 0$) there exists a constant $L > 0$ such that

$$\int_0^{t_0} \|CT(t)x\|^2 \, dt \le L\|x\|^2, \qquad x \in D(A).$$

*Received by the editors February 17, 2003; accepted for publication (in revised form) October 23, 2003; published electronically June 4, 2004.
http://www.siam.org/journals/sicon/43-1/42323.html
†Department of Mathematics, University of Dortmund, D-44221 Dortmund, Germany (birgit.jacob@math.uni-dortmund.de).
‡Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands (h.j.zwart@math.utwente.nl).

Here $(T(t))_{t\geq 0}$ is the C_0-semigroup generated by A. If the C_0-semigroup is exponentially stable, then t_0 can be replaced by ∞. Now an interesting question is if there are simple conditions on C (and A) such that C is an admissible observation operator.

Dual to the concept of admissible observation operator is the concept of admissible control operator. An operator B is said to be an admissible control operator if $\dot{x}(t) = Ax(t) + Bu(t)$ has a continuous (weak) solution for every locally square integrable input u. It is well known that C is an admissible observation operator for A if and only if C^* is an admissible control operator for A^*; see [20] for a proof of this statement. Here $*$ denotes the adjoint operator. Because of this duality any result for admissible observation operators has an equivalent counterpart for admissible control operators, and vice versa. Hence if we refer to a paper which only deals with control operators, we trust that the reader can make the equivalent statement for observation operators. Basically, it boils down to replacing B by C^* and replacing the infinitesimal generator by its dual one.

In Weiss [21] it is shown that if C is admissible, then there exists a constant $M > 0$ such that

$$\|C(sI - A)^{-1}\| \leq \frac{M}{\sqrt{\mathrm{Re}(s)}} \tag{1.2}$$

for all s in some right-half plane. He conjectured in [21] (see also [22]) that this condition is also sufficient. The sufficiency of condition (1.2) was proved for surjective semigroups in Weiss [21], for normal, analytic semigroups in Weiss [21, 22], for the right shift semigroup with scalar output in Partington and Weiss [15], for contraction semigroups with scalar output by Jacob and Partington [6], and for analytic contraction semigroups by Le Merdy [12]. Recently, Zwart, Jacob, and Staffans [26] and Jacob, Partington, and Pott [8] showed that in general estimate (1.2) is not sufficient. Their observation operator is infinite-dimensional. Here we use techniques similar to those in [26] to show that (1.2) is not sufficient for scalar outputs. Note that in [5] a necessary and sufficient condition has been obtained. This condition involves all powers of the resolvent, as in the Hille–Yosida theorem. Some sufficient conditions for admissibility can be found in [24].

Apart from the well-posedness of the abstract differential equation (1.1) one would like to characterize other properties in terms of the pair (A, C). One property that has received a lot of attention is the property of exact observability. Assuming that the observation operator C is admissible, the system (1.1) is said to be exactly observable if there is a bounded mapping from the output trajectory to the initial condition, that is, for some $t_0 > 0$ (and hence any $t_0 > 0$) there exists a constant $l > 0$ such that

$$\int_0^{t_0} \|CT(t)x\|^2 \, dt \geq l\|x\|^2, \qquad x \in D(A).$$

If the C_0-semigroup is exponentially stable, then t_0 can be replaced by ∞. Note that admissibility gives that the mapping from initial condition to output trajectory is bounded. If the state space H is finite-dimensional, and thus A and C are just matrices, then it is well known that (1.1) is exactly observable if and only if

$$\mathrm{rank} \begin{bmatrix} C \\ sI - A \end{bmatrix}$$

is full for all complex s. For infinite-dimensional systems, Russell and Weiss [17], proposed the following test for exact observability of an exponentially stable system:

$$\|(sI - A)x_0\|^2 + |\mathrm{Re}(s)|\|Cx_0\|^2 \geq m|\mathrm{Re}(s)|^2\|x_0\|^2 \tag{1.3}$$

for all complex s with negative real part, for all $x_0 \in D(A)$, and for some positive m independent of s and x_0. In [17] they proved that this condition is always necessary, and that for A and C bounded this condition is sufficient as well. In the same paper they showed that if A has a Riesz basis of eigenfunctions and an extra condition on the eigenvalues is satisfied, then (1.3) is sufficient. In Zhou and Yamamoto [23] it was shown that (1.3) is sufficient if A is skew adjoint and C is bounded. For Riesz spectral systems with finite-dimensional output space Y inequality (1.3) is sufficient as well; see Jacob and Zwart [9, 10]. Grabowski and Callier [5] proved that if m in (1.3) is equal to one, then this estimate implies exact observability. In section 3 we show that for general m estimate (1.3) is not sufficient. Note that in our counterexample the output is one-dimensional and that A generates an analytic semigroup. In [11] we give a refined version of this conjecture.

We conclude this paper with a section on left-invertibility of C_0-semigroups. It is known that uniform left-invertibility of the semigroup implies uniform left-invertibility of the generator on the open left-half plane. We show that in general the inverse implication does not hold.

2. General results. Let H be a separable Hilbert space with a conditional basis $\{\varphi_n\}_{n\in\mathbb{N}}$. Since $\{\varphi_n\}_{n\in\mathbb{N}}$ is a conditional basis, we have that for every $x \in H$ there exists a unique sequence of complex numbers α_n such that

$$x = \lim_{k\to\infty} \sum_{n=1}^{k} \alpha_n \varphi_n. \tag{2.1}$$

Hence, we can write

$$x = \sum_{n=1}^{\infty} \alpha_n \varphi_n.$$

Using (2.1) it is not hard to see that the following holds (see also Singer [18, pages 18–20]).

LEMMA 2.1. *If $\{\varphi_n\}_{n\in\mathbb{N}}$ is a conditional basis, then the following mappings are uniformly bounded:*

$$P_n x = \sum_{k=1}^{n} \alpha_k \varphi_k \tag{2.2}$$

and

$$\tilde{P}_n x = \alpha_n \varphi_n, \tag{2.3}$$

where $x = \sum_{n=1}^{\infty} \alpha_n \varphi_n$.

Furthermore, if $\inf_{n\in\mathbb{N}} \|\varphi_n\| > 0$, then

$$\sup_{n\in\mathbb{N}} |\alpha_n| \le \kappa \|x\| \tag{2.4}$$

for some $\kappa > 0$ independent of x.

The following two properties of a conditional basis are important for the construction of our counterexamples.

DEFINITION 2.2. *Let $\{\varphi_n\}_{n\in\mathbb{N}}$ be a conditional basis.*

1. $\{\varphi_n\}_{n\in\mathbb{N}}$ *is* Besselian *if there exists a constant* $c > 0$ *such that*

$$\sum_{k=1}^{n} |a_k|^2 \leq c \left\| \sum_{k=1}^{n} a_k \varphi_k \right\|^2$$

for all finite sequences of scalars $a_1, \ldots, a_n$.

2. $\{\varphi_n\}_{n\in\mathbb{N}}$ *is* Hilbertian *if there exists a constant* $c > 0$ *such that*

$$\left\| \sum_{k=1}^{n} a_k \varphi_k \right\|^2 \leq c \sum_{k=1}^{n} |a_k|^2$$

for all finite sequences of scalars $a_1, \ldots, a_n$.

Equivalently, $\{\varphi_n\}_{n\in\mathbb{N}}$ is Besselian if and only if there exists a bounded linear operator S such that $v_n := S\varphi_n$ is an orthonormal basis for H. More information on conditional bases can be found in Singer [18].

For diagonal operators on a conditional basis of H there is the following nice result, which can be found in Benamara and Nikolski [1, Lemma 3.2.5].

LEMMA 2.3. *Let* $\{\varphi_n\}_n$ *be a conditional basis of* H. *If* Q *is defined as*

$$Q\varphi_n = q_n \varphi_n$$

with $\{q_n\}_{n\in\mathbb{N}} \subset \mathbb{C}$, *and the total variation of the sequence* $\{q_n\}$ *is finite, i.e.,*

$$Var(q_n) := \sum_{n=1}^{\infty} |q_{n+1} - q_n| < \infty,$$

then Q *can be extended to a linear bounded operator on* H, *and*

$$\|Q\| \leq K(Var(q_n) + \limsup |q_n|), \tag{2.5}$$

where K *is the supremum of* $\|P_n\|$; *see Lemma* 2.1.

In order to calculate the total variation, the following observation is useful. If f is a continuous function which is nondecreasing or nonincreasing on the interval (a, b), and if the sequence $\{q_n\}_n \subset (a, b)$ is nondecreasing or nonincreasing, then

$$Var(f(q_n)) \leq |f(a) - f(b)|.$$

In [26] the following useful result can be found.

LEMMA 2.4. *Let* $\{\mu_n\}_n \subset (-\infty, -1]$ *be a monotonically decreasing sequence with* $\lim_{n\to\infty} \mu_n = -\infty$. *Furthermore, let* $\{\varphi_n\}_{n\in\mathbb{N}}$ *be a conditional basis for the Hilbert space* H.

For $t \geq 0$, *we define* $(T(t))_{t\geq 0}$ *by*

$$T(t)\varphi_n := e^{\mu_n t}\varphi_n, \qquad n \in \mathbb{N}. \tag{2.6}$$

The operator valued function $(T(t))_{t\geq 0}$ *defines an analytic, exponentially stable* C_0*-semigroup on* H.

3. Counterexample on admissibility. In this section we show that the conjecture of George Weiss for admissibility of scalar observation operators (see [21, 22]) does not hold. That means we construct an exponentially stable C_0-semigroup $(T(t))_{t\geq 0}$ on H with infinitesimal generator A and an operator $C \in \mathcal{L}(D(A), \mathbb{C})$ such that

$$\|C(sI - A)^{-1}\| \leq \frac{M}{\sqrt{\mathrm{Re}(s)}}$$

for all s in some right-half plane and some constant $M > 0$, but C is not an admissible observation operator for $(T(t))_{t\geq 0}$.

Let $\{e_n\}_{n\in\mathbb{N}}$ be a conditional basis on H which has the following properties:

1. $\inf_{n\in\mathbb{N}} \|e_n\| > 0$.
2. $\{e_n\}_{n\in\mathbb{N}}$ is not Besselian.

Such Hilbert spaces and bases do exist; see, for example, Singer [18, page 351, example 11.2].

We define the sequence μ_n as

$$\mu_n := -4^n, \qquad n \in \mathbb{N}, \tag{3.1}$$

and the C_0-semigroup $(T(t))_{t\geq 0}$ as

$$T(t)e_n = e^{\mu_n t} e_n. \tag{3.2}$$

By Lemma 2.4 we know that $(T(t))_{t\geq 0}$ is an exponentially stable analytic semigroup. By A we denote the infinitesimal generator of $(T(t))_{t\geq 0}$. It is easy to see that A satisfies

$$Ae_n = \mu_n e_n, \qquad n \in \mathbb{N}.$$

For $x \in D(A)$, $x = \sum_{n=1}^{\infty} x_n e_n$, we further define

$$Cx = \sum_{n=1}^{\infty} \sqrt{-\mu_n}\, x_n. \tag{3.3}$$

First of all we show that C is a bounded linear operator from the domain of A into $\mathbb{C}$.

PROPOSITION 3.1. *Let C be given as in* (3.3) *and let A be the infinitesimal generator of the C_0-semigroup* (3.2). *Then we have $C \in \mathcal{L}(D(A), \mathbb{C})$.*

Proof. It is enough to show that there exists a constant $c > 0$ such that

$$|CA^{-1}x| \leq c, \qquad x \in H, \|x\| = 1.$$

Let $x \in H$ with $\|x\| = 1$. Then there exist scalars x_n, $n \in \mathbb{N}$, such that

$$x = \sum_{n=1}^{\infty} x_n e_n.$$

Using that $\inf_{n\in\mathbb{N}} \|e_n\| > 0$, we get from Lemma 2.1 that $\sup_{n\in\mathbb{N}} |x_n| \leq \kappa < \infty$. Note that κ is independent of $x \in H$ with $\|x\| = 1$. Now we have

$$|CA^{-1}x| = \left| \sum_{n=1}^{\infty} \frac{x_n}{\sqrt{-\mu_n}} \right| \leq \kappa \sum_{n=1}^{\infty} 2^{-n} = \kappa.$$

Thus the proposition is proved. □

Next we show that C satisfies the estimate (1.2).

PROPOSITION 3.2. *For C given by* (3.3) *and A as the infinitesimal generator of the semigroup* (3.2) *the following holds. There exists a constant $M > 0$ such that*

$$\|C(sI - A)^{-1}\| \leq \frac{M}{\sqrt{\operatorname{Re}(s)}}, \qquad s \in \mathbb{C}_+.$$

Proof. Let s be an element of $\mathbb{C}_+$, and let $x \in H$ have norm one. We have the following estimate:

$$\begin{aligned}\sqrt{\operatorname{Re}(s)}|C(sI - A)^{-1}x| &= \sqrt{\operatorname{Re}(s)}\left|\sum_{k=1}^{\infty}\frac{2^k}{s+4^k}x_k\right| \\ &\leq \sqrt{\operatorname{Re}(s)}\sum_{k=1}^{\infty}\frac{2^k}{|\operatorname{Re}(s)+4^k|}|x_k| \\ &\leq \kappa\sqrt{\operatorname{Re}(s)}\sum_{k=1}^{\infty}\frac{2^k}{\operatorname{Re}(s)+4^k},\end{aligned}$$

where we have used Lemma 2.1. Note that κ is independent of x. In order to estimate this last expression we introduce the monotonically decreasing sequence $a_k := \frac{1}{\operatorname{Re}(s)+k^2}$. Then for $N \geq 2^K$ we have

$$\begin{aligned}\sum_{k=1}^{N} a_k &\geq a_1 + a_2 + (a_3 + a_4) + \cdots + (a_{2^{K-1}+1} + \cdots + a_{2^K}) \\ &\geq a_2 + 2a_4 + \cdots + 2^{K-1}a_{2^K} \\ &= \frac{1}{2}\sum_{k=1}^{K} 2^k a_{2^k},\end{aligned}$$

and so

$$\sum_{k=1}^{\infty}\frac{2^k}{\operatorname{Re}(s)+4^k} \leq 2\sum_{k=1}^{\infty}\frac{1}{\operatorname{Re}(s)+k^2}.$$

Using this in our estimate of $\sqrt{\operatorname{Re}(s)}|C(sI - A)^{-1}x|$, we obtain that

$$\begin{aligned}\sqrt{\operatorname{Re}(s)}|C(sI - A)^{-1}x| &\leq 2\kappa\sqrt{\operatorname{Re}(s)}\sum_{k=1}^{\infty}\frac{1}{\operatorname{Re}(s)+k^2} \\ &\leq 2\kappa\sqrt{\operatorname{Re}(s)}\int_0^{\infty}\frac{1}{\operatorname{Re}(s)+t^2}\,dt \\ &\leq 2\kappa\sqrt{\operatorname{Re}(s)}\left(\frac{1}{\sqrt{\operatorname{Re}(s)}}\arctan\left(\frac{t}{\sqrt{\operatorname{Re}(s)}}\right)\Bigg|_0^{\infty}\right) \\ &\leq 2\kappa\frac{\pi}{2} = \kappa\pi,\end{aligned}$$

which proves our assertion. □

PROPOSITION 3.3. *If C given by* (3.3) *is an admissible observation operator for the C_0-semigroup given by* (3.2)*, then $\{e_n\}$ is Besselian.*

Proof. If C is an admissible observation operator for $(T(t))_{t\geq 0}$, then there would exist a constant $L > 0$ such that

$$\int_0^\infty |CT(t)x|^2 dt \leq L\|x\|^2, \qquad x \in D(A).$$

Now take a finite sequence of α_k's and consider

$$x := \sum_{k=1}^{n} \alpha_k e_k.$$

Then the above estimate gives

$$\int_0^\infty \left| \sum_{k=1}^{n} \sqrt{-\mu_k} e^{\mu_k t} \alpha_k \right|^2 dt \leq L\|x\|^2.$$

However, from Nikolski and Pavlov [14] (see also Jacob and Zwart [10]), we know that there exists a constant $L_1 > 0$, independent of x, such that

$$\int_0^\infty \left| \sum_{k=1}^{n} \sqrt{-\mu_k} e^{\mu_k t} \alpha_k \right|^2 dt \geq L_1 \sum_{k=1}^{n} |\alpha_k|^2.$$

Thus we have that for any finite sequence

$$\|x\|^2 \geq \frac{L_1}{L} \sum_{k=1}^{n} |\alpha_k|^2,$$

which shows that $\{e_n\}$ is Besselian. □

Thus we have disproved the scalar admissibility conjecture of George Weiss.

4. Counterexample on exact observability. In this section we use the operators A and C constructed in section 3 with different assumptions on the basis to settle another question about operator semigroups.

We disprove the conjecture of Russell and Weiss [17] on exact observability. That means we construct an exponentially stable C_0-semigroup $(T(t))_{t\geq 0}$ with infinitesimal generator A and an operator $C \in \mathcal{L}(D(A), \mathbb{C})$ such that

$$\|(sI - A)x_0\|^2 + |\mathrm{Re}(s)|\|Cx_0\|^2 \geq m|\mathrm{Re}(s)|^2\|x_0\|^2, \quad s \in \mathbb{C}_-, x_0 \in D(A),$$

for some constant $m > 0$, but the pair (A, C) is not exactly observable.

Let $\{e_n\}_{n\in\mathbb{N}}$ be a conditional basis on H which is Besselian, normalized—that is, $\|e_n\| = 1$, but not Hilbertian. Such Hilbert spaces and bases do exist; see, for example, Singer [18, page 351, example 11.2].

We define the sequence μ_n as

$$\mu_n := -4^n, \qquad n \in \mathbb{N}, \tag{4.1}$$

and the C_0-semigroup as

$$T(t)e_n = e^{\mu_n t} e_n. \tag{4.2}$$

By Lemma 2.4 we know that this is an exponentially stable analytic C_0-semigroup. By A we denote the infinitesimal generator of $(T(t))_{t\geq 0}$. It is easy to see that A satisfies

$$Ae_n = \mu_n e_n, \qquad n \in \mathbb{N}.$$

Since $\{e_n\}_{n\in\mathbb{N}}$ is Besselian, we know that there exists a bounded linear operator S such that $v_n := Se_n$ is an orthonormal basis for H. On this new basis we define

$$\tilde{A}v_n = \mu_n v_n.$$

It is easy to see that $\tilde{A}$ generates a C_0-semigroup $(\tilde{T}(t))_{t\geq 0}$, and that

$$ST(t) = \tilde{T}(t)S. \tag{4.3}$$

Now define the operator $\tilde{C}$ as

$$\tilde{C}v_n = \sqrt{-\mu_n}.$$

It is easy to see that we can extend $\tilde{C}$ as a bounded operator from the domain of $\tilde{A}$ to $\mathbb{C}$. We denote this extension again by $\tilde{C}$. We shall prove that $\tilde{C}$ is an admissible observation operator for $(\tilde{T}(t))_{t\geq 0}$. Since $(\tilde{T}(t))_{t\geq 0}$ has an orthonormal basis of eigenfunctions, we can use the result of Weiss [19], which tells us that $\tilde{C}$ is admissible if and only if

$$\sum_{-\mu_n \in R(h,\omega)} |\mu_n| \leq \beta h,$$

where

$$R(h,\omega) := \{s \in \mathbb{C}_+ \mid \operatorname{Re}(s) \leq h, |\operatorname{Im}(s) - \omega| \leq h\}$$

and β is independent of h. Using the definition of μ_n this is easy to prove. Now we define for $x \in D(A)$,

$$Cx = \tilde{C}Sx. \tag{4.4}$$

From this and (4.3) we see that for $x \in D(A)$

$$CT(t)x = \tilde{C}\tilde{T}(t)Sx.$$

Since S is bounded and since $\tilde{C}$ is admissible for $(\tilde{T}(t))_{t\geq 0}$, we obtain that C is an admissible output operator for $(T(t))_{t\geq 0}$.

In several steps we shall prove that the pair (A, C) satisfies the estimate of Russell and Weiss, but that it is not exactly observable. In our proof we follow closely the proof of Theorem 4.4 of Russell and Weiss [17]. As in [17] we define $N : \mathbb{C}_- \to \mathbb{N}$ as the integer such that

$$|s - \mu_{N(s)}| = \min_{k\in\mathbb{N}} |s - \mu_k|. \tag{4.5}$$

This number is well defined if the real part of s is unequal to $(\mu_k + \mu_{k+1})/2$ for all k. We define the set for which this mapping is well defined as $\mathbb{C}_g$.

LEMMA 4.1. *There exists a constant $c > 0$ such that, for all $s \in \mathbb{C}_g$, we have that*

$$\left|\frac{\operatorname{Re}(s)}{s-\mu_k}\right| \leq c, \qquad s \in \mathbb{C}_g, k \neq N(s),$$

and

$$\left|\frac{\operatorname{Re}(s)}{\operatorname{Re}(s)-\mu_k}\right| \leq c, \qquad s \in \mathbb{C}_g, k \neq N(s).$$

Proof. In Weiss and Russell [17] it is shown that the first estimate holds. Since $\{\mu_k\}$ is a real sequence, it is easy to see that $N(s) = N(\operatorname{Re}(s))$. Taking s to be real in the first inequality, and using this observation, proves the second inequality. □

For $s \in \mathbb{C}_g$, we define

$$V(s) := \overline{\operatorname{span}_{n \neq N(s)}\{e_n\}}. \tag{4.6}$$

Clearly, $V(s)$ is again a Hilbert space and in Singer [18, page 26, Proposition 4.1] it is shown that $\{e_n\}_{n\neq N(s)}$ is a conditional basis of $V(s)$. By $P_{V(s)}$ we denote the projection from H onto $V(s)$ given by

$$P_{V(s)} := I - \tilde{P}_{N(s)}.$$

Using Lemma 2.1 we see that the projections $P_{V(s)}$ are uniformly bounded. For $s \in \mathbb{C}_g$, we introduce the notation

$$e_n^s := \begin{cases} e_n, & n < N(s), \\ e_{n+1}, & n \geq N(s), \end{cases} \tag{4.7}$$

and

$$\mu_n^s := \begin{cases} \mu_n, & n < N(s), \\ \mu_{n+1}, & n \geq N(s). \end{cases} \tag{4.8}$$

The constant K in Lemma 2.3 is given by $K := \sup_{n\in\mathbb{N}} \|P_n\|$. Let $K(s)$ be the corresponding constant for $V(s)$ with conditional basis $\{e_n^s\}$, for $s \in \mathbb{C}_g$. Then it follows easily that $K(s) \leq K$.

Let $s \in \mathbb{C}_g$. We denote by A_s the part of A in $V(s)$, that is,

$$A_s x := Ax, \qquad x \in D(A_s),$$

and $D(A_s) := D(A) \cap V(s)$. Note that $V(s)$ is a $T(t)$-invariant subspace. Thus it is easy to see that C_s, defined by

$$C_s x := Cx, \qquad x \in D(A_s),$$

is an admissible observation operator for $(T_s(t))_{t\geq 0}$. Here $(T_s(t))_{t\geq 0}$ is the C_0-semigroup generated by A_s. Now we shall prove two important estimates.

LEMMA 4.2. *Let A_s, C_s, and $V(s)$ denote the objects defined above. The following two estimates hold.*

1. *There exists a constant $M > 0$ such that*

$$\|(sI - A_s)^{-1}\|_{V(s)} \leq \frac{M}{|\operatorname{Re}(s)|}, \qquad s \in \mathbb{C}_g.$$

2. *There exists a constant $d > 0$ such that*

$$\|C_s(sI - A_s)^{-1}\| \leq \frac{d}{\sqrt{|\operatorname{Re}(s)|}}, \qquad s \in \mathbb{C}_g.$$

Proof. Part 1. Let $s = s_r + is_i \in \mathbb{C}_g$. Clearly,

$$(sI - A_s)^{-1} e_n^s = \frac{1}{s - \mu_n^s} e_n^s, \qquad n \in \mathbb{N}.$$

This is an operator of the form as discussed in Lemma 2.3, and thus we have to show that $1/(s - \mu_n^s)$ is of bounded variation. We begin with the following simple observation:

$$\begin{aligned} \left| \frac{1}{s - \mu_{n+1}^s} - \frac{1}{s - \mu_n^s} \right| &= \left| \frac{\mu_{n+1}^s - \mu_n^s}{(s - \mu_{n+1}^s)(s - \mu_n^s)} \right| \\ &\leq \left| \frac{\mu_{n+1}^s - \mu_n^s}{(s_r - \mu_{n+1}^s)(s_r - \mu_n^s)} \right| \\ &= \left| \frac{1}{s_r - \mu_{n+1}^s} - \frac{1}{s_r - \mu_n^s} \right|, \end{aligned} \tag{4.9}$$

where we have used the fact that μ_n^s is real.

Next we define

$$h : \mathbb{R}_- \setminus \{s_r\} \to \mathbb{R}, \qquad h(x) := \frac{1}{s_r - x}.$$

Then we have $h(-\infty) = 0$, $h(0) = \frac{1}{s_r}$, and h is monotonically increasing on $(-\infty, s_r)$ and on $(s_r, 0)$. Combining the above results with Lemma 2.3 we get the following estimate for $\|(sI - A_s)^{-1}\|$:

$$\begin{aligned} &\|(sI - A_s)^{-1}\| \\ &\leq K \left(\operatorname{Var} \left(\frac{1}{s - \mu_n^s} \right) + \left| \lim_{n \to \infty} \frac{1}{s - \mu_n^s} \right| \right) = K \sum_{n=1}^{\infty} \left| \frac{1}{s - \mu_{n+1}^s} - \frac{1}{s - \mu_n^s} \right| \\ &\leq K \sum_{n=1}^{\infty} \left| \frac{1}{s_r - \mu_{n+1}^s} - \frac{1}{s_r - \mu_n^s} \right| \\ &\leq K \left[\left[0 + \frac{1}{s_r - \mu_{N(s)+1}} \right] + \left[\frac{1}{s_r - \mu_{N(s)+1}} - \frac{1}{s_r - \mu_{N(s)-1}} \right] \right. \\ &\quad \left. + \left[\frac{1}{s_r} - \frac{1}{s_r - \mu_{N(s)-1}} \right] \right] \\ &\leq \frac{(4c+1)K}{|\operatorname{Re}(s)|}, \end{aligned}$$

where we have used Lemmas 2.3 and 4.1 and (4.9). Since c and K are independent of s we have proved the statement.

Part 2. In order to prove this statement we follow Lemma 4.6 of Russell and Weiss [17]. Let $s \in \mathbb{C}_g$. Using the resolvent identity, we have

$$C_s(sI - A_s)^{-1} = C_s(-\bar{s}I - A_s)^{-1}[I - (\bar{s} + s)(sI - A_s)^{-1}].$$

Since C_s is an admissible observation operator for $(T_s(t))_{t\geq 0}$ there exists a constant $\tilde{d} > 0$, independent of s, such that

$$\|C_s(-\bar{s}I - A_s)^{-1}\| \leq \frac{\tilde{d}}{\sqrt{|\operatorname{Re}(s)|}}$$

(see, e.g., Weiss [22]). Combining this with Part 1, the statement is proved. □

Now we can prove the estimate of Russell and Weiss [17].

LEMMA 4.3. *For C defined by* (4.4) *and A as the infinitesimal generator of* (4.2) *the following holds. There exists a constant $m > 0$ such that, for every $s \in \mathbb{C}_-$ and every $x \in D(A)$, we have*

$$\frac{1}{|\operatorname{Re}(s)|^2}\|(sI - A)x\|^2 + \frac{1}{|\operatorname{Re}(s)|}\|Cx\|^2 \geq m\|x\|^2. \tag{4.10}$$

Proof. The proof of this lemma is divided into two steps. First, we show that the estimate holds for $s \in \mathbb{C}_- \setminus \mathbb{C}_g$. Second, we prove the estimate for $s \in \mathbb{C}_g$.

Part 1. If s is not in $\mathbb{C}_g$, then there exists an $k_0 \in \mathbb{N}$ such that $\operatorname{Re}(s) = (\mu_{k_0+1} + \mu_{k_0})/2$. It is easy to see that

$$(sI - A)^{-1}e_n = \frac{1}{s - \mu_n}e_n.$$

We use Lemma 2.3 to estimate the norm of this operator. Using (4.9) we see that it is sufficient to show that $\{\frac{1}{\operatorname{Re}(s)-\mu_n}\}$ is of bounded variation. Similar to the proof of Part 1 of Lemma 4.2, we obtain that

$$\|(sI - A)^{-1}\| \leq K\sum_{n=1}^{\infty}\left|\frac{1}{\operatorname{Re}(s) - \mu_{n+1}} - \frac{1}{\operatorname{Re}(s) - \mu_n}\right|.$$

Now we have that $\operatorname{Re}(s) = (\mu_{k_0+1} + \mu_{k_0})/2$, and thus we obtain

$$\begin{aligned}
&\|(sI - A)^{-1}\| \\
&\leq K\left[\left[0 + \frac{1}{\operatorname{Re}(s) - \mu_{k_0+1}}\right] + \left[\frac{1}{\operatorname{Re}(s) - \mu_{k_0+1}} - \frac{1}{\operatorname{Re}(s) - \mu_{k_0}}\right]\right. \\
&\quad \left. + \left[\frac{1}{\operatorname{Re}(s)} - \frac{1}{\operatorname{Re}(s) - \mu_{k_0}}\right]\right] \\
&\leq K\left[\frac{8}{\mu_{k_0} - \mu_{k_0+1}} + \frac{1}{|\operatorname{Re}(s)|}\right].
\end{aligned}$$

Now the sequence $\{\mu_n\} = \{-4^n\}$ satisfies

$$\frac{1}{\mu_n - \mu_{n+1}} = \frac{5/3}{|\mu_n + \mu_{n+1}|}.$$

So we see that

$$\|(sI - A)^{-1}\| \leq \frac{40K}{3|\mu_{k_0} + \mu_{k_0+1}|} + \frac{K}{|\operatorname{Re}(s)|} = \frac{23K}{3|\operatorname{Re}(s)|}.$$

This is equivalent to

$$|\operatorname{Re}(s)|^{-1}\|(sI - A)x\| \geq \frac{3}{23K}\|x\|,$$

and so (4.10) holds for $s \in \mathbb{C}_- \setminus \mathbb{C}_g$.

Part 2. In order to prove this statement we follow Theorem 4.4 of Russell and Weiss.

If (4.10) would not hold, then there would exist sequences $\{s_n\}$ and $\{z^n\}$ such that $s_n \in \mathbb{C}_g$, $z^n \in D(A)$, $\|z^n\| = 1$, and

$$\frac{1}{|\operatorname{Re}(s_n)|^2}\|(s_nI - A)z^n\|^2 + \frac{1}{|\operatorname{Re}(s_n)|}|Cz^n|^2 = \varepsilon_n^2, \tag{4.11}$$

where $\varepsilon_n \geq 0$ and $\varepsilon_n \to 0$.

Now define

$$q^n := \frac{1}{|\operatorname{Re}(s_n)|}(s_nI - A_{s_n})P_{V(s_n)}z^n$$

and the scalar α_n such that

$$\alpha_n e_{N(s_n)} = \tilde{P}_{N(s_n)}z^n = (I - P_{V(s_n)})z^n.$$

Thus we have that

$$\frac{1}{|\operatorname{Re}(s_n)|}(s_nI - A)z^n = \frac{s_n - \mu_{N(s_n)}}{|\operatorname{Re}(s_n)|}\alpha_n e_{N(s_n)} + q_n.$$

Now we have that

$$\|q^n\| = \left\|P_{V(s_n)}\frac{1}{|\operatorname{Re}(s_n)|}(s_nI - A)z^n\right\| \leq K\frac{1}{|\operatorname{Re}(s_n)|}\|(s_nI - A)z^n\| \leq K\varepsilon_n \tag{4.12}$$

by (4.11). For α_n, we obtain

$$\begin{aligned}\left|\frac{s_n - \mu_{N(s_n)}}{\operatorname{Re}(s_n)}\alpha_n\right| &= \left\|\frac{s_n - \mu_{N(s_n)}}{\operatorname{Re}(s_n)}\alpha_n e_{N(s_n)}\right\| = \left\|\frac{s_n - \mu_{N(s_n)}}{\operatorname{Re}(s_n)}\tilde{P}_{N(s_n)}z^n\right\| \\ &= \frac{1}{|\operatorname{Re}(s_n)|}\|\tilde{P}_{N(s_n)}(s_n - A)z^n\| \\ &\leq 2K\frac{1}{|\operatorname{Re}(s_n)|}\|(s_n - A)z^n\| \leq 2K\varepsilon_n. \qquad (4.13)\end{aligned}$$

By definition of q^n, we have that

$$P_{V(s_n)}z^n = |\operatorname{Re}(s_n)|(s_nI - A_{s_n})^{-1}q^n.$$

Using (4.12) and Lemma 4.2, we get

$$\|P_{V(s_n)}z^n\| \leq MK\varepsilon_n,$$

whence $P_{V(s_n)}z^n \to 0$. Since $\|z^n\| = 1$, it follows that $\|(I - P_{V(s_n)})z^n\| \to 1$, i.e.,

$$\lim_{n\to\infty}|\alpha_n| = 1. \tag{4.14}$$

Together with (4.13) this implies that

$$\lim_{n\to\infty}\left|\frac{s_n - \mu_{N(s_n)}}{\operatorname{Re}(s_n)}\right| = 0.$$

It is now easy to see that

$$\lim_{n\to\infty}\left|\frac{\mu_{N(s_n)}}{\operatorname{Re}(s_n)}\right| = 1. \tag{4.15}$$

Now we turn our attention to the second term of (4.11). We have

$$\begin{aligned} Cz^n &= C(I - P_{V(s_n)})z^n + CP_{V(s_n)}z^n \\ &= \alpha_n Ce_{N(s_n)} + C_{s_n}(s_nI - A_{s_n})^{-1}(s_nI - A_{s_n})P_{V(s_n)}z^n \\ &= \alpha_n\sqrt{-\mu_{N(s_n)}} + |\operatorname{Re}(s_n)|C_{s_n}(s_nI - A_{s_n})^{-1}q^n. \end{aligned}$$

Thus we can estimate the norm of this number as

$$|Cz^n| \geq |\alpha_n\sqrt{-\mu_{N(s_n)}}| - |\operatorname{Re}(s_n)||C_s(s_nI - A_{s_n})^{-1}q^n|.$$

Hence using Lemma 4.2, Part 2, we obtain that

$$\frac{1}{\sqrt{|\operatorname{Re}(s_n)|}}|Cz^n| \geq |\alpha_n|\left|\frac{\mu_{N(s_n)}}{\operatorname{Re}(s_n)}\right|^{\frac{1}{2}} - d\|q^n\|. \tag{4.16}$$

By (4.12) and (4.14)–(4.16), we conclude that there exists a positive number κ such that for n sufficiently large,

$$\frac{1}{|\operatorname{Re}(s_n)|}|Cz^n|^2 \geq \kappa.$$

On the other hand, (4.11) implies that for each $n \in \mathbb{N}$,

$$\frac{1}{|\operatorname{Re}(s_n)|}|Cz^n|^2 \leq \varepsilon_n^2,$$

which is a contradiction. Therefore, (4.10) must be true. □

So we know that the system (A, C) as defined in the beginning of this section satisfies the estimate of Russell and Weiss. Suppose now that the pair would be exactly observable. Then there would exist a constant $l > 0$ such that

$$\int_0^\infty |CT(t)x|^2 dt \geq l\|x\|^2, \qquad x \in D(A).$$

Now take a finite sequence of α_k's and consider

$$x := \sum_{k=1}^n \alpha_k e_k.$$

Then the above estimate gives

$$\int_0^\infty \left|\sum_{k=1}^n \sqrt{-\mu_k}e^{\mu_k t}\alpha_k\right|^2 dt \geq l\|x\|^2.$$

However, from Nikolski and Pavlov [14] (see also Russell and Weiss [17]) we know

that there exists a constant $l_1 > 0$ such that

$$\int_0^\infty \left| \sum_{k=1}^n \sqrt{-\mu_k} e^{\mu_k t} \alpha_k \right|^2 dt \leq l_1 \sum_{k=1}^n |\alpha_k|^2.$$

Thus we have that for any finite sequence,

$$\|x\|^2 \leq \frac{l_1}{l} \sum_{k=1}^n |\alpha_k|^2.$$

However, this implies that $\{e_n\}$ is Hilbertian, providing the contradiction.

Thus we have disproved the conjecture of Russell and Weiss on exact observability.

5. On left-invertibility of C_0-semigroups. We consider a bounded C_0-semigroup $(T_e(t))_{t\geq 0}$ with infinitesimal generator A_e on a separable Hilbert space Z. A natural question is whether uniform left-invertibility of the C_0-semigroup, that is,

$$\|T_e(t)x\| \geq c_1 \|x\|, \quad x \in Z, \tag{5.1}$$

for some $c_1 > 0$, is equivalent to uniform left-invertibility of $sI - A_e$ on the open left-half plane, that is,

$$\|(sI - A_e)x\| \geq c_2 |\operatorname{Re}(s)| \, \|x\|, \quad x \in D(A_e), s \in \mathbb{C}_-, \tag{5.2}$$

for some constant $c_2 > 0$.

In van Neerven [13] it is shown that (5.1) implies (5.2). Van Neerven considered only the case of a semigroup of isometries, but the general case can be proved in a similar way. If $(T_e(t))_{t\geq 0}$ can be extended to a group or if $\mathbb{C}_-$ is contained in the resolvent set of A, then (5.2) implies (5.1); see van Casteren [2, 3] or Zwart [25].

We now show that in general (5.2) does not imply (5.1). Consider the operators A and C of section 3, and let $(T(t))_{t\geq 0}$ denote the exponentially stable C_0-semigroup generated by A. We now define the semigroup $(T_e(t))_{t\geq 0}$ on $H \oplus L^2(0,\infty)$ by

$$T_e(t) \begin{pmatrix} x \\ f \end{pmatrix} := \begin{pmatrix} T(t)x \\ CT(t-\cdot)x|_{[0,t]} + f(\cdot - t)|_{[t,\infty)} \end{pmatrix}.$$

In Grabowski and Callier [5] it is shown that $(T_e(t))_{t\geq 0}$ is a uniformly bounded C_0-semigroup on $H \oplus L^2(0,\infty)$, and that the infinitesimal generator A_e of $(T_e(t))_{t\geq 0}$ is given by

$$A_e \begin{pmatrix} x \\ f \end{pmatrix} := \begin{pmatrix} Ax \\ -\dot{f} \end{pmatrix}, \quad \begin{pmatrix} x \\ f \end{pmatrix} \in D(A_e),$$

$$D(A_e) := \left\{ \begin{pmatrix} x \\ f \end{pmatrix} \mid x \in D(A), f, \dot{f} \in L^2(0,\infty), \right.$$
$$\left. f \text{ is absolutely continuous and } f(0) = Cx \right\}.$$

Next we calculate the norm of $\|(sI - A_e)\binom{x}{f}\|$. For $s = s_r + is_i \in \mathbb{C}_-$ we have

$$\begin{aligned}
&\left\|(sI - A_e)\begin{pmatrix} x \\ f \end{pmatrix}\right\|^2 \\
&\quad = \|(sI - A)x\|^2 + \|sf + \dot{f}\|^2_{L^2(0,\infty)} \\
&\quad = \|(sI - A)x\|^2 + |s|^2\|f\|^2_{L^2(0,\infty)} + \|\dot{f}\|^2_{L^2(0,\infty)} \\
&\qquad +2s_r \operatorname{Re}(\langle f, \dot{f}\rangle_{L^2(0,\infty)}) + is_i(\langle f, \dot{f}\rangle_{L^2(0,\infty)} - \langle \dot{f}, f\rangle_{L^2(0,\infty)}) \\
&\quad = \|(sI - A)x\|^2 + \|is_i f + \dot{f}\|^2_{L^2(0,\infty)} + s_r^2\|f\|^2_{L^2(0,\infty)} + 2s_r \operatorname{Re}(\langle f, \dot{f}\rangle_{L^2(0,\infty)}) \\
&\quad = \|(sI - A)x\|^2 + \|is_i f + \dot{f}\|^2_{L^2(0,\infty)} + s_r^2\|f\|^2_{L^2(0,\infty)} \\
&\qquad +s_r \int_0^\infty \frac{d}{dt}\langle f(t), f(t)\rangle \, dt \\
&\quad = \|(sI - A)x\|^2 + \|is_i f + \dot{f}\|^2_{L^2(0,\infty)} + s_r^2\|f\|^2_{L^2(0,\infty)} - s_r\|Cx\|^2,
\end{aligned}$$

because $f(0) = Cx$ and $f, \dot{f} \in L^2(0,\infty)$. Thus

$$\begin{aligned}
&\left\|(sI - A_e)\begin{pmatrix} x \\ f \end{pmatrix}\right\|^2 \\
&\geq \|(sI - A)x\|^2 + |\operatorname{Re}(s)|^2\|f\|^2_{L^2(0,\infty)} + |\operatorname{Re}(s)|\|Cx\|^2 \\
&\geq c_2|\operatorname{Re}(s)|^2 \left\|\begin{pmatrix} x \\ f \end{pmatrix}\right\|^2 \qquad \text{(using Lemma 4.3)},
\end{aligned}$$

where c_2 is independent of x and f. This shows that (5.2) holds. Assuming (5.1) holds as well, we get

$$\left\|T_e(t)\begin{pmatrix} x \\ f \end{pmatrix}\right\| \geq c_1 \left\|\begin{pmatrix} x \\ f \end{pmatrix}\right\|, \qquad t \geq 0, x \in H, f \in L^2(0,\infty),$$

for some constant $c_1 > 0$. Thus

$$\|T(t)x\|^2 + \|CT(\cdot)x\|^2_{L^2(0,t)} = \left\|T_e(t)\begin{pmatrix} x \\ 0 \end{pmatrix}\right\| \geq c_1\|x\|^2, \qquad x \in H, t \geq 0. \tag{5.3}$$

Using that $(T(t))_{t\geq 0}$ is exponentially stable, we get $\lim_{t\to\infty}\|T(t)x\|^2 = 0$, and so letting t to infinity in (5.3) gives

$$\|CT(\cdot)x\|_{L^2(0,\infty)} \geq \sqrt{c_1}\|x\|, \qquad x \in H,$$

which says that the pair (A, C) is exactly observable. However, this is in contradiction with section 3, where we showed that the pair (A, C) is not exactly observable. Thus (5.2) holds, but (5.1) is not valid.

We conclude this section with a positive result; it shows that (5.2) implies (5.1) if the constant c_2 satisfies $c_2 \geq 1$.

PROPOSITION 5.1. *Let $(T_e(t))_{t\geq 0}$ be a bounded C_0-semigroup with infinitesimal generator A_e on a separable Hilbert space Z. If (5.2) holds with $c_2 \geq 1$, then (5.1) holds as well.*

Proof. If $c_2 \geq 1$, then it is easy to see that (5.2) implies that

$$\|(sI - A_e)x\| \geq |\operatorname{Re} s|\|x\|, \quad s \in \mathbb{C}_-,$$

for all $x \in D(A)$. Choosing $s < 0$ and taking the square of the above equation gives

$$\|(sI - A_e)x\|^2 \geq s^2\|x\|^2.$$

Using the fact that Z is a Hilbert space gives that the above inequality is equivalent to

$$s^2\|x\|^2 - 2s\,\mathrm{Re}\langle x, A_e x\rangle + \|A_e x\|^2 \geq s^2\|x\|^2,$$

which is equivalent to

$$-2s\,\mathrm{Re}\langle x, A_e x\rangle + \|A_e x\|^2 \geq 0.$$

Since this must hold for all negative s, we see that

$$\mathrm{Re}\langle x, A_e x\rangle \geq 0.$$

We now consider the function $f(t) := \|T_e(t)x\|^2$. Taking the derivative of f gives

$$\dot{f}(t) = 2\,\mathrm{Re}\langle T_e(t)x, A_e T_e(t)x\rangle \geq 0.$$

Hence f is nondecreasing, and thus

$$\|T_e(t)x\|^2 = f(t) \geq f(0) = \|x\|^2.$$

Since x was arbitrary, we have shown the result. □

REFERENCES

[1] N.-E. Benamara and N. Nikolski, *Resolvent test for similarity to a normal operator*, Proc. London Math. Soc., 78 (1999), pp. 585–626.

[2] J. A. van Casteren, *Operators similar to unitary or selfadjoint ones*, Pacific J. Math., 104 (1983), pp. 241–255.

[3] J. A. van Casteren, *Boundedness properties of resolvents and semigroups of operators*, in Linear Operators (Warsaw, 1994), Banach Center Publ. 38, Polish Acad. Sci., Warsaw, 1997, pp. 59–74.

[4] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[5] P. Grabowski and F. M. Callier, *Admissible observation operators, semigroup criteria of admissibility*, Integral Equations Operator Theory, 25 (1996), pp. 182–198.

[6] B. Jacob and J. R. Partington, *The Weiss conjecture on admissibility of observation operators for contraction semigroups*, Integral Equations Operator Theory, 40 (2001), pp. 231–243.

[7] B. Jacob and J. R. Partington, *Admissibility of control and observation operators for semigroups: A survey*, in Proceedings of the IWOTA 2002, J. A. Ball, J. W. Helton, M. Klaus, and L. Rodman, eds., Birkhäuser Verlag, 2004, to appear.

[8] B. Jacob, J. R. Partington, and S. Pott, *Admissible and weakly admissible observation operators for the right shift semigroup*, Proc. Edinburgh Math. Soc., 45 (2002), pp. 353–362.

[9] B. Jacob and H. Zwart, *Exact observability of diagonal systems with a finite-dimensional output operator*, Systems Control Lett., 43 (2001), pp. 101–109.

[10] B. Jacob and H. Zwart, *Exact observability of diagonal systems with a one-dimensional output operator*, Appl. Math. Comput. Sci., 11 (2001), pp. 1277–1283.

[11] B. Jacob and H. Zwart, *A Hautus test for infinite-dimensional systems*, in Unsolved Problems in Mathematical Systems and Control Theory, V. Blondel and A. Megretski, eds., Princeton University Press, Princeton, NJ, 2004, to appear.

[12] C. Le Merdy, *The Weiss conjecture for bounded analytic semigroups*, J. London Math. Soc., 67 (2003), pp. 715–738.

[13] J. van Neerven, *The Asymptotic Behaviour of Semigroups of Linear Operators*, Oper. Theory Adv. Appl. 88, Birkhäuser, Basel, 1996.

[14] N. K. Nikol'skiĭ and B. S. Pavlov, *Bases of eigenvectors of completely nonunitary contractions and the characteristic function*, Math. USSR-Izvestija, 4 (1970), pp. 91–134.

[15] J. R. Partington and G. Weiss, *Admissible observation operators for the right shift semigroup*, Math. Control Signals Systems, 13 (2000), pp. 179–192.

[16] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, 1983.

[17] D. L. Russell and G. Weiss, *A general necessary condition for exact observability*, SIAM J. Control Optim., 32 (1994), pp. 1–23.

[18] I. Singer, *Bases in Banach Spaces* I, Springer-Verlag, Berlin, 1970.

[19] G. Weiss, *Admissibility of input elements for diagonal semigroups on* l^2, Systems Control Lett., 10 (1988), pp. 79–82.

[20] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.

[21] G. Weiss, *Two conjectures on the admissibility of control operators*, in Estimation and Control of Distributed Parameter Systems, F. Kappel and W. Desch, eds., Birkhäuser Verlag, Basel, 1991, pp. 367–378.

[22] G. Weiss, *A powerful generalization of the Carleson measure theorem?*, in Open Problems in Mathematical Systems Theory and Control, V. Blondel, E. Sontag, M. Vidyasagar, and J. Willems, eds., Springer-Verlag, London, 1999, pp. 267–272.

[23] Q. Zhou and M. Yamamoto, *Hautus condition on the exact controllability of conservative systems*, Internat. J. Control, 67 (1997), pp. 371–379.

[24] H. Zwart, *Sufficient Conditions for Admissibility*, Memorandum 1547, Department of Applied Mathematics, University of Twente, 2000, available online from http://www.math.utwente.nl/publications.

[25] H. Zwart, *On the invertibility and bounded extension of* C_0*-semigroups*, Semigroup Forum, 63 (2001), pp. 153–160.

[26] H. Zwart, B. Jacob, and O. Staffans, *Weak admissibility does not imply admissibility for analytic semigroups*, Systems Control Lett., 48 (2003), pp. 341–350.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 154–173

INPUT CONSTRAINED ADAPTIVE TRACKING WITH APPLICATIONS TO EXOTHERMIC CHEMICAL REACTION MODELS*

ACHIM ILCHMANN†, MOSALAGAE THUTO‡, AND STUART TOWNLEY§

Abstract. We consider input constrained adaptive output feedback control for a class of nonlinear systems which are prototype models for controlled exothermic chemical reactions. Our objective is set-point control of the output, i.e., the temperature of the reaction. In the context of chemical reactions, practical considerations lead us to work in the presence of input constraints. We adopt an approach based on modified λ-tracking controllers, whereby prespecified asymptotic tracking accuracy, quantified by $\lambda > 0$ set by the designer, is ensured. The adaptive control strategy does not require any knowledge of the system's parameters and does not invoke an internal model. Only a feasibility assumption in terms of the reference temperature and the input constraints is assumed.

Key words. adaptive control, exothermic chemical reaction models, global stabilization, input saturation, tracking

AMS subject classifications. 93C20, 93D40

DOI. 10.1137/S0363012901391081

1. Introduction.

In this paper, we consider input constrained adaptive output feedback control for a class of nonlinear systems which arise as prototype models for controlled exothermic chemical reactions. The output of the system is the reaction temperature, and primarily we control the rate of change of reaction temperature. Secondary control is achieved via dilution, specifically by feedrate control of reactants. Our objective is set-point control of the output, i.e., the temperature of the reaction. In the context of chemical reactions, since the rate of conversion of product into reactant should be economically profitable, this set-point temperature is often close to a hyperbolic equilibrium of the open-loop system. Additional practical considerations lead us also to work in the presence of input constraints. We adopt an approach based on modified λ-tracking controllers [7]. We are motivated by results obtained by Viel, Jadot, and Bastin [12] for similar prototype chemical reaction models. Our aims are two-fold: to show that the λ-tracking approach can be developed for this relevant class of nonlinear systems, and moreover to show that input constraints are allowed. Of particular interest is the interplay between the input constraints, the specific nature of the nonlinearities in chemical reaction models, and the set-point to be tracked.

In chemical engineering, the analysis and control of exothermic continuous stirred tank reactors (ExCSTRs) originated in [2]. They have subsequently been used extensively as models in several industries including continuous polymerization reactors, distillation columns, biochemical fermentation, and biological processes. More recently, for the prototype class of chemical reaction models used in this paper, various

*Received by the editors June 15, 2001; accepted for publication (in revised form) September 11, 2003; published electronically June 4, 2004. This work was supported by the British Council and the Belgian Fonds National de la Recherche Scientifique.

http://www.siam.org/journals/sicon/43-1/39108.html

†Institute of Mathematics, Technical University Ilmenau, Weimarer Straße 25, 98693 Ilmenau, Germany (ilchmann@mathematik.tu-ilmenau.de).

‡Department of Mathematics, University of Botswana, Private Bag 0022, Gaborone, Botswana (THUTOMV@mopipi.ub.bw).

§Department of Mathematical Sciences, University of Exeter, North Park Road, Exeter EX4 4QE, UK (townley@maths.ex.ac.uk).

nonadaptive control theory approaches have been developed for the set-point control of temperature. Specifically, in [12] a state feedback controller, with observer, was proposed for globally stabilizing the temperature of ExCSTRs; in [9] (adaptive) dynamic output PI type controllers were derived, and similar stabilization results were obtained in [1].

Whilst we are motivated by the issues raised and the results in [12, 9, 1], we adopt a different approach based on adaptive λ-tracking. This means that asymptotically a prespecified, arbitrarily small accuracy $\lambda > 0$ of the tracking error is ensured; see [6, 7]. This λ-tracking technique is well suited to classes of systems with "strict relative-degree" one, which include models for temperature control in the prototype exothermic reactions. It is therefore reasonable to expect that λ-trackers would be well-suited in this context of exothermic reactions. However, their direct application is not so straightforward because of the input constraints and also the need to find alternatives to the "minimum phase assumptions" typical in λ-tracking. In fact, instead of "minimum phase assumptions" we need a certain feasibility assumption which essentially captures the interplay between the input constraints, the specific nonlinearity in the exothermic reaction model, and the set-point (temperature) to be tracked. In the case of global set-point control, we also need to accommodate more ad hoc, non–relative-degree one, control action via dilution rates. To some extent, in modifying the λ-tracking technique, we are guided by the developments in [12]. However, our results actually go further in that we tolerate disturbances to the temperature measurement and also parameters of the system model are not invoked in the controller. We also overcome some of the drawbacks in the previous approaches in [12], and also [9] and [1], which need state feedback, or have complicated controller structure, or else require the system to be minimum phase (i.e., have exponentially stable zero dynamics).

We consider the following class of nonlinear systems:

$$(1)\qquad \begin{cases} \dot{x}(t) = C\, r\big(x(t), T(t)\big) + d\,[x^{\mathrm{in}} - x(t)], & x(0) = x^0 \in \mathbb{R}^n_{\geq 0}, \\ \dot{T}(t) = b^T\, r\big(x(t), T(t)\big) - q\, T(t) \;+\; u(t), & T(0) = T^0 \in \mathbb{R}_{>0}. \end{cases}$$

In (1), $n \in \mathbb{N}$ and the constants and variables represent the following for $m \in \mathbb{N}$ with $n > m$:

$x(t) \in \mathbb{R}^n_{\geq 0}$	concentrations of n chemical species,
$T(t) \in \mathbb{R}_{>0}$	temperature of the reactor,
$u(t) \in \mathbb{R}_{\geq 0}$	control, a combination of the temperatures of reactant feed and coolant,
$x^{\mathrm{in}} \in \mathbb{R}^n_{\geq 0}$	constant feed concentrations,
$C = [c_1, \ldots, c_m] \in \mathbb{R}^{n \times m}$	stoichiometric matrix,
$b \in \mathbb{R}^m_{\geq 0}$	coefficients of the exothermicity,
$d > 0$	dilution rate,
$q > 0$	heat transfer rate between heat exchanger and reactor.

The function

$$(2)\qquad r(\cdot,\cdot) : \mathbb{R}^n_{\geq 0} \times \mathbb{R}_{>0} \to \mathbb{R}^m_{\geq 0}$$

is locally Lipschitz with $r(0, T) = 0$ for all $T > 0$ and models the reaction kinetics.

In the context of chemical reactions, practical considerations lead us to assume that the control input $u(\cdot)$ is constrained so that there exist $\underline{u}$ and $\overline{u}$ with $0 < \underline{u} < \overline{u}$

so that

$$\underline{u} \le u(t) \le \overline{u} \qquad \text{for all} \qquad t \ge 0\,. \tag{3}$$

Remark 1.

(i) Nonlinear systems of the form (1) have been used extensively in the last thirty years as simplified models for ExCSTR models, both mathematically and in industrial applications. Their relevance was established in [3].

(ii) The values of $\underline{u}$ and $\overline{u}$ will depend on the specific application. In our work, and also in [12] and [9], they are fixed numbers which then feature strongly in the assumptions needed so as to prove convergence for the control schemes.

To make sense of (1) as a model for exothermic reactions, we make the following assumptions.

(A1) $\mathbb{R}^n_{\ge 0} \times \mathbb{R}_{>0}$ is positively invariant under (1) for any bounded, nonnegative, locally integrable $u(\cdot) : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$.

(A2) There exists $\gamma \in \mathbb{R}^n_{>0}$ such that $\gamma^T c_i \le 0$ for all columns $c_1, \dots, c_m$ of the stoichiometric matrix C.

(A3) For $T^* > 0$ there exist $0 \le \underline{u} < \overline{u}$ such that

$$\underline{u} < q T^* - b^T r(x, T^*) < \overline{u} \quad \text{for all} \quad x \in \Omega(\gamma, x^{\text{in}}) := \left\{ x \in \mathbb{R}^n_{\ge 0} \,\middle|\, \gamma^T x < \gamma^T x^{\text{in}} \right\}.$$

Remark 2.

(i) The system (1) and assumptions (A1)–(A3) capture the essential features of ExCSTRs. They give rise to a class of nonlinear systems for which a λ-tracking approach seems plausible, whilst the interplay between the nonlinearity, input constraints, and the feasibility assumption provides novelty in controller design and convergence proofs.

(ii) The assumption (A1) is natural for exothermic reactions. Indeed, concentrations and temperature should not become zero once they are positive. In fact, since $r(\cdot,\cdot)$ is nonnegative, if $u(\cdot)$ is nonnegative, then it is clear that $T(t) > 0$ whenever $T^0 > 0$. It is easy to show that the remainder of (A1) holds automatically when $n = 2$, i.e., in the case of a single reaction. For multiple reactions, there are various conditions (see, e.g., [8, Proposition 6]) in terms of specific rates which imply that (A1) holds.
(A1) has been formulated for the closed positive orthant $\mathbb{R}^n_{\ge 0}$ of the concentrations and the open half line for the temperature. The latter is natural since the reactor should not operate with zero or negative temperature; the former could also be assumed for the open positive orthant $\mathbb{R}^n_{>0}$; the analysis goes through without any changes.

(iii) (A2) holds if (1) satisfies the law of conservation of mass, which means that there exists $\gamma \in \mathbb{R}^n_{>0}$ with $\gamma^T C = 0$. This can be found implicitly in [4], and it is also assumed in [12]. If C does not represent exactly the stoichiometric relationships between all species, then conservation of mass need not be satisfied. Nevertheless, the reaction model might still be relevant provided that all essential reactions are obeyed. This approach was adopted in [3] and also in [8]. In [8] a concept of a noncyclic process was developed and shown to ensure dissipativity of mass and hence that (A2) is satisfied.

(iv) (A3) is simply a feasibility assumption arising because of the saturation of the nonnegative input $u(\cdot)$ at $\underline{u}$ and $\overline{u}$. Assumption (A3) coincides with (H3) in [12].

Note that, by continuity of $r(\cdot,\cdot)$, assumption (A3) implies, for some $\underline{T}$, $\overline{T}$, and small enough $\rho > 0$, the assumption

(A3′) For $T^* > 0$ there exist $0 < \underline{T} < T^* < \overline{T}$, $\rho > 0$, $0 < \underline{u} < \overline{u}$, such that

$$0 < \underline{u} + \rho < qT - b^T r(x,T) < \overline{u} - \rho \qquad \text{for all} \quad (x,T) \in \Omega(\gamma, x^{\text{in}}) \times [\underline{T}, \overline{T}].$$

We will work with (A3′) rather than with the weaker (A3) for the following two reasons: The explicit introduction of ρ makes the exposition in the proofs clearer, and in some of the results we need to use explicit knowledge of $[\underline{T}, \overline{T}]$ so that (A3′) holds for a given ρ.

The control objective is to regulate the temperature $T(t)$ toward a prespecified neighborhood of a given reference temperature T^*. In specific applications, T^* would correspond to a desirable, but possibly unstable, set-point temperature.

The actual error between T^* and $T(t)$ is denoted by

$$\hat{e}(t) = T^* - T(t),$$

and, since the temperature measurement may be corrupted by disturbances, we denote by $e(t)$ the measured error, i.e.,

$$e(t) \;=\; T^* - T(t) + \xi(t).$$

We assume that the disturbance signal $\xi(\cdot) : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a continuous bounded function.

To achieve the control objective, we use a λ-tracking controller

$$\boxed{\begin{aligned} e(t) &= T^* - T(t) + \xi(t), \\ u(t) &= \operatorname{sat}_{[\underline{u},\overline{u}]}\big(\beta(t)\, e(t) \,+\, u^*\big), \\ \dot{\beta}(t) &= \kappa \begin{cases} (|e(t)| - \lambda)^l & \text{if} \quad |e(t)| > \lambda, \\ 0 & \text{if} \quad |e(t)| \leq \lambda, \end{cases} \qquad \beta(0) = \beta^0, \end{aligned}} \tag{4}$$

and variations thereof. Here $\lambda > 0$ specifies the tolerance of the tracking error; $l \geq 1, \kappa, \beta^0 > 0$ are design parameters, and $u^* \in (\underline{u}, \overline{u})$ is a constant offset. Significantly, the controller involves a saturation function

$$\operatorname{sat}_{[\underline{u},\overline{u}]}(\eta) \;:=\; \begin{cases} \underline{u} & \text{if} \quad \eta < \underline{u}, \\ \eta & \text{if} \quad \eta \in [\underline{u}, \overline{u}], \\ \overline{u} & \text{if} \quad \eta > \overline{u}. \end{cases}$$

Remark 3. Note the simplicity of the adaptive λ-tracker. It consists of a proportional error feedback with saturation and a time-varying proportional gain $\beta(\cdot)$ determined adaptively by the error measurement alone. However, the design parameters should be carefully chosen when the feedback controller is applied to a real process. The upper bound $\overline{u}$ depends not only on the feasibility condition (A3′) but also on the physical limitations of the actuator. When both conditions are compatible, i.e., the actuator limit is higher than the bound in (A3′), one should choose $\overline{u}$ close to the actuator upper bound to avoid unnecessary cut off by the saturation bound.

To specify λ appropriately, one needs to know in advance an estimate of the upper bound for the magnitude of the measurement accuracy and disturbance signal. The power l in the gain adaptation influences the speed of the adaptation. If the difference $(|e(t)| - \lambda)$ is smaller than 1, then a bigger $l \geq 1$ gives a slower increase in $\beta(t)$; if the difference is bigger than 1, then the bigger l is the faster $\beta(t)$ increases. Similar effects can be achieved by varying κ or the initial gain β^0. The constant u^* is an input reference, an appropriate choice for which might be known from experiments with constant feedback. Note also that in applications any information specific to the chemical reaction of interest would be used to make additional modifications to the λ-tracking controller so as to fine-tune the performance.

Although our emphasis is on the adaptive controller (4), we also consider the nonadaptive version

$$\boxed{u(t) \;=\; \mathrm{sat}_{[\underline{u},\overline{u}]}\big(\beta(t)\, e(t) \,+\, u^*\big), \qquad \beta(\cdot) : \mathbb{R}_{\geq 0} \to [\beta^*, \infty) \quad \text{continuous.}} \tag{5}$$

Although the gain $\beta(t) \geq \beta^*$ in this nonadaptive controller might be conservatively too large, this nonadaptive controller is useful because it is even simpler than the already simple (4). We give explicit lower bounds for β^* in terms of weak conditions on the system data.

Throughout the paper we assume that the saturation bounds, the offset, the temperature set-point, and λ satisfy

$$0 < \underline{u} < u^* < \overline{u}, \qquad 0 < \lambda < \overline{T} - T^*, \qquad 0 < \underline{T} < T^* < \overline{T}. \tag{6}$$

The paper is organized as follows. In section 2 we consider local (adaptive and nonadaptive) λ-set-point control in the sense that the initial temperature T^0 belongs to $(0, \overline{T})$. We prove additional properties of the closed-loop system in the special case of a single reactant and a single product. In section 3 we consider the global tracking problem in the sense that we assume only $T^0 > 0$. This problem is solved by introducing a feedback control for the feedrate of reactants which has the effect of reducing the concentration of the reactants if the temperature of the reaction is too high. We make some conclusions in section 4. To help the presentation flow, we prove most of the results in the appendix.

2. λ-set-point control for $T^0 \in (0, \overline{T})$. In this section we consider local λ-set-point control in the sense that the initial temperature T^0 is constrained in the interval $(0, \overline{T})$. We present two feedback strategies which force the temperature into a λ-neighborhood of the given setpoint. The first is nonadaptive, whilst the second is adaptive.

PROPOSITION 4. *Suppose* (6), (A1), (A2), (A3′) *hold, and the continuous disturbance satisfies*

$$\sup_{t \geq 0} \{|\xi(t)|\} =: \|\xi\|_\infty < \overline{T} - T^* \,. \tag{7}$$

If the initial data of (1) *satisfy* $(x^0, T^0) \in \Omega(\gamma, x^{\mathrm{in}}) \times (0, \overline{T})$*, then the feedback* (5) *with*

$$\beta^* \geq [u^* - \underline{u}]/[\overline{T} - T^* - \|\xi\|_\infty] \tag{8}$$

applied to (1) *yields a unique solution*

$$(x(\cdot), T(\cdot)) : \mathbb{R}_{\geq 0} \to \Omega(\gamma, x^{\mathrm{in}}) \times (0, \overline{T}), \quad t \mapsto (x(t), T(t)). \tag{9}$$

If (8) *is strengthened to*

$$\beta^* \geq \max\left\{\frac{u^* - \underline{u}}{\overline{T} - T^* - \|\xi\|_\infty}, \frac{\overline{u} - u^*}{\lambda}, \frac{u^* - \underline{u}}{\lambda}\right\}, \tag{10}$$

then there exists $t' \geq 0$ *such that*

$$T(t) \in [T^* - \lambda - \|\xi\|_\infty, T^* + \lambda + \|\xi\|_\infty] \qquad \textit{for all } \ t \geq t'. \tag{11}$$

Proposition 4 is proved in the appendix.

Remark 5. In Proposition 4, it is ensured that the set $\Omega(\gamma, x^{\text{in}}) \times (0, \overline{T})$ (where $\Omega(\gamma, x^{\text{in}})$ denotes the generalized triangle as defined in (A3)) remains positively invariant under the closed-loop system (1), (5); more importantly, after some finite time, the temperature $T(t)$ is within the $(\lambda + \|\xi\|_\infty)$-neighborhood of the reference temperature. The width $\lambda > 0$ of the strip around the reference temperature is prespecified, but the neighborhood is corrupted by $\|\xi\|_\infty$. The condition in (7) requires that the amplitude of the measurement disturbance must be sufficiently small when compared to $\overline{T} - T^*$. Note also that the feedback gain $\beta(\cdot)$ must be large enough.

The following remark provides some intuition behind the dynamics of the closed-loop system (1), (5).

Remark 6. Consider the closed-loop system (1), (5). For any initial condition $(x^0, T^0) \in \mathbb{R}^n_{\geq 0} \times \mathbb{R}_{>0}$, there exists a unique continuously differentiable solution on a maximally extended interval $[0, \omega)$, $\omega \in (0, \infty]$. This is a standard result of the theory of ordinary differential equations following from (2).

In the following we show that $\Omega(\gamma, x^{\text{in}}) \times (0, \overline{T})$, where $\Omega(\gamma, x^{\text{in}})$ denotes the generalized triangle as defined in (A3), is invariant under (1), (5). Therefore, boundedness of $(x(\cdot), T(\cdot))$ yields $\omega = \infty$; i.e., finite escape time cannot occur.

(i) Suppose (A1), (A2) hold. We show that for any initial data $(x^0, T^0) \in \Omega(\gamma, x^{\text{in}}) \times (0, \overline{T})$, the $x(t)$ component of the solution (1), (5) remains in $\Omega(\gamma, x^{\text{in}})$ for all $t \in [0, \omega)$. In particular, $x(\cdot)$ is bounded on $[0, \omega)$.
To see this, we note from (A1) that we need only to show that $\gamma^T x(t) < \gamma^T x^{\text{in}}$ for all $t \in [0, \omega)$. This follows from integration of

$$\frac{d}{dt}\gamma^T x(t) \ = \ \gamma^T C r(x(t), T(t)) + d\,\gamma^T [x^{\text{in}} - x(t)],$$

which yields, by invoking (A2) and $\gamma^T x(0) < \gamma^T x^{\text{in}}$, for all $t \in [0, \omega)$,

$$\gamma^T x(t) \leq e^{-dt}\gamma^T x(0) + d\int_0^t e^{-d(t-\tau)} d\tau\, \gamma^T x^{\text{in}} = \gamma^T x^{\text{in}} - e^{-dt}[\gamma^T x^{\text{in}} - \gamma^T x(0)] \leq \gamma^T x^{\text{in}}.$$

(ii) From Remark 2(i), if $T^0 > 0$, then $T(t) > 0$ for all $t \in [0, \omega)$. Now suppose that (A1), (A2), (A3′), (7), and (8) hold. Now to see that $T(t) < \overline{T}$ for all $t \in [0, \omega)$, first note that from (i) we have that $x(t) \in \Omega(\gamma, x^{\text{in}})$ for all $t \in [0, \omega)$. Seeking a contradiction, suppose there exists $t' \in [0, \omega)$ such that $T(t') = \overline{T}$ and $T(t) < \overline{T}$ for all $t \in [0, t')$.
Then by (8) we have that

$$\beta(t')e(t') \ \leq \ \beta(t')[T^* - \overline{T} + \|\xi\|_\infty] \ \leq \ \underline{u} - u^*,$$

and hence $u(t') = \underline{u}$. Using the feasibility condition (A3′) yields

$$\dot{T}(t') \ = \ b^T r(x(t'), \overline{T}) - q\overline{T} + \underline{u} \ < \ -\rho,$$

and this contradicts the assumption. It follows that if $T^0 \in (0, \overline{T})$, then $T(t) \in (0, \overline{T})$ for all $t \in [0, \omega)$.

In the following theorem, we show that it is possible to determine a sufficiently large $\beta(\cdot)$ in (5) adaptively.

THEOREM 7. *Suppose* (6), (A1), (A2), (A3′) *hold, and the continuous disturbance satisfies*

$$\sup_{t\geq 0}\{|\xi(t)|\} \;=:\; \|\xi\|_\infty \;<\; \lambda/2\,. \tag{12}$$

Then an application of the λ*-tracker* (4) *to any system* (1) *yields, for any initial data*

$$(x^0, T^0) \in \Omega(\gamma, x^{\text{in}}) \times (0, \overline{T}), \qquad \beta^0 \geq [u^* - \underline{u}]/[\overline{T} - T^* - \|\xi\|_\infty], \tag{13}$$

a closed-loop system with unique solution

$$\big(x(\cdot), T(\cdot), \beta(\cdot)\big) \;:\; \mathbb{R}_{\geq 0} \;\longrightarrow\; \Omega(\gamma, x^{\text{in}}) \times (0, \overline{T}) \times \mathbb{R}_{>0} \tag{14}$$

defined on the whole time axis $\mathbb{R}_{\geq 0}$ *and, moreover,*

(i) $\lim_{t\to\infty} \beta(t) = \beta_\infty \in \mathbb{R}_{\geq 0}$*, i.e., adaptation of the gain is convergent,*

(ii) $\lim_{t\to\infty} \operatorname{dist}(|T^* - T(t)|,\ [0, \lambda + \|\xi\|_\infty]) = 0$*, i.e., the temperature* $T(t)$ *tends to the* $[\lambda + \|\xi\|_\infty]$*-strip* $\big[T^* - [\lambda + \|\xi\|_\infty],\, T^* + [\lambda + \|\xi\|_\infty]\big]$ *as* $t \to \infty$*.*

Theorem 7 is proved in the appendix.

Note that the only information needed for the λ-tracker (4) to work is that the initial gain parameter $\beta(0)$ is sufficiently large as determined from knowledge of the upper feasibility bound $\overline{T}$ and $\|\xi\|_\infty$; see (13). This has advantages when compared to the nonadaptive controller (5) in Proposition 4, which requires the stronger condition (10). The nonadaptive result in Proposition 4 guarantees that the temperature $T(t)$ remains in the $[\lambda + \|\xi\|_\infty]$-strip after some finite time, but this time is unknown, whereas Theorem 7 ensures that $T(t)$ approaches the $[\lambda + \|\xi\|_\infty]$-strip asymptotically.

To conclude this section, we consider the special case of (1) with only a single reaction. Specifically, we assume a model for a single reaction of the form

$$\left\{\begin{array}{rcrcl} \dot{x}_1(t) &=& -k\big(T(t)\big)\, x_1(t) &+& d\,[x_1^{\text{in}} - x_1(t)], \\ \dot{x}_2(t) &=& k\big(T(t)\big)\, x_1(t) &-& d\, x_2(t), \\ \dot{T}(t) &=& b\, k\big(T(t)\big)\, x_1(t) &-& q\, T(t) \;+\; u(t). \end{array}\right. \tag{15}$$

Here $b > 0$ denotes the exothermicity of a reaction $A \longrightarrow B$, $x^{\text{in}} = (x_1^{\text{in}}, 0)^T$, where x_1^{in} is the constant feed rate of reactant A, and the reaction kinetics are given by a locally Lipschitz function $k(\cdot) : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $k(0) = 0$. A typical example of $k(\cdot)$ is the Arrhenius law $k(T) = k_0 e^{-\frac{E}{RT}}$ (extended to zero by continuity), where k_0 is a constant, E is the activation energy, and R is the Joule constant. The function $k(\cdot)$ and the positive constants d, q, and b are typically unknown.

In this case $\gamma = (1, 1)^T$ and the feasibility assumption (A3′) becomes the following:

(A3″) There exist $\rho > 0$ and $0 < \underline{T} < T^* < \overline{T}$ such that
$$0 < \underline{u} + \rho < q\,T - b\,k(T)\,x_1 < \overline{u} - \rho \quad \text{for all} \quad (x_1, T) \in [0, x_1^{\text{in}}] \times [\underline{T}, \overline{T}].$$

In [12] it is shown that the nonadaptive feedback law (5) with "sufficiently large" and constant $\beta(\cdot) \equiv \beta^*$ ensures that $(x_1(t), x_2(t))$ tends to an asymptotically stable equilibrium of the closed-loop reactor dynamics. The corresponding result for λ-tracking is stated as follows.

PROPOSITION 8. *Suppose* (6), (13), $\xi(\cdot) \equiv 0$, *and* (A3″) *hold. Define*

$$x_1^* := \frac{d\, x_1^{\text{in}}}{k(T^*) + d} \qquad \textit{and} \qquad x_2^* := \frac{k(T^*)\, x_1^{\text{in}}}{k(T^*) + d}.$$

Then the solution of the closed-loop system (4), (15), *parametrized by* λ, *satisfies*

$$\lim_{\lambda \to 0} \limsup_{t \to \infty} (|x_1(t) - x_1^*|) = 0 \qquad \textit{and} \qquad \lim_{\lambda \to 0} \limsup_{t \to \infty} (|x_2(t) - x_2^*|) = 0; \tag{16}$$

i.e., the narrower the λ*-strip (i.e., smaller* λ*) is, then the closer* $(x_1(t), x_2(t))$ *is, eventually, to* (x_1^*, x_2^*).

Note that (x_1^*, x_2^*) *is an equilibrium of* (15) *for* $T(\cdot) \equiv T^*$ *and so* $(x_1^*, x_2^*) \in \partial\Omega\big((1,1), (x_1^{\text{in}}, 0)\big)$.

The proof of Proposition 8 is given in the appendix.

In the remainder of this section, we illustrate previous results by some simulations. In the simulations we use a prototype model for a single exothermic chemical reaction as was also used in [12]. By using the same model, we can at least check that the performance of the λ-tracker is not out of line with a controller which actually relies on more system information. Specifically we consider (15) with reaction kinetics modelled by the Arrhenius law $k(T) = k_0 e^{-k_1 T}$. As in [12], we use the following system parameters:

(17)
$$k_0 = e^{25},\ d = 1.1,\ q = 1.25\,[\text{min}^{-1}],\ k_1 = 8700\,[\text{K}],\ x_i^{\text{in}} = 1\,[\text{mol/l}],\ b = 209.2\,[\text{Kl/mol}].$$

These parameter values are consistent with a laboratory-scale reaction vessel of approximately 100 liters [5].

The objective is to regulate the temperature to a neighborhood of $T^* = 337.1$[K]. Our constraints for the input $u(\cdot)$ are similar to those in [12]. Specifically we suppose that

$$\underline{u} = 295, \quad \overline{u} = 505. \tag{18}$$

It is easy to see that the feasibility assumption (A3″) is satisfied in this case if

$$\underline{T} = 240, \quad \overline{T} = 339.65\ [\text{K}], \quad \rho = 5. \tag{19}$$

We assume in this simulation that the error is disturbance free, i.e., $\xi \equiv 0$, and aim for a tracking error of within 1%. This leads us to choose the following parameters in the λ-tracker (4):

$$\lambda = 2.85, \quad u^* = 330, \quad T^* = 337.1\,[\text{K}], \quad l = 2. \tag{20}$$

In the simulations we choose $\beta^0 = 12$, which satisfies (13), and we consider three different initial conditions $T^0 = 270$, $T^0 = 320$, and $T^0 = 390$. As in [12], we choose $x_1(0) = 0.02$ and $x_2(0) = 1.07$ for the initial conditions of the single reactor (15).

For the two initial conditions $T^0 = 270$ and $T^0 = 320$, we see from Figure 1 that λ-tracking of T^* by $T(t)$ is achieved in 1 minute. Note that in both cases the transient behavior of the input hits the saturation values only for a short period at the beginning of the simulation. Otherwise the input behaves smoothly. The simulation results are similar to those in [12].

The λ-tracker (4) does not work for $T^0 = 390$, which is outside the interval $[0, \overline{T}]$. As shown by the dotted line in Figure 1, a thermal runaway occurs and the

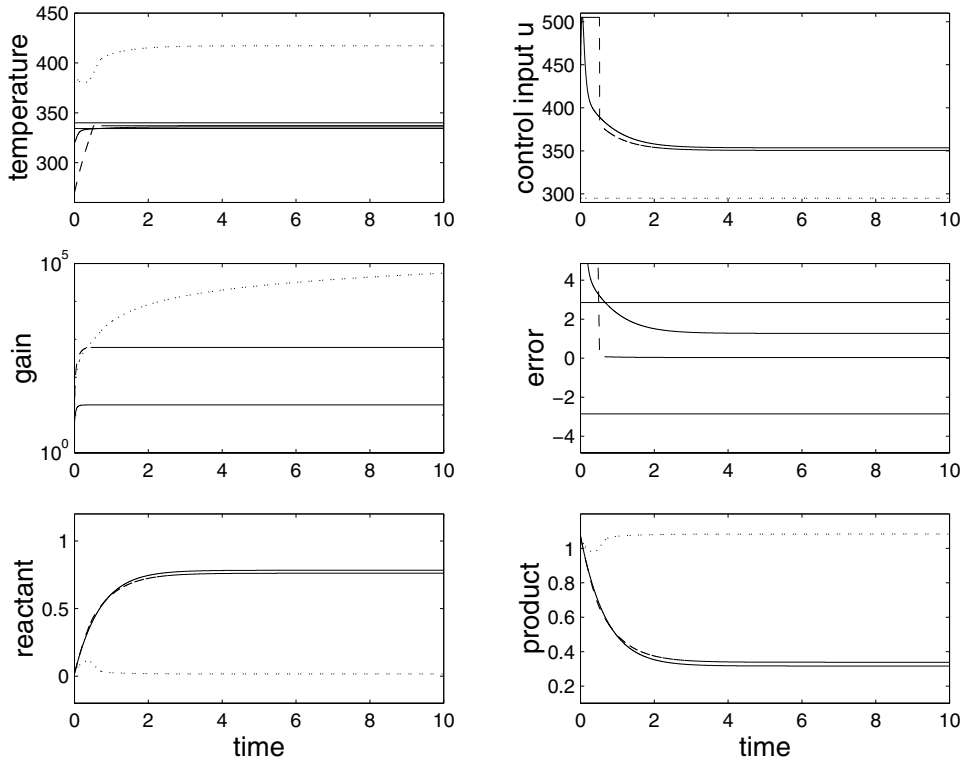


FIG. 1. *Closed-loop behavior of the adaptive λ-tracker* (4) *for local set-point control without measurement disturbance with parameters* (20) *applied to the single reaction* (15) *with parameters* (17)*, input constraints* (18)*, feasibility bounds* (19)*,* $T^0 = 270$ *(dashed),* $T^0 = 320$ *(solid),* $T^0 = 390$ *(dotted). The latter exhibits a thermal runaway.*

temperature is attracted to a stable but undesirably high temperature. As a result, the reaction becomes overheated, the reactant burns out, and there is a rapid growth of the product. Furthermore, the control input saturates at its lower limit throughout the simulation and the gain increases unboundedly.

3. Global tracking. The main result of the previous section, i.e., Theorem 7, has the shortcoming in that it is local in the sense that the initial temperature must lie inside $(0, \overline{T})$. This shortcoming can, under adverse temporary disturbances to the reaction, lead to a problem of thermal runaway in that the reaction dynamics are attracted to an undesirable equilibrium. See the simulations in Figure 1. Due to the given input saturations, it may even be impossible to reduce the temperature of the reaction from such equilibria by any type of control of the temperature alone. To overcome this problem, we borrow an idea from [12] and introduce an additional input action which has a cooling effect if the temperature is too large. To see the idea, consider the modification of the single reaction model (15) of the form

$$\left\{\begin{array}{rcrcl} \dot{x}_1(t) &=& -k\big(T(t)\big)\, x_1(t) &+& d\,[v(t) - x_1(t)], \\ \dot{x}_2(t) &=& k\big(T(t)\big)\, x_1(t) &-& d\, x_2(t), \\ \dot{T}(t) &=& b\, k\big(T(t)\big)\, x_1(t) &-& q\, T(t) \;+\; u(t), \end{array}\right. \tag{21}$$

with constant feedrate x_1^{in} replaced by $v(\cdot)$, an additional open-loop control of the feedrate of reactant. In [12] a choice of v as feedback control is

$$v(T) \;=\; \begin{cases} x_1^{\text{in}} & \text{if} \quad T \in (0, \overline{T}), \\ 0 & \text{if} \quad T \in [\overline{T}, \infty)\,. \end{cases} \tag{22}$$

The additional feedback (22) has the following beneficial effect: if $T(t) \geq \overline{T}$, then $\dot{x}_1(t) \leq -d\, x_1(t)$ and hence $x_1(\cdot)$ decreases; if $T(t) \geq \overline{T}$ is maintained, then $x_1(t)$ is eventually small enough to yield a decrease in temperature.

It is not clear to us whether the resulting discontinuous closed-loop system has a solution. It seems that the discontinuity should be harmless if the intervals in (22) are replaced by $(0, \overline{T}]$ and $(\overline{T}, \infty)$. However, since we also assume that the temperature measurement is corrupted by measurement disturbance, this discontinuity will be difficult to handle rigorously. To circumvent this technical difficulty, we replace the discontinuity in (22) by a simple piecewise linear control for $v(\cdot) : \mathbb{R} \to [0, x_1^{\text{in}}]$ given by

$$v(\beta e) \;=\; \begin{cases} 0 & \text{if} \quad \beta e \in (-\infty, \underline{u} - u^*], \\ (\beta e + u^* - \underline{u})\, x_1^{\text{in}}/\delta & \text{if} \quad \beta e \in (\underline{u} - u^*, \underline{u} - u^* + \delta), \\ x_1^{\text{in}} & \text{if} \quad \beta e \in [\underline{u} - u^* + \delta, \infty). \end{cases} \tag{23}$$

Here $\delta > 0$ would be small.

The additional feedrate control action (23) can also be introduced for multiple reactions as follows. We divide the state $x(t)$ into two substates $x_1(t)$ and $x_2(t)$ so that all reactants are collected in x_1. Applying a permutation of coordinates to (1) yields a system of the form

$$\begin{aligned} \dot{x}_1(t) &= C_1\, r\big(x(t), T(t)\big) + d\,[x_1^{\text{in}} - x_1(t)], \\ \dot{x}_2(t) &= C_2\, r\big(x(t), T(t)\big) + d\,[x_2^{\text{in}} - x_2(t)], \\ \dot{T}(t) &= b^T\, r\big(x(t), T(t)\big) - q\,T(t) + u(t), \end{aligned}$$

where $C_1 \in \mathbb{R}^{(n-m)\times m}$, $C_2 \in \mathbb{R}^{m\times m}$, $x_1^{\text{in}} \in \mathbb{R}^{n-m}_{\geq 0}$, and $x_2^{\text{in}} \in \mathbb{R}^{m}_{\geq 0}$. Since x_1 represents the reactants of the chemical reactor, it follows that each entry of C_1 is nonpositive, i.e., $C_1 \in \mathbb{R}^{(n-m)\times m}_{\leq 0}$. In this multireaction global case, the assumption (2) on the reaction kinetics must be strengthened to

$$\text{(A4)} \quad \|r(x,T)\| \leq \hat{r}(x_1)\, T \qquad \text{for all} \quad (x,T) = (x_1^T, x_2^T)^T, T) \in \Omega(\gamma, x^{\text{in}}) \times \mathbb{R}_{>0}$$

for some locally Lipschitz function $\hat{r} : \mathbb{R}^{n-m}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\lim_{x_1 \to 0} \hat{r}(x_1) = 0$.

Remark 9. Note that (A4) encompasses the class of functions considered in [12], where $b^T\, r(x,T) = \sum_{i=1}^m b_i\, k_i(T)\, \varphi_i(x)$, each $b_i > 0$, each function $T \mapsto k_i(T)$ is positive, bounded, and globally Lipschitz, and each function $x \mapsto \varphi_i(x)$ is nonnegative and continuous and vanishes if any component of x is zero for $i = 1, \ldots, m$, respectively.

The constant concentration of reactants in the feed flow x_1^{in} is replaced by an $(n-m)$-dimensional feedback term $v(\beta(\cdot)e(\cdot))$ given by (23) and the overall model becomes (compare [12, eq. (20)])

$$\begin{cases} \dot{x}_1(t) &= C_1\, r\big(x(t), T(t)\big) + d\,[v(\beta(t)e(t)) - x_1(t)], \\ \dot{x}_2(t) &= C_2\, r\big(x(t), T(t)\big) + d\,[x_2^{\text{in}} - x_2(t)], \\ \dot{T}(t) &= b^T\, r\big(x(t), T(t)\big) - q\,T(t) + u(t). \end{cases} \tag{24}$$

To proceed, we need to ensure that if the control $u(\cdot)$ and concentration of reactant $\nu(\cdot)$ in the feed of (24) are nonnegative, then the solution takes values in the positive orthant. To do this, we replace (A1) with the following:

(A1′) $\mathbb{R}^n_{\geq 0} \times \mathbb{R}_{>0}$ is positively invariant under

$$\begin{aligned} \dot{x}_1(t) &= C_1\, r\big(x(t), T(t)\big) + d\,[\nu(t) - x_1(t)], \\ \dot{x}_2(t) &= C_2\, r\big(x(t), T(t)\big) + d\,[x_2^{\text{in}} - x_2(t)], \\ \dot{T}(t) &= b^T\, r\big(x(t), T(t)\big) - q\,T(t) \;+\; u(t) \end{aligned}$$

for any bounded, nonnegative, locally integrable functions $u(\cdot) : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $\nu(\cdot) : \mathbb{R} \to [0, x_1^{\text{in}}]$.

As we pointed out in Remark 2(i), $T(t) > 0$ for all $t \geq 0$ is immediate and is only included in (A1′) for a less technical presentation. For the same reason, we have stated (A1′) for $\nu(\cdot)$, whereas it is only needed for $t \mapsto v(\beta(t)e(t))$.

Note that the comments we made for Assumption (A1) in Remark 2 apply here also. If we are in the situation described in Remark 9, then (A1′) holds.

We are now in a position to state the main result of this paper.

THEOREM 10 (adaptive tracking with measurement disturbance). *Suppose* (6), (A1′), (A2), (A3′), (A4) *hold and that the continuous disturbance satisfies*

$$\sup_{t\geq 0} \{|\xi(t)|\} \;=:\; \|\xi\|_\infty \;<\; \min\{T^* - \underline{T},\, \lambda/2\}\,. \tag{25}$$

Then an application of the λ*-tracker* (4) *combined with* (23) *to any system* (24) *yields, for any initial data* $(x^0, T^0, \beta^0) \in \Omega(\gamma, x^{\text{in}}) \times \mathbb{R}^2_{>0}$*, a closed-loop system with unique solution* $\big(x(\cdot), T(\cdot), \beta(\cdot)\big) : \mathbb{R}_{\geq 0} \longrightarrow \Omega(\gamma, x^{\text{in}}) \times \mathbb{R}^2_{>0}$ *defined on the whole time axis* $\mathbb{R}_{\geq 0}$*. Moreover,*

(i) $\lim_{t\to\infty} \beta(t) = \beta_\infty \in \mathbb{R}_{\geq 0}$*, i.e., the gain adaptation is convergent,*

(ii) $\lim_{t\to\infty} \operatorname{dist}(|T^* - T(t)|,\ [0, \lambda + \|\xi\|_\infty]) = 0$*; i.e., the temperature* $T(t)$ *tends to the* $[\lambda + \|\xi\|_\infty]$*-strip* $[T^* - [\lambda + \|\xi\|_\infty],\ T^* + [\lambda + \|\xi\|_\infty]]$ *as* $t \to \infty$*.*

The proof of Theorem 10 relies on the following high-gain lemma. This lemma is of interest in its own right, as it also gives insight into essential structural properties of the system class (24). It also shows that for sufficiently large gain, after some finite time the error enters and remains in the λ-strip.

LEMMA 11. *Suppose* (6), (A1′), (A2), (A3′), (A4), (25) *hold. Then an application to any system* (24) *of the nonadaptive feedback*

$$u(t) \;=\; sat_{[\underline{u},\overline{u}]}(\beta(t)e(t) + u^*), \qquad e(t) = T^* - T(t) + \xi(t), \tag{26}$$

combined with (23)*, yields, for any continuous* $\beta(\cdot) : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ *satisfying, for some* $t' \geq 0$*,*

$$\beta(t) \;\geq\; \beta' \;:=\; \max\left\{\frac{\overline{u} - u^*}{\lambda - 2\|\xi\|_\infty}, \frac{u^* - \underline{u}}{\lambda - 2\|\xi\|_\infty}\right\} \qquad \textit{for all} \qquad t \geq t' \tag{27}$$

and any initial data $(x^0, T^0) \in \Omega(\gamma, x^{\text{in}}) \times \mathbb{R}_{>0}$*, a closed-loop system with unique solution*

$$\big(x(\cdot), T(\cdot)\big) \;:\; \mathbb{R}_{\geq 0} \quad\longrightarrow\quad \Omega(\gamma, x^{\text{in}}) \times \mathbb{R}_{>0}$$

on the whole time axis $\mathbb{R}_{\geq 0}$*. Moreover, there exists a time* $t_1 \geq t'$ *such that*

$$e(t) \;\in\; (-\lambda, \lambda) \qquad \textit{for all} \qquad t \geq t_1. \tag{28}$$

Theorem 10 and Lemma 11 are proved in the appendix.

A simple consequence of Lemma 11 is the following theorem, which shows that tracking can be achieved by the nonadaptive feedback (5) if the constant gain parameter β^* is sufficiently large (depending on the feasibility bounds). This feedback is simpler than (19) in [12], and we give an explicit lower bound for the gain in terms of weak conditions on the system.

THEOREM 12 (nonadaptive tracking with measurement disturbance). *Suppose* (6), (A1′), (A2), (A3′), (A4), (25) *hold, and* $\beta^* \geq \beta'$ *as defined in* (27). *Then an application of the nonadaptive output feedback*

$$u(t) \;=\; sat_{[\underline{u},\overline{u}]}\big(\beta^*\, e(t) \,+\, u^*\big), \qquad e(t) \;=\; T^* - T(t) + \xi(t), \tag{29}$$

combined with (23) *to any system* (24) *yields, for any initial data* $(x^0, T^0) \in \Omega(\gamma, x^{\text{in}}) \times \mathbb{R}_{>0}$, *a closed-loop system with a unique solution*

$$\big(x(\cdot), T(\cdot)\big) \;:\; \mathbb{R}_{\geq 0} \quad \longrightarrow \quad \Omega(\gamma, x^{\text{in}}) \times \mathbb{R}_{>0}$$

on the whole time axis $\mathbb{R}_{\geq 0}$, *and moreover, there exists a time* $t_1 \geq t'$ *such that* (28) *is satisfied.*

Remark 13. If $\xi(\cdot) \equiv 0$, then (25) holds trivially, and so the adaptive gain feedback controller (4) can be applied without restriction, whereas the constant gain feedback controller (29) needs $\beta^* \geq \beta'$. If $\xi(\cdot) \not\equiv 0$, then in applying either the nonadaptive or the adaptive controller, we need to check conditions involving $\underline{T}$ and $\overline{T}$. Although this suggests that we might just as well use the simpler nonadaptive controller, in practice the adaptive gain is less conservative and the adaptive controller produces better results.

Figure 2 shows that the problem of thermal runaway above, exhibited for the local λ-set-point controller with $T^0 = 390$, is overcome by incorporating into the λ-tracker (4) the additional feedrate control via (23). Indeed, when $T^0 = 390$ the input v is switched off, i.e., $v(0) = 0$, and consumption of reactant is increased. This causes the temperature to drop, and λ-tracking is achieved. On the other hand, for $T^0 = 270$ and $T^0 = 320$, $v(\cdot) \equiv x_1^{\text{in}}$ and the response curves are the same as in Figure 1.

To illustrate the effectiveness of the controller in the presence of temperature measurement disturbances, we consider a disturbance signal

$$\xi(t) = \frac{1}{12} q_1(t), \tag{30}$$

where $q_1(\cdot)$ is the first component of the Lorenz equation

$$\begin{aligned}
\dot{q}_1(t) &= 10[q_2(t) - q_1(t)], & q_1(0) &= 1,\\
\dot{q}_2(t) &= 28q_1(t) - q_2(t) - q_1(t)q_3(t), & q_2(0) &= 0,\\
\dot{q}_3(t) &= q_1(t)q_2(t) - \tfrac{8}{3}q_3(t), & q_3(0) &= 3.
\end{aligned}$$

This Lorenz equation is known [11] to exhibit chaotic but bounded behavior. In this case $|\xi(t)| \leq 1.42$ for all $t \geq 0$. Hence, $\xi(\cdot)$ satisfies (25) for the data given in (18), (19), and (20).

Looking at Figures 3 and 4, we see that the error $T^* - T(t)$ is forced into the $[\lambda + \|\xi\|_\infty]$-strip $[-4.27, 4.27]$ despite the chaotic behavior of the disturbance signal (30). Since the constant gain in (5) is at all time equal to $\beta' = 6619$, unlike the adaptive gain in (4), which can be less than $\beta' = 6619$, the error in Figure 4 tends

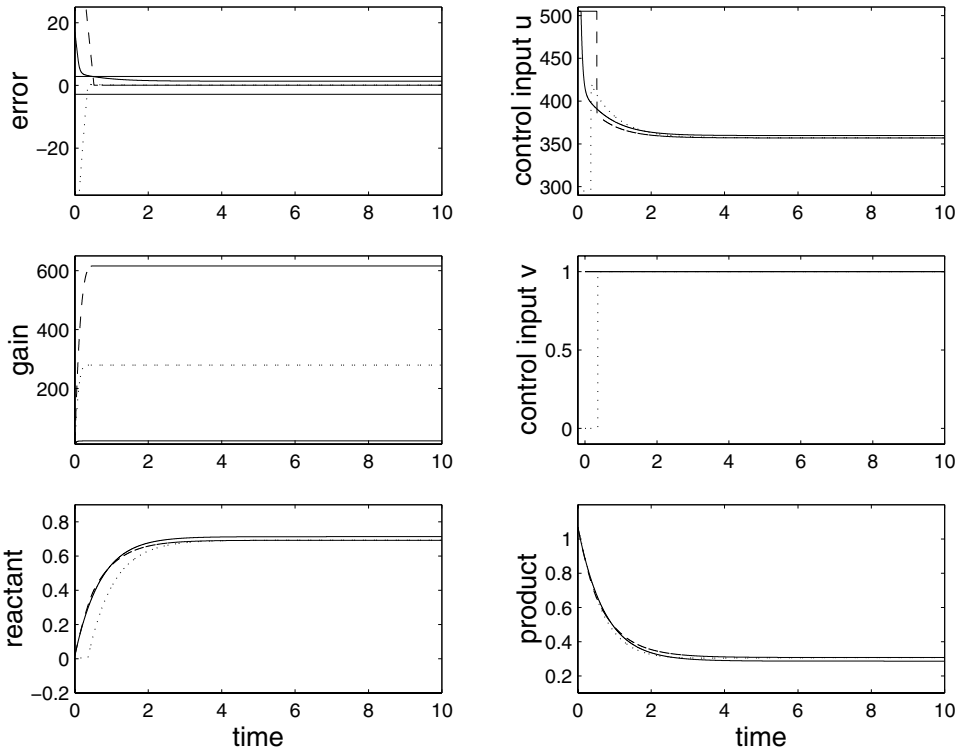


FIG. 2. *Closed-loop behavior of the adaptive λ-tracker* (4) *combined with cooling action* (23) *for global setpoint control without disturbance with parameters* (20) *applied to the single reaction* (21) *with parameters* (17), *input constraints* (18), *feasibility bounds* (19), $T^0 = 270$ *(dashed),* $T^0 = 320$ *(solid),* $T^0 = 390$ *(dotted). Here thermal runaway is overcome.*

to a smaller strip than that of the error in Figure 4. Note that there is considerably more control action for the fixed gain controller than for the adaptive gain controller, even after the control objective has been met. This increased control action is bang-bang in nature, leads to a repeated switching on and off of the control action, and is therefore undesirable from a practical point of view. This observation provides some justification for the use of the adaptive gain controller in preference to the fixed gain controller. In this group of simulations, we have omitted the graphs corresponding to the initial temperature $T^0 = 270$ since they are similar to those for $T^0 = 320$. Moreover, we have replaced the graph of the product in Figure 3 (which is close to that in Figure 4) by that of an error in a longer simulation time to show that λ-tracking is indeed achieved.

4. Conclusion. In the present paper we have developed a λ-tracking approach to the set-point control of the temperature for a class of nonlinear systems arising as models in chemical reactor control. The novelty in this development is the need to carefully consider the interplay between the reaction dynamics, input constraints, and feasibility. The application of λ-trackers requires only limited information concerning the system. In addition, the λ-trackers quite readily tolerate bounded temperature measurement disturbances. In many respects they generalize the controllers developed by [12]. It is worth noting that the minimum phase assumption usually needed for

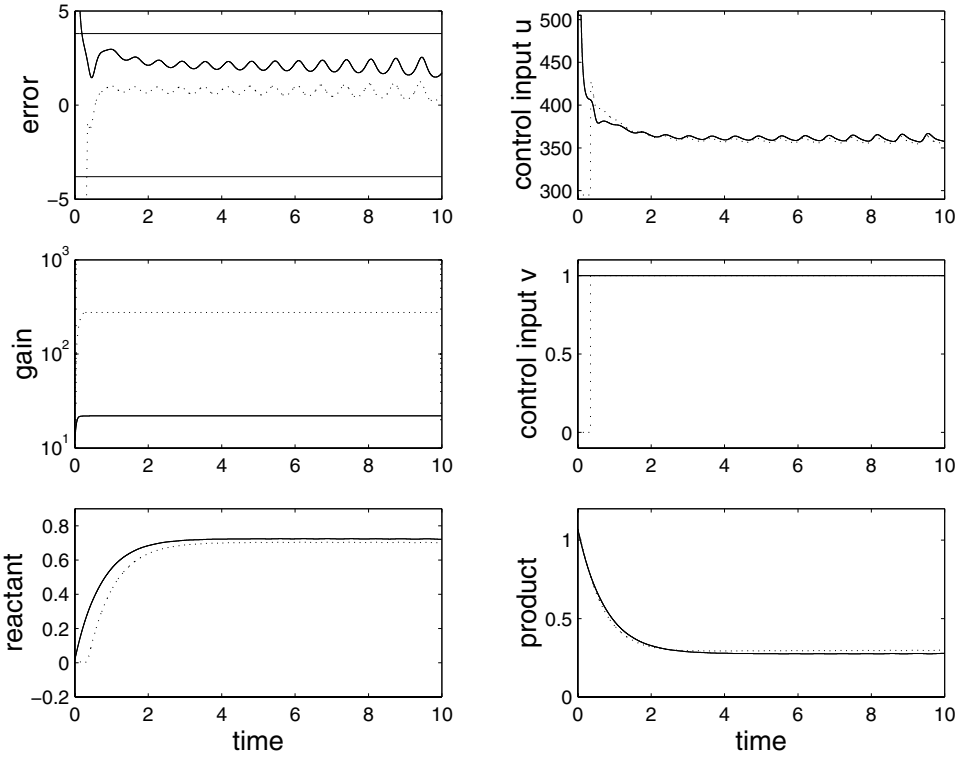


FIG. 3. *Closed-loop behavior of the adaptive λ-tracker* (4) *combined with cooling action* (23) *for global setpoint control with measurement disturbance with parameters* (20) *and disturbance signal given by* (30)*, applied to the single reaction* (21) *with parameters* (17)*, input constraints* (18)*, feasibility bounds* (19)*,* $T^0 = 320$ *(solid),* $T^0 = 390$ *(dotted).*

λ-tracking is not needed here. Instead, we exploit the natural property of chemical reactions that the internal state, i.e., the concentrations, is bounded.

Appendix. Proofs. For the sake of presentation, we define, for arbitrary $\Lambda > 0$, the distance function

$$d_\Lambda(\eta) \;:=\; \max\{|\eta| - \Lambda, 0\} \qquad \text{for all } \; \eta \in \mathbb{R}.$$

Note that for every solution (x, T) of (1) or (21) on $\mathbb{R}_{\geq 0}$ and $\hat{e}(t) = T^* - T(t)$, differentiation of

$$V_\Lambda(t) \;:=\; d_\Lambda(\hat{e}(t))^2 \qquad \text{for all } t \geq 0, \tag{31}$$

along (1) or (21) satisfies

$$\dot{V}_\Lambda(t) \;=\; \begin{cases} 2\sqrt{V_\Lambda(t)}\,\big[-b^T r\big(x(t),T(t)\big) + q\,T(t) - u(t)\big]\,, & \hat{e}(t) > 0, \\ 0, & \hat{e}(t) = 0, \\ -2\sqrt{V_\Lambda(t)}\,\big[-b^T r\big(x(t),T(t)\big) + q\,T(t) - u(t)\big]\,, & \hat{e}(t) < 0, \end{cases} \qquad \text{for all } t \geq 0. \tag{32}$$

Proof of Proposition 4. Existence and uniqueness of the solution (9) follow from Remark 6.

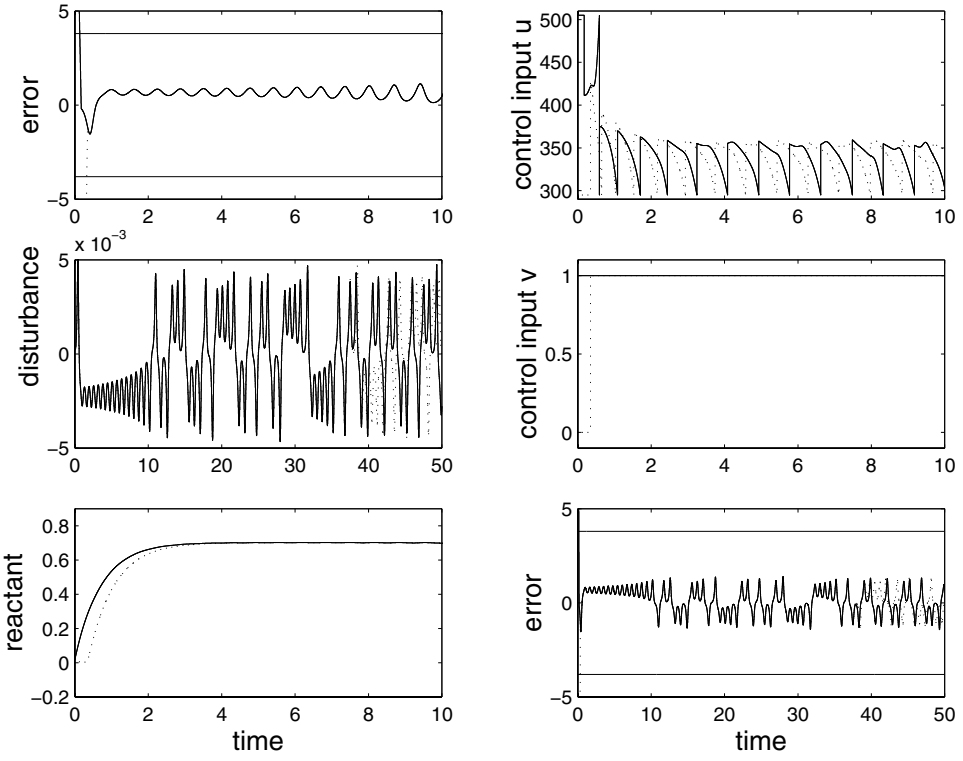


FIG. 4. *Closed-loop behavior of the constant gain controller* (5) *combined with cooling action* (23) *for global set-point control with measurement disturbance with parameters* (20) *and disturbance signal given by* (30)*, applied to the single reaction* (21) *with parameters* (17)*, input constraints* (18)*, feasibility bounds* (19)*,* $T^0 = 320$ *(solid),* $T^0 = 390$ *(dotted).*

Set $\Lambda := \lambda + \|\xi\|_\infty$ and consider, for all $t \geq 0$, the evolution of the actual error $\hat{e}(t) = T^* - T(t)$ with respect to V_Λ as in (31):

$$
\begin{aligned}
\hat{e}(t) \in [-\Lambda, \Lambda] \quad &\Longrightarrow \quad V_\Lambda(t) = 0 \ ; \\
\hat{e}(t) > \Lambda \quad &\Longrightarrow \quad \beta(t)\, e(t) + u^* \ = \ \beta(t)\, [\hat{e}(t) + \xi(t)] + u^* \\
&\qquad \overset{(5)}{>} \ \beta^* \, [\Lambda - \|\xi\|_\infty] \ + u^* = \beta^* \lambda + u^* \ \overset{(10)}{\geq} \ \bar{u} \ \overset{(5)}{=} \ u(t) \\
&\qquad \overset{(A3') \,\&\, (32)}{\Longrightarrow} \quad \dot{V}_\Lambda(t) \leq -2\rho \sqrt{V_\Lambda(t)} \, ; \\
\hat{e}(t) < -\Lambda \quad &\Longrightarrow \quad \beta(t)\, e(t) + u^* \ = \ \beta(t)\, [\hat{e}(t) + \xi(t)] + u^* \\
&\qquad < \ \beta(t)\, [-\Lambda + \|\xi\|_\infty] + u^* \ \overset{(5)}{\leq} \ -\beta^* \lambda + u^* \ \overset{(10)}{\leq} \ \underline{u} \ \overset{(5)}{=} \ u(t) \\
&\qquad \overset{(A3') \,\&\, (32)}{\Longrightarrow} \quad \dot{V}_\Lambda(t) \leq -2\rho \sqrt{V_\Lambda(t)} \, .
\end{aligned}
$$

Summarizing, we have, for all $t \geq 0$, $\dot{V}_\Lambda(t) \leq -2\rho \sqrt{V_\Lambda(t)}$, and so there exists $t' \geq 0$ such that $\hat{e}(t) \in [-\Lambda, \Lambda]$ for all $t \geq t'$, whence (11).

Proof of Theorem 7. Existence and uniqueness of the initial value problem (1), (4), (13) on a maximally extended interval $[0, \omega)$, $\omega \in (0, \infty]$, follows from the

theory of ordinary differential equations. Monotonicity of $t \mapsto \beta(t)$ and (13) yield $\beta(t) \geq \beta^*$ with β^* satisfying (8). Therefore Proposition 4 applies and $\Omega(\gamma, x^{\text{in}}) \times (0, \overline{T})$ is positively invariant. Now the gain adaptation (4) yields that $\beta(\cdot)$ cannot exhibit a finite escape time on $[0, \omega)$ and hence $\omega = \infty$.

For the remainder of the proof, consider the unique solution $\big(x(\cdot), T(\cdot), \beta(\cdot)\big)$ of (14).

We show assertion (i). Seeking a contradiction, suppose that β is unbounded. Then

$$\text{(33)} \qquad \text{there exists } \hat{t} \geq 0 \ : \ \text{for all } t \geq \hat{t} \quad : \qquad \beta(t) \ \geq \ \frac{\max\big\{u^* - \underline{u},\, \overline{u} - u^*\big\}}{\lambda - 2\,\|\xi\|_\infty}\,.$$

Set $\Lambda := \lambda - \|\xi\|_\infty$. By (12) $\Lambda > \|\xi\|_\infty$. Now consider, for all $t \geq \hat{t}$, the evolution of the actual error $\hat{e}(t) = T^* - T(t)$ with V_Λ as in (31):

$$\begin{aligned}
\hat{e}(t) \in [-\Lambda, \Lambda] \qquad &\Longrightarrow \qquad V_\Lambda(t) = 0\ ; \\
\hat{e}(t) > \Lambda \qquad &\Longrightarrow \qquad \beta(t)\, e(t) + u^* \ = \ \beta(t)\,[\hat{e}(t) + \xi(t)] + u^* \\
&\qquad\qquad > \ \beta(t)\,[\Lambda - \|\xi\|_\infty] + u^* = \ \beta(t)\,[\lambda - 2\|\xi\|_\infty] + u^* \\
&\qquad\qquad \overset{(33)}{\geq} \ \beta^*\big[\lambda - 2\|\xi\|_\infty\big] + u^* \ \geq \ \bar{u} \ \overset{(5)}{=} \ u(t) \\
&\overset{(A3')\,\&\,(32)}{\Longrightarrow} \qquad \dot{V}_\Lambda(t) \leq -2\rho\,\sqrt{V_\Lambda(t)}; \\
\hat{e}(t) < -\Lambda \qquad &\Longrightarrow \qquad \beta(t)\, e(t) + u^* \ = \ \beta(t)\,[\hat{e}(t) + \xi(t)] + u^* \\
&\qquad\qquad < \ \beta(t)\,[-\Lambda + \|\xi\|_\infty] + u^* \\
&\qquad\qquad \overset{(33)}{\leq} \ -\beta^*\big[\lambda - 2\|\xi\|_\infty\big] + u^* \ \leq \ \underline{u} \overset{(5)}{=} u(t) \\
&\overset{(A3')\,\&\,(32)}{\Longrightarrow} \qquad \dot{V}_\Lambda(t) \leq -2\rho\,\sqrt{V_\Lambda(t)}.
\end{aligned}$$

Summarizing, we have, for all $t \geq \hat{t}$, $\dot{V}_\Lambda(t) \leq -2\rho\,\sqrt{V_\Lambda(t)}$, and so there exists $t' \geq \hat{t}$ such that $\hat{e}(t) \in [-\Lambda, \Lambda]$ for all $t \geq t'$, whence $\beta(t) = 0$ for all $t \geq t'$, which contradicts the supposition of unboundedness of β. Therefore, β is bounded and assertion (i) follows by monotonicity of β.

Finally, we show assertion (ii). It is easy to see that

$$\kappa \int_0^t d_{\lambda + \|\xi\|_\infty}\big(\hat{e}(\tau)\big)^l \,\mathrm{d}\tau \ \leq \ \kappa \int_0^t d_\lambda\big(e(\tau)\big)^l \,\mathrm{d}\tau \ = \ \beta(t) - \beta^0 \qquad \text{for all } t \geq 0,$$

and hence assertion (i) yields $d_{\lambda+\|\xi\|_\infty}\big(\hat{e}(\cdot)\big)^l \in \mathcal{L}^1([0,\infty); \mathbb{R})$. Since continuity of $\eta \mapsto d_\lambda(\eta)$, together with boundedness and uniform continuity of $t \mapsto \hat{e}(t)$, yields uniform continuity of the composition $t \mapsto d_\lambda(\hat{e}(t))$, we may apply Barbălat's lemma (see, e.g., [10]) to conclude

$$\lim_{t\to\infty} \operatorname{dist}\Big(|\hat{e}(t)|,\ [0, \lambda + \|\xi\|_\infty]\Big) = 0.$$

This proves assertion (ii) and completes the proof of the theorem.

Proof of Proposition 8. By Theorem 7, there exist $t_0 \geq 0$ such that, for all $t \geq t_0$, $T(t) \in [T^* - 2\lambda, T^* + 2\lambda]$, and so, for all $t \geq t_0$,

$$\begin{aligned}
\text{(34)} \quad \inf\Big\{k(T)\Big|\ T \in [T^* - 2\lambda, T^* + 2\lambda]\Big\} \ &=: \ k_1(\lambda) \ \leq \ k(T(t)) \\
&\leq \ k_2(\lambda) \ := \ \sup\Big\{k(T)\Big|\ T \in [T^* - 2\lambda, T^* + 2\lambda]\Big\},
\end{aligned}$$

whence, by the continuity of $k(\cdot)$,

$$\lim_{\lambda\to 0} k_1(\lambda) = \lim_{\lambda\to 0} k_2(\lambda) = k(T^*). \tag{35}$$

Integrating the first equation in (15) yields

$$x_1(t) = e^{-\int_{t_0}^t (k(T(\tau))+d)\,\mathrm{d}\tau}\, x_1(t_0) + \int_{t_0}^t e^{-\int_s^t (k(T(\tau))+d)\,\mathrm{d}\tau} d\, x_1^{\mathrm{in}}\,\mathrm{d}s.$$

So, applying (34), we obtain

$$\frac{d\, x_1^{\mathrm{in}}}{k_2(\lambda)+d} \;\le\; \liminf_{t\to\infty} x_1(t) \;\le\; \limsup_{t\to\infty} x_1(t) \;\le\; \frac{d\, x_1^{\mathrm{in}}}{k_1(\lambda)+d}. \tag{36}$$

Therefore the first equation in (16) follows from (36) and (35). Integrating the second equation in (15) yields

$$x_2(t) = e^{-d\,(t-t_0)}\, x_2(t_0) + \int_{t_0}^t e^{-d(t-s)} k\big(T(s)\big)\, x_1(s)\,\mathrm{d}s.$$

Now applying (34) and (36), we obtain

$$\frac{k_1(\lambda)\, x_1^{\mathrm{in}}}{k_2(\lambda)+d} \;\le\; \liminf_{t\to\infty} x_2(t) \;\le\; \limsup_{t\to\infty} x_2(t) \;\le\; \frac{k_2(\lambda)\, x_1^{\mathrm{in}}}{k_1(\lambda)+d}, \tag{37}$$

and so the second equation in (16) follows from (37) and (35). This completes the proof.

Proof of Lemma 11. We proceed in several steps.

Step 1. The right-hand side of the closed-loop system is locally Lipschitz, and so the existence and uniqueness of the solution on a maximally extended interval $[0,\omega)$, $\omega \in (0,\infty]$, follow from the theory of ordinary differential equations.

Step 2. We show positive invariance of $\Omega(\gamma, x^{\mathrm{in}}) \times (0,\infty)$. Note that (A2) yields, for $x^{\mathrm{in}} = (x_1^{\mathrm{in}^T}, x_2^{\mathrm{in}^T})^T$, $\frac{d}{dt}\gamma^T x(t) \le -d\gamma^T x(t) + d\gamma^T x^{\mathrm{in}}$, and hence by integration

$$\gamma^T x(t) \le e^{-dt}\gamma^T x(0) + \gamma^T x^{\mathrm{in}}[1 - e^{-dt}] \qquad \text{for all } \; t \in [0,\omega).$$

If $x(0) \in \Omega(\gamma, x^{\mathrm{in}})$, then $\gamma^T x(0) < \gamma^T x^{\mathrm{in}}$, and so this inequality together with assumption (A1′) proves $x(t) \in \Omega(\gamma, x^{\mathrm{in}})$ for all $t \in [0,\omega)$. To see that $T(t)$ is positive, note that if we had $T(t) = 0$, then, by (1) and (6), $\dot{T}(t) \ge u(t) \ge \underline{u} > 0$.

Step 3. We show $\omega = \infty$. Since $x(\cdot)$ and $u(\cdot)$ are bounded, (A4) ensures that the right-hand side of $\dot{T}$ in (24) is affine linearly bounded in T and hence $T(\cdot)$ cannot escape in finite time. Applying the boundedness of $x(\cdot)$ and the maximality of ω yields the claim.

Step 4. We show that there exists $t_1 \ge t'$ such that $T(t) \in (0, \overline{T})$ for all $t \ge t_1$. Recall that, by Step 2, $T(t) > 0$ for all $t \in [0,\omega)$.

(4a) We claim that $T(s) \le \overline{T}$, for some $s \in [0,\omega)$, implies $T(t) \in (0,\overline{T})$ for all $t \in (s,\omega)$. This follows from (24) and (A3′), which give, in the case of $T(t) = \overline{T}$, that

$$\dot{T}(t) \;=\; b^T\, r\big(x(t), \overline{T}\big) - q\,\overline{T} + \underline{u} \;<\; -\rho\,.$$

(4b) It remains to be shown that if $T(t') > \overline{T}$, then $T(t) = \overline{T}$ for some $t > t'$. Seeking a contradiction, suppose

$$T(t) \;>\; \overline{T} \qquad \text{for all} \qquad t \ge t'\,. \tag{38}$$

Then (38) together with (27) gives

$$\beta(t)e(t) + u^* \;\le\; \beta(t)\,[T^* + \|\xi\|_\infty - \overline{T}] + u^* \;\le\; \underline{u} \qquad \text{for all} \qquad t \ge t', \tag{39}$$

and hence, by (23), $v(\beta(t)e(t)) = 0$ for all $t \ge t'$. Therefore, (24) and the fact that all entries of C_1 are nonpositive yield

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\,\|x_1(t)\|^2 = 2\,x_1(t)^T\,\big[C_1\,r\big(x(t),T(t)\big) - d\,x_1(t)\big] \le -2\,d\,\|x_1(t)\|^2,$$

and it follows that

$$\|x_1(t)\| \;\le\; e^{-d(t-t')}\|x_1(t')\| \qquad \text{for all} \quad t \ge t'\,. \tag{40}$$

By (A3′), we may choose $\varepsilon \in (0,q)$ sufficiently small so that

$$-\,[q-\varepsilon]\,\overline{T} + \underline{u} \;<\; -\,\rho/2\,. \tag{41}$$

By (A4) and (40), there exists $t_1 \ge t'$ such that

$$\hat{r}(x_1(t)) \;\le\; \varepsilon/\|b\| \qquad \text{for all} \quad t \ge t_1\,. \tag{42}$$

Finally, applying (39), (24), (A4), (42), (38), and (41) yields

$$\begin{aligned}\dot{T}(t) \;&\le\; b\,r\big(x(t),T(t)\big) - qT(t) + \underline{u} \;\le\; \|b\|\,\hat{r}(x_1(t))\,T(t) - qT(t) + \underline{u}\\ &\le\; -\,[q-\varepsilon]\,T(t) + \underline{u} \;\le\; -\,[q-\varepsilon]\,\overline{T} + \underline{u} \;<\; -\,\rho/2 \qquad \text{for all} \qquad t \ge t_1\,.\end{aligned}$$

It then follows that there exists $t_2 \ge t_1$ such that $T(t_2) = \overline{T}$, which contradicts (38). This completes the proof of Step 4.

Step 5. Finally, we prove the existence of some $t_1 \ge t'$ such that (28) holds. Note that it suffices to show that there exists $t_2 \ge t_1$, t_1 as in Step 4, such that the actual error satisfies

$$\hat{e}(t) \in (-\lambda + \|\xi\|_\infty, \lambda - \|\xi\|_\infty) \qquad \text{for all} \qquad t \ge t_2\,, \tag{43}$$

since then (25) yields (28).

Set $\Lambda := \lambda - \|\xi\|_\infty$ and consider, for all $t \ge t_1$, the evolution of the actual error $\hat{e}(t) = T^* - T(t)$ with respect to V_Λ as in (31). Then, for all $t \ge t_1$,

$$\begin{aligned}
&\hat{e}(t) \in [-\Lambda,\Lambda] \quad &&\Longrightarrow \quad V_\Lambda(t) = 0\;;\\
&\hat{e}(t) > \Lambda \quad &&\Longrightarrow \quad \beta(t)\,e(t) + u^* \;=\; \beta(t)\,[\hat{e}(t) + \xi(t)] + u^*\\
&&&\qquad > \beta(t)\,[\Lambda - \|\xi\|_\infty] + u^* \;=\; \beta(t)\,[\lambda - 2\|\xi\|_\infty] + u^*\\
&&&\qquad \overset{(25,27)}{\ge} \beta'\,[\lambda - 2\|\xi\|_\infty] + u^* \overset{(5,27)}{\ge} \bar{u} \;=\; u(t)\\
&&&\overset{(A3')\,\&\,(32)}{\Longrightarrow} \quad \dot{V}_\Lambda(t) \le -2\rho\,\sqrt{V_\Lambda(t)};\\
&\hat{e}(t) < -\Lambda \quad &&\Longrightarrow \quad \beta(t)\,e(t) + u^* \;=\; \beta(t)\,[\hat{e}(t) + \xi(t)] + u^*\\
&&&\qquad <\; \beta(t)\,[-\Lambda + \|\xi\|_\infty] + u^* \;= \beta(t)\,[\lambda - 2\|\xi\|_\infty] + u^*\\
&&&\qquad \overset{(25,27)}{\le} -\beta'\,[\lambda - 2\|\xi\|_\infty] + u^* \overset{(5,27)}{\le} \underline{u} \;=\; u(t)\\
&&&\overset{(A3')\,\&\,(32)}{\Longrightarrow} \quad \dot{V}_\Lambda(t) \le -2\rho\,\sqrt{V_\Lambda(t)}.
\end{aligned}$$

Summarizing, we have, for all $t \geq t_1$, $\dot{V}_\Lambda(t) \leq -2\rho\sqrt{V_\Lambda(t)}$, and so there exists $t_2 \geq \hat{t}$ such that $\hat{e}(t) \in [-\Lambda, \Lambda]$ for all $t \geq t_2$, whence (43). This completes the proof of the lemma.

Proof of Theorem 10.

Steps 1–3. These steps are the same as in the proof of Lemma 11. The only addition to the proofs is that $\beta(\cdot)$ does not have a finite escape time if $T(\cdot)$ does not have a finite escape time.

Step 4. We prove (i). Note that $t \mapsto \beta(t)$ is monotonically nondecreasing. Then either (27) is not satisfied, in which case (i) is immediate, or (27) is satisfied. However, the latter yields by Lemma 11 that (28) holds, and thus the "dead zone" in the adaptation law (4) guarantees boundedness of $\beta(\cdot)$.

Step 5. We prove boundedness of $T(\cdot)$. Seeking a contradiction, suppose that $T(\cdot)$ is unbounded. Then there exists a sequence of disjoint intervals $I_m = (a_m, b_m)$, $m \in N$, with

$$T(a_m) = T^0 + T^* + \|\xi\|_\infty + \lambda + m < T(t) < T(b_m) = T^0 + T^* + \|\xi\|_\infty + \lambda + m + 1$$

for all $t \in I_m$. It follows that

$$|e(t)| = |T(t) - T^* - \xi(t)| \geq T(t) - T^* - \|\xi\|_\infty > \lambda + 1\,,$$

and hence $d_\lambda(e(t)) = |e(t)| - \lambda > 1$ for all $t \in \bigcup_{m\in\mathbb{N}} I_m$. Furthermore, we have, with $d := c[T^0 + T^* + \|\xi\|_\infty + \lambda + 1] + \overline{u}$ and for all $m \in \mathbb{N}$,

$$1 \;=\; T(b_m) - T(a_m) \;=\; \int_{a_m}^{b_m} \dot{T}(t)dt \;<\; \int_{a_m}^{b_m} [cT(b_m) + \overline{u}]dt \;=\; [c\,m + d]\,[b_m - a_m]\,.$$

This leads to the contradiction

$$\infty \;=\; \sum_{m\in\mathbb{N}} \frac{1}{c\,m + d} \;<\; \sum_{m\in\mathbb{N}} |b_m - a_m| \;<\; \sum_{m\in\mathbb{N}} \int_{a_m}^{b_m} d_\lambda(|e(t)|)^l dt \;\leq\; \frac{\beta_\infty}{\kappa} \;<\; \infty\,.$$

Step 6. Since all variables of the closed-loop system are bounded, the proof of (ii) is identical to Step 6 in the proof of Theorem 7. This completes the proof of the theorem.

Acknowledgments. We are indebted to an anonymous referee for several constructive comments which have helped to improve the paper. We add special thanks to Petia Georgieva (Bulgarian Academy of Sciences, Sofia) for a number of remarks on the modelling of exothermic chemical reactions.

REFERENCES

[1] J. Alvarez-Ramirez and R. Fermat, *Robust PI stabilization of a class of chemical reactors*, Systems Control Lett., 38 (1999), pp. 219–225

[2] R. Aris and N. Admudson, *An analysis of chemical reactor stability and control:* I, II, III, Chemical Engrg. Sci., 7 (1958), pp. 121–155.

[3] G. Bastin and D. Dochain, *On-Line Estimation and Adaptive Control of Bioreactors*, Elsevier, Amsterdam, 1990.

[4] G. R. Gavalas, *Nonlinear Differential Equations of Chemically Reacting Systems*, Springer-Verlag, Berlin, 1968.

[5] P. Georgieva, *Private communication*, Bulgarian Academy of Sciences, Sofia, 2002.

[6] A. Ilchmann and E. P. Ryan, *Universal λ-tracking for nonlinearly-perturbed systems in the presence of noise*, Automatica J. IFAC, 30 (1994), pp. 337–346.
[7] A. Ilchmann, E. P. Ryan, and C. J. Sangwin, *Systems of controlled functional differential equations and adaptive tracking*, SIAM J. Control Optim., 40 (2002), pp. 1746–1764.
[8] A. Ilchmann and M.-F. Weirig, *Modelling of general biotechnological processes*, Math. Comput. Model. Dyn. Syst., 5 (1999), pp. 152–178.
[9] F. Jadot, G. Bastin, and V. Viel, *Robust global stabilization of stirred tank reactors by saturated output feedback*, Eur. J. Control, 5 (1999), pp. 361–371.
[10] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.
[11] C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*, Springer-Verlag, New York, 1982.
[12] F. Viel, F. Jadot, and G. Bastin, *Global stabilization of exothermic chemical reactors under input constraints*, Automatica J. IFAC, 33 (1997), pp. 1437–1448.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 174–193

$C^{2,\alpha}$ EXISTENCE RESULT FOR A CLASS OF SHAPE OPTIMIZATION PROBLEMS*

ARIAN NOVRUZI†

Abstract. In this paper we give a sufficient condition for the existence of regular optimal domains for a class of shape optimization problems. This consists of finding a $C^{2,\alpha}$ domain minimizing locally a shape functional depending on the perimeter of the domain and on a general term, which in most PDE applications represents the energy associated with the state equation, under the constraint that the measure of the domain is given. The proof of this result is based on another existence result for $C^{2,\alpha}$ solutions for a class of free boundary problems that are critical domains for the shape functionals considered previously. A key point is the introduction of a special domain transformation, which has a separate term responsible for the domain translation and another which is basically only responsible for "pure" deformation of the domain. As an application, a typical example involving the Dirichlet problem in R^N is considered.

Key words. shape optimization, optimal domain, existence result

AMS subject classifications. 35R35, 31B20, 65N21

DOI. 10.1137/S0363012902382401

1. Introduction.

In this paper we present a sufficient condition for the existence of $C^{2,\alpha}$ *optimal domains* which minimize locally a shape functional $E(\Omega) = e(\Omega) + \frac{1}{\sigma^2}P(\Omega)$ under the constraint that the measure $m(\Omega)$ of Ω is given, where $\Omega \subset R^N$ is a $C^{2,\alpha}$ open set, $e(\Omega)$ is a shape functional satisfying some hypotheses given later, $P(\Omega)$ is the perimeter of Ω, and $\sigma \in R\backslash\{0\}$.

Let $N \geq 2$, $0 < \alpha < 1$, $\mathcal{O} = \{\Omega \subset R^N,\ C^{2,\alpha}$ *open simply connected bounded set*$\}$, and let us consider the following problem:

$$\text{Find } \Omega_\sigma \in \mathcal{O} \mid E(\Omega_\sigma) = \min_{\text{loc}}\{E(\Omega),\ \Omega \in \mathcal{O},\ m(\Omega) = m_0\}, \quad m_0 > 0 \text{ given.} \tag{1.1}$$

The existence of solutions of (1.1) is proven in two steps: first we prove that there exist $C^{2,\alpha}$ *critical domains*, which are solutions of free boundary problems associated with (1.1), and second, based on this result, we prove that these critical domains are optimal domains.

For a particular $e(\Omega)$ and in two- and three-dimensional cases this problem has been studied in [3], [6], [7], [8], [9], and [11]. In [6] is given a sufficient existence condition for C^2 critical domains in dimension two with a particular $e(\Omega)$.

In this paper we prove an existence result for the existence of optimal domains of E for $|\sigma| << 1$, under some "reasonable assumptions" on $e(\Omega)$. Compared with the work presented in [6], in this paper there are two main contributions:

(i) the existence result is proven for both the minimization problem (1.1) and the free boundary problem rather than only for the free boundary problem associated with (1.1), and

(ii) the results are in dimension $N \geq 2$.

*Received by the editors March 11, 2002; accepted for publication (in revised form) April 30, 2003; published electronically June 15, 2004.

http://www.siam.org/journals/sicon/43-1/38240.html

†Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Ave., Ottawa, ON K1N 6N5, Canada (novruzi@uottawa.ca).

To solve (1.1), in addition to some conditions on $e(\Omega)$ required to solve the free boundary problem (hypotheses (1.2), (1.5), (1.6)), we need an interesting new condition concerning H^1 continuity of second shape derivative of $e(\Omega)$ (hypothesis (1.10)).

Usually, the necessary condition for the existence of Ω_σ is obtained by differentiating the functional $E(\Omega)$ with respect to the shape (we assume that the reader is familiar with shape derivatives; see, for example, [12], [13]); if not, most information is given here: for given $\Omega_0 \in \mathcal{O}$, let $\Gamma_0 = \partial\Omega_0$, $\Theta := C^{2,\alpha}(\Gamma_0; R^N)$, and $\theta \in \Theta$ with $\|\theta\|_\Theta$ small enough. Let us consider the domain $\Omega_\theta = \{x + \overset{*}{\theta}(x),\ x \in \Omega_0\} \in \mathcal{O}$, where $\overset{*}{\cdot} : \Theta \mapsto C^{2,\alpha}(\overline{\Omega}_0; R^N)$ is a $C^{2,\alpha}$ linear extension operator; see, for example, [4]. Then the shape derivative of e at Ω_θ is defined as the usual derivative at θ of the function $\theta \mapsto e(\Omega_\theta)$, considered as a function from Θ into R. Its value at θ evaluated at $\xi \in \Theta$ is denoted by $\partial_\theta e(\Omega_\theta)(\xi)$.

We make this general assumption for the functional $e(\Omega)$:

$$\forall\theta \in \Theta \ \text{ small}, \exists\ b(\theta) \in C^{0,\alpha}(\Gamma_\theta) \mid \forall\xi \in \Theta,\ \partial_\theta e(\theta)(\xi) = \int_{\partial\Omega_\theta} b(\theta)(\check{\xi} \cdot \nu_\theta), \tag{1.2}$$

where ν_θ is the outward normal unit vector to $\partial\Omega_\theta$, $\check{\xi} = \xi \circ (I + \theta)^{-1}$, and $\cdot$ is the inner product in R^N.

We will look for Ω_σ of the form Ω_θ, $\theta \in \Theta$. Knowing that (see, for example, [5], [9], [13])

$$\partial_\theta P(\Omega_\theta)(\xi) = \int_{\partial\Omega_\theta} \mathcal{H}_\theta(\check{\xi} \cdot \nu_\theta), \quad \partial_\theta m(\Omega_\theta)(\xi) = \int_{\partial\Omega_\theta} (\check{\xi} \cdot \nu_\theta),$$

where $\mathcal{H}_\theta$ is the mean curvature of $\partial\Omega_\theta$ (defined as the sum of principal curvatures), the necessary condition for the optimal domain Ω_σ is the existence of a θ_σ such that

$$\begin{cases} b(\theta_\sigma) + \frac{1}{\sigma^2}\mathcal{H}_{\theta_\sigma} + \Lambda_\sigma = 0 \quad \textit{on } \partial\Omega_\sigma, \\ m(\Omega_{\theta_\sigma}) = m_0, \end{cases} \tag{1.3}$$

where $\Lambda_\sigma \in R$ is a Lagrange multiplier due to the constraint $m(\Omega) = m_0$.

Let us explain first how we find a necessary condition for the existence of solutions of (1.3) in terms of $e(\Omega)$. From (1.1) it follows that $\mathbf{x} \in R^N \mapsto e(\Omega_{\sigma,\mathbf{x}})$, where $\Omega_{\sigma,\mathbf{x}} = (I + \mathbf{x})(\Omega_\sigma)$, has a minimum at $\mathbf{x} = 0$. Formally, it is expected that when σ tends to 0, Ω_σ should tend to a ball $B_{\mathbf{x}_0}$ of measure m_0. This allows thinking that the map $\mathbf{x} \to e(B_\mathbf{x})$, where $B_\mathbf{x} = \{\mathbf{y} \in R^N, |\mathbf{y} - \mathbf{x}| < r_0\}$, $m(B_\mathbf{x}) = m_0$, should also have a minimum at $\mathbf{x}_0$. To prove this let us make the following general assumptions:

(i) for any $\sigma^2 << 1$ there exists a Ω_σ solution of (1.1) in $\{B_\mathbf{x},\ \mathbf{x} \in R^N\}$;
(ii) $\Omega_\sigma \subset B_R := \{\mathbf{x} \in R^N, |\mathbf{x}| < R\}$ for some $R > 0$;
(iii) $\Omega \mapsto e(\Omega)$ is semicontinuous for $BV(B_R)$ weak topology; and
(iv) $\inf\{e(\Omega),\ \Omega \subset B_R,\ \Omega \in \mathcal{O}\} > -\infty$.

Then we may prove the following.

LEMMA 1.1. *Under assumptions* (i)–(iv) *there exists* $\mathbf{x}_0 \in R^N$ *such that* $\mathbf{x} \in R^N \to e(B_\mathbf{x}) \in R$ *has a minimum at* $\mathbf{x}_0$.

Proof. As $m(\Omega_\sigma) = m_0$, from isoperimetric inequality it follows that $P(\Omega_\sigma) = P(B_\mathbf{0}) + \epsilon_\sigma$, where $\epsilon_\sigma \geq 0$ and $B_\mathbf{0}$ is the ball with its center at origin and measure m_0. From (i) we have $E(\Omega_\sigma) \leq E(B_\mathbf{x})$. It follows that

$$e(\Omega_\sigma) + \frac{\epsilon_\sigma}{\sigma^2} \leq e(B_\mathbf{x}) \ \ \forall\mathbf{x} \in R^N. \tag{1.4}$$

From (iv) one obtains $\lim_{\sigma\to 0}\epsilon_\sigma = 0$. Thus $\{\Omega_\sigma, \sigma^2 << 1\}$ is bounded in $BV(B_R)$. Then there exists a subsequence of $\{\Omega_\sigma, \sigma^2 << 1\}$ denoted again by $\{\Omega_\sigma, \sigma^2 << 1\}$ such that $\lim_{\sigma\to 0}\Omega_\sigma = \Omega_0$ in $BV(B_R)$ weakly. As Ω_σ converges in L^1 we have $m(\Omega_0) = m_0$. From semicontinuity of the $BV(B_R)$ norm and $\lim_{\sigma\to 0}\epsilon_\sigma = 0$ it follows that $P(\Omega_0) \le P(B_{\mathbf{0}})$. Then from isoperimetric inequality (for Ω_0 it becomes equality) it follows that $\Omega_0 = B_{\mathbf{x}_0} \subset B_R$ for any $\mathbf{x}_0 \in R^N$. On the other hand from (iii) we have $e(B_{\mathbf{x}_0}) \le \liminf_{\sigma\to 0} e(\Omega_\sigma)$. Taking $\mathbf{x} = \mathbf{x}_0$ in (1.4) and letting σ tend to 0 gives $\lim_{\sigma\to 0}\frac{\epsilon_\sigma}{\sigma^2} = 0$. Letting once again σ tend to zero in (1.4) yields $e(B_{\mathbf{x}_0}) \le e(B_{\mathbf{x}})$ for all $\mathbf{x} \in R^N$. This proves the lemma. □

COROLLARY 1.2. *Assuming* (i)–(iv) *and* C^2 *regularity for* $e(B_{\mathbf{x}})$ *we have* $\nabla e(B_{\mathbf{x}})|_{\mathbf{x}_0} = 0$, *where* $\nabla = (\partial_1, \dots, \partial_N)$, ∂_i *denotes the* x_i *partial derivative. Moreover, if* $e(B_{\mathbf{x}})$ *has a minimum at* $\mathbf{x}_0$, *we have also* $\nabla^2 e(B_{\mathbf{x}})|_{\mathbf{x}_0} \ge 0$, *where* $\nabla^2 = [\partial_i\partial_j]$. *Lemma* 1.1 *may be useful to find counterexamples for the existence of optimal domains of* (1.1). *For example, if* $\nabla^2 e(B_{\mathbf{x}})|_{\mathbf{x}_0}$ *has any negative eigenvalue, then* (1.1) *cannot have solutions for* $\sigma^2 << 1$ *satisfying* (i)–(iv).

Let ν_0 denote the outward unitary normal vector to Γ_0. We will make these assumptions for the functional $e(\Omega)$:

$$\theta \mapsto b(\theta)\circ(I+\theta) \text{ is a } C^1 \text{ function near 0 from } \Theta \text{ into } C^{0,\alpha}(\Gamma_0), \tag{1.5}$$

$$\nabla e(B_{\mathbf{x}})|_{\mathbf{x}_0} = (0,\dots,0), \quad \nabla^2 e(B_{\mathbf{x}})|_{\mathbf{x}_0}, \text{ is positive.} \tag{1.6}$$

While (1.5) is a quite general assumption often satisfied, (1.6) has to be checked for any particular $e(\Omega)$. In section 4 we give an example for which we show that the hypothesis (1.6) is satisfied under some conditions that may be checked.

We think of the solutions Ω_σ of (1.1) and (1.3) as perturbations of the domain $\Omega_0 = B_{\mathbf{x}_0}$. While the perimeter $P(\Omega_\theta)$ is invariant by translation, the term $e(\Omega_\theta)$ is not. But $e(\Omega_\theta)$ has a minimum at $\theta = 0$ in the set $\{\theta \in R^N\}$ (from hypothesis (1.6)). This suggests for us to consider domain transformations having an explicit translation like

$$T : \begin{cases} R\times C^{2,\alpha}(\Gamma_0) & \longrightarrow \quad C^{2,\alpha}(\Gamma_0; R^N), \\ (\sigma,\varphi) & \longmapsto \quad I + u(\sigma,\varphi) := I + \mathbf{t}(\varphi) + \sigma\varphi\nu_0, \end{cases} \tag{1.7}$$

where $\mathbf{t}(\varphi) = \int_{\Gamma_0}\varphi\nu_0\, d\Gamma_0$.

It is clear that for (σ,φ) small in $R\times C^{2,\alpha}(\Gamma_0)$, $\Omega_{u(\sigma,\varphi)}$ is a $C^{2,\alpha}$ open set, denoted by $\Omega(\sigma,\varphi)$. Let $\Gamma(\sigma,\varphi) = \partial\Omega(\sigma,\varphi)$ be its boundary and $d\Gamma(\sigma,\varphi)$ its element area. Moreover, let $\nu(\sigma,\varphi)$ denote the outward unitary normal vector to $\Gamma(\sigma,\varphi)$, let $\mathcal{H}(\sigma,\varphi)$ denote the curvature of $\Gamma(\sigma,\varphi)$, and let $\mathcal{H}_0$ denote the curvature of Γ_0.

Let us also set $\mathbf{E}(\sigma,\varphi) = E(\Omega(\sigma,\varphi))$, $\mathbf{e}(\sigma,\varphi) = e(\Omega(\sigma,\varphi))$, $\mathbf{b}(\sigma,\varphi) = b(\Omega(\sigma,\varphi))$, $\mathbf{P}(\sigma,\varphi) = P(\Omega(\sigma,\varphi))$, and $\mathbf{m}(\sigma,\varphi) = m(\Omega(\sigma,\varphi))$.

REMARK 1.3. *A main part of the paper deals with the solution of* (1.3). *The difficulty in solving it comes from the singularity in* σ^{-2} *of* (1.3). *Let us first point out that from* (1.3) *it is expected that the leading term of* Λ_σ *should be* $-\frac{\mathcal{H}_0}{\sigma^2}$ *(this follows from the regularity of* $b(\theta)$ *and* $\mathcal{H}_\theta$*); thus* $\Lambda_\sigma = \frac{\Lambda}{\sigma} - \frac{\mathcal{H}_0}{\sigma^2}$ *for any* $\Lambda \in R$. *Then one may prove (we will see throughout this paper) that the first and second* φ *derivatives of* $\mathbf{E}(\sigma,\varphi) + \left(\frac{\Lambda}{\sigma} - \frac{\mathcal{H}_0}{\sigma^2}\right)\mathbf{m}(\sigma,\varphi)$ *are no longer singular. This is the first main advantage of having* σ *in* (1.7) *(see also Remark* 2.8).

In the case when (1.1) has a solution in the set $\{\Omega(\sigma,\varphi),\ \varphi \in C^{2,\alpha}(\Gamma_0)\}$, a priori, this is not a solution of (1.1). A key point is to show that for $(\sigma,\varphi) \in R\times C^{2,\alpha}(\Gamma_0)$, if $\mathbf{t}(\varphi) + \sigma\varphi\nu_0$ is small in Θ, then for every $\xi \in \Theta$ small, there exists a small $\phi \in$

$C^{2,\alpha}(\Gamma_0)$ such that $\Omega_{u(\sigma,\varphi)+\xi} = \Omega(\sigma, \varphi + \phi)$ (Proposition 2.6). Thus if $\Omega(\sigma, \varphi_\sigma)$ for any $(\sigma, \varphi_\sigma) \in R \times C^{2,\alpha}(\Gamma_0)$ is the solution of (1.1) in the set $\{\Omega(\sigma, \varphi), \varphi \in C^{2,\alpha}(\Gamma_0)\}$, then $\Omega(\sigma, \varphi_\sigma)$ is the solution of (1.1) too.

The main results we prove are the following two theorems.

THEOREM 1.4. *Under the hypotheses* (1.2), (1.5), *and* (1.6) *there exist* $\sigma^* > 0$ *and a unique* C^1 *map* $\sigma \in (-\sigma^*, \sigma^*) \longmapsto (\varphi_\sigma, \Lambda_\sigma) \in C^{2,\alpha}(\Gamma_0) \times R$ *such that for all* $\sigma \in (-\sigma^*, \sigma^*)\backslash\{0\}$ *we have*

$$\mathbf{b}(\sigma, \varphi_\sigma) + \frac{1}{\sigma^2}\mathcal{H}(\sigma, \varphi_\sigma) + \left(\frac{\Lambda_\sigma}{\sigma} - \frac{\mathcal{H}_0}{\sigma^2}\right) = 0 \quad on \quad \Gamma(\sigma, \varphi_\sigma), \tag{1.8}$$

$$\mathbf{m}(\sigma, \varphi_\sigma) - m_0 = 0. \tag{1.9}$$

Thus $\left(\Omega(\sigma, \varphi_\sigma), \frac{\Lambda_\sigma}{\sigma} - \frac{\mathcal{H}_0}{\sigma^2}\right)$ *is the solution of* (1.3).

REMARK 1.5.

(i) *We emphasize that* (1.8) *is equivalent to* $\partial_\varphi \mathbf{L}(\sigma, \varphi_\sigma, \Lambda_\sigma) = 0$ *(Remark* 3.1, *Lemma* 3.2) *where*

$$\mathbf{L}(\sigma, \varphi, \Lambda) = \mathbf{e}(\sigma, \varphi) + \frac{1}{\sigma^2}\mathbf{P}(\sigma, \varphi) + \left(\frac{\Lambda}{\sigma} - \frac{\mathcal{H}_0}{\sigma^2}\right)\mathbf{m}(\sigma, \varphi).$$

The form of $\mathbf{L}(\sigma, \varphi, \Lambda)$ *is* $\mathbf{E}(\sigma, \varphi)$ *plus the last term, which is related to the constraint that the measure is fixed. The coefficient* $\frac{\Lambda}{\sigma} - \frac{\mathcal{H}_0}{\sigma^2}$ *is the Lagrange multiplier in* (1.3). *Its coefficient near* σ^{-2} *is such that* $\partial_\varphi \mathbf{L}$ *is a* C^1 *function (see the discussion before Theorem* 1.4 *and Lemma* 3.3).

(ii) *Rather than the positiveness of* $\nabla^2 e(B_\mathbf{x})|_{\mathbf{x}_0}$, *what is essential for the existence of critical points is* $\det(\nabla^2 e(B_\mathbf{x})|_{\mathbf{x}_0}) \neq 0$. *Indeed, to prove Theorem* 1.4 *it is required that only* $\nabla^2 e(B_\mathbf{x})|_{\mathbf{x}_0}$ *have nonzero eigenvalues.*

THEOREM 1.6. *Let us assume that the hypotheses* (1.2), (1.5), *and* (1.6) *are satisfied and the maps*

$$\theta \mapsto \partial_\theta^2 e(\Omega_\theta), \quad \theta \mapsto \partial_\theta^2 P(\Omega_\theta), \quad \theta \mapsto \partial_\theta^2 m(\Omega_\theta)$$
$$\text{are continuous from } \Theta \text{ to } H^1(\Gamma_0; R^N) \times H^1(\Gamma_0; R^N). \tag{1.10}$$

Then there exist $\sigma^* > 0$ *and a unique* C^1 *map* $\sigma \in (-\sigma^*, \sigma^*) \longmapsto (\varphi_\sigma, \Lambda_\sigma) \in C^{2,\alpha}(\Gamma_0) \times R$ *such that for all* $\sigma \in (-\sigma^*, \sigma^*)\backslash\{0\}$, $\Omega(\sigma, \varphi_\sigma)$ *is the solution of* (1.1).

REMARK 1.7. *Although the condition* (1.10) *seems a "nonnatural hypothesis," because we have chosen* $C^{2,\alpha}$ *as the framework and* (1.10) *involves* H^1 *norms, it is quite reasonable in a shape optimization context. Indeed, in* [10] *is given the precise structure of shape derivatives. In particular, there is shown that the second shape derivative applied to* (ξ, η) *depends only on the* $C^1(\Gamma_0)$ *norm of* ξ *and* η, *which makes it quite reasonable to expect a control of* $\partial_\theta^2 e(\Omega_\theta)(\xi, \eta)$ *through the* $H^1(\Gamma_0)$ *norm of* ξ *and* η. *The perimeter* $P(\Omega_\theta)$ *and the measure* $m(\Omega_\theta)$ *satisfy the condition* (1.10), *thus the condition* (1.10) *directly concerns* $e(\Omega_\theta)$.

In section 4 *we will give an example satisfying all the hypotheses of Theorem* 1.6. *A condition similar to* (1.10) *for a class of energy functionals associated with a general second order elliptic operator is proven in* [2].

2. Some shape derivatives results. In this section we deal with some results that will be very useful throughout this paper. We start with the derivative of $\mathcal{H}(\sigma, \varphi) \circ T(\sigma, \varphi)$ and of $\frac{1}{\sigma}(\mathcal{H}(\sigma, \varphi) \circ T(\sigma, \varphi) - \mathcal{H}_0)$. In order to make full use of a particular structure of the transformation (1.7) we will make the computations by hand. Later on we show some results related to the domain transformations.

Throughout this paper we set $\mathbf{x}_0 = \mathbf{0}$, $\Omega_0 = B_{r_0}(\mathbf{x}_0)$, and $\Gamma_0 = \partial\Omega_0$. For $\sigma \neq 0$ given and $\varphi \in C^{2,\alpha}(\Gamma_0)$, let $\mathbf{y} : R^N \mapsto R$ be defined by

$$\mathbf{y}(y) = |y - \mathbf{t}(\varphi)| - \sigma\varphi\left(r_0 \frac{y - \mathbf{t}(\varphi)}{|y - \mathbf{t}(\varphi)|}\right) - r_0.$$

It is clear that for $y \in \Gamma(\sigma, \varphi)$, $\mathbf{y}(y) = 0$ and $\nu(\sigma, \varphi)(y) = \frac{\nabla_y \mathbf{y}(y)}{|\nabla_y \mathbf{y}(y)|}$. Thus we obtain

$$\nu(\sigma,\varphi)(y) = \frac{z - \sigma r_0 \overline{\nabla}\varphi(x)}{|z - \sigma r_0 \overline{\nabla}\varphi(x)|}, \quad \text{where} \quad z = y - \mathbf{t}(\varphi), \quad x = r_0 \frac{z}{|z|}, \quad x \in \Gamma_0, \tag{2.1}$$
$$\overline{\nabla} = (\overline{\partial}_1, \ldots, \overline{\partial}_N), \quad \overline{\partial}_i\varphi = \partial_i\varphi - (\nabla\varphi \cdot \nu_0)\nu_0^i, \quad \text{(the tangential gradient)}.$$

It is clear also that for $\sigma\varphi$ small, $\nu(\sigma, \varphi)$ is a $C^{1,\alpha}(\Gamma_0)$ well-defined function. Indeed, we have $|z - \sigma r_0 \overline{\nabla}\varphi(x)| = [(r_0 + \sigma\varphi)^2 + r_0^2\sigma^2|\overline{\nabla}\varphi|^2]^{1/2}$, which shows that for $\sigma\varphi$ small, for example, $\|\sigma\varphi\|_{C^{2,\alpha}(\Gamma_0)} \leq \frac{r_0}{2}$, the denominator in the expression of $\nu(\sigma, \varphi)(y)$ is greater than $\frac{r_0}{2}$ and of course implies $C^{1,\alpha}$ regularity of $\nu(\sigma, \varphi)$.

LEMMA 2.1. *The curvature of* $\Gamma(\sigma, \varphi)$ *is given by*

$$\mathcal{H}(\sigma,\varphi) \circ T(\sigma,\varphi) = \mathcal{H}_0 \frac{r_0}{|(r_0 + \sigma\varphi)\nu_0 - \sigma r_0 \overline{\nabla}\varphi|} - \frac{r_0}{|r_0 + \sigma\varphi|}\left[\frac{\sigma r_0 \ \underline{\Delta}\varphi}{|(r_0 + \sigma\varphi)\nu_0 - \sigma r_0 \overline{\nabla}\varphi|}\right.$$
$$\left. - \frac{\sigma^3 r_0^3 \ \overline{\nabla}\varphi \cdot \overline{\nabla}^2\varphi \cdot \overline{\nabla}\varphi}{|(r_0 + \sigma\varphi)\nu_0 - \sigma r_0 \overline{\nabla}\varphi|^3}\right], \tag{2.2}$$

where $\overline{\nabla}^2\varphi$ *is the matrix* $[\overline{\partial}_i\overline{\partial}_j\varphi]$ *and* $\underline{\Delta}\varphi = tr[\overline{\nabla}^2\varphi]$ *is the Laplace–Beltrami operator on* Γ_0.

Proof. Because $\mathcal{H}(\sigma, \varphi) = \sum_{i=1}^N \overline{\partial}_i \nu^i(\sigma, \varphi)$, where

$$\overline{\partial}_i \nu^i(\sigma,\varphi) = \partial_{y_i} \overset{*}{\nu}{}^i(\sigma,\varphi) - (\nu(\sigma,\varphi) \cdot \nabla_y \overset{*}{\nu}(\sigma,\varphi))\nu^i(\sigma,\varphi),$$

we have to compute the terms $\overline{\partial}_{y_i}\nu^i(\sigma, \varphi)$. It is easy to verify the relations

$$\overline{\partial}_{y_i} z_i = \delta_{ij} - \nu^i(\sigma,\varphi)\nu^j(\sigma,\varphi), \qquad \overline{\partial}_{y_i}\phi(x) = \frac{r_0}{|z|}\overline{\partial}_i\phi(x), \qquad \phi \in C^1(\Gamma_0),$$

where δ_{ij} is the Kronecker's symbol. For simplicity, let's set $\phi = \sigma r_0 \varphi$. Then

$$\begin{aligned}
\mathcal{H}(\sigma,\varphi) \circ T(\sigma,\varphi) &= \overline{\partial}_{y_i}\nu^i(\sigma,\varphi) \\
&= \frac{\overline{\partial}_{y_i} z_i - \overline{\partial}_{y_i}\overline{\partial}_i\phi(x)}{|z - \overline{\nabla}\phi(x)} \\
&\quad - \frac{(z_i - \overline{\partial}_i\phi(x))(z_j - \overline{\partial}_j\phi(x)}{|z - \overline{\nabla}\phi(x)|^3}(\delta_{ij} - \nu^i(\sigma,\varphi)\nu^j(\sigma,\varphi) - \frac{r_0}{|z|}\overline{\partial}_i\overline{\partial}_j\phi(x)) \\
&= \frac{N-1}{|z - \overline{\nabla}\phi(x)|} - \frac{r_0}{|z|}\frac{\underline{\Delta}\phi(x)}{|z - \overline{\nabla}\phi(x)|} + \frac{r_0}{|z|}\frac{\overline{\nabla}\phi(x) \cdot \overline{\nabla}^2\phi(x) \cdot \overline{\nabla}\phi(x)}{|z - \overline{\nabla}\phi(x)|^3},
\end{aligned}$$

because $x \cdot \overline{\nabla}^2\phi(x) = 0$. From $\mathcal{H}_0 = \frac{N-1}{r_0}$, $|z| = |r_0 + \sigma\varphi|$, and $|z - \overline{\nabla}\varphi| = |(r_0 + \sigma\varphi)\nu_0 - \sigma r_0 \overline{\nabla}\varphi|$ we obtain immediately the expression for $\mathcal{H}(\sigma, \varphi) \circ T(\sigma, \varphi)$. ∎

As a simple corollary of this lemma we have the following.

COROLLARY 2.2. *The map* $(\sigma,\varphi)\mapsto \mathcal{H}(\sigma,\varphi)\circ T(\sigma,\varphi)$ *from* $R\times C^{2,\alpha}(\Gamma_0)$ *into* $C^{0,\alpha}(\Gamma_0)$ *is of class* C^1 *near* $(0,0)$.

Proof. First, we point out that $\mathcal{H}(\sigma,\varphi)\circ T(\sigma,\varphi)\in C^{0,\alpha}(\Gamma_0)$. This is a simple consequence of the fact that $\mathcal{H}(\sigma,\varphi)\circ T(\sigma,\varphi)$ is a simple algebraic expression depending on $C^{0,\alpha}(\Gamma_0)$ functions. We emphasize that its denominator is strictly greater than, say, $r_0/2$ for $\sigma\varphi$ small. Then its differentiability with respect to (σ,φ) is clear for $\sigma\varphi$ small. □

COROLLARY 2.3. *The map* $(\sigma,\varphi)\mapsto \mathcal{K}(\sigma,\varphi):=\frac{\mathcal{H}(\sigma,\varphi)\circ T(\sigma,\varphi)-\mathcal{H}_0}{\sigma}$ *from* $R\times C^{2,\alpha}(\Gamma_0)$ *into* $C^{0,\alpha}(\Gamma_0)$ *is of class* C^1 *near* $(0,0)$. *Moreover, its derivative with respect to* φ *at* $(0,0)$ *is given by*

$$\partial_\varphi \mathcal{K}(0,0)(\phi)=-\underline{\Delta}\phi-\frac{\mathcal{H}_0^2}{N-1}\phi. \tag{2.3}$$

Proof. By direct computations and using (2.2) one may easily find that

$$\begin{aligned}\mathcal{K}(\sigma,\varphi)=&-\mathcal{H}_0\frac{2r_0\varphi+\sigma\varphi^2+\sigma r_0^2|\overline{\nabla}\varphi|^2}{|(r_0+\sigma\varphi)\nu_0-\sigma r_0\overline{\nabla}\varphi|(r_0+|(r_0+\sigma\varphi)\nu_0-\sigma r_0\overline{\nabla}\varphi|)}\\ &-\frac{r_0^2}{|r_0+\sigma\varphi|}\left[\frac{\underline{\Delta}\varphi}{|(r_0+\sigma\varphi)\nu_0-\sigma r_0\overline{\nabla}\varphi|}-\frac{\sigma^2 r_0^2\ \overline{\nabla}\varphi\cdot\overline{\nabla}^2\varphi\cdot\overline{\nabla}\varphi}{|(r_0+\sigma\varphi)\nu_0-\sigma r_0\overline{\nabla}\varphi|^3},\right],\end{aligned} \tag{2.4}$$

which shows that $\mathcal{K}(\sigma,\varphi)$ is of class C^1 near $(0,0)$, for example, for $\|\sigma\varphi\|_{C^{2,\alpha}(\Gamma)}<r_0/2$. The expression of $\mathcal{K}(\sigma,\varphi)$ also gives immediately the value of $\partial_\varphi\mathcal{K}(0,0)(\varphi)$. □

As we mentioned in the introduction, we will prove first that the problem (1.1) has a solution $\Omega(\sigma,\varphi_\sigma)$ in the set of domains $\{\Omega(\sigma,\varphi),\varphi\in C^{2,\alpha}(\Gamma_0)\}$. Then a key point is to prove that $\Omega(\sigma,\varphi)$ is the solution of (1.1) in Θ. For this it is sufficient to show that for any $\xi\in\Theta$ small, there exists a $\phi\in C^{2,\alpha}(\Gamma_0)$ such that $\Omega_{u(\sigma,\varphi_\sigma)+\xi}=\Omega(\sigma,\varphi_\sigma+\phi)$. Let us prove first the following important lemma.

LEMMA 2.4. *Let* $\sigma\neq 0$, $\sigma\le\frac{P(\Omega_0)}{2N}$. *There exist* Θ_0, *a small neighborhood of* 0 *in* Θ *not depending on* σ, *and two* C^1 *functions*

$$\theta\in\Theta_0\mapsto\varphi=\boldsymbol{\varphi}(\theta)\in C^{2,\alpha}(\Gamma_0)\quad\text{and}\quad\theta\in\Theta_0\mapsto g=\mathbf{g}(\theta)\in C^{2,\alpha}(\Gamma_0;\Gamma_0),$$

with g *invertible, such that*

$$(I+\theta)\circ g=I+\mathbf{t}(\varphi)+\sigma\varphi\nu_0\quad\text{on }\Gamma_0. \tag{2.5}$$

Moreover, there exist $C>0$ *not depending on* σ *such that for* $\theta\in\Theta_0$, *we have*

$$\|\mathbf{g}(\theta)\|_\Theta\le C, \tag{2.6}$$

$$\|\mathbf{g}^{-1}(\theta)-I\|_\Theta\le C\|\theta\|_\Theta. \tag{2.7}$$

Proof. Let $t\in R^N$, and for $y\in\Gamma_0$ set $x=r_0\frac{(I+\theta)(y)-t}{|(I+\theta)(y)-t|}\in\Gamma_0$ and $g(\theta,t):\Gamma_0\mapsto\Gamma_0$, $g(\theta,t)(x)=y$. For $(\theta,t)\in\Theta_1\times\mathcal{U}_1$, a small neighborhood of $(0,0)$ in $\Theta\times R^N$, the function $g(\theta,t)$ is a $C^{2,\alpha}(\Gamma_0;\Gamma_0)$ well-defined invertible function. Indeed, for $(\theta,t)=(0,0)$ we have $g(0,0)(x)=x$. Then the inversibility of $g(\theta,t)$ for (θ,t) small follows from C^1 regularity of $(\theta,t)\mapsto g(\theta,t)$.

Now let us introduce the function $\varphi(\theta,t):\Gamma_0\mapsto R$ by

$$\sigma\varphi(\theta,t)(x)=|(I+\theta)\circ g(\theta,t)(x)-t|-r_0. \tag{2.8}$$

We look for t such that $t = \mathbf{t}(\varphi(\theta, t))$. For this, let us consider the equation

$$\sigma t = \int_{\Gamma_0} (|(I+\theta)\circ g(\theta,t)(x) - t| - r_0)\nu_0(x)dx. \tag{2.9}$$

The function

$$\begin{aligned} f : (\theta, t) \in \Theta \times R^N \mapsto \sigma t - \int_{\Gamma_0} (|(I+\theta)\circ g(\theta,t)(x) - t| - r_0)\nu_0(x)dx \in R^N \\ = \sigma t - \int_{\Gamma_0} (I+\theta)\circ g(\theta,t)(x) - t - xdx \end{aligned}$$

is of class C^1, $f(0,0) = 0$, and for $s \in R^N$, $\partial_t f(0,0)(s) = (\sigma + \frac{P(\Omega_0)}{N})s$.

Let us prove the expression of $\partial_t f(0,0)(s)$. For $\theta = 0$, we have $\frac{g(0,t)-t}{|g(0,t)-t|} = \nu_0$. This gives $(|g(0,t) - t|\nu_0 + t)^2 = r_0^2$ because $|g(\theta,t)| = r_0$. Then we may find that

$$|g(0,t) - t| = [(\nu_0 \cdot t)^2 - t^2 + r_0^2]^{1/2} - (\nu_0 \cdot t).$$

It follows that

$$g(0,t) = t + [(\nu_0 \cdot t)^2 - t^2 + r_0^2]^{1/2}\nu_0 - (\nu_0 \cdot t)\nu_0, \quad \partial_t g(0,0)(s) = s - (\nu_0 \cdot s)\nu_0,$$

and we obtain

$$\begin{aligned} \partial_t f(0,0)(s) &= \sigma s - \int_{\Gamma_0} \partial_t g(0,0)(s)(x) - s dx \\ &= \sigma s + \int_{\Gamma_0} (\nu_0(x)\cdot s)\nu_0(x)dx = \left(\sigma + \frac{P(\Omega_0)}{N}\right)s. \end{aligned}$$

From the implicit function theorem it follows that we have the existence of Θ_2, a neighborhood of 0 in Θ, and a map $\theta \in \Theta_2 \mapsto t_\theta \in R^N$ such that $t_\theta|_{\theta=0} = \mathbf{0}$ and t_θ satisfies (2.9). Taking σ small, say, $|\sigma| < \frac{P(\Omega_0)}{2N}$, allows us to choose Θ_2 independent of σ.

Now, putting $t = t_\theta$ in (2.8), multiplying by $\nu_0(x)$, and integrating on Γ_0, from (2.9) it follows that $\mathbf{t}(\varphi(\theta, t_\theta)) = t_\theta$. Finally, let $\Theta_0 = \Theta_1 \cap \Theta_2$. For $\theta \in \Theta_0$ let's set $\mathbf{g}(\theta) = g(\theta, t_\theta)$, $\boldsymbol{\varphi}(\theta) = \varphi(\theta, t_\theta)$. These functions satisfy the lemma. Indeed, for θ and $t = t_\theta$, from the construction of $g(\theta,t)$, $\varphi(\theta,t)$, and (2.8) we have

$$\begin{aligned} (I+\theta)\circ g(\theta,t_\theta) &= |(I+\theta)\circ g(\theta,t_\theta) - t_\theta|\nu_0 + t_\theta = (r_0 + \sigma\varphi(\theta,t_\theta))\nu_0 + t_\theta \\ &= I + \mathbf{t}(\varphi(\theta,t_\theta)) + \sigma\varphi(\theta,t_\theta)\nu_0. \end{aligned}$$

The C^1 regularity of the functions $\boldsymbol{\varphi}$ and $\mathbf{g}$ follows immediately from the formula giving $g(\theta,t_\theta)$, $\varphi(\theta,t_\theta)$ and from the C^1 regularity of $\theta \mapsto t_\theta$.

The estimation (2.6) follows from the fact that $\mathbf{g}^{-1}(\theta) = r_0\frac{I+\theta-t_\theta}{|I+\theta-t_\theta|}$ and that $\theta \mapsto t_\theta$ is a uniformly C^1 function (with respect to σ) from Θ to itself.

For (2.7) we write

$$\begin{aligned} \mathbf{g}^{-1}(\theta) - I &= \frac{1}{|I+\theta-t_\theta|}\left[r_0(\theta - t_\theta) + \frac{r_0^2 - |r_0\nu_0 + \theta - t_\theta|^2}{r_0 + |r_0\nu_0 + \theta - t_\theta|}I\right] \\ &= \frac{1}{|I+\theta-t_\theta|}\left[r_0(\theta - t_\theta) - \frac{\theta^2 + t_\theta^2 + 2r_0(\nu_0\cdot\theta) - 2r_0(\nu_0\cdot t_\theta) - 2(\theta\cdot t_\theta)}{r_0 + |r_0\nu_0 + \theta - t_\theta|}I\right]. \end{aligned}$$

This equality, together with the facts $t_\theta|_{\theta=0} = 0$ and C^1 continuity of $\theta \to t_\theta$, gives $|t_\theta| \le C\|\theta\|_\Theta$. It follows that $\|\mathbf{g}^{-1}(\theta) - I\|_\Theta \le C\|\theta\|_\Theta$, which proves (2.7). □

COROLLARY 2.5. *For any $\varphi_\sigma \in C^{2,\alpha}(\Gamma_0)$, let's set $\theta = u(\sigma, \varphi_\sigma)$. Then $\boldsymbol{\varphi}(\theta) = \varphi_\sigma$.*

Proof. First we show that $t_\theta = \mathbf{t}(\varphi_\sigma)$. Indeed, as we have $g(\theta, \mathbf{t}(\varphi_\sigma)) = I$, it follows that $\mathbf{t}(\varphi_\sigma)$ satisfies (2.9). From uniqueness of t satisfying (2.9), we obtain $t_\theta = \mathbf{t}(\varphi_\sigma)$. Then from the formula defining $\varphi(\theta, t)$, by direct verification, we obtain $\boldsymbol{\varphi}(\theta) = \varphi(\theta, t_\theta) = \varphi_\sigma$. □

PROPOSITION 2.6. *Let's assume that for any σ small, there exists $\varphi_\sigma \in C^{2,\alpha}(\Gamma_0)$ such that $\Omega(\sigma, \varphi_\sigma)$ is the solution of* (1.1) *in $\{\Omega(\sigma, \varphi), \varphi \in C^{2,\alpha}(\Gamma_0)\}$. If $u(\sigma, \varphi_\sigma) \in \Theta_0$, where Θ_0 is given by Lemma* 2.4, *then $\Omega(\sigma, \varphi_\sigma)$ is the solution of* (1.1) *in Θ.*

Proof. As σ is small, it follows that $u(\sigma, \varphi_\sigma) \in \Theta_0$. Then there exists a ball $B(0, \epsilon) = \{\|\theta\|_\Theta < \epsilon\}$ such that $u(\sigma, \varphi_\sigma) + B(0, \epsilon) \subset \Theta_0$. From Lemma 2.4, $\Omega_{u(\sigma,\varphi_\sigma)+\xi}$ $= \Omega(\sigma, \boldsymbol{\varphi}(u(\sigma, \varphi_\sigma) + \xi))$ for all $\xi \in B(0, \epsilon)$. Then the proposition follows from the C^1 regularity of $\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}(u(\sigma, \varphi_\sigma)) = \varphi_\sigma$ (Corollary 2.5). □

LEMMA 2.7. *For $\theta \in \Theta_0$, let φ and g be as in Lemma* 2.4, *$(I + \theta) \circ g = I + \mathbf{t}(\varphi) + \sigma\varphi\nu_0$. Then there exist two constants C_1, C_2 not depending on σ such that*

$$\|\theta\|_\Theta \le \delta \Longrightarrow \|\mathbf{t}(\varphi) + \sigma\varphi\nu_0\|_\Theta \le C_1\delta, \tag{2.10}$$

$$\|\mathbf{t}(\varphi) + \sigma\varphi\nu_0\|_\Theta \le \delta \Longrightarrow |\mathbf{t}(\varphi)| + \|\sigma\varphi\|_{C^{2,\alpha}(\Gamma_0)} \le C_2\delta. \tag{2.11}$$

Proof. Let us first prove (2.10). From (2.8) we have $\mathbf{t}(\varphi) + \sigma\varphi\nu_0 = (I + \theta) \circ \mathbf{g}(\theta) - I = \theta \circ \mathbf{g}(\theta) - \mathbf{g}(\theta) \circ (\mathbf{g}^{-1}(\theta) - I)$. From (2.6) and (2.7), we get

$$\|\mathbf{t}(\varphi) + \sigma\varphi\nu_0\|_\Theta \le C\|\mathbf{g}(\theta)\|_\Theta\|\theta\|_\Theta + \|\mathbf{g}(\theta)\|_\Theta\|\mathbf{g}^{-1}(\theta) - I\|_\Theta \le C\|\theta\|_\Theta,$$

which proves (2.10).

Now, let's prove (2.11). As each component of ν_0 becomes 0 for any $|\mathbf{x}| = r_0$, it follows that $|\mathbf{t}(\varphi)| \le \delta$. This implies $\|\sigma\varphi\nu_0\|_\Theta \le 2\delta$, and

$$\|\sigma\varphi\|_{C^{2,\alpha}(\Gamma_0)} = \|\sigma\varphi\nu_0 \cdot \nu_0\|_{C^{2,\alpha}(\Gamma_0)} \le \nu_0\|_\Theta\|\sigma\varphi\nu_0\|_\Theta \le 2\|\nu_0\|_\Theta\delta,$$

which proves (2.11) with $C_2 = 1 + 2\|\nu_0\|_\Theta$. □

REMARK 2.8. *Proposition* 2.6 *and Lemma* 2.7 *are very important, in particular for proving Theorem* 1.6. *Indeed, let $\Omega(\sigma, \varphi_\sigma)$ be the solution of* (1.3). *To prove that $\Omega(\sigma, \varphi_\sigma)$ is the solution of* (1.1), *from Lemma* 2.4 *it follows that it is enough to consider perturbations of the form $I + \mathbf{t}(\varphi) + \sigma\varphi\nu_0$. Moreover, from Proposition* 2.6 *and Lemma* 2.7, *it is enough to consider perturbation of $\Omega(\sigma, \varphi_\sigma)$ of the form $I + \mathbf{t}(\varphi) + \sigma\varphi\nu_0$ with φ satisfying $|\mathbf{t}(\varphi)| + \|\sigma\varphi\|_{C^{2,\alpha}(\Gamma_0)} \le \delta$. Thus, φ may be large but $\sigma\varphi$ must be of order δ. Thus, if*

$$\varphi = \sum_{i,j} \varphi_{ij}u_{ij} = \sum_{j=1,n} \varphi_{1j}u_{1j} + \sum_{i\ne1,j} \varphi_{ij}u_{ij} =: \varphi_1 + \varphi_1^c,$$

where u_{ij} are the eigenfunctions of $\underline{\Delta}$, from

$$\|\varphi\|^2_{H^1(\Gamma_0)} = \|\varphi_1\|^2_{H^1(\Gamma_0)} + \|\varphi_1^c\|^2_{H^1(\Gamma_0)},$$

$$\|\varphi_1\|^2_{H^1(\Gamma_0)} = \left(1 + \frac{\mathcal{H}_0^2}{N-1}\right) \sum_{j=1,N} \varphi_{1j}^2 = \frac{N}{|\Gamma_0|}\left(1 + \frac{\mathcal{H}_0^2}{N-1}\right)|\mathbf{t}(\varphi)|^2, \tag{2.12}$$

it follows that

$$\|\varphi_1\|^2_{H^1(\Gamma_0)} + \|\sigma\varphi_1^c\|^2_{H^1(\Gamma_0)} \le C\delta. \tag{2.13}$$

From (2.5) *and* (2.12) *it follows that* φ_1 *is responsible for the domain translation* $\mathbf{t}(\varphi)$. *The "pure" deformation of the domain is basically due to* φ_1^c *because the contribution of* φ_1 *is of order* $\sigma\delta$ *(for the* $H^1(\Gamma_0)$ *norm).*

Writing the domain perturbation as the sum of two terms involving φ_1 *and* φ_1^c *with* φ_1^c *multiplied by* σ *is very important. Indeed, if one considers* $E(\Omega_{\theta+\varphi\nu_0})-E(\Omega_\theta)$ *(instead of* $E(\Omega_{\theta+\mathbf{t}(\varphi)+\sigma\varphi\nu_0})-E(\Omega_\theta)$*), its main part is*

$$\frac{1}{\sigma^2}\int_{\Gamma_0}|\overline{\nabla}\varphi|^2-\frac{\mathcal{H}_0^2}{N-1}\varphi^2+\frac{o(1)}{\sigma^2}\|\phi\|_{H^1(\Gamma_0)}+\mathbf{t}(\varphi)\cdot\nabla^2 e(B_{\mathbf{x}_0})\cdot\mathbf{t}(\varphi).$$

The integral term has no φ_1 *contribution, but it is positive (greater than* $C\|\varphi_1^c\|^2_{H^1(\Gamma_0)}$*). The term with* φ_1 *appears only near to the* $o(1)$ *term and the* $e(\Omega)$ *derivative. In general, it is difficult to control the residue* $o(1)\|\varphi\|^2_{H^1(\Gamma_0)}$*, and here the transformation* $\mathbf{t}(\varphi)+\sigma\varphi\nu_0$ *is of interest, because in this case one would have* $o(1)\|\varphi\|_{H^1(\Gamma_0)}$ *instead of* $\frac{o(1)}{\sigma^2}\|\varphi\|_{H^1(\Gamma_0)}$. *The* $\|\varphi_1\|_{H^1(\Gamma_0)}$ *part of the* $o(1)$ *term may be controlled by the* $e(\Omega)$ *derivative term (and here the hypothesis* (1.6) *intervenes crucially), while the* $\|\varphi_1^c\|_{H^1(\Gamma_0)}$ *part of the* $o(1)$ *term is controlled by the integral term.*

LEMMA 2.9. *Let* $K_0=\{\varphi,\partial_\theta m(\Omega_0)(\varphi\nu_0)=0\}$. *There exists a* C^2 *function* $q: dom(q)\subset K_0\mapsto C^{2,\alpha}(\Gamma_0)$ *where* $dom(q)$ *is an open neighborhood of* 0 *in* K_0 *such that* $m(\Omega_{q(\varphi)\nu_0})=m_0$, $q(\varphi)=\varphi+\beta(\varphi)$ *with* β *a* C^2 *function, and* $\beta(0)=0$. *Moreover,*

$$\partial_\varphi\beta(0)(\phi)=0, \tag{2.14}$$

$$\partial_\theta m(\Omega_{q(\varphi)\nu_0})(\partial^2_\varphi\beta(\varphi)(\phi,\phi))+\partial^2_\theta m(\Omega_{q(\varphi)\nu_0})(\partial_\varphi\beta(\varphi)(\phi),\partial_\varphi\beta(\varphi)(\phi))=0, \tag{2.15}$$

$$\varphi\mapsto\partial_\varphi\beta \quad \textit{is continuous from} \quad C^{2,\alpha}(\Gamma_0) \textit{ to } H^1(\Gamma_0),$$

$$\varphi\mapsto\partial^2_\varphi\beta \quad \textit{is continuous from} \quad C^{2,\alpha}(\Gamma_0) \textit{ to } H^1(\Gamma_0)\times H^1(\Gamma_0).$$

Proof. The existence of β and q as well as (2.14) and (2.15) are classical results and may be proven easily by direct computations and applying the implicit function theorem to $(\varphi,t)\in K_0\times R\mapsto m(\Omega_{(\varphi+t)\nu_0})-m_0\in R$ at $(0,0)$ (although a proof of it is presented in [9]).

Here, we will not present the complete proof. Instead, we will prove the regularity for $\partial_\varphi\beta$ and $\partial^2_\varphi\beta$. Indeed, differentiating one and two times with respect to φ, the equation $m(\Omega_{q(\varphi)\nu_0})-m_0=0$ gives

$$\begin{aligned}
0&=\int_{\partial\Omega_{q(\varphi)\nu_0}}(\phi+\partial_\varphi\beta(\varphi)(\phi))(\nu_0\cdot\nu_{q(\varphi)\nu_0})\\
&=\int_{\Gamma_0}(\phi+\partial_\varphi\beta(\varphi)(\phi))(\nu_0\cdot\nu_{q(\varphi)\nu_0})Jac_\Gamma(I+q(\varphi)\nu_0),\\
0&=\int_{\Gamma_0}\partial^2_\varphi\beta(\varphi)(\phi,\phi)(\nu_0\cdot\nu_{q(\varphi)\nu_0})Jac_\Gamma(I+q(\varphi)\nu_0)\\
&\quad+\int_{\Gamma_0}(\phi+\partial_\varphi\beta(\varphi)(\phi))\partial_\varphi[(\nu_0\cdot\nu_{q(\varphi)\nu_0})Jac_\Gamma(I+q(\varphi)\nu_0)](\phi).
\end{aligned}$$

It follows that

$$\partial_\varphi\beta(\varphi)(\phi)=\frac{\int_{\Gamma_0}\phi(\nu_0\cdot\nu_{q(\varphi)\nu_0})Jac_\Gamma(I+q(\varphi)\nu_0)d\Gamma_0}{\int_{\Gamma_0}(\nu_0\cdot\nu_{q(\varphi)\nu_0})Jac_\Gamma(I+q(\varphi)\nu_0)},$$

$$\partial^2_\varphi\beta(\varphi)(\phi,\phi)=\frac{\int_{\Gamma_0}(\phi+\partial_\varphi\beta(\varphi)(\phi))\partial_\varphi[(\nu_0\cdot\nu_{q(\varphi)\nu_0})Jac_\Gamma(I+q(\varphi)\nu_0)](\phi)}{\int_{\Gamma_0}(\nu_0\cdot\nu_{q(\varphi)\nu_0})Jac_{\Gamma_0}(I+q(\varphi)\nu_0)}.$$

From the C^1 continuity of $\varphi \mapsto \partial_\varphi \beta$ and of the maps $\theta \mapsto \nu_\theta \circ (I+\theta)$ and $\theta \mapsto Jac_\Gamma(I+\theta)$, the regularity claimed for the maps $\varphi \mapsto \partial_\varphi \beta(\varphi)$ and $\varphi \mapsto \partial^2_\varphi \beta(\varphi)$ follows. □

3. Proof of the main results. Let F be the function

$$F : R \times C^{2,\alpha}(\Gamma_0) \times R \longmapsto C^{0,\alpha}(\Gamma_0) \times R,$$
$$(\sigma, \varphi, \Lambda) \longmapsto \big(F_1(\sigma,\varphi,\Lambda), F_2(\sigma,\varphi,\Lambda)\big),$$

where F_1, F_2 are defined by

$$(3.1)\quad F_1(\sigma,\varphi,\Lambda) = \nu_0 \cdot \int_{\Gamma(\sigma,\varphi)} \mathbf{b}(\sigma,\varphi)\nu(\sigma,\varphi)d\Gamma + [\sigma \mathbf{b}(\sigma,\varphi)\circ T(\sigma,\varphi) + \mathcal{K}(\sigma,\varphi) - \Lambda](\nu(\sigma,\varphi)\circ T(\sigma,\varphi)\cdot\nu_0) J_{\Gamma_0} T(\sigma,\varphi),$$

$$(3.2)\quad F_2(\sigma,\varphi,\Lambda) = [\mathbf{m}(\sigma,\varphi) - \mathbf{m}(0,0)]/\sigma.$$

Here $J_{\Gamma_0}T(\sigma,\varphi) = |^t[DT(\sigma,\varphi)]^{-1}\cdot\nu_0| JT(\sigma,\varphi)$ is the tangential Jacobian of $T(\sigma,\varphi)$ and $JT(\sigma,\varphi)$ is the usual Jacobian of $T(\sigma,\varphi)$.

REMARK 3.1. *Note that F_1/σ differs from the left-hand side of* (1.8) *basically by the term $\sigma^{-1}\nu_0 \cdot \int_\Gamma \mathbf{b}(\sigma,\varphi)\nu(\sigma,\varphi)d\Gamma$. F_1 is the $C^{0,\alpha}(\Gamma)$ representation of $\partial_\varphi \mathbf{L}(\sigma,\varphi,\Lambda)$. The function $F_1(\sigma,\varphi,\Lambda)$ is defined explicitly, while $F_2(\sigma,\varphi,\Lambda)$ initially is defined implicitly. In* (3.6) *is given the explicit form of $F_2(\sigma,\varphi,\Lambda)$. It is clear that in particular $F(0,\varphi,\Lambda)$ is well defined. See Remark* 3.4 *for more discussion about $F(0,\varphi,\Lambda)$.*

LEMMA 3.2. *For $(\sigma,\varphi) \in (R\backslash\{0\}) \times C^{2,\alpha}(\Gamma_0)$ small such that $u(\sigma,\varphi) \in \Theta_0$, $F_1(\sigma,\varphi,\Lambda) = 0$ is equivalent to* (1.8).

Proof. It is enough to prove that $\partial_\varphi \mathbf{L}(\sigma,\varphi,\Lambda) = 0$ implies

$$\partial_\theta E(\Omega_{u(\sigma,\varphi)}) + \left(\frac{\Lambda}{\sigma} - \frac{\mathcal{H}_0}{\sigma^2}\right)\partial_\theta m(\Omega_{u(\sigma,\varphi)}) = 0.$$

(See the introduction for the difference between ∂_θ and ∂_φ.)

Let $s \in R$ be small such that $u(\sigma,\varphi) + s\xi \in \Theta_0$. From Lemma 2.4, $\Omega_{u(\sigma,\varphi)+s\xi} = \Omega(\sigma, \boldsymbol{\varphi}(u(\sigma,\varphi)+s\xi))$. But

$$\partial_\theta E(\Omega_{u(\sigma,\varphi)})(\xi) = \partial_s \mathbf{E}(\sigma, \boldsymbol{\varphi}(u(\sigma,\varphi)+s\xi)) = \partial_\varphi \mathbf{E}(\sigma,\varphi)(\partial_\theta \boldsymbol{\varphi}(u(\sigma,\varphi))(\xi)),$$
$$\partial_\theta m(\Omega_{u(\sigma,\varphi)})(\xi) = \partial_s \mathbf{m}(\sigma, \boldsymbol{\varphi}(u(\sigma,\varphi)+s\xi)) = \partial_\varphi \mathbf{m}(\sigma,\varphi)(\partial_\theta \boldsymbol{\varphi}(u(\sigma,\varphi))(\xi)),$$

which leads to

$$\partial_\theta E(\Omega_{u(\sigma,\varphi)})(\xi) + \left(\frac{\Lambda}{\sigma} - \frac{\mathcal{H}_0}{\sigma^2}\right)\partial_\theta m(\Omega_{u(\sigma,\varphi)})(\xi) = \partial_\varphi \mathbf{L}(\sigma,\varphi,\Lambda)(\partial_\theta \boldsymbol{\varphi}(u(\sigma,\varphi))(\xi)) = 0$$

and finishes the proof. □

LEMMA 3.3. *The function $F : R \times C^{2,\alpha}(\Gamma_0) \times R \to C^{0,\alpha}(\Gamma_0) \times R$ is of class C^1 near $(0,0,0)$. Moreover, the derivative of F with respect to (φ,Λ) at $(0,0,0)$ is given by*

$$(3.3)\quad \partial_{\varphi,\Lambda} F_1(0,0,0)(\varphi,\lambda) = -\underline{\Delta}\varphi - \frac{\mathcal{H}_0^2}{N-1}\varphi + \lambda + \nu_0 \cdot \nabla^2 e(B_x)|_{\mathbf{x}_0} \cdot \mathbf{t}(\varphi),$$

$$(3.4)\quad \partial_{\varphi,\Lambda} F_2(0,0,0)(\varphi,\lambda) = \int_{\Gamma_0} \varphi d\Gamma_0.$$

Proof. Taking into account (3.1), the C^1 regularity of F_1 near $(0,0)$ follows from the hypothesis (1.5), the regularity result of Corollary 2.3, and the classical regularity results of the maps $(\sigma,\varphi)\mapsto \nu(\sigma,\varphi)\circ T(\sigma,\varphi)$ and $J_{\Gamma_0}T(\sigma,\varphi)$; see, for example, [12].

Let $F_{11}(\sigma,\varphi)$ be the first term of $F_1(\sigma,\varphi,\Lambda)$. It is easy to verify that

$$\begin{aligned} F_{11}(0,\varphi) &= \sum_{i=1,N} \nu_0^i \int_{\Gamma(0,\varphi)} \mathbf{b}(0,\varphi)\nu^i(0,\varphi) \\ &= \frac{N}{P(\Omega_0)} \sum_{i=1,N} \nu_0^i \int_{\Gamma(0,\varphi)} \mathbf{b}(0,\varphi)(\nu(0,\varphi)\cdot \mathbf{t}(\nu_0^i)) \\ &= \frac{N}{P(\Omega_0)} \sum_{i=1,N} \nu_0^i \partial_\varphi \mathbf{e}(0,\varphi)(\nu_0^i). \end{aligned}$$

As we also have $\mathbf{e}(0,\varphi)=e(B_{\mathbf{x}_0+\mathbf{t}(\varphi)})$, it follows that

$$\begin{aligned} \partial_\varphi F_{11}(0,0)(\varphi) &= \frac{N}{P(\Omega_0)} \sum_{i=1,N} \nu_0^i\ \partial_\varphi^2 \mathbf{e}(0,0)(\nu_0^i,\varphi) \\ &= \frac{N}{P(\Omega_0)} \sum_{i=1,N} \nu_0^i\ \mathbf{t}(\nu_0^i)\cdot \nabla^2 e(B_x)|_{\mathbf{x}_0}\cdot \mathbf{t}(\varphi) = \nu_0\cdot \nabla^2 e(B_x)|_{\mathbf{x}_0}\cdot \mathbf{t}(\varphi). \end{aligned}$$

The expression of $\partial_{\varphi,\Lambda}F_1(0,0,0)$ is completed then by Corollary 2.3.

The C^1 regularity of F_2 is a classical result. However, the proof is simple. Indeed, we have

$$\mathbf{m}(\sigma,\varphi)=\int_{\Omega_0} JT(\sigma,\varphi))d\Omega_0 = \int_{\Omega_0} det[\delta_{ij}+\sigma\partial_j(\overset{*}{\varphi}\overset{*}{\nu}{}_0^i)]d\Omega_0. \tag{3.5}$$

Thus we obtain

$$F_2(\sigma,\varphi,\Lambda)=\int_{\Omega_0}\sum_{i=1,N}\partial_i(\overset{*}{\varphi}\overset{*}{\nu}{}_0^i)d\Omega_0+\sum_{k=2,N}\sigma^{k-1}\int_{\Omega_0}Q_k d\Omega_0, \tag{3.6}$$

where Q_k is a homogeneous polynomial of order N depending on the first derivatives of $\overset{*}{\varphi}\overset{*}{\nu}{}_0^i$. Then it is clear that F_2 is a C^1 function and

$$\partial_\varphi F_2(0,0,0)(\varphi)=\int_{\Omega_0}\sum_{i=1,N}\partial_i(\overset{*}{\varphi}\overset{*}{\nu}{}_0^i)d\Omega_0=\int_{\Gamma_0}\varphi d\Gamma_0,$$

which achieves the proof of the lemma. □

Remark 3.4. *In the next theorem we will apply the implicit function theorem near* $(0,0,0)$ *to the function* F. *Let us emphasize that* $F_1(0,0,0)=0$ *is ensured by the first part of* (1.6) *and* (2.4), *while* $F_2(0,0,0)=0$ *is guaranteed by* (3.6). *Also, from* (2.4) *and* (3.6), *it follows that*

$$F(0,\varphi,\Lambda)=\begin{pmatrix} -\underline{\Delta}\varphi-\frac{\mathcal{H}_0^2}{N-1}\varphi-\Lambda+\nu_0\cdot\int_{\Gamma_{\mathbf{t}(\varphi)}}\mathbf{b}(0,\varphi)\nu_0 d\Gamma_0 \\ \int_{\Gamma_0}\varphi \end{pmatrix}.$$

This implies $F(0,\varphi,\Lambda)\neq(0,0)$ *iff* $(\varphi,\Lambda)\neq 0$. *Indeed, if* $F(0,\varphi,\Lambda)=(0,0)$, *then* $\varphi=\boldsymbol{\alpha}\cdot\nu_0$, *for any* $\boldsymbol{\alpha}\in R^N$. *It follows that* $\Lambda=\nu_0\cdot\int_{\Gamma_{\mathbf{t}(\varphi)}}\mathbf{b}(0,\varphi)\nu_0 d\Gamma_0$. *From the second part of* (1.6), $\nu_0\cdot\int_{\Gamma_{\mathbf{t}(\varphi)}}\mathbf{b}(0,\varphi)\nu_0 d\Gamma_0\neq 0$ *iff* $\mathbf{t}(\varphi)\neq\mathbf{0}$. *As* $\mathbf{t}(\varphi)=\boldsymbol{\alpha}\frac{P(\Omega_0)}{N}$ *it follows that* $\boldsymbol{\alpha}=\mathbf{0}$ *and* $(\varphi,\Lambda)=(0,0)$.

Proof of Theorem 1.4. As we mentioned in the introduction, the main tool of the proof is the implicit function theorem applied to the function F near $(0,0,0)$. The only thing we have to prove is that the application $(\phi,\lambda) \in C^{2,\alpha}(\Gamma_0) \times R \mapsto \partial_{\varphi,\Lambda} F(0,0,0)(\phi,\lambda) \in C^{0,\alpha}(\Gamma_0) \times R$ defines an isomorphism. Then it would follow that there exist $\sigma^* > 0$ and a unique C^1 map

$$\Phi : (-\sigma^*, \sigma^*) \mapsto (C^{2,\alpha}(\Gamma_0), R), \quad \Phi(\sigma) = (\varphi_\sigma, \Lambda_\sigma),$$

such that $F(\sigma, \varphi_\sigma, \Lambda_\sigma) = 0$. Lemma 3.2 shows that for $\sigma \neq 0$, $(\varphi_\sigma, \Lambda_\sigma)$ satisfies Theorem 1.4. Thus $\Omega(\sigma, \varphi_\sigma)$ is the solution of the free boundary problem (1.3).

Let $(f,\mu) \in C^{0,\alpha}(\Gamma_0) \times R$ and look for the $(\varphi,\lambda) \in C^{2,\alpha}(\Gamma_0) \times R$ solution of

$$\partial_{\varphi,\Lambda} F(0,0,0)(\varphi,\lambda) = (f,\mu). \tag{3.7}$$

Let us prove first that this equation can have only one solution. Indeed, the eigenvalues of $\underline{\Delta}$ (if $\underline{\Delta}$ is considered as an operator in $H^2(\Gamma_0)$) are $\{\lambda_k = -k\frac{(N+k-2)}{(N-1)^2}\mathcal{H}_0^2, k = 0,1,\dots\}$. Let $\mathcal{U}_k = \{u_{ki},\ i = 1,\dots,i_k\}$ be the orthonormal family (with respect to $L^2(\Gamma_0)$ norm) of eigenfunctions associated to λ_k and $\mathcal{U} = \cup_{k=0}^{\infty}\mathcal{U}_k$. From [1], for example, we know completely $\mathcal{U}$: $\mathcal{U}_k$ is the trace on Γ_0 of kth order harmonic homogeneous polynomials in R^N. For example, we have

$$\mathcal{U}_0 = \left\{\frac{1}{P(\Omega_0)}\right\},\ \mathcal{U}_1 = \left\{\frac{x_1}{r_0}\sqrt{\frac{N}{P(\Omega_0)}}, \dots, \frac{x_N}{r_0}\sqrt{\frac{N}{P(\Omega_0)}}\right\},\ \dots .$$

As $\mathcal{U}$ is a base for $L^2(\Gamma_0)$, f can be written in the form $f = \sum_{k,i} f_{ki}u_{ki}$. Writing φ, the solution of (3.7), in the form $\varphi = \sum_{k,i}\varphi_{ki}u_{ki}$ and replacing it in (3.7) allows us to identify its coefficients

$$\begin{aligned}
&\varphi_{00} = \mu,\\
&\sum_{i=1,N} \partial^2_{ki} e(B_{\mathbf{x}})|_{\mathbf{x}_0}\varphi_{1i} = \frac{N}{P(\Omega_0)}f_{1i}, \qquad k = 1,\\
&\varphi_{ki} = \frac{1}{\mathcal{H}_0^2}\frac{(N-1)^2}{k(N+k-2)(N-1)}f_{ki}, \quad k \geq 2,\ i = 1,\dots,i_k,\\
&\lambda = \frac{f_0}{P(\Omega_0)} + \frac{\mu}{P(\Omega_0)}\frac{\mathcal{H}_0^2}{N-1}.
\end{aligned}$$

This system defines uniquely all the coefficients φ_{ki} and λ in terms of f_{ki} and μ, which together with the second part of (1.6) (in fact, only nonzero eigenvalues of $\nabla^2 e(B_{\mathbf{x}})|_{\mathbf{x}_0}$ are required) proves the uniqueness of the solution of (3.7). We emphasize that the hypothesis (1.6) is essential for the proof of the uniqueness.

Now let us prove the existence of the solution of (3.7). Let $C_0^{0,\alpha}(\Gamma_0)$, respectively, $C_0^{2,\alpha}(\Gamma_0)$, be the space of $C^{0,\alpha}(\Gamma_0)$, respectively, $C^{2,\alpha}(\Gamma_0)$, functions having zero integral over Γ_0. It is well known that $\underline{\Delta}$ defines an isomorphism from $C_0^{2,\alpha}(\Gamma_0)$ into $C_0^{0,\alpha}(\Gamma_0)$ and its inverse $\underline{\Delta}^{-1}$ is compact from $C_0^{0,\alpha}(\Gamma_0)$ into itself. If $f_0 = f - \fint_{\Gamma_0} f$ and φ_0 is the solution in $C_0^{2,\alpha}(\Gamma_0)$ of

$$-\underline{\Delta}\varphi_0 - \frac{\mathcal{H}_0^2}{N-1}\varphi_0 + \nu_0 \cdot \nabla^2 e(B_{\mathbf{x}})|_{\mathbf{x}_0} \cdot \mathbf{t}(\varphi_0) = f_0, \tag{3.8}$$

then

$$\varphi = \varphi_0 + \frac{\mu}{P(\Omega_0)}, \quad \lambda = \mu \frac{\mathcal{H}_0^2}{(N-1)P(\Omega_0)} + \fint_{\Gamma_0} f$$

is the solution of (3.7). Thus, to prove that (3.7) has a solution is sufficient to prove that (3.8) has a solution. Multiplying (3.8) by $-\underline{\Delta}^{-1}$ we obtain

$$\varphi_0 + A\varphi_0 = -\underline{\Delta}^{-1} f_0,$$

where $A : C_0^{0,\alpha}(\Gamma_0) \mapsto C_0^{0,\alpha}(\Gamma_0)$ is given by

$$A\varphi_0 = \underline{\Delta}^{-1} \left(\frac{\mathcal{H}_0^2}{N-1} \varphi_0 - \nu_0 \cdot \nabla^2 e(B_x)|_{\mathbf{x}_0} \cdot \mathbf{t}(\varphi_0) \right).$$

But A is compact and A has no eigenvalue equal to -1 (from injectivity). Then $I + A$ is invertible, which proves the existence of a $C_0^{0,\alpha}(\Gamma_0)$ solution for (3.8). From the regularity results related to $\underline{\Delta}$ we have $\mathrm{Im}(A) \subset C_0^{2,\alpha}(\Gamma_0)$, which proves that $\varphi_0 \in C^{2,\alpha}(\Gamma_0)$ and that $\partial_{\varphi,\Lambda} F(0,0,0)$ defines an isomorphism from $C^{2,\alpha}(\Gamma_0) \times R$ into $C^{0,\alpha}(\Gamma_0) \times R$ and ends the proof. □

Proof of Theorem 1.6. The conditions of Theorem 1.4 are fulfilled and so for $\sigma \neq$ we have $(\Omega(\sigma, \varphi_\sigma), \Lambda_\sigma)$, a solution of (1.3). If σ is small enough, then $\mathbf{t}(\varphi_\sigma) + \sigma\varphi_\sigma\nu_0 \in \Theta_0$, where Θ_0 is given by Lemma 2.4. By virtue of this lemma, instead of θ perturbations of $\Omega_\sigma := \Omega(\sigma, \varphi_\sigma)$, one can consider only perturbations of the form $\mathbf{t}(\varphi) + \sigma\varphi\nu_0$. From Lemma 2.7 it is sufficient to consider $\varphi \in C^{2,\alpha}(\Gamma_0)$ with $|\mathbf{t}(\varphi)| + \|\sigma\varphi\|_{C^{2,\alpha}(\Gamma_0)} \leq \delta$ for any $\delta > 0$ small. In order to preserve the measure, one has to consider the perturbations of the form $I + \mathbf{t}(\varphi) + (\sigma\varphi + \beta(\sigma\varphi))\nu_0 =: I + \mathbf{u}(\varphi)$ with $\varphi \in K_0 := \{\partial_\theta m(\Omega_0)(\varphi\nu_0) = 0\}$; see Lemma 2.9.

It is easy to prove that $\varphi_\sigma^0 = \varphi_\sigma - \fint_{\Gamma_0} \varphi_\sigma \in K_0$ and $\Omega_\sigma = \Omega_{\mathbf{u}(\varphi_\sigma^0)}$. Then it is enough to show that

$$E(\Omega_\sigma) = \min\{E(\Omega_{\mathbf{u}(\varphi_\sigma^0+\varphi)}), \quad \varphi \in C^{2,\alpha}(\Gamma_0) \cap K_0, \quad |t(\varphi)| + \|\sigma\varphi\|_{C^{2,\alpha}} \text{ small enough}\}. \tag{3.9}$$

Let us first estimate $E(\Omega_{\mathbf{u}(\varphi_\sigma^0+\varphi)}) - E(\Omega_{\mathbf{u}(\varphi_\sigma^0)})$. Let $G(\Omega)$ denote $e(\Omega)$, $P(\Omega)$, or $m(\Omega)$. Then, by virtue of (1.2), (1.5), and (1.10), we have

$$\begin{aligned}
&G(\Omega_{\mathbf{u}(\varphi_\sigma^0+\varphi)}) - G(\Omega_{\mathbf{u}(\varphi_\sigma^0)}) \\
&= \partial_\theta G(\Omega_{\mathbf{u}(\varphi_\sigma^0)})(\partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi)) \\
&+ \partial_\theta^2 G(\Omega_{\mathbf{u}(\varphi_\sigma^0+s\varphi)})(\partial_\varphi \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi), \partial_\varphi \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi)) \\
&+ \partial_\theta G(\Omega_{\mathbf{u}(\varphi_\sigma^0+s\varphi)})(\partial_\varphi^2 \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi, \varphi)) \\
&= \partial_\theta G(\Omega_\sigma)(\partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi)) \\
&+ \partial_\theta^2 G(\Omega_\sigma)(\partial_\varphi \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi), \partial_\varphi \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi)) + \partial_\theta G(\Omega_\sigma)(\partial_\varphi^2 \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi, \varphi)) \\
&+ o(1)(\|\partial_\varphi \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi)\|_{H^1(\Gamma_0)}^2 + \|\partial_\varphi^2 \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi, \varphi)\|_{H^1(\Gamma_0)}),
\end{aligned}$$

where $\lim o(1) = 0$ when $|\mathbf{t}(\varphi)| + \|\sigma\varphi\|_{C^{2,\alpha}(\Gamma_0)} \to 0 = 0$, $s \in (0,1)$. From the continuity of $\partial_\varphi \beta$, $\partial_\varphi^2 \beta$ given by Lemma 2.9, it follows that

$$\begin{aligned}
\partial_\varphi \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi) &= \partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi) + o_{\sigma\varphi}(1)\|\sigma\varphi\|_{H^1(\Gamma_0)}, \\
\partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi) &= \mathbf{t}(\varphi) + \sigma\varphi\nu_0 + \partial_\varphi \beta(\sigma\varphi_\sigma^0)(\sigma\varphi) \\
&= \mathbf{t}(\varphi) + \sigma\varphi\nu_0 + o_\sigma(1)\|\sigma\varphi\|_{H^1(\Gamma_0)}, \\
\partial_\varphi^2 \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi, \varphi) &= \partial_\varphi^2 \mathbf{u}(\varphi_\sigma^0)(\varphi, \varphi) + o_{\sigma\varphi}(1)\|\sigma\varphi\|_{H^1(\Gamma_0)}^2,
\end{aligned}$$

where $\lim o_\sigma(1) = 0$ when $\sigma \to 0$ and $\lim o_{\sigma\varphi}(1) = 0$ when $\|\sigma\varphi\|_{C^{2,\alpha}(\Gamma_0)} \to 0$. Then

$$\begin{aligned}
G(\Omega_{\mathbf{u}(\varphi_\sigma^0+\varphi)}) - G(\Omega_{\mathbf{u}(\varphi_\sigma^0)}) &= \partial_\theta G(\Omega_\sigma)(\partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi)) \\
&+ \partial_\theta^2 G(\Omega_\sigma)(\partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi), \partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi)) + \partial_\theta G(\Omega_\sigma)(\partial_\varphi^2 \mathbf{u}(\varphi_\sigma^0)(\varphi,\varphi)) \\
&+ o_{\sigma\varphi}(1)\|\sigma\varphi\|^2_{H^1(\Gamma_0)} \\
&+ o(1)(\|\partial_\varphi \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi)\|^2_{H^1(\Gamma_0)} + \|\partial_\varphi^2 \mathbf{u}(\varphi_\sigma^0 + s\varphi)(\varphi,\varphi)\|_{H^1(\Gamma_0)}) \\
&= \partial_\theta G(\Omega_\sigma)(\partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi)) \\
&+ \partial_\theta^2 G(\Omega_\sigma)(\partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi), \partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi)) + \partial_\theta G(\Omega_\sigma)(\partial_\varphi^2 \mathbf{u}(\varphi_\sigma^0)(\varphi,\varphi)) \\
&+ o(1)(|\mathbf{t}(\varphi)|^2 + \|\sigma\varphi\|^2_{H^1(\Gamma_0)}).
\end{aligned}$$

Consequently, taking into account (1.8) and the fact that $P(\Omega)$ and $m(\Omega)$ do not depend on $\mathbf{t}(\varphi)$, we may write

$$\begin{aligned}
&E(\Omega_{\mathbf{u}(\varphi_\sigma^0+\varphi)}) - E(\Omega_{\mathbf{u}(\varphi_\sigma^0)}) \\
&\quad= \partial_\theta^2 \left[e(\Omega_\sigma) + \frac{1}{\sigma^2} P(\Omega_\sigma) + \left(\frac{\Lambda_\sigma}{\sigma} - \frac{\mathcal{H}_0}{\sigma^2} \right) m(\Omega_\sigma) \right] (\partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi), \partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi)) \\
&\quad+ o(1)(|\mathbf{t}(\varphi)|^2 + \sigma^{-2}\|\sigma\varphi\|^2_{H^1(\Gamma_0)}) \\
&\quad= \partial_\theta^2 \left[e(\Omega_0) + \frac{1}{\sigma^2} P(\Omega_0) + \left(\frac{\Lambda_\sigma}{\sigma} - \frac{\mathcal{H}_0}{\sigma^2} \right) m(\Omega_0) \right] (\partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi), \partial_\varphi \mathbf{u}(\varphi_\sigma^0)(\varphi)) \\
&\quad+ (o(1) + o_\sigma(1))(|\mathbf{t}(\varphi)|^2 + \sigma^{-2}\|\sigma\varphi\|^2_{H^1(\Gamma_0)}) \\
&\quad= \partial_\theta^2 e(\Omega_0)(\mathbf{t}(\varphi) + \sigma\varphi\nu_0, \mathbf{t}(\varphi) + \sigma\varphi\nu_0) + \frac{1}{\sigma^2}\partial_\theta^2 [P(\Omega_0) - \mathcal{H}_0 m(\Omega_0)](\sigma\varphi\nu_0, \sigma\varphi\nu_0) \\
&\quad+ (o(1) + o_\sigma(1))(|\mathbf{t}(\varphi)|^2 + \sigma^{-2}\|\sigma\varphi\|^2_{H^1(\Gamma_0)}) \\
&\quad= \int_{\Gamma_0} |\overline{\nabla}\varphi|^2 - \frac{\mathcal{H}_0^2}{N-1}\varphi^2 d\Gamma_0 + \partial_\theta^2 e(\Omega_0)(\mathbf{t}(\varphi) + \sigma\varphi\nu_0, \mathbf{t}(\varphi) + \sigma\varphi\nu_0) \\
&\quad+ (o(1) + o_\sigma(1))(|\mathbf{t}(\varphi)|^2 + \sigma^{-2}\|\sigma\varphi\|^2_{H^1(\Gamma_0)}). \qquad (3.10)
\end{aligned}$$

Let us write $\varphi = \sum_{k,l} \varphi_{kl} u_{kl}$. As $\varphi \in K_0$, it follows that $\varphi_{00} = 0$. Moreover, from the estimations

$$\begin{aligned}
\partial_\theta^2 e(\Omega_0)(\mathbf{t}(\varphi), \sigma\varphi\nu_0) &= o_\sigma(1)(|\mathbf{t}(\varphi)|^2 + \|\varphi\|^2_{H^1(\Gamma_0)}), \\
\partial_\theta^2 e(\Omega_0)(\sigma\varphi\nu_0, \sigma\varphi\nu_0) &= o_\sigma(1)\|\varphi\|^2_{H^1(\Gamma_0)}, \\
\frac{k(N+k-2)}{(N-1)^2} - \frac{1}{N-1} &\geq \frac{1}{2}\frac{k(N+k-2)}{(N-1)^2}, \quad k \geq 2, \\
\left|\mathbf{t}(\varphi) \cdot \nabla^2 e(B_x)|_{\mathbf{x}_0} \cdot \mathbf{t}(\varphi)\right| &\geq \min\{\lambda_i, i = 1, \ldots, N\}|\mathbf{t}(\varphi)|^2 \\
&= \min\{\lambda_i, i = 1, \ldots, N\} \sum_{l=1,N} \left(\sum_{k=1,N} \varphi_{1k} \sqrt{\frac{N}{P(\Omega_0)}} \mathbf{t}_l(\nu_0^k) \right)^2 \\
&= \lambda_1 P(\Omega_0) \frac{N-1}{N}(N-1) \sum_l \varphi_{1l}^2,
\end{aligned}$$

where $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_N$ are eigenvalues of $\nabla^2 e(B_x)|_{\mathbf{x}_0}$, one can obtain

$$\begin{aligned}
\int_{\Gamma_0} |\overline{\nabla}\varphi|^2 - \frac{\mathcal{H}_0^2}{N-1}\varphi^2 d\Gamma_0 + \mathbf{t}(\varphi)\cdot\nabla^2 e(B_x)|_{\mathbf{x}_0}\cdot\mathbf{t}(\varphi) \\
+ 2\partial_\theta^2 e(\Omega_0)(\mathbf{t}(\varphi), \sigma\varphi\nu_0) + \partial_\theta^2 e(\Omega_0)(\sigma\varphi, \sigma\varphi\nu_0) \\
= \sum_{k\geq 2,l} \varphi_{kl}^2 \mathcal{H}_0^2 \left(\frac{k(N+k-2)}{(N-1)^2} - \frac{1}{N-1}\right) + \mathbf{t}(\varphi)\cdot\nabla^2 e(B_x)|_{\mathbf{x}_0}\cdot\mathbf{t}(\varphi) \\
+ (o(1) + o_\sigma(1))(|\mathbf{t}(\varphi)|^2 + \|\varphi\|^2_{H^1(\Gamma_0)}), \\
\geq \frac{1}{2\mathcal{H}_0^2}\sum_{k\geq 2,l} \frac{k(N+k-2)}{(N-1)^2}\mathcal{H}_0^2\varphi_{kl}^2 + \lambda_1\frac{P(\Omega_0)}{\mathcal{H}_0^2}\frac{N-1}{N}\sum_l (N-1)\mathcal{H}_0^2\varphi_{1l}^2 \\
+ (o(1) + o_\sigma(1))(|\mathbf{t}(\varphi)|^2 + \|\varphi\|^2_{H^1(\Gamma_0)}), \\
= \frac{1}{2\mathcal{H}_0^2}\|\varphi_1^c\|^2_{H^1(\Gamma_0)} + \lambda_1\frac{P(\Omega_0)}{\mathcal{H}_0^2}\frac{N-1}{N}\|\varphi_1\|^2_{H^1(\Gamma_0)} \\
+ (o(1) + o_\sigma(1))(|\mathbf{t}(\varphi)|^2 + \|\varphi\|^2_{H^1(\Gamma_0)}), \\
\geq C_0\|\varphi\|^2_{H^1(\Gamma_0)} + (o(1) + o_\sigma(1))\|\varphi\|^2_{H^1(\Gamma_0)},
\end{aligned}$$

where $C_0 = \min\{\frac{1}{2\mathcal{H}_0^2}, \lambda_1\frac{P(\Omega_0)}{\mathcal{H}_0^2}\frac{N-1}{N}\}$. From (3.10) and (2.12) it follows that

$$E(\Omega_{\mathbf{u}(\varphi_\sigma^0+\varphi)}) - E(\Omega_{\mathbf{u}(\varphi_\sigma^0)}) \geq C_0\|\varphi\|^2_{H^1(\Gamma_0)} + (o_{\sigma\varphi}(1) + o_\sigma(1))\|\varphi\|^2_{H^1(\Gamma_0)},$$

which proves (3.9) for σ, $|\mathbf{t}(\varphi)| + \|\sigma\varphi\|_{C^{2,\alpha}(\Gamma_0)}$ small, and proves the theorem. □

4. Applications. We will consider an application which involves the Laplacian in R^N. We show that the hypotheses (1.2), (1.5), and (1.10) are satisfied. We give a sufficient condition on data for the hypothesis (1.6).

Let us consider $f \in C^{1,\alpha}(R^N)$, $N \geq 2$. For $\theta \in \Theta$ let's set $e(\Omega_\theta) = -\frac{1}{2}\int_{\Omega_\theta} |\nabla u(\theta)|^2$, where $u(\theta) \in C^{2,\alpha}(\overline{\Omega}_\theta)$ is the solution of

$$\left\{\begin{array}{rl} -\Delta u(\theta) = f & \text{in } \Omega_\theta, \\ u(\theta) = 0 & \text{on } \Gamma_\theta. \end{array}\right. \tag{4.1}$$

The main result we prove in this section follows.

THEOREM 4.1. *Let $f \in C^{1,\alpha}(R^N)$ be a radial function and $\mathbf{x}_0 = 0$.*

(i) *If $[\frac{N}{r_0^N}\int_0^{r_0} s^{N-1}f(s)ds - f(r_0)]\int_0^{r_0} s^{N-1}f(s)ds \neq 0$, then (1.3) has a unique $C^{2,\alpha}$ solution for σ small.*

(ii) *If $[\frac{N}{r_0^N}\int_0^{r_0} s^{N-1}f(s)ds - f(r_0)]\int_0^{r_0} s^{N-1}f(s)ds > 0$, then (1.1) has a unique $C^{2,\alpha}$ solution for σ small.*

To prove the theorem we need some preliminary results, most of them quite classical results in shape optimization theory. Let us point out that throughout this chapter, $\overset{*}{\cdot}$ denotes an extension operator from Θ to $C^{2,\alpha}(\overline{\Omega}_0)$, or (explicitly) from $H^{1/2}(\Gamma_0)$ to $H^1(\Omega_0)$.

PROPOSITION 4.2. *The shape functional $\theta \to u(\theta)\circ(I + \overset{*}{\theta})$ is differentiable from Θ into $C^{2,\alpha}(\overline{\Omega}_0)$. Moreover, $u(\theta)(\xi)$, the first shape derivative of $u(\theta)$, satisfies*

$u(\theta)(\xi) \in C^{1,\alpha}(\overline{\Omega}_0)$ *and*

$$u(\theta)(\xi) \circ (I + \overset{*}{\theta}) + \xi \cdot \nabla u(\theta) \circ (I + \overset{*}{\theta}) = \partial_\theta (u(\theta) \circ (I + \overset{*}{\theta}))(\xi), \tag{4.2}$$

$$\Delta u(\theta)(\xi) = 0 \qquad in \ \ \Omega_\theta, \tag{4.3}$$

$$u(\theta)(\xi) = -(\overset{\vee}{\xi} \cdot \nu_\theta) \frac{\partial u(\theta)}{\partial \nu_\theta} \quad on \ \ \Gamma_\theta, \tag{4.4}$$

$$\int_{\Omega_\theta} |\nabla u(\theta)(\xi)|^2 \leq C \|\xi\|^2_{H^{1/2}(\Gamma_0)}, \tag{4.5}$$

$$\int_{\Omega_\theta} |\nabla u(\theta + \zeta)(\xi) \circ (I + \overset{\vee}{\zeta}) - \nabla u(\theta)(\xi)|^2 = o_\zeta(1) \|\xi\|^2_{H^{1/2}(\Gamma_0)}, \tag{4.6}$$

where $\overset{\vee}{\zeta} = \overset{*}{\zeta} \circ (I + \overset{*}{\theta})^{-1}$ *and* $\lim_{\|\zeta\|_\Theta \to 0} o_\zeta(1) = 0$.

REMARK 4.3. *The estimation* (4.6) *is crucial in proving that* $e(\Omega)$ *satisfies* (1.10)*; see Lemma* 4.5.

Proof of Proposition 4.2. The differentiability of $u(\theta) \circ (I + \overset{*}{\theta})$ is a rather classical result. The proof is based on the implicit function theorem and the uniqueness of (4.1). A proof in a Sobolev spaces framework is shown in [12], [13].

Although the proof is quite standard, let us prove that the result is true even in Hölder spaces. Let $F(\theta, v)$ be defined by

$$\begin{array}{rccl} F: & \Theta \times C_0^{2,\alpha}(\overline{\Omega}_0) & \mapsto & C^\alpha(\overline{\Omega}_0), \\ & (\theta, v) & \mapsto & -\displaystyle\sum_{i,j,k=1,N} M_{i,j}(\theta) \frac{\partial}{\partial x_j} \left(M_{i,k}(\theta) \frac{\partial v}{\partial x_k} \right) + f \circ (I + \overset{*}{\theta}), \end{array}$$

where $C_0^{2,\alpha}(\overline{\Omega}_0) = C^{2,\alpha}(\overline{\Omega}_0) \cap \{u = 0 \text{ on } \Gamma_0\}$ and $[M_{i,j}(\theta)] = {}^t[\nabla(I + \overset{*}{\theta})]^{-1}$ (let us point out that the function $F(\theta, v)$ is obtained by transporting (4.1) in Ω_0 and replacing $u(\theta) \circ (I + \overset{*}{\theta})$ by v). The function F is linear in v and derivatives of v up to second order, while $\theta \mapsto M_{i,j}(\theta)$ is a C^1 function from Θ to $C^{1,\alpha}(\overline{\Omega}_0; R^N)$ in a small neighborhood of $0 \in \Theta$ (as a rational function in θ derivatives and with a denominator greater than, let's say, $1/2$), as well as $\theta \to f \circ (I + \overset{*}{\theta})$. Thus F is a C^1 function in a small neighborhood of $0 \in \Theta$. Moreover, $F(0, u(0)) = 0$, $\partial_v F(0, u(0))(v) = -\Delta v$, which defines an isomorphism from $C_0^{2,\alpha}(\overline{\Omega}_0)$ to $C^\alpha(\overline{\Omega}_0)$. From the implicit function theorem we have the existence of a C^1 map $\theta \mapsto v(\theta)$ from Θ to $C_0^{2,\alpha}(\overline{\Omega}_0)$ such that $F(\theta, v(\theta)) = 0$. It follows that $v(\theta) \circ (I + \overset{*}{\theta})^{-1}$ is the solution of (4.1) and from the uniqueness of its solutions we get $v(\theta) = u(\theta) \circ (I + \overset{*}{\theta})$ and thus the C^1 regularity for $\theta \mapsto u(\theta) \circ (I + \overset{*}{\theta})$.

From Sobolev space embeddings we have that $\theta \to u(\theta) \circ (I + \overset{*}{\theta})$ is a C^1 map from Θ to $H^2(\Omega_0)$. From [12, Lemma 2.1] it follows that $u(\theta)(\xi) \in H^1(\Omega_\theta)$, satisfying (4.2). The conditions of [12, Theorems 3.1 and 3.2] are satisfied, and (4.3) and (4.4) follow. The estimation (4.5) is proven, for example, in [4].

To prove (4.6), let us introduce $\tilde{u}(\theta)(\xi) = \frac{1}{\|\xi\|_{H^{1/2}(\Gamma_0)}} u(\theta)(\xi)$. The map $\theta \mapsto \tilde{u}(\theta)(\xi) \circ (I + \overset{*}{\theta})$ from Θ to $H^1(\Omega_0)$ is of class C^1. Indeed, as $\tilde{u}(\theta)(\xi)$ satisfies

$$\Delta \tilde{u}(\theta)(\xi) = 0 \quad \text{in} \ \ \Omega_\theta, \tag{4.7}$$

$$\tilde{u}(\theta)(\xi) = -\frac{\overset{\vee}{\xi} \cdot \nu_\theta}{\|\xi\|_{H^{1/2}(\Gamma_0)}} \frac{\partial u(\theta)}{\partial \nu_\theta} \quad \text{on} \ \ \Gamma_\theta, \tag{4.8}$$

let us introduce the function

$$\begin{aligned}\tilde{F}: \quad \Theta \times H_0^1(\Omega_0) &\mapsto H^{-1}(\Omega_0),\\ (\theta, v) &\mapsto \int_{\Omega_0} \nabla(v+g(\theta))\cdot[M_{ij}(\theta)]\cdot {}^t[M_{ij}(\theta)]\cdot\nabla\varphi \; Jac(I+\overset{*}{\theta}),\end{aligned}$$

where $\varphi \in H_0^1(\Omega_0)$, $g(\theta) = -\frac{\overset{*}{\xi}\cdot\overset{*}{\nu}_\theta}{\|\xi\|_{H^{1/2}(\Gamma_0)}}(\overset{*}{\nu}_\theta\cdot\nabla u(\theta))\circ(I+\overset{*}{\theta})$ and $\overset{*}{\cdot}$ is a $H^1(\Omega_0)$ extension of $\cdot$, thus, in particular, $\|\overset{*}{\xi}\|_{H^1(\Omega_0)} \le C\|\xi\|_{H^{1/2}(\Gamma_0)}$ (note that the function $\tilde{F}$ is obtained basically by transporting the weak form of (4.7) in Ω_0).

We have $\tilde{F}(0,\tilde{u}(0)(\xi)-g(0))=0$ and the function $\tilde{F}$ is of class C^1 near $(0,\tilde{u}(0)(\xi)-g(0))$. Indeed, it is linear in v. Moreover, from the first part of the proposition, $\theta\mapsto g(\theta)$ is a C^1 map from Θ to $H^1(\Omega_0)$ and $\partial_v\tilde{F}(0,\tilde{u}(0)(\xi)-g(0))(v)=\int_{\Omega_0}\nabla v\cdot\nabla\varphi$, which defines an isomorphism from $H^1(\Omega_0)$ into $H^{-1}(\Omega_0)$. The conditions of the implicit function theorem are satisfied. Then there exists a C^1 map $\theta\mapsto v(\theta)$ from a small open neighborhood of $0\in\Theta$ into $H^1(\Omega_0)$ such that $\tilde{F}(\theta,v(\theta))=0$. This neighborhood may be chosen independent of ξ because $\|g(\theta)\|_{H^1(\Omega_0)}\le C$, $\|\partial_\theta[g(\theta)](\eta)\|_{H^1(\Omega_0)} < C\|\eta\|_\Theta$ with C independent of ξ, and so $\tilde{F}(\theta,v)$ is a uniformly C^1 function (in ξ) from $\Theta\times H^1(\Omega_0)$ to $H^{-1}(\Omega_0)$. Writing $\tilde{F}(\theta,v(\theta))$ as an integral in Ω_θ, from the uniqueness result for (4.7), (4.8), it follows that $v(\theta)=\tilde{u}(\theta)(\xi)\circ(I+\overset{*}{\theta})-g(\theta)$, which proves that $\theta\mapsto\tilde{u}(\theta)(\xi)\circ(I+\overset{*}{\theta})$ is a C^1 function from Θ to $H^1(\Omega_0)$. It follows that for $\|\theta\|_\Theta$ small we have

$$\int_{\Omega_0}|\nabla(\tilde{u}(\theta+\zeta)(\xi)\circ(I+\overset{*}{\theta}+\overset{*}{\zeta}))-\nabla(\tilde{u}(\theta)(\xi)\circ(I+\overset{*}{\theta}))|^2 = o_\zeta(1), \qquad \lim_{\|\zeta\|_\Theta\to 0} o_\zeta(1)=0.$$

On the other hand,

$$\begin{aligned}\nabla(\tilde{u}(\theta+\zeta)(\xi))\circ(I+\check{\zeta}) &= {}^t[\nabla(I+\check{\zeta})]^{-1}\nabla(\tilde{u}(\theta+\zeta)(\xi)\circ(I+\check{\zeta}))\\ &= {}^t[\nabla(I+\check{\zeta})]^{-1}\cdot{}^t[\nabla(I+\overset{*}{\theta})]^{-1}\nabla(\tilde{u}(\theta+\zeta)(\xi)\circ(I+\overset{*}{\theta}+\overset{*}{\zeta})),\\ [\nabla(I+\check{\zeta})]^{-1} &= I_N+o_\zeta(1), \quad I_N \text{ the identity matrix},\\ [\nabla(I+\overset{*}{\theta})]^{-1} &= I_N+o_\theta(1), \quad \lim_{\|\theta\|_\Theta\to 0} o_\theta(1)=0,\\ |[\nabla(I+\overset{*}{\theta})]^{-1}| &= 1+o_\theta(1).\end{aligned}$$

It follows that

$$\begin{aligned}&\int_{\Omega_\theta}|\nabla\tilde{u}(\theta+\zeta)(\xi)\circ(I+\check{\zeta})-\nabla\tilde{u}(\theta)(\xi)|^2\\ &=\int_{\Omega_\theta}|{}^t[\nabla(I+\check{\zeta})]^{-1}\nabla\tilde{u}(\theta+\zeta)(\xi)\circ(I+\check{\zeta})-\nabla\tilde{u}(\theta)(\xi)|^2\\ &=\int_{\Omega_\theta}|\nabla(\tilde{u}(\theta+\zeta)(\xi)\circ(I+\check{\zeta}))-\nabla\tilde{u}(\theta)(\xi)|^2+o_\zeta(1)\\ &=\int_{\Omega_\theta}{}^t[\nabla(I+\overset{*}{\theta})]^{-1}|\nabla(\tilde{u}(\theta+\zeta)(\xi)\circ(I+\overset{*}{\theta}+\overset{*}{\zeta}))-\nabla(\tilde{u}(\theta)(\xi)\circ(I+\overset{*}{\theta}))|^2+o_\zeta(1)\\ &\le C\int_{\Omega_0}|\nabla(\tilde{u}(\theta+\zeta)(\xi)\circ(I+\overset{*}{\theta}+\overset{*}{\zeta}))-\nabla(\tilde{u}(\theta)(\xi)\circ(I+\overset{*}{\theta}))|^2+o_\zeta(1)\\ &\le (C+1)o_\zeta(1).\end{aligned}$$

The estimation (4.6) is proved because $u(\theta)(\xi)=\|\xi\|_{H^{1/2}(\Gamma_0)}\tilde{u}(\theta)(\xi)$. □

PROPOSITION 4.4. *The functions $\theta \to e(\Omega_\theta)$, $\theta \to P(\Omega_\theta)$, and $\theta \to m(\Omega_\theta)$ are near $0 \in \Theta$ two times differentiable. Their derivatives are given by*

$$\partial_\theta e(\Omega_\theta)(\xi) = -\frac{1}{2}\int_{\Gamma_\theta} |\nabla u(\theta)|^2 (\check{\xi}\cdot\nu_\theta), \tag{4.9}$$

$$\begin{aligned}\partial^2_\theta e(\Omega_\theta)(\xi,\eta) = &-\frac{1}{2}\int_{\Gamma_\theta} (\check{\xi}\cdot\nu_\theta)(\check{\eta}\cdot\nu_\theta)\left[\mathcal{H}_\theta|\nabla u(\theta)|^2 + \partial_{\nu_\theta}|\nabla u(\theta)|^2\right]\\ &-\frac{1}{2}\int_{\Gamma_\theta}\left((\check{\xi}\cdot\nu_\theta)\partial_{\nu_\theta}u(\theta)(\eta) + (\check{\eta}\cdot\nu_\theta)\partial_{\nu_\theta}u(\theta)(\xi)\right)\partial_{\nu_\theta}u(\theta)\\ &+\frac{1}{2}\int_{\Gamma_\theta}\left(\nu_\theta\cdot\overline{\nabla}\check{\xi}\cdot\check{\eta} + \nu_\theta\cdot\overline{\nabla}\check{\eta}\cdot\check{\xi} + \check{\xi}\cdot\overline{\nabla}\nu_\theta\cdot\check{\eta}\right)|\nabla u(\theta)|^2,\end{aligned} \tag{4.10}$$

$$\partial_\theta P(\Omega_\theta)(\xi) = \int_{\Gamma_\theta} \mathcal{H}_\theta(\check{\xi}\cdot\nu_\theta), \tag{4.11}$$

$$\partial^2_\theta P(\Omega_\theta)(\xi,\eta) = \int_{\Gamma_\theta} (\overline{\nabla}\cdot\check{\xi})(\overline{\nabla}\cdot\check{\eta}) - tr[(\overline{\nabla}\check{\xi})\cdot(\overline{\nabla}\check{\eta})] + (\nu_\theta\cdot(\overline{\nabla}\check{\xi}))\cdot(\nu_\theta\cdot(\overline{\nabla}\check{\eta})), \tag{4.12}$$

$$\partial_\theta m(\Omega_\theta)(\xi) = \int_{\Gamma_\theta} (\check{\xi}\cdot\nu_\theta), \tag{4.13}$$

$$\partial^2_\theta m(\Omega_\theta)(\xi,\eta) = \int_{\Gamma_\theta} (\check{\xi}\cdot\nu_\theta)(\overline{\nabla}\cdot\check{\eta}) - \check{\xi}\cdot\overline{\nabla}\check{\eta}\cdot\nu_\theta, \tag{4.14}$$

where $\check{\xi} = \xi\circ(I+\theta)^{-1}$, $\check{\eta} = \eta\circ(I+\theta)^{-1}$.

Proof. The differentiability and the derivative (4.9) follow from [12, Theorem 3.3], taking $C(u(\theta)) = |\nabla u(\theta)|^2$. The derivative (4.10) follows from [12, Theorem 5.1], which says that (under some regularity assumptions)

$$\partial_\theta\left(\int_{\Gamma_\theta} G(u(\theta))\right)(\xi) = \int_{\Gamma_\theta} \partial_\theta G(u(\theta))(\xi) + (\check{\xi}\cdot\nu_\theta)\left[\mathcal{H}_\theta G(u(\theta)) + \partial_{\nu_\theta}G(u(\theta))\right].$$

The expression (4.10) then follows by taking in the previous formula $G(u(\theta)) = |\nabla u(\theta)|^2(\check{\xi}\cdot\nu_\theta)$ (a proof for three dimensions is given in [9]).

The derivatives (4.11) and (4.12) are proven in [5, Theorem 10.4]. Finally, the derivatives (4.13) and (4.14) follow by elementary calculus (a proof of them is presented in [9]). □

LEMMA 4.5. *The functionals $e(\Omega_\theta)$, $P(\Omega_\theta)$, and $m(\Omega_\theta)$ satisfy the hypotheses* (1.2), (1.5), *and* (1.10).

Proof. The proposition is true for $P(\Omega_\theta)$ and $m(\Omega_\theta)$. The proof follows immediately from Proposition 4.4 and (4.12), (4.14). Let us prove that $e(\Omega_\theta)$ satisfies Lemma 4.5. Proposition 4.4 proves the hypothesis (1.2) and gives $b(\theta) = -\frac{1}{2}|\nabla u(\theta)|^2$. The hypothesis (1.5) is satisfied by virtue of Proposition 4.2. For the hypothesis (1.10), from the form of $\partial^2_\theta e(\Omega_\theta)$, it is enough to show only that $\theta \mapsto \int_{\Gamma_\theta}(\check{\xi}\cdot\nu_\theta)\partial_{\nu_\theta}u(\theta)(\eta)$ is a continuous map from Θ to R. As $\int_{\Gamma_\theta}(\check{\xi}\cdot\nu_\theta)\partial_{\nu_\theta}u(\theta)(\eta) = -\int_{\Omega_\theta}\nabla u(\theta)(\xi)\cdot\nabla u(\theta)(\eta)$ it suffices to show that $\int_{\Omega_\theta}|\nabla u(\theta+\zeta)(\xi\circ(I+\check{\zeta}) - \nabla u(\theta)(\xi)|^2 = o_\zeta(1)\|\xi\|_{H^1(\Gamma_0)}$, which is proved in Proposition 4.2. □

PROPOSITION 4.6. *Let $N \geq 2$ and let us assume that f is a radial function. Then $\nabla e(B_{\mathbf{x}})|_{\mathbf{x}_0} = \mathbf{0}$ and the second derivative matrix of $e(B_{\mathbf{x}})$ at 0 is*

$$\nabla^2 e(B_{\mathbf{x}})|_{\mathbf{x}=\mathbf{0}} = \left(\frac{m(\Omega_0)}{r_0^N}\left[\frac{N}{r_0^N}\int_0^{r_0} s^{N-1}f(s)ds - f(r_0)\right]\int_0^{r_0} s^{N-1}f(s)ds\right)\mathrm{Id},$$

where Id *is the identity matrix.*

Proof. It is clear that the derivatives of $e(B_{\mathbf{x}})$ are intimately related to the shape derivatives of $e(\Omega_\theta)$. Indeed, we have

$$\partial_i e(B_{\mathbf{x}})|_{\mathbf{x}_0} = \partial_\theta e(\Omega_0)(\xi_i), \qquad i = 1, \dots, N, \tag{4.15}$$

$$\partial^2_{ij} e(B_{\mathbf{x}})|_{\mathbf{x}_0} = \partial^2_\theta e(\Omega_0)(\xi_i, \xi_j), \ i, j = 1, \dots, N, \tag{4.16}$$

where $\{\xi_1, \xi_2, \dots, \xi_N\}$ is the standard base of R^N. The solution $u(0)(\xi_i)$ of problems (4.3) and (4.4) is given by $u(0)(\xi_i) = -\frac{x_i}{r_0}\partial_{\nu_0} u(0)$. Also, for $\theta = 0$, the solution of (4.1) is a radial function, denoted by $u(0)$. We have

$$\begin{aligned}
u(0)(\cdot) &= -\int_0^{|\cdot|} t^{1-N} \int_0^t s^{N-1} f(t) ds dt + \text{constant},\\
\partial_r u(0)(\cdot) &= -|\cdot|^{1-N} \int_0^{|\cdot|} s^{N-1} f(s) ds,\\
\partial^2_r u(0)(\cdot) &= (N-1)|\cdot|^{-N} \int_0^{|\cdot|} s^{N-1} f(s) ds - f(|\cdot|).
\end{aligned}$$

From (4.15) and (4.9) it follows that

$$\partial_i e(B_{\mathbf{x}})|_{\mathbf{x}_0} = -\frac{1}{2}\int_{\Gamma_0} (\xi_i \cdot \nu_0)|\nabla u(0)|^2 = -\frac{1}{2}\int_{\Gamma_0} |\nabla u(0)|^2 \nu_0^i = 0.$$

From (4.16) and (4.10) it follows that

$$\begin{aligned}
\partial^2_{ij} e(B_{\mathbf{x}})|_{\mathbf{x}_0} &= -\frac{1}{2}\int_{\Gamma_0} (\xi_i \cdot \nu_0)(\xi_j \cdot \nu_0)\left[\mathcal{H}_0 |\nabla u(0)|^2 + \partial_{\nu_0}|\nabla u(0)|^2\right]\\
&\quad -\frac{1}{2}\int_{\Gamma_0} \left[(\xi_i \cdot \nu_0)\partial_{\nu_0} u(0)(\xi_j) + (\xi_j \cdot \nu_0)\partial_{\nu_0} u(0)(\xi_i)\right]\partial_{\nu_0} u(0)\\
&\quad +\frac{1}{2}\int_{\Gamma_0} \left[\nu_0 \cdot \overline{\nabla}\xi_i \cdot \xi_j + \nu_0 \cdot \overline{\nabla}\xi_j \cdot \xi_i + \xi_i \cdot \overline{\nabla}\nu_0 \cdot \xi_j\right]|\nabla u(0)|^2\\
&= -\frac{1}{2}\int_{\Gamma_0} \nu_0^i \nu_0^j \left[\mathcal{H}_0 |\partial_r u(0)|^2 + \partial_r |\partial_r u(0)|^2\right]\\
&\quad +\frac{1}{2}\int_{\Gamma_0} 2\nu_0^i \nu_0^j r_0^{-1} |\partial_r u(0)|^2 + \overline{\partial}_i \nu_0^j |\partial_r u(0)|^2.
\end{aligned}$$

But we have

$$\int_{\Gamma_0} \nu_0^i \nu_0^j = \delta_{ij}\frac{m(\Omega_0)}{r_0}, \qquad \int_{\Gamma_0} \overline{\partial}_i \nu_0^j = \int_{\Gamma_0} \mathcal{H}_0 \nu_0^i \nu_0^j = \delta_{ij}\mathcal{H}_0 \frac{m(\Omega_0)}{r_0}.$$

From the last equation of $\partial^2_{ij} e(B_{\mathbf{x}})$ we obtain

$$\begin{aligned}
\partial^2_{ij} e(B_{\mathbf{x}})|_{\mathbf{x}_0} &= -\frac{1}{2}\delta_{ij}\frac{m(\Omega_0)}{r_0}\left[\mathcal{H}_0 |\partial_r u(0)|^2 + \partial_r |\partial_r u(0)|^2 - 2r_0^{-1}|\partial_r u(0)|^2 - \mathcal{H}_0 |\partial_r u(0)|^2\right]\\
&= -\delta_{ij}\frac{m(\Omega_0)}{r_0}\left[\partial^2_r u(0) - r_0^{-1}\partial_r u(0)\right]\partial_r u(0)\\
&= \delta_{ij}\frac{m(\Omega_0)}{r_0}\left[N r_0^{-N}\int_0^{r_0} s^{N-1} f(s) ds - f(r_0)\right] r_0^{1-N}\int_0^{r_0} s^{N-1} f(s) ds,
\end{aligned}$$

which proves the proposition. ∎

Proof of Theorem 4.1. Item (i), respectively, item (ii), follows from Theorem 1.4, respectively, Theorem 1.6—the hypotheses are satisfied from Lemma 4.5 and Proposition 4.6. □

Acknowledgments. I am deeply grateful to professor Michel Pierre for several comments and suggestions, as well as to one of the anonymous referees for his constructive report.

REFERENCES

[1] M. Berger, P. Gauduchon, and E. Mazet, *Le Spectre d'une Variété Rimannienne*, Lecture Notes in Math., Springer-Verlag, Berlin, Heidelberg, New York, 1971.
[2] M. Dambrine, *Hessiennes de formes et stabilité des formes critiques*, Doctorat de l'École Normale Supérieure de Cachan, France, 2000.
[3] J. Descloux, *On the Two-Dimensional Magnetic Shaping Problem without Surface Tension*, École Polytechnique Fédérale de Lausanne, Switzerland, 1990.
[4] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
[5] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkäuser-Verlag, Basel, 1984.
[6] M. Hayouni and A. Novruzi, *Sufficient condition for the existence of solutions of a free boundary problem*, Quart. Appl. Math., 60 (2002), pp. 425–435.
[7] A. Henrot and M. Pierre, *About critical points of the energy in an electromagnetic shaping problem*, in Boundary Control and Boundary Variation, Lecture Notes in Control and Inform. Sci. 178, J.P. Zolésio, ed., Springer-Verlag, Berlin, 1992, pp. 238–252.
[8] A. Henrot and M. Pierre, *About existence of equilibria in electromagnetic casting*, Quart. Appl. Math., 49 (1991), pp. 563–575.
[9] A. Novruzi, *Contribution en Optimisation de Formes et Applications*, Ph.D. thesis, Université Henri Poincaré Nancy 1, France, 1997.
[10] A. Novruzi and M. Pierre, *Structure of shape derivatives*, J. Evol. Equ., 2 (2002), pp. 365–382.
[11] M. Pierre and J.-R. Roche, *Numerical simulation of tridimensional electromagnetic shaping of liquid metals*, Numer. Math., 65 (1993), pp. 203–217.
[12] J. Simon, *Differentiation with respect to the domain in boundary value problems*, Numer. Funct. Anal. Optim., 2 (1980), pp. 649–687.
[13] J. Sokolowski and J. P. Zolésio, *Introduction to Shape Optimization. Shape Sensitivity Analysis*, Springer-Verlag, Berlin, 1992.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 194–215

SMALL BV SOLUTIONS OF HYPERBOLIC NONCOOPERATIVE DIFFERENTIAL GAMES*

ALBERTO BRESSAN† AND WEN SHEN†

Abstract. The paper is concerned with an n-person differential game in one space dimension. We state conditions for which the system of Hamilton–Jacobi equations for the value functions is strictly hyperbolic. In the positive case, we show that the weak solution of a corresponding system of conservation laws determines an n-tuple of feedback strategies. These yield a Nash equilibrium solution to the noncooperative differential game.

Key words. noncooperative differential games, Nash equilibrium, system of Hamilton–Jacobi equations, hyperbolic system of conservation laws, BV solutions, optimal control theory, discontinuous ODE

AMS subject classifications. 91A23, 91A10, 49N70, 49N90, 49N35, 49L20, 93B52, 35L65, 34A36

DOI. 10.1137/S0363012903425581

1. Introduction.

This paper is concerned with the global existence of a Nash equilibrium solution for a noncooperative n-person differential game. The evolution of the system is governed by a differential equation of the form

$$\dot{x}(t) = \sum_{i=1}^{n} f_i(x, u_i), \tag{1.1}$$

say, with initial data

$$x(\tau) = y. \tag{1.2}$$

Here the real valued map $t \mapsto u_i(t)$ is the control implemented by the ith player. Together with (1.1) we consider the cost functionals

$$J_i = J_i(\tau, y, u_1, \ldots, u_n) \doteq \int_{\tau}^{T} h_i(x(t), u_i(t))\, dt + g_i(x(T)), \tag{1.3}$$

consisting of a running cost h_i and a terminal cost g_i. The goal of the ith player is to minimize J_i. An n-tuple of feedback strategies

$$U_i^* = U_i^*(t, x), \quad i = 1, \ldots, n,$$

is called a Nash equilibrium solution if the following holds. For each i, if the ith player chooses an alternative strategy U_i while every other player sticks to his previous strategy U_j^*, $j \neq i$, then the cost for the ith player does not decrease:

$$J_i(\tau, y, U_1^*, \ldots, U_{i-1}^*, U_i, U_{i+1}^*, \ldots, U_n^*) \geq J_i(\tau, y, U_1^*, \ldots, U_{i-1}^*, U_i^*, U_{i+1}^*, \ldots, U_n^*). \tag{1.4}$$

*Received by the editors April 4, 2003; accepted for publication (in revised form) September 30, 2003; published electronically June 15, 2004.
http://www.siam.org/journals/sicon/43-1/42558.html

†SISSA, Via Beirut 4, 34014 Trieste, Italy (bressan@sissa.it, shen@sissa.it). Current address: Department of Mathematics, Penn State University, University Park, State College, PA 16802 (bressan@math.psu.edu, shen_w@math.psu.edu).

Assume that a value function $V = (V_1, \dots, V_n)$ exists, so that $V_i(t,x)$ is the minimum cost for the ith player, when everyone plays optimally but no cooperation is allowed. Under suitable regularity conditions (see [F1, p. 292]), the function V provides a solution to the system of Hamiltonian equations

$$\frac{\partial}{\partial t} V_i + H_i(x, \nabla V_1, \dots, \nabla V_n) = 0, \tag{1.5}$$

with terminal data

$$V_i(T,x) = g_i(x). \tag{1.6}$$

The Hamiltonian functions H_i are defined as follows. Assume that, for any given gradient vectors $p_1, \dots, p_n$, there exist optimal control values $u_j^*(x,p_j)$, $j = 1, \dots, n$, such that

$$p_j \cdot f_j(x, u_j^*(x,p_j)) + h_j(x, u^*(x,p_j)) = \min_{\omega \in \mathbb{R}} \{p_j \cdot f_j(x,\omega) + h_j(x,\omega)\}. \tag{1.7}$$

Then

$$\begin{aligned} H_i(x,p_1,\dots,p_m) &= p_i \cdot \sum_{j=1}^{m} f_j(x, u_j^*(x,p_j)) + h_i(x, u_i^*(x,p_i)) \\ &= p_i \cdot \sum_{j \neq i} f_j(x, u_j^*(x,p_j)) + \min_{\omega \in \mathbb{R}} \{p_i \cdot f_i(x,\omega) + h_i(x,\omega)\}. \end{aligned} \tag{1.8}$$

In the case of a two-person, zero-sum differential game, the value function is obtained from the scalar Bellman–Isaacs equation [F1, I]. The analysis can thus rely on comparison principles and on the well-developed theory of viscosity solutions for Hamilton–Jacobi equations; see, for example, [BC]. On the other hand, in the case of noncooperative n-person games, one has to study a highly nonlinear system of Hamilton–Jacobi equations. Little is yet known in this direction, except for particular examples as in [CR, O]. Instead, local existence results are known within the class of open-loop strategies [F1, VZ]. They apply to the case where the players cannot use any additional information on the state $x(t)$ of the system, for $t > 0$.

In the one-dimensional case, differentiating (1.5) with respect to (w.r.t.) x, one obtains a system of conservation laws for the gradient functions $p_i \doteq V_{i,x}$, namely,

$$p_{i,t} + H_i(x,p)_x = 0. \tag{1.9}$$

In recent years, considerable progress has been achieved in the understanding of weak solutions to hyperbolic systems of conservation laws in one space dimension. In particular, entropy admissible solutions with small total variation are known to be unique and depend continuously on the initial data [B3, BLY]. Moreover, they can be obtained as the unique limits of vanishing viscosity approximations [BB].

The aim of the present paper is to investigate the relevance of these new results in the field of conservation laws toward the existence and stability of Nash equilibrium solutions, in the context of differential games. In particular, our main goals are

- to identify the situations where the hyperbolic theory is applicable, and
- in the favorable case, to derive the existence and properties of a Nash equilibrium solution.

The hyperbolicity of the system is clearly a crucial assumption, in order that the Cauchy problem for the value functions be well posed. In section 3 we show that hyperbolicity holds provided that the derivatives of the cost functions g_i all have the same sign. In practice, this means that all players wish to steer the system in the same direction.

Granted that the system is strictly hyperbolic, the known results on systems of conservation laws can be applied. The theorem of Glimm [G] or its more general versions [BB, ILF, L] provide then the existence of a global solution to the Hamilton–Jacobi equations for terminal data g_i whose gradients have sufficiently small total variation.

To obtain an existence result for solutions of differential games, one has to show that, for each single player, the feedback strategy corresponding to the solution of the Hamilton–Jacobi system is indeed an optimal one. We remark that, if the value functions V_i were smooth, the optimality would be an immediate consequence of the equations. The main technical difficulty stems from the nondifferentiability of these value functions.

In the literature on control theory, sufficient conditions for optimality have been obtained along two main directions. On one hand, there is the "regular synthesis" approach developed by Boltianskii [Bo], Brunovskii [Br], and Sussmann and Piccoli [PS]. In this case, one typically requires that the value function be piecewise $\mathcal{C}^1$ and satisfy the Hamilton–Jacobi equations outside a finite or countable number of smooth manifolds $\mathcal{M}_i$. On the other hand, one can use the Crandall–Lions theory of viscosity solutions, and show that the value function is the unique solution of the Hamilton–Jacobi equation in the viscosity sense [BC].

None of these approaches is applicable in the present situation because of lack of regularity. Indeed, each player now has to solve an optimal control problem for a system whose dynamics (determined by the feedbacks used by all other players) is discontinuous. Our proof of optimality will strongly rely on the special structure of BV solutions of hyperbolic systems of conservation laws. In particular, we show that the solution has bounded directional variation along a cone Γ bounded away from all characteristic directions. As a consequence, the value functions V_i always admit a directional derivative in the directions of the cone Γ. For trajectories whose speed remains inside Γ, the optimality can thus be tested directly from the equations. An additional argument, using Clarke's generalized gradients [C], will rule out the optimality of trajectories whose speed falls outside the above cone of directions.

It is interesting to observe that the entropy admissibility conditions play no role in our analysis. For example, a solution of the system of conservation laws consisting of a single, nonentropic shock still determines a Nash equilibrium solution, provided that the amplitude of the shock is small enough. There is, however, a way to distinguish entropy solutions from all others that is also in the context of differential games. Indeed, entropy solutions are precisely the ones obtained as vanishing viscosity limits [BB]. They can thus be derived from a stochastic differential game of the form

$$dx = \sum_{i=1}^{n} f_i(x, u_i)\, dt + \varepsilon\, dw,$$

letting the white noise parameter $\varepsilon \to 0$. Here dw formally denotes the differential of a Brownian motion. For a discussion of stochastic differential games we refer to [F2].

In general, a weak solution of the hyperbolic system of conservation laws uniquely determines a family of discontinuous feedback controls $U_i^* = U_i^*(t,x)$. Inserting these controls in (1.1) we obtain the ODE

$$\dot{x} = \sum_{i=1}^{n} f_i(x, U_i^*(t,x)). \tag{1.10}$$

In spite of the right-hand side being discontinuous, we show that the solution of the Cauchy problem is unique and depends continuously on the initial data. Indeed, every trajectory of (1.10) crosses transversally all lines of discontinuity in the functions f_i. Because of the bound on the total variation, the uniqueness result in [B1] can thus be applied.

Our analysis will be concerned with the spatially homogeneous case, where the functions f_i, h_i do not depend on x. In the last section we shall see what results remain valid in the nonhomogeneous case, and discuss other possible extensions.

2. The basic framework. Consider an n-person differential game on the real line, having the special form

$$\dot{x} = f_0 + \sum_i u_i, \qquad x(\tau) = y. \tag{2.1}$$

Here the controls u_i can be any measurable, real valued functions, while f_0 is a fixed real number. The cost functionals take the form

$$J_i = J_i(\tau, y, u_1, \dots, u_n) = \int_\tau^T h_i(u_i(t))\,dt + g_i(x(T)). \tag{2.2}$$

To simplify the problem, for the time being we thus assume that the system has the simple dynamics (2.1) and that the running costs h_i do not depend on the space variable x. In section 6 we shall discuss how to extend the results to more general situations.

A key assumption, used throughout the paper, is that the cost functions h_i are smooth and strictly convex, with a positive second derivative

$$\frac{\partial^2}{\partial \omega^2} h_i(\omega) > 0 \tag{2.3}$$

at every point ω. Each player seeks a feedback strategy $u_i = U_i^*(t,x)$, which minimizes his own cost. Call $V_i = V_i(\tau, y)$ the value function for the ith player, and consider the spatial derivative $p_i = V_{i,x}$. The Hamiltonian functions H_i are defined as

$$H_i(p_1, \dots, p_n) = p_i \cdot \left(f_0 + \sum_j u_j^*(p_j) \right) + h_i(u_i^*(p_i)), \tag{2.4}$$

where the controls $u_j^* = u_j^*(p_j)$ provide the solutions to the following minimization problems:

$$p_j u_j^* + h_j(u_j^*) = \min_\omega \{p_j \cdot \omega + h_j(\omega)\} \doteq \phi_j(p_j). \tag{2.5}$$

At a point of minimum the first derivative vanishes. For every p_j we thus have

$$p_j + \frac{\partial h_j}{\partial u_j}\big(u^*(p_j)\big) = 0. \tag{2.6}$$

The Hamiltonian functions in (2.4) can also be written as

$$H_i(p_1,\dots,p_n) = p_i\left(f_0 + \sum_{j\neq i} u_j^*(p_j)\right) + \phi_i(p_i). \tag{2.7}$$

The corresponding Hamilton–Jacobi equation for V_i takes the form

$$V_{i,t} + H_i(V_{1,x},\dots,V_{n,x}) = 0. \tag{2.8}$$

To determine the value functions V_i, the above system has to be solved backward in time, with data given at the terminal time $t = T$:

$$V_i(T,x) = g_i(x), \quad i = 1,\dots,n. \tag{2.9}$$

In turn, the gradients $p_i \doteq V_{i,x}$ of the value functions satisfy the system of conservation laws

$$\frac{\partial}{\partial t}p_i + \frac{\partial}{\partial x}H_i(p_1,\dots,p_n) = 0 \tag{2.10}$$

with the terminal data

$$p_i(T,x) = g_i'(x). \tag{2.11}$$

Computing the Jacobian matrix $A(p)$ of this system, with entries $A_{ij} = \partial H_i/\partial p_j$, we find

$$A_{ii} = f_0 + \sum_j u_j^*(p_j) = \dot{x}, \tag{2.12}$$

$$A_{ij} = p_i\frac{\partial u_j^*(p_j)}{\partial p_j} = p_i\frac{\partial^2\phi_j}{\partial p_j^2}, \quad i \neq j. \tag{2.13}$$

Indeed, by (2.6) the functions $\phi_j \doteq p_j u_j^*(p_j) + h_j(u_j^*(p_j))$ defined at (2.5) satisfy

$$\frac{\partial\phi_j}{\partial p_j} = u_j^* + p_j\frac{\partial u_j^*}{\partial p_j} + \frac{\partial h_j}{\partial u_j}\frac{\partial u_j^*}{\partial p_j} = u_j^*. \tag{2.14}$$

It will be convenient to write second derivatives as $h_j'' \doteq \partial^2 h_j/\partial u_j^2$, $\phi_j'' \doteq \partial^2\phi_j/\partial p_j^2$. Differentiating w.r.t. p_j the identity (2.6), we find

$$1 + h_j''(u_j^*(p_j))\cdot\frac{\partial u_j^*}{\partial p_j}(p_j) = 0. \tag{2.15}$$

Using (2.14) and (2.15) one obtains

$$\phi_j''(p_j) = \frac{\partial u_j^*}{\partial p_j}(p_j) = -\frac{1}{h_j''(u_j^*(p_j))}. \tag{2.16}$$

Of course, this relation is well known from the theory of Legendre transforms.

3. Hyperbolicity conditions. In order that the Cauchy problem (2.10)–(2.11) be well posed, the system of conservation laws should be hyperbolic. It is thus important to determine in which cases the Jacobian matrix $A(p)$ has n distinct real eigenvalues. According to (2.12)–(2.13), this matrix takes the form

$$A(p) = \begin{pmatrix} \dot{x} & p_1\phi_2'' & p_1\phi_3'' & \cdots & p_1\phi_n'' \\ p_2\phi_1'' & \dot{x} & p_2\phi_3'' & \cdots & p_2\phi_n'' \\ p_3\phi_1'' & p_3\phi_2'' & \dot{x} & \cdots & p_2\phi_n'' \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_n\phi_1'' & p_n\phi_2'' & p_n\phi_3'' & \cdots & \dot{x} \end{pmatrix}. \tag{3.1}$$

We recall that, by (2.16) and (2.3), all second derivatives ϕ_j'' are strictly negative. The next lemma provides sufficient conditions on $p_1, \dots, p_n$ by which the system of conservation laws (2.10) is hyperbolic.

LEMMA 3.1. *Assume that all p_j have the same sign, i.e., either $p_j > 0$ for all j, or $p_j < 0$ for all j. Moreover, assume that there are no distinct indices $i \neq j \neq k$ such that*

$$p_i\phi_i'' = p_j\phi_j'' = p_k\phi_k''. \tag{3.2}$$

Then the matrix $A(p)$ in (3.1) *has n real distinct eigenvalues, and the system in* (2.10) *is strictly hyperbolic. Furthermore, all of these eigenvalues are different from $\dot{x}$.*

Proof. To fix the ideas, consider the case where $p_i > 0$ for all $i = 1, \dots, n$. The case where $p_i < 0$ is entirely similar.

Let $B \doteq A - \dot{x}I$, where I is the $n \times n$ identity matrix, and call $\lambda(A)$ the eigenvalues of a matrix A. Since

$$\lambda(B) = \lambda(A) - \dot{x}, \tag{3.3}$$

it suffices to show that B has n distinct real eigenvalues, all different from zero.

First we show that B has no zero eigenvalue. This is clear because

$$\begin{aligned} \det(B) &= \begin{vmatrix} 0 & p_1\phi_2'' & \cdots & p_1\phi_n'' \\ p_2\phi_1'' & 0 & \cdots & p_2\phi_n'' \\ \vdots & \vdots & \ddots & \vdots \\ p_n\phi_1'' & p_n\phi_2'' & \cdots & 0 \end{vmatrix} = \prod_{i=1}^{n} p_i\phi_i'' \cdot \begin{vmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{vmatrix} \\ &= (-1)^{n-1}(n-1)\prod_{i=1}^{n} p_i\phi_i'' \neq 0. \end{aligned}$$

As customary, the bars $|\cdot|$ around a matrix are used here to denote its determinant.

The eigenvalues of B are the zeros of the characteristic polynomial $\det(B - \lambda I)$. We observe that

$$\det(B - \lambda I) = \begin{vmatrix} -\dfrac{\lambda}{p_1\phi_1''} & 1 & \cdots & 1 \\ 1 & -\dfrac{\lambda}{p_2\phi_2''} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & -\dfrac{\lambda}{p_n\phi_n''} \end{vmatrix} \cdot \prod_{i=1}^{n} p_i\phi_i''.$$

By assumption, the numbers $c_i \doteq -p_i \phi_i''$ are all strictly positive. Up to a permutation of indices, which does not affect the determinant, we can assume that $0 < c_1 \leq c_2 \leq \cdots \leq c_n$. The polynomial $\det(B - \lambda I)$ has the same zeros as $q(\lambda) \doteq \det \widetilde{B}(\lambda)$, where

$$\widetilde{B}(\lambda) \doteq \begin{pmatrix} \frac{\lambda}{c_1} & 1 & \cdots & 1 \\ 1 & \frac{\lambda}{c_2} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & \frac{\lambda}{c_n} \end{pmatrix}. \tag{3.4}$$

The leading term of the polynomial $q(\lambda)$ is easily computed:

$$q(\lambda) = \prod_{i=1}^{n} \frac{\lambda}{c_i} + \mathcal{O}(1) \cdot \lambda^{n-1}.$$

For some constant $M > 0$ sufficiently large, we clearly have $\mathrm{sign}(q(\lambda)) = +1$ for $\lambda > M$ and $\mathrm{sign}(q(\lambda)) = (-1)^n$ for all $\lambda < -M$. Moreover, when $\lambda = 0$ we have $\mathrm{sign}(q(0)) = (-1)^{n-1}$.

Two cases need to be considered, depending on whether all c_i are distinct, or whether two of them coincide.

First case. Assume that all the c_i's are distinct, say $0 < c_1 < c_2 < \cdots < c_n$. Let us compute the determinant of $\widetilde{B}(\lambda)$ at the point $\lambda = c_i$. In this case, the ith row of $\widetilde{B}(\lambda)$ is identically 1, and we can subtract it from all the other rows, thus obtaining

$$\begin{aligned} q(\lambda) &= \begin{vmatrix} \frac{c_i}{c_1} & 1 & \cdots & 1 \\ 1 & \frac{c_i}{c_2} & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \ddots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & \frac{c_i}{c_n} \end{vmatrix} \quad \begin{matrix} \\ \\ \\ \leftarrow: i\text{th row} \\ \\ \\ \end{matrix} = \begin{vmatrix} \frac{c_i}{c_1} - 1 & 0 & \cdots & 0 \\ 0 & \frac{c_i}{c_2} - 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \ddots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{c_i}{c_n} - 1 \end{vmatrix} \\ &= \prod_{j \neq i} \left(\frac{c_i}{c_j} - 1 \right). \end{aligned}$$

Since

$$\mathrm{sign} \left(\frac{c_i}{c_j} - 1 \right) = \begin{cases} 1 & \text{if } i > j, \\ -1 & \text{if } i < j, \end{cases}$$

we conclude that

$$\mathrm{sign}(q(\lambda)) = (-1)^{n-i} \quad \text{when } \lambda = c_i.$$

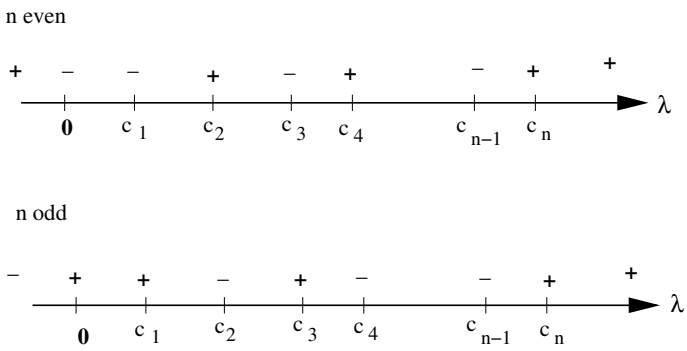


FIG. 3.1. *The sign of* $\det(\bar{B})(\lambda)$ *at various points. This shows the location of the real eigenvalues.*

As shown in Figure 3.1, the function $\lambda \mapsto q(\lambda)$ thus changes sign inside each of the intervals:

$$]-\infty, 0[\,,\ ]c_1, c_2[\,,\ \ldots\ ,\ ]c_{i-1}, c_i[\,,\ \ldots\ ,\ ]c_{n-1}, c_n[\,.$$

By continuity, there exist n distinct real zeros, with

$$\lambda_1 < 0 < c_1 < \lambda_2 < c_2 < \cdots < c_{i-1} < \lambda_i < c_i < \cdots < \lambda_n < c_n.$$

Notice that we must have $|\lambda_1| = \sum_{i=2}^{n} |\lambda_i|$ because the trace of $\widetilde{B}$ is zero.

Second case. Assume that two (but not three consecutive) numbers c_i coincide, say, $c_i = c_{i+1}$. We claim that the polynomial $q(\lambda)$ still has n distinct zeros, and the $(i+1)$th zero is $\lambda_{i+1} = c_i = c_{i+1}$.

When $\lambda = c_i = c_{i+1}$, the matrix $\widetilde{B}(\lambda)$ has both the ith and the $(i+1)$th rows identically equal to 1. Hence the determinant is zero. This shows that $\lambda_{i+1} = c_i = c_{i+1}$ is a zero of $q(\lambda)$.

To prove that it is a single root, we need to check that the derivative $q'(\lambda)$ does not vanish at $\lambda = c_i$. A direct computation yields

$$q'(\lambda) = \begin{vmatrix} \frac{1}{c_1} & 1 & \cdots & 1 \\ 0 & \frac{\lambda}{c_2} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & \frac{\lambda}{c_n} \end{vmatrix} + \begin{vmatrix} \frac{\lambda}{c_1} & 0 & \cdots & 1 \\ 1 & \frac{1}{c_2} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & \frac{\lambda}{c_n} \end{vmatrix} + \cdots + \begin{vmatrix} \frac{\lambda}{c_1} & 1 & \cdots & 0 \\ 1 & \frac{\lambda}{c_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & \frac{1}{c_n} \end{vmatrix}$$

$$\doteq \det B_1(\lambda) + \det B_2(\lambda) + \cdots + \det B_n(\lambda).$$

When $\lambda = c_i = c_{i+1}$ we have $\det B_j(\lambda) = 0$ for all $j \neq i, i+1$, because the ith and $(i+1)$th rows of the matrix B_j are identical (all entries are 1 except the jth entry which is 0). Moreover, $\det B_i(\lambda) = \det B_{i+1}(\lambda)$ because $B_i(\lambda)$ can be obtained from $B_{i+1}(\lambda)$ by first exchanging ith and $(i+1)$th rows and then exchanging ith and $(i+1)$th columns. Now we compute $\det B_i(\lambda)$. The entries of the $(i+1)$th row are all 1 except the ith entry which is zero. We subtract this row from all the other rows

and obtain

$$\det B_i(\lambda) = \begin{vmatrix} \frac{c_i}{c_1} & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cdots & \frac{1}{c_i} & 1 & \cdots & 1 \\ 1 & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 0 & 1 & \cdots & \frac{c_i}{c_n} \end{vmatrix} \begin{matrix} \\ \\ \leftarrow: i\text{th row}, \\ \leftarrow: (i+1)\text{th row}, \\ \\ \\ \end{matrix}$$

$$= \begin{vmatrix} \frac{c_i}{c_1} - 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \frac{1}{c_i} & 0 & \cdots & 0 \\ 1 & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \frac{c_i}{c_n} - 1 \end{vmatrix} \begin{matrix} \\ \\ \leftarrow: i\text{th row}, \\ \leftarrow: (i+1)\text{th row}, \\ \\ \\ \end{matrix}$$

$$= \frac{1}{c_i} \prod_{j \neq i, i-1} \left(\frac{c_i}{c_j} - 1 \right) \neq 0.$$

Observing that $q'(c_i) = 2 \cdot \det B_i(c_i) \neq 0$, we conclude that $\lambda = c_i$ is a single root.

It now remains to prove that we still have n distinct real zeros, i.e., that the coincidence of the two numbers c_i and c_{i+1} does not destroy any of the other sign changes in the polynomial $q(\lambda)$. In particular, there is still a zero inside each interval $]c_{i-1}, c_i[$ and $]c_{i+1}, c_{i+2}[$. From the previous computation we have that

$$\begin{aligned} \operatorname{sign}(q'(c_i)) &= (-1)^{n-i-1}, \\ \operatorname{sign}(q(c_{i-1})) &= (-1)^{n-i+1}, \\ \operatorname{sign}(q(c_{i+2})) &= (-1)^{n-i-2}. \end{aligned}$$

Therefore, $q'(c_i)$ and $q(c_{i-1})$ have the same sign, while $q'(c_i)$ and $q(c_{i+2})$ have opposite signs. Looking at Figure 3.2 it is clear that there is one root within the interval $]c_{i-1}, c_i[$, another inside the interval $]c_{i+1}, c_{i+2}[$, while $\lambda = c_i = c_{i+1}$ is still another root.

In the special case $i = 1$ (or $i = n - 1$), it can be checked in the same way that $c_1 = c_2$ (or $c_{n-1} = c_n$) is a single root, while another root lies in the interval $]c_2, c_3[$ (or in the interval $]c_{n-2}, c_{n-1}[$, respectively).

At last, we need to consider the case where more than one couple of numbers c_i, c_{i+1} coincide. We claim that, in all cases, the polynomial $q(\lambda)$ still has n distinct roots. Indeed, in the case where the two coinciding pairs $c_i = c_{i+1}$, $c_j = c_{j+1}$ are not adjacent (i.e., $j \neq i + 2$), the previous analysis applies. On the other hand, in the case where, say, $c_{i-2} = c_{i-1} < c_i = c_{i+1}$, the analysis of the signs of $q'(c_{i-1})$, $q'(c_i)$, $q(c_{i+2})$, and $q(c_{i-3})$ (see Figure 3.3) yields the desired results.

The proof of Lemma 3.1 is now completed. □

Remark 1. If three or more of the numbers c_j coincide, say, $c_{i-1} = c_i = c_{i+1}$, then c_i becomes a multiple zero of $\det(B - \lambda I)$. In this case, the system of conservation laws will still be hyperbolic, but no longer strictly hyperbolic.

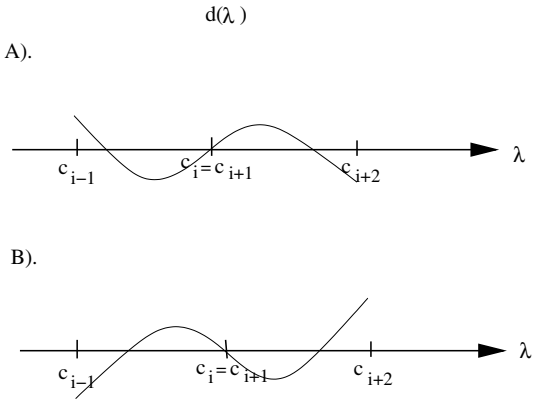


FIG. 3.2. *Checking that $\lambda = c_i = c_{i+1}$ is a single root.*

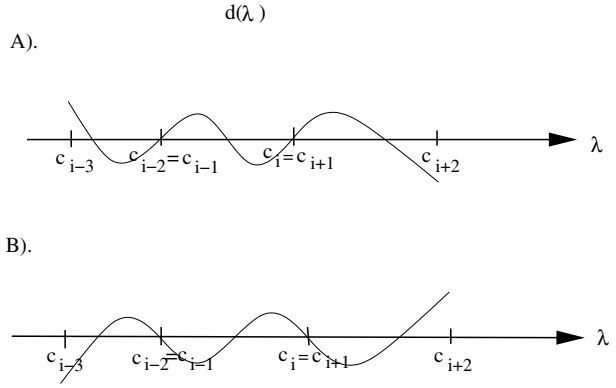


FIG. 3.3. *Two coinciding pairs next to each other.*

Remark 2. In the case of 2×2 systems, the condition $p_1 p_2 \geq 0$ is necessary for the hyperbolicity of the system. However, when $n \geq 3$, the system (2.10) can be strictly hyperbolic also at points where $p_1 < 0 < p_2 < p_3$. For example, let $n = 3$, $\phi_i'' = -1$ for all i, $p_1 = -1$, $p_2 = 5$, and $p_3 = 20$. Then the characteristic polynomial $q(\lambda) = \det \bar{B}(\lambda)$ is

$$q(\lambda) = \lambda^3 - 75\lambda + 200.$$

One can easily check that

$$q(-10) = -50, \quad q(0) = 200, \quad q(5) = -50, \quad q(10) = 450.$$

Therefore, there are three distinct real eigenvalues:

$$\lambda_1 \in]-10, 0[, \quad \lambda_2 \in]0, 5[, \quad \lambda_3 \in]5, 10[.$$

4. Review of hyperbolic systems and discontinuous ODE. In this section we collect some results on hyperbolic conservation laws and discontinuous ODEs, which will be used in what follows. Consider the Cauchy problem for a system of

conservation laws

$$v_t + F(v)_x = 0, \qquad v(0,x) = \bar v(x). \tag{4.1}$$

In the case where the system is strictly hyperbolic, the global existence of weak solutions with small BV initial data is well known.

PROPOSITION 4.1. *Assume that the flux function $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ is smooth and that, at some point v^*, the Jacobian matrix $A(v^*) = DF(v^*)$ has n real distinct eigenvalues. Then there exists $\delta > 0$ for which the following holds. If*

$$\|\bar v(\cdot) - v^*\|_{\mathbf{L}^\infty} < \delta, \qquad \textit{Tot. Var.}\{\bar v\} < \delta, \tag{4.2}$$

then the Cauchy problem (4.1) *admits a unique entropy weak solution $v = v(t,x)$ defined for all $t \geq 0$, obtained as the limit of vanishing viscosity approximations.*

In the case where each characteristic field is either linearly degenerate or genuinely nonlinear, the existence of a global weak solution was proved by Glimm [G]. The more general case was later covered in [L, ILF] using the Glimm scheme and in [AM] using wave-front tracking. The convergence of vanishing viscosity approximations was recently proved in [BB], together with the uniqueness and Lipschitz continuous dependence of solutions on the initial data in the $\mathbf{L}^1$ distance. We remark that, for each time t, the function $v(t,\cdot)$ has small total variation. Its pointwise values can be uniquely assigned by the convention

$$v(t,x) = \lim_{y \to x+} v(t,y). \tag{4.3}$$

For applications to game theory, we shall need some additional properties of these weak solutions. By assumption, the matrix $A(v^*)$ has distinct eigenvalues $\lambda_1^* < \lambda_2^* < \cdots < \lambda_n^*$. By continuity, there exists $\varepsilon > 0$ such that, for all v in the ε-neighborhood

$$\Omega_\varepsilon^* \doteq \{v;\ |v - v^*| \leq \varepsilon\},$$

the characteristic speeds range inside disjoint intervals

$$\lambda_j(v) \in [\lambda_j^-, \lambda_j^+]. \tag{4.4}$$

Moreover, if $v^-, v^+ \in \Omega_\varepsilon^*$ are two states connected by a j-shock, the speed $\lambda_j(v^-, v^+)$ of the shock remains inside the interval $[\lambda_j^-, \lambda_j^+]$.

Now consider an open cone of the form

$$\Gamma \doteq \{(t,x);\ t > 0,\ a < x/t < b\}. \tag{4.5}$$

Following [B1] we define the *directional variation* of the function $(t,x) \mapsto v(t,x)$ along the cone Γ as

$$\sup \left\{ \sum_{i=1}^N |v(t_i,x_i) - v(t_{i-1},x_{i-1})| \right\}, \tag{4.6}$$

where the supremum is taken over all finite sequences $(t_0,x_0), (t_1,x_1), \ldots, (t_N,x_N)$ such that

$$(t_i - t_{i-1}, x_i - x_{i-1}) \in \Gamma \quad \text{for every } i = 1, \ldots, N. \tag{4.7}$$

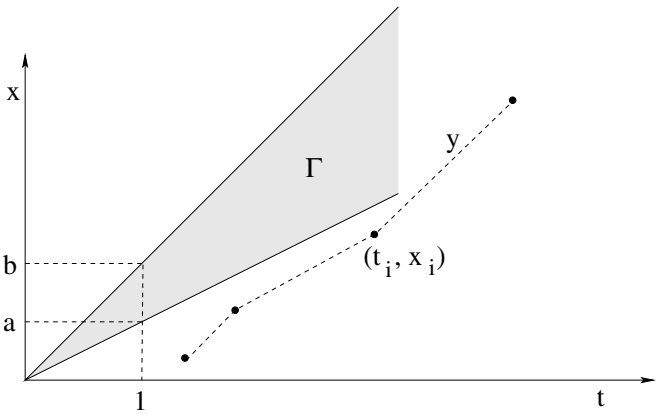


FIG. 4.1. *Directional variation along the cone* Γ.

(See Figure 4.1.) We now show that the weak solution $v = v(t,x)$ has bounded directional variation along a suitable cone Γ.

LEMMA 4.2. *Let $v = v(t,x)$ be an entropy weak solution of* (4.1) *taking values inside the domain Ω^*_ε. Assume that $\lambda^+_{k-1} < a < b < \lambda^-_k$ for some k. Then v has bounded directional variation along the cone Γ in* (4.5).

Proof. Fix any finite sequence of points (t_i, x_i), $i = 0, \dots, N$, satisfying (4.7). It is not restrictive to assume that $t_0 = 0$. Call $T \doteq t_N$ and define $y : [0,T] \mapsto \mathbb{R}$ as the polygonal line with nodes at the points (t_i, x_i) (Figure 4.1). Clearly y is almost everywhere differentiable, with $\dot y(t) \in]a,b[$. From the theory of conservation laws, it is well known that the entropy weak solution v can be obtained as the limit of a sequence of front tracking approximate solutions v_ν. For each $\nu \geq 1$, one can derive a uniform bound on the total variation of the map $t \mapsto v_\nu(t, y(t))$. Indeed, call $V^y(t)$ the total strength of all wave-fronts in $v_\nu(t,\cdot)$ approaching $y(t)$ at time t, i.e.,

$$V^y(t) \doteq \sum_{\alpha \in \mathcal{A}(y)} |\sigma_\alpha| \,. \tag{4.8}$$

Here σ_α denotes the strength of the wave-front in $v_\nu(t,\cdot)$ located at x_α. Observing that $\lambda^+_k < \dot y < \lambda^-_{k+1}$, the above summation will include the following fronts:

- the fronts of a family $k_\alpha \leq k$ located at a point $x_\alpha > y$, and
- the fronts of a family $k_\alpha > k$ located at a point $x_\alpha < y$.

We now call

$$Q(t) = \sum_{\alpha,\beta \in \mathcal{A}} |\sigma_\alpha||\sigma_\beta| \tag{4.9}$$

the *interaction potential* of $v_\nu(t,\cdot)$, i.e., the sum of products of all couples of approaching waves in $v_\nu(t,\cdot)$. Assuming that the total variation of the solution remains small, for some constant C_0 the positive functional

$$\Upsilon(t) \doteq V^y(t) + C_0 Q(t)$$

is nonincreasing in time. Moreover, at each time τ where a wave-front of strength σ_α crosses $y(\cdot)$, we have

$$\Upsilon(\tau+) - \Upsilon(\tau-) = -|\sigma_\alpha|.$$

Therefore, the total strength of all wave-fronts in v_ν which cross the polygonal line $y(\cdot)$ is bounded by

$$V^y(0) + C_0 Q(0) = \mathcal{O}(1) \cdot \text{Tot.Var.}\{\bar v\}.$$

This proves that the total variation of the maps $t \mapsto v_\nu\big(t, y(t)\big)$ is uniformly bounded for all $\nu \geq 1$. To get the desired estimate for the solution v, we now let $\nu \to \infty$. If we have the pointwise convergence $v_\nu(t_i, x_i) \to v(t_i, x_i)$ for every $i = 0, \ldots, N$, we can immediately conclude

$$\begin{aligned}\sum_{i=1}^N |v(t_i, x_i) - v(t_{i-1}, x_{i-1})| &\leq \limsup_{\nu\to\infty} |v_\nu(t_i, x_i) - v_\nu(t_{i-1}, x_{i-1})| \\ &= \mathcal{O}(1) \cdot \text{Tot.Var.}\{\bar v\},\end{aligned}$$

proving our claim. However, if v is discontinuous at some point (t_i, x_i), the pointwise convergence may not hold. To achieve the result also in the general case we observe that, for each time τ, we have the convergence $v_\nu(\tau, x) \to v(\tau, x)$ for a.e. $x \in \mathbb{R}$. Using the right continuity of the functions $v(t_i, \cdot)$, we can find points x_i' sufficiently close to x_i such that

$$|v(t_i, x_i') - v(t_i, x_i)| < 1/N, \qquad (t_i - t_{i-1}, x_i' - x_{i-1}') \in \Gamma,$$

and such that $v_\nu(t_i, x_i') \to v(t_i, x_i')$ for every i. This yields the estimate

$$\begin{aligned}&\sum_{i=1}^N |v(t_i, x_i) - v(t_{i-1}, x_{i-1})| \\ &\quad\leq \sum_{i=1}^N \{|v(t_i, x_i') - v(t_{i-1}, x_{i-1}')| + |v(t_i, x_i') - v(t_{i-1}, x_i)| \\ &\qquad\qquad + |v(t_i, x_{i-1}') - v(t_{i-1}, x_{i-1})|\} \\ &\quad\leq 2 + \limsup_{\nu\to\infty} \sum_{i=1}^N |v_\nu(t_i, x_i) - v_\nu(t_{i-1}, x_{i-1})| \\ &\quad= 2 + \mathcal{O}(1) \cdot \text{Tot.Var.}\{\bar v\},\end{aligned}$$

proving the lemma. □

Together with Γ we now consider a strictly smaller cone, say,

$$\Gamma' \doteq \{(t, x);\ t > 0,\ a' < x/t < b'\}, \tag{4.10}$$

with $a < a' < b' < b$. A standard theorem in real analysis states that a BV function of one real variable admits left and right limits at every point. We now prove an analogous result for functions with bounded directional variation.

LEMMA 4.3. *Let $v = v(t, x)$ be a function with bounded directional variation along the cone Γ in (4.5), and consider the smaller cone $\Gamma' \subset \Gamma$ in (4.10), with $a < a' < b' < b$. Then at every point $P = (t, x)$ there exist the directional limits*

$$v^+(P) \doteq \lim_{Q\to P,\ Q-P\in\Gamma'} v(Q), \qquad v^-(P) \doteq \lim_{Q\to P,\ P-Q\in\Gamma'} v(Q). \tag{4.11}$$

Proof. If the first limit does not exist, we can find two sequences $Q'_\nu \to P$, $Q''_\nu \to P$ with $Q'_\nu - P \in \Gamma'$, $Q''_\nu - P \in \Gamma'$ for every $\nu \geq 1$, along which the function v converges to distinct limits:

$$v(Q'_\nu) \to v', \qquad v(Q''_\nu) \to v'',$$

with $v' \neq v''$. Since Γ' is strictly smaller than Γ, by induction we can select two subsequences

$$Q'_{\nu(1)}, Q'_{\nu(2)}, \ldots, \qquad Q''_{\nu(1)}, Q''_{\nu(2)}, \ldots,$$

such that

$$Q'_{\nu(j)} - Q''_{\nu(j)} \in \Gamma, \qquad Q''_{\nu(j)} - Q'_{\nu(j+1)} \in \Gamma$$

for every j. In this case

$$\lim_{N\to\infty} \sum_{i=1}^{N} |v(Q'_{\nu(j)}) - v(Q''_{\nu(j)})| = \infty,$$

in contrast with the assumption of bounded directional variation. This proves the existence of the first limit in (4.11). The second one is entirely similar. □

Next, we recall some results on differential equations with discontinuous right-hand sides. Let $f = f(t,x)$ be a bounded function. By a *Carathéodory solution* of the ODE

$$\dot{x}(t) = f(t, x(t)), \tag{4.12}$$

we mean an absolutely continuous function $t \mapsto x(t)$ which satisfies the equation (4.12) at a.e. time t.

In the case where f is discontinuous, it is well known that the Cauchy problem may have no Carathéodory solutions. One can then relax the concept of solution, introducing multivalued regularizations of f. For example, consider the multifunction

$$F(t,x) \doteq \bigcap_{\varepsilon>0} \overline{co}\{f(s,y);\ |s-t| \leq \varepsilon,\ |y-x| \leq \varepsilon\}, \tag{4.13}$$

where $\overline{co}$ denotes the convex closure of a set. Following [H], by a *Krasovskii solution* of (4.12) we mean an absolutely continuous function $t \mapsto x(t)$ which satisfies the differential inclusion

$$\dot{x}(t) \in F(t, x(t)) \tag{4.14}$$

at a.e. time t. Another concept of solution, proposed by Filippov, relates to the multifunction

$$F^*(t,x) \doteq \bigcap_{\varepsilon>0} \bigcap_{\mathrm{meas}(\mathcal{N})=0} \overline{co}\{f(s,y);\ |s-t| \leq \varepsilon,\ |y-x| \leq \varepsilon,\ (s,y) \notin \mathcal{N}\}, \tag{4.15}$$

obtained as in (4.13), neglecting the behavior of f on sets of measure zero. An absolutely continuous function $t \mapsto x(t)$ which satisfies almost everywhere the differential inclusion

$$\dot{x}(t) \in F^*(t, x(t)) \tag{4.16}$$

is called a *Filippov solution* of (4.12). Notice that $F^* \subseteq F$. Moreover, the multifunction $F^*(t,x)$ is not affected if the function f is modified on sets of measure zero.

Under the only assumption that f is bounded, it is well known that the multifunctions F, F^* are both upper semicontinuous, with compact convex values [AC]. Hence the Cauchy problem

$$\dot{x}(t) = f(t, x(t)), \qquad x(s) = y \tag{4.17}$$

admits at least one solution, according to the definitions of Filippov and of Krasovskii. In the case where the function f has directionally bounded variation, a much stronger result can be proved.

LEMMA 4.4. *Assume that the function f has bounded directional variation along the cone Γ in* (4.5). *Moreover, assume that*

$$a < a' \leq f(t,x) \leq b' < b$$

for all t, x. Then the Cauchy problem (4.17) *has a unique Carathéodory solution, Lipschitz continuously depending on the initial data (s,y). Such a solution is also the unique Krasovskii and Filippov solution of the Cauchy problem.*

Proof. The existence, uniqueness, and continuous dependence of the Carathéodory solution was proved in [B1]. For directionally continuous vector fields, the equivalence between Carathéodory, Filippov, and Krasovskii solutions was shown in [B2, p. 26]. □

We conclude this section by proving a simple result from nonsmooth analysis.

LEMMA 4.5. *Consider a Lipschitz continuous function $V = V(t,x)$ and call $(\phi, \psi) \doteq (V_t, V_x)$ its partial derivatives, defined at a.e. point (t,x). Let Γ be the cone at* (4.5). *Assume that, at a given point $(\bar{t}, \bar{x})$, there exists the directional limit*

$$(\bar{\phi}, \bar{\psi}) = \lim_{\substack{(t,x)\to(\bar{t},\bar{x}) \\ (t-\bar{t},x-\bar{x})\in\Gamma}} (\phi(t,x), \psi(t,x))\,. \tag{4.18}$$

Moreover, consider a continuous function $t \mapsto x(t)$ which is differentiable at $t = \bar{t}$ and assume

$$x(\bar{t}) = \bar{x}, \qquad \dot{x}(\bar{t}) \in \,]a, b[.$$

Then the composite function admits the one-sided derivative

$$\lim_{h\to 0+} \frac{V(\bar{t}+h, x(\bar{t}+h)) - V(\bar{t},\bar{x})}{h} = \bar{\phi} + \bar{\psi}\cdot\dot{x}(\bar{t}). \tag{4.19}$$

Proof. From the theory of generalized gradients [C] it follows that, for $h > 0$ small,

$$\begin{aligned} V(\bar{t}+h, x(\bar{t}+h)) - V(\bar{t},\bar{x}) \in\; & h\cdot\overline{co}\{\phi(t,x);\ \bar{t} < t < t+h,\ (t-\bar{t}, x-\bar{x})\in\Gamma\} \\ & + h\cdot\overline{co}\{\psi(t,x);\ \bar{t} < t < t+h,\ (t-\bar{t}, x-\bar{x})\in\Gamma\}\cdot[x(t)-\bar{x}]. \end{aligned}$$

Letting $h \to 0+$ and using (4.18) one obtains (4.19). □

5. Optimal feedback strategies. The analysis in section 3 has identified conditions which ensure that the system of conservation laws (2.10) is strictly hyperbolic in a neighborhood of a point $p = (p_1, p_2, \ldots, p_n)$. In this case, assuming that the terminal condition (2.11) has small total variation, one can apply Glimm's theorem and obtain the global existence of a weak solution. We shall now prove that the components of this solution determine a family of feedback strategies $u_i = U_i^*(t, x)$, which provide a Nash equilibrium solution to the noncooperative differential game.

THEOREM 5.1. *Consider the differential game* (2.1)–(2.2)*, where the cost functions* h_i *are smooth and satisfy the convexity assumption* (2.3)*. In connection with the functions* ϕ_j *at* (2.5)*, let* $p^* \doteq (p_1^*, \ldots, p_n^*)$ *be a point where the assumptions of Lemma* 3.1 *are satisfied. Then there exists* $\delta > 0$ *such that the following holds. If*

$$\|g_i' - p_i^*\|_{\mathbf{L}^\infty} < \delta, \qquad \textit{Tot. Var.}\{g_i'(\cdot)\} < \delta, \tag{5.1}$$

then for any $T > 0$ *the terminal value problem* (2.10)–(2.11) *has a weak solution* $p : [0, T] \times \mathbb{R} \mapsto \mathbb{R}^n$*. The (possibly discontinuous) feedback controls* $U_j^*(t, x) \doteq u_j^*(p(t, x))$ *defined at* (2.5) *provide a Nash equilibrium solution to the differential game. The trajectories* $t \mapsto x(t)$ *are Lipschitz continuous functions of the initial data* (τ, y)*.*

Proof. The proof will be given in several steps.

Step 1. By the assumptions, the system of conservation laws

$$\frac{\partial}{\partial t} v_i - \frac{\partial}{\partial x} H_i(v_1, \ldots, v_n) = 0 \tag{5.2}$$

is strictly hyperbolic in a neighborhood of the point p^*. Given the initial data $v_i(0, x) = g_i'(x)$ with sufficiently small total variation, by Proposition 4.1 the Cauchy problem admits a weak solution $v = v(t, x)$, defined for all $t \geq 0$. Reversing time, we thus obtain a weak solution $p(t, x) \doteq v(T - t,\, x)$ of the terminal value problem (2.10)–(2.11). For each time t, the map $x \mapsto p(t, x)$ has small total variation. Its pointwise values can be uniquely assigned by the convention

$$p(t, x) = \lim_{y \to x+} p(t, y). \tag{5.3}$$

Step 2. By strict hyperbolicity and continuity, there exists $\varepsilon > 0$ such that, for all p in the ε-neighborhood

$$\Omega_\varepsilon^* \doteq \{p;\ |p - p^*| \leq \varepsilon\} \subset \mathbb{R}^n, \tag{5.4}$$

the following holds. The characteristic speeds for (2.10) range inside disjoint intervals

$$\lambda_j(p) \in [\lambda_j^-, \lambda_j^+]. \tag{5.5}$$

Moreover, the speed $\dot{x}$ at (2.1) remains bounded away from all characteristic speeds. Namely, there exists an index $k \in \{1, \ldots, n-1\}$ and numbers $\epsilon > 0$ and $a < b$ such that

$$\lambda_k^+ < a - \epsilon < b + \epsilon < \lambda_{k+1}^- \tag{5.6}$$

and

$$f_0 + \sum_j u_j^*(p_j) \in [a + \epsilon, b - \epsilon] \tag{5.7}$$

whenever $p \in \Omega^*_\varepsilon$.

Together with the cone Γ at (4.5) we now define

$$\Gamma^+_\epsilon \doteq \{(t,x);\ t>0,\ a-\epsilon \leq x/t \leq b+\epsilon\},$$
$$\Gamma^-_\epsilon \doteq \{(t,x);\ t>0,\ a+\epsilon \leq x/t \leq b-\epsilon\}.$$

Clearly, $\Gamma^-_\epsilon \subset \Gamma \subset \Gamma^+_\epsilon$. By Lemma 4.2, each $p_i = p_i(t,x)$ has bounded variation in the direction of the cone Γ. By the assumptions, the maps $p_j \mapsto u^*_j(p_j)$ in (2.5) are locally Lipschitz continuous. Hence, for $i = 1, \ldots, n$, all the composed maps $(t,x) \mapsto u^*_i(p_i(t,x))$ also have bounded directional variation along the cone Γ. By (5.7) we can thus apply Lemma 4.4, showing that the Cauchy problem for the ODE

$$\dot x(t) = f_0 + \sum_j u^*_j(p_j(t,x)) \tag{5.8}$$

has a unique Carathéodory (equivalently, Filippov or Krasovskii) solution, depending Lipschitz continuously on initial data (τ, y) in (1.2).

Step 3. We now construct the value functions V_i, corresponding to the feedback strategies $U^*_j(t,x) \doteq u^*_j(p_j(t,x))$. For $j = 1, \ldots, n$, define the cost functions

$$h^*_j(t,x) \doteq h_j(u^*_j(p_j(t,x))).$$

Given a point (τ, ξ), let $t \mapsto x(t;\tau,\xi)$ be the trajectory of (5.8) passing through (τ,ξ). For each $i = 1, \ldots, n$ we define

$$V_i(\tau,\xi) \doteq \int_\tau^T h^*_i(t, x(t;\tau,\xi))\, dt + g_i(x(T;\tau,\xi)). \tag{5.9}$$

By the same arguments as in [B1] one can show that the functions $V_i = V_i(t,x)$ are Lipschitz continuous, and hence almost everywhere differentiable. At every point where the differential exists, by construction one has

$$\frac{\partial V_i}{\partial t} + \dot x \cdot \frac{\partial V_i}{\partial x} = -h^*_i(t,x). \tag{5.10}$$

To derive further properties of the gradient of V_i, fix a time τ and any two points $\xi' < \xi''$. We claim that

$$V_i(\tau,\xi'') - V_i(\tau,\xi') = \int_{\xi'}^{\xi''} p_i(\tau,x)\, dx. \tag{5.11}$$

Indeed, let $x'(\cdot)$ and $x''(\cdot)$ be the two trajectories of (5.8) which start from the initial points $x'(\tau) = \xi'$ and $x''(\tau) = \xi''$, respectively. Consider the region $\Delta \subset \mathbb{R}^2$ defined as

$$\Delta \doteq \{(t,x);\ \tau \leq t \leq T,\ x'(t) \leq x \leq x''(t)\}.$$

Applying the divergence theorem to the vector field $\mathbf{v} \doteq (p_i, H_i)$ on the domain Δ and using the conservation equation (2.10), we obtain

$$\begin{aligned}\int_{x'(T)}^{x''(T)} p_i(T,x)\, dx = &\int_{\xi'}^{\xi''} p_i(\tau,x)\, dx + \int_\tau^T \{[p_i \cdot \dot x' + h^*_i(t,x'(t))] - p_i \cdot \dot x'\}\, dt \\ &- \int_\tau^T \{[p_i \cdot \dot x'' + h^*_i(t,x''(t))] - p_i \cdot \dot x''\}\, dt.\end{aligned}$$

Observing that $p_i(T,x) = g_{i,x}(x) = V_{i,x}(T,x)$, we conclude

$$g_i(x''(T)) - g_i(x'(T)) = \int_{\xi'}^{\xi''} p_i(\tau,x)\,dx + \int_\tau^T [h_i^*(t,x'(t)) - h_i^*(t,x''(t))]\,dt.$$

The above two equalities yield (5.11). Since ξ', ξ'' are arbitrary, this in turn implies

$$V_{i,x} = p_i \tag{5.12}$$

at a.e. point (t,x). Together with (5.10), this yields

$$V_{i,t} = -p_i \cdot \left(f_0 + \sum_j u_j^*(p_j) \right) - h_i^*. \tag{5.13}$$

Therefore, the value functions $(V_1, \ldots, V_n)$ satisfy almost everywhere the system of Hamilton–Jacobi equations

$$V_{i,t} + V_{i,x} \cdot \left(f_0 + \sum_j u_j^*(V_{j,x}) \right) + h_i(u_i^*(V_{i,x})) = 0. \tag{5.14}$$

We recall that $u_j^* = u_j^*(p_j)$ are the optimal control values defined at (2.5).

Step 4. We now conclude the proof, showing that the feedback strategies $U_j^*(t,x) \doteq u_j^*(p_j(t,x))$ represent a Nash equilibrium solution. Fix an index $i \in \{1, \ldots, n\}$ and consider the optimal control problem for the ith player:

$$\min_{z(\cdot)} \left\{ \int_\tau^T h_i(z(t))\,dt + g_i(T, x(T)) \right\}, \tag{5.15}$$

$$\dot{x}(t) = f_0 + \sum_{j \neq i} U_j^*(t,x) + z(t), \qquad x(\tau) = y. \tag{5.16}$$

We claim that the minimum cost is precisely $V_i(\tau,y)$. Consider any absolutely continuous trajectory $x : [\tau,T] \mapsto \mathbb{R}$ with $x(\tau) = y$. It suffices to show that, for a.e. $t \in [\tau,T]$,

$$\frac{d}{dt} V_i(t,x(t)) \geq -h_i(z(t)), \tag{5.17}$$

where the control function $z(\cdot)$ implemented by the ith player is

$$z(t) \doteq \dot{x}(t) - f_0 - \sum_{j \neq i} U_j^*(t,x).$$

Indeed, if (5.17) holds, then

$$\int_\tau^T h_i(z(t))\,dt + g_i(T,x(T)) \geq \int_\tau^T \left\{ -\frac{d}{dt} V_i(t,x(t)) \right\} dt + V_i(T,x(T)) = V_i(\tau,y),$$

as claimed.

We now give a proof of (5.17), assuming that the total variation of the functions $g_j'(\cdot)$ is sufficiently small. In connection with the vector $p^* = (p_1^*, \ldots, p_n^*)$ considered in (5.1), define the constant controls

$$\omega_j^* \doteq u_j^*(p_j^*).$$

Moreover, recalling (5.13), set

$$q_i^* \doteq -p_i^* \cdot \left(f_0 + \sum_j \omega_j^* \right) - h_i(\omega_i^*).$$

Choose $\varepsilon_1 > 0$ small enough so that, if $|u_j - \omega_j^*| \leq \varepsilon_1$ for all j, then

$$\dot{x} = f_0 + \sum_j u_j \in [a + \epsilon, b - \epsilon]. \tag{5.18}$$

Observe that our definitions imply

$$q_i^* + p_i^* \cdot \left(f_0 + \sum_{j \neq i} \omega_j^* \right) + \min_\omega \{ p_i^* \omega + h_i(\omega) \} = 0.$$

By the strict convexity of the cost function h_i at (2.5), there exists $\delta' > 0$ such that

$$q_i^* + p_i^* \cdot \left(f_0 + \sum_{j \neq i} \omega_j^* \right) + p_i^* \omega + h_i(\omega) > \delta'$$

whenever $|\omega - \omega_i^*| \geq \varepsilon_1$. By continuity, there exists $\varepsilon_2 > 0$ such that, if

$$|q_j - q_j^*| \leq \varepsilon_2, \quad |p_j - p_j^*| \leq \varepsilon_2, \quad j = 1, \ldots, n, \tag{5.19}$$

$$|\omega - \omega_i^*| \geq \varepsilon_1, \quad |u_j - \omega_j^*| \leq \varepsilon_2, \quad j \neq i, \tag{5.20}$$

then

$$q_i + p_i \cdot \left(f_0 + \sum_{j \neq i} u_j \right) + p_i \omega + h_i(\omega) > 0. \tag{5.21}$$

Choosing $\delta > 0$ in (5.1) sufficiently small, we can assume that the partial derivatives $p_j = V_{j,x}$, $q_j = V_{j,t}$ satisfy almost everywhere all the bounds in (5.19). Moreover, for $j \neq i$, the functions $u_j = u_j^*(p_j)$ satisfy the bounds in (5.20). Observe that a.e. time $\bar{t} \in [0, T]$ is in the Lebesgue set of all three measurable functions

$$z(t), \qquad \frac{d}{dt} x(t), \qquad \frac{d}{dt} V_i(t, x(t))$$

(see [Fo, p. 92]). Choose any such Lebesgue point $\bar{t}$ and call $\bar{x} = x(\bar{t})$. To prove (5.17) we consider two alternatives.

Case 1. $|z(\bar{t}) - \omega_i^*| \leq \varepsilon_1$. In this case (5.18) holds. Define the "one-sided" partial derivatives of V_i at $(\bar{t}, \bar{x})$:

$$(\bar{\phi}_i, \bar{\psi}_i) = \lim_{\substack{(t,x) \to (\bar{t}, \bar{x}) \\ (t - \bar{t}, x - \bar{x}) \in \Gamma}} (V_{i,t}(t,x), V_{i,x}(t,x)). \tag{5.22}$$

Notice that these directional limits exist because of (5.12)–(5.13) and the directional continuity of all functions p_j. Since (5.14) holds almost everywhere, we have

$$\bar{\phi}_i + \bar{\psi}_i \left(f_0 + \sum_{j \neq i} u_j^*(\bar{\psi}_j) \right) + \min_\omega \{ \bar{\psi}_i \omega + h_i(\omega) \} = 0. \tag{5.23}$$

By the assumptions, the function $t \mapsto V_i(t, x(t))$ is differentiable at $t = \bar{t}$. Its derivative can be computed by taking the one-sided limit in (4.19). Using Lemma 4.5 together with (5.23) we obtain

$$\begin{aligned}\frac{d}{dt} V_i(t, x(t))\bigg|_{t=\bar{t}} &= \bar{\phi}_i + \bar{\psi}_i \dot{x}(t) \\ &= -h_i(z(t)) + \{\bar{\psi}_i z(t) + h_i(z(t))\} - \min_{\omega}\{\bar{\psi}_i\, \omega + h_i(\omega)\} \\ &\geq -h_i(z(t)).\end{aligned}$$

Hence (5.17) holds.

Case 2. $|z(\bar{t}) - \omega_i^*| > \varepsilon_1$. In this case, by a nonsmooth version of the chain rule [C], there exist numbers

$$\begin{aligned}\phi_i &\in \overline{co}\{V_{i,t}(t,x);\ \ t \in [0,T],\ \ x \in \mathbb{R}\}, \\ \psi_i &\in \overline{co}\{V_{i,x}(t,x);\ \ t \in [0,T],\ \ x \in \mathbb{R}\}\end{aligned}$$

such that, at $t = \bar{t}$,

$$\frac{d}{dt} V_i(t, x(t))\bigg|_{t=\bar{t}} = \phi_i + \psi_i \dot{x}(\bar{t}). \tag{5.24}$$

The previous assumptions now imply

$$|\phi_i - q_j^*| \leq \varepsilon_2, \qquad |\psi_i - p_i^*| \leq \varepsilon_2.$$

Hence, by (5.21),

$$\frac{d}{dt} V_i(t, x(t))\bigg|_{t=\bar{t}} = \phi_i + \psi_i \left(f_0 + \sum_{j \neq i} u_j^*(p_j) \right) + \psi_i z(\bar{t}) > -h_i(z(\bar{t})), \tag{5.25}$$

showing that (5.17) holds also in this case. This completes the proof of Theorem 5.1. We remark that, if the other players adopt the feedback strategies $u_j = U_j^*(t,x)$, the choice $u_i = U_i^*(t,x)$ is the unique optimal strategy for the ith player. □

6. Concluding remarks. In this final section we point out some possible extensions of our previous results. Consider a differential game with the more general form

$$\dot{x} = f_0 + \sum_{i=1}^{n} f_i(\tilde{u}_i) \tag{6.1}$$

and cost functionals

$$J_i \doteq \int_{\tau}^{T} \tilde{h}_i(\tilde{u}_i(t))\, dt + g_i(x(T)). \tag{6.2}$$

Assume that each f_i is a homeomorphism from a (possibly unbounded) open interval $]a_i, b_i[$ into $\mathbb{R}$, with a smooth inverse $f_i^{-1} : \mathbb{R} \mapsto]a_i, b_i[$. Then the reparametrization of the control functions $u_i \doteq f_i(\tilde{u}_i)$ puts the system (6.1)–(6.2) in the standard form (2.1)–(2.2), with $h_i(\omega) \doteq \tilde{h}_i(f_i^{-1}(\omega))$. Of course, the key assumption (2.3) must now be carefully checked.

If the functions f_i in (6.1) and the running costs h_i in (6.2) also depend on x, then the corresponding system of conservation laws (1.9) will also depend on the space variable x. Assuming that the system is strictly hyperbolic, one can then use the results in [DH], and obtain the local existence of weak solutions, on a time interval $[0, T]$ suitably small. A similar analysis as in previous sections would now provide the existence of a Nash equilibrium solution in feedback form, but only locally in time.

Another possible extension of our results is to the case where the data $g_i'(\cdot)$ have large total variation. Using the local existence theorem of Schochet [Sc], one can still construct a weak solution to the system of conservation laws (2.10), at least on a short time interval $[0, T]$. For large BV solutions, however, checking that the feedbacks $U_i^*(t, x) \doteq u^*(p_i(t, x))$ at (2.5) yield a Nash equilibrium solution to the differential game will require a more accurate analysis. Furthermore, it is not clear whether, for large BV initial data, the solution to the system of conservation laws (5.2) can blow up in finite time. For general hyperbolic systems this can indeed happen [J]. The particular form of the flux functions H_i, however, may prevent such a blow up. To understand the matter, a more detailed analysis is again required.

The basic assumption in Theorem 5.1 was the hyperbolicity of the Hamiltonian system, in a neighborhood of the reference point p^*. When this condition is violated, searching for a Nash equilibrium in feedback form leads to an elliptic Cauchy problem. It is well known that this is ill posed [Lx]. Indeed, by elementary Fourier analysis one checks that even the constant solutions are linearly unstable. It thus appears that, in the elliptic regime, the model provided by noncooperative games must be revised. A concept of a "partially cooperative" solution should be considered, in order to recover the well posedness of the problem. This will be the content of a forthcoming paper [BS].

REFERENCES

[AM] F. Ancona and A. Marson, *Well-Posedness for General 2×2 Systems of Conservation Laws*, Mem. Amer. Math. Soc., 169 (801) (2004).

[AC] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[BC] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[BB] S. Bianchini and A. Bressan, *Vanishing viscosity solutions of nonlinear hyperbolic systems*, Ann. of Math, to appear.

[Bo] V. G. Boltyanskii, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control Optim., 4 (1966), pp. 326–361.

[B1] A. Bressan, *Unique solutions for a class of discontinuous differential equations*, Proc. Amer. Math. Soc., 104 (1988), pp. 772–778.

[B2] A. Bressan, *Upper and lower semicontinuous differential inclusions: A unified approach*, in Nonlinear Controllability and Optimal Control, H. Sussmann, ed., M. Dekker, New York, 1990, pp. 21–31.

[B3] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One Dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.

[BLY] A. Bressan, T. P. Liu, and T. Yang, *L^1 stability estimates for $n \times n$ conservation laws*, Arch. Ration. Mech. Anal., 149 (1999), pp. 1–22.

[BS] A. Bressan and W. Shen, *Semi-cooperative strategies for differential games*, Internat. J. Game Theory, to appear.

[Br] P. Brunovsky, *Existence of regular syntheses for general problems*, J. Differential Equations, 38 (1980), pp. 317–343.

[CR] G. Q. Chen and A. Rustichini, *The Riemann solution to a system of conservation laws, with application to a nonzero sum game*, Contemp. Math., 100 (1988), pp. 287–297.

[C] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[DH] C. Dafermos and L. Hsiao, *Hyperbolic systems of balance laws with inhomogeneity and absorption*, Indiana Univ. Math. J., 31 (1982), pp. 471–491.

[Fo] G. B. Folland, *Real Analysis. Modern Techniques and Their Applications*, Wiley, New York, 1984.

[F1] A. Friedman, *Differential Games*, Wiley-Interscience, New York, 1971.

[F2] A. Friedman, *Stochastic differential games*, J. Differential Equations, 11 (1972), pp. 79–108.

[G] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[H] O. Hajek, *Discontinuous differential equations* I, J. Differential Equations, 32 (1979), pp. 149–170.

[ILF] T. Iguchi and P. G. LeFloch, *Existence theory for hyperbolic systems of conservation laws with general flux-functions*, Arch. Ration. Mech. Anal., 168 (2003), pp. 165–244.

[I] R. Isaacs, *Differential Games*, Wiley, New York, 1965.

[J] H. K. Jenssen, *Blowup for systems of conservation laws*, SIAM J. Math. Anal., 31 (2000), pp. 894–908.

[Lx] P. Lax, *Asymptotic solution of oscillatory initial value problems*, Duke Math. J., 24 (1957), pp. 627–646.

[L] T. P. Liu, *Admissible solutions of hyperbolic conservation laws*, Mem. Amer. Math. Soc., 30 (1981).

[O] G. J. Olsder, *On open- and closed-loop bang-bang control in nonzero-sum differential games*, SIAM J. Control Optim., 40 (2001), pp. 1087–1106.

[PS] B. Piccoli and H. J. Sussmann, *Regular synthesis and sufficiency conditions for optimality*, SIAM J. Control Optim., 39 (2000), pp. 359–410.

[Sc] S. Schochet, *Sufficient conditions for local existence via Glimm's scheme for large BV data*, J. Differential Equations, 89 (1991), pp. 317–354.

[VZ] E. M. Vaisbord and V. I. Zhukovskii, *Introduction to Multi-players Differential Games and Their Applications*, Gordon and Breach Science Publishers, New York, 1988.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 216–239

OPTIMALITY CONDITIONS FOR NONCONVEX MULTISTATE CONTROL PROBLEMS IN THE COEFFICIENTS*

JUAN CASADO-DÍAZ†, JULIO COUCE-CALVO†, AND JOSÉ D. MARTÍN-GÓMEZ†

Abstract. The purpose of this paper is to attain some optimality conditions for the identification of a diffusion matrix (material) under several restrictions. Assuming that the set of such diffusion matrices is closed for the H-convergence, we give a method to obtain admissible directions which applies to a not-necessarily convex control set. Our results permit obtaining the diffusion matrix from the state functions.

Key words. control in the coefficients, admissible directions, nonconvex control set

AMS subject classification. 49K20

DOI. 10.1137/S0363012902411714

1. Introduction. The problem we consider in the present paper is related to the choice of an optimal material under several conditions, or the identification of a material from a finite number of observations. In a mathematical setting, we have the model problem

$$\min_{A\in M(\Omega)} J(y_1,\dots,y_k), \tag{1.1}$$

where $y_i = y_i(A)$, $1 \le i \le k$, are the solutions of the equations

$$\begin{cases} -\operatorname{div} A\nabla y_i = f_i & \text{in } \mathcal{D}'(\Omega), \\ y_i \in H_0^1(\Omega), & 1 \le i \le k. \end{cases} \tag{1.2}$$

Here Ω is a bounded open set of $\mathbf{R}^N$, J is a smooth objective functional in $H_0^1(\Omega)^k$, $f_1,\dots,f_k$ are k fixed elements of $H^{-1}(\Omega)$, and $M(\Omega)$ is a given set of measurable functions with values in the space of symmetric matrices of order N. The elements of $M(\Omega)$ are uniformly elliptic and bounded. Clearly, other generalizations can be considered: J depending on A, other boundary conditions, etc. A physical example is the identification of a material. For this purpose, we apply a finite number k of external conditions (in our case they are represented by f_i) and in each case we realize a measure of the corresponding state. For example, we give the value z_i of the state in a subset $\omega \subset \Omega$. Then the problem can be formulated as

$$\min \sum_{i=1}^{k} \int_{\omega} |y_i - z_i|^2 \, dx,$$

where y_i are the solutions of (1.2).

Assuming J is sequential lower semicontinuous for the weak topology of $H_0^1(\Omega)^k$ and $M(\Omega)$ is closed for the H-convergence or the G-convergence, because we are

*Received by the editors July 22, 2002; accepted for publication (in revised form) November 24, 2003; published electronically June 15, 2004. This work has been partially supported by the projects BFM 2002-00672 of the D.G.I. of the "Ministerio de Ciencia y Tecnología" of Spain and FQM309 of the "Junta de Andalucía."

http://www.siam.org/journals/sicon/43-1/41171.html

†Depto. de Ecuaciones Diferenciales y Análisis Numérico, Fac. de Matemáticas, Universidad de Sevilla, C. Tarfia s/n, 41012 Sevilla, Spain (jcasadod@us.es, jcouce@us.es, jdmartin@us.es).

working with symmetric matrices and thus the two concepts are equivalent (see, e.g., [22], [19], [18], [4], [25]), it is well known that (1.1) has, at least, a solution (see, e.g., [25], [17]). If $M(\Omega)$ does not satisfy this last condition and J is sequential continuous for the weak topology of $H_0^1(\Omega)^k$, we can obtain a relaxed problem replacing $M(\Omega)$ by its H-closure. Thus, it is natural to assume $M(\Omega)$ is H-closed. The calculus of the H-closure of a set is a very difficult problem and there are a lot of works in this field (see, e.g., [23], [14], [13], [8], [25], [3], [16] and the references in them). In this paper we are interested in obtaining some necessary conditions which must satisfy the optimal solution of (1.1). For $k = 1$, the problem has been studied in [13], [20], [8], [25], [3]. For $k > 1$, there are few results to our knowledge (see [25], [6], [7], [3]).

The paper is organized as follows:

In section 3 we give a definition of admissible direction (see Definition 3.1). Then we prove that if A is a solution of (1.1), $y_1, \dots, y_k$ the corresponding state functions, $p_1, \dots, p_k$ the solutions of (3.2) (the adjoint states), and C the matrix defined by (3.3), we have

$$\int_\Omega H : C \, dx \leq 0 \tag{1.3}$$

for every admissible direction H. Related results can be found, for example, in [8], [13], [25], [3]. However, in these papers, the admissible directions are of the form $H = B - A$, with B in $M(\Omega)$, which needs some convexity assumptions. When $k \leq N - 1$ (in particular $k = 1$) and $M(\Omega)$ is obtained by homogenization, mixing a finite number of matrices with fixed proportions, a result of Tartar (see [25]) shows that although $M(\Omega)$ is not convex, for every $\xi_1, \dots, \xi_k \in \mathbf{R}^N$, the set

$$\{(B\xi_1, \dots, B\xi_k) \,/\, B \in M(\Omega)\} \subset L^\infty(\Omega)^k \tag{1.4}$$

is convex, and thus the directions $H = B - A$ can still be considered. However, this is not true for $k \geq N$ (or, in principle, for other choices of $M(\Omega)$ even if $k \leq N - 1$). This is the reason we have given a more general definition of admissible direction.

In section 4, assuming $M(\Omega)$ local (see Definition 4.1) and closed for the H-convergence, we give an original method to find admissible directions following our definition. As a consequence, we obtain the main result of the paper, Theorem 4.5, where we prove that for every $A, B \in M(\Omega)$, $l \in \{1, \dots, N\}$, $W \subset \mathbf{R}^N$ linear subspace of dimension l, and every bounded measurable set T of W, with l-dimensional positive measure, the matrix H defined by

$$H(x)e_i = (B(x) - A(x)) \left(e_i + \frac{1}{|T|_\ell} \int_T \nabla_z^W \hat{w}_i(x, z) \, d_\ell(z) \right) \tag{1.5}$$

is an admissible direction in A. Here $e_1, \dots, e_N$ is the standard basis of $\mathbf{R}^N$, ∇_z^W denotes the gradient with respect to W, and $\hat{w}_i$ is the solution of the partial differential problem given by (4.4). This direction has the difficulty that $\hat{w}_i$ (and then H) cannot be explicitly obtained. However, we think that it can be interesting, for example, to apply a descent method in order to solve numerically problem (1.1), where we can obtain $\hat{w}_i$ numerically. Related to this point, an interesting question, one that we want to study in the future, is the optimal choice of W and T to obtain the steepest descent direction. A criterion to determine this direction (see Remark 4.9) can be to calculate the maximum of $H : C$ on the set of matrices H obtained by (1.5).

Although, as we have said above, it is not possible in general to obtain $\hat{w}_i$ explicitly, we show in Theorem 4.12 that this can be carried out for a particular choice of

T (which probably is not optimal). This permits us to obtain a family of admissible direction, depending on the subspace W chosen. Essentially, they are of the form $B - A$ plus a term which has a growth of order two in $B - A$ for every $B \in M(\Omega)$. When the dimension of W is equal to 1, the corresponding admissible direction comes just from a lamination. In this case the expression of H is known and it can be found, for example, in [25] (it can also be obtained from the results in [10]), but to our knowledge its utility to obtain optimality conditions for problem (1.1) has not been exploited. Most of the consequences we obtain in the present paper using Theorem 4.12 use only, in fact, $l = 1$. However, we show that in some cases (see Remark 4.17) it is better to use a subspace of dimension greater than one. Using Theorem 4.12 we prove in Corollary 4.18 that for every $B \in M(\Omega)$, the condition

$$C : (A - B) \geq 0 \tag{1.6}$$

(which is the condition we find if the admissible directions are of the form $B-A$) is still true on the set where C has a nonpositive eigenvalue or where $\mathrm{Ker}(A - B) \neq \{0\}$. In particular (see Corollary 4.20) the condition (1.6) holds a.e. in Ω for every $B \in M(\Omega)$ when $k \leq N-1$. When $M(\Omega)$ comes from the mixture of a finite number of materials with fixed proportions, this result can also be obtained from the convexity of the set defined by (1.4), but we note that our set $M(\Omega)$ is more general.

In section 5 we study the case where $M(\Omega)$ is invariable by rotations, which is a natural assumption in the applications. Then we show that condition (1.3) implies that C and A are mutually diagonalizable a.e. in Ω. Moreover, assuming further hypotheses (in particular if $M(\Omega)$ is H-closed and $N \geq 3$), we prove in Proposition 5.4 that the eigenvalues of A and C are mutually ordered.

As application of the results stated above, it is possible to obtain, in some situations, the matrix A from C and then to reduce the set of optimality conditions to a nonlinear partial differential system with variables y_i, p_i, $1 \leq i \leq k$. The main problem to carrying out this point is that in general, the H-closure of a given set is unknown. In section 6, we apply our results to two examples: The first one is the mixture of two homogeneous isotropic materials, which has also been studied in [3] (see also [25] for $k = 1$). In this case $M(\Omega)$ is convex. In second problem we consider a polycrystal in dimension 2, where $M(\Omega)$ is not convex.

2. Notation. For a linear subspace $W \subset \mathbf{R}^N$, we define $\mathcal{L}(W, W)$ as the space of the linear applications from W into W and by $\mathcal{L}^s(W, W)$ the subspace of the symmetric applications. When $W = \mathbf{R}^N$ we write $\mathcal{M}_N = \mathcal{L}(\mathbf{R}^N, \mathbf{R}^N)$, $\mathcal{M}_N^s = \mathcal{L}^s(\mathbf{R}^N, \mathbf{R}^N)$.

The orthogonal projection of $\mathbf{R}^N$ into W is denoted by P^W.

For a matrix $A \in \mathcal{M}_N$, we define $A_W \in \mathcal{L}(W, W)$ by $A_W = P^W A_{|_W}$.

The orthogonal subspace of W is denoted by $W^\perp$.

For $u : W \to \mathbf{R}$, we denote $\nabla^W u : W \to W$ the gradient of u with respect to W, i.e., $\nabla^W u$ is defined by

$$\nabla^W u \, \xi = D_\xi u \quad \forall \xi \in W,$$

where $D_\xi u$ is the derivative of u in the direction ξ.

We denote by $\{e_1, \ldots, e_N\}$ the standard basis of $\mathbf{R}^N$.

The group of the orthogonal matrices in $\mathbf{R}^N$ of determinant 1 is denoted by $\mathcal{O}_N$.

The scalar product of two matrices $A, B \in \mathcal{M}_N$ is written $A : B$.

The tensorial product of two vectors $\xi, \eta \in \mathbf{R}^N$ is denoted as $\xi \otimes \eta$.

For a bounded open set $\Omega \subset \mathbf{R}^N$, we denote by $M(\Omega)$ a fixed subset of the space $L^\infty(\Omega, \mathcal{M}_N^s)$ such that there exist $\alpha, \beta > 0$ which satisfy

$$B(x)\xi\xi \geq \alpha|\xi|^2, \quad |B(x)\xi| \leq \beta|\xi| \quad \forall\, B \in M(\Omega) \text{ a.e. } x \in \Omega. \tag{2.1}$$

For a matrix $A \in M(\Omega)$, $K_A(M(\Omega))$ and $\bar{K}_A(M(\Omega))$ are the cones of admissible directions of $M(\Omega)$ in A; see Definition 3.1.

For $T \subset \mathbf{R}^N$ and $\ell \in (0, N]$, we denote by $|T|_\ell$ the ℓ-dimensional Hausdorff measure of T. The integral of a function $u : T \to \mathbf{R}$, with respect to the ℓ-dimensional Hausdorff measure, is written

$$\int_T u(z)\, d_\ell z.$$

When $\ell = N$, we simplify the notation by writing $|T|$ and

$$\int_T u(z)\, dz.$$

We use the subindex $\sharp$ to mean periodicity. For example, for a cube $Y \subset \mathbf{R}^N$, $H^1_\sharp(Y)$ is the space of functions of $H^1_{loc}(\mathbf{R}^N)$ which are Y-periodic.

3. Optimality conditions. In this section we introduce the definition of admissible direction. Using it, we obtain the first optimality result for the control problem (1.1).

DEFINITION 3.1. *For $A \in M(\Omega)$, let us define the cone of admissible directions $\bar{K}_A(M(\Omega))$ as the closure in the weak-$*$ topology of $L^\infty(\Omega, \mathcal{M}_N^s)$ of the set $K_A(M(\Omega))$, where $K_A(M(\Omega))$ is the set of $H \in L^\infty(\Omega, \mathcal{M}_N^s)$ such that there exist a constant $c > 0$ and $A_\varepsilon \in M(\Omega)$, $\varepsilon > 0$, such that*

$$\begin{cases} \|A_\varepsilon - A\|_{L^\infty(\Omega, \mathcal{M}_N^s)} \leq c\varepsilon, \\ \lim\limits_{\varepsilon \to 0} \dfrac{A_\varepsilon - A}{\varepsilon} = H \quad \text{a.e. in } \Omega. \end{cases} \tag{3.1}$$

THEOREM 3.2. *We consider $J : H_0^1(\Omega)^k \to \mathbf{R}$, Fréchet derivable, $f_1, \dots, f_k \in H^{-1}(\Omega)^k$. Let $A \in M(\Omega)$ be a solution of (1.1) and $y_1, \dots, y_k$ the solutions of (1.2). We define the adjoint states $p_1, \dots, p_k$ as the solutions of*

$$\begin{cases} -\operatorname{div}(A\nabla p_i) = \partial_i J(y_1, \dots, y_k) & \text{in } \mathcal{D}'(\Omega), \\ p_i \in H_0^1(\Omega), & 1 \leq i \leq k, \end{cases} \tag{3.2}$$

and the matrix $C \in L^1(\Omega, \mathcal{M}_N^s)$ by

$$C = \frac{1}{2} \sum_{i=1}^{k} (\nabla y_i \otimes \nabla p_i + \nabla p_i \otimes \nabla y_i). \tag{3.3}$$

Then we have

$$\int_\Omega H : C\, dx \leq 0 \quad \forall\, H \in \bar{K}_A(M(\Omega)). \tag{3.4}$$

Proof. Let us first prove the result for $H \in K_A(M(\Omega))$. For $\varepsilon > 0$ small enough, we define $y^*_{i,\varepsilon}$, $1 \le i \le k$, as the solution of

$$\begin{cases} -\operatorname{div}((A+\varepsilon H)\nabla y^*_{i,\varepsilon}) = f_i & \text{in } \mathcal{D}'(\Omega), \\ y^*_{i,\varepsilon} \in H^1_0(\Omega). \end{cases} \tag{3.5}$$

Then it is easy to check that for $1 \le i \le k$, we have

$$\lim_{\varepsilon\to 0} \frac{y^*_{i,\varepsilon} - y_i}{\varepsilon} = \dot{y}_i \quad \text{in } H^1_0(\Omega), \tag{3.6}$$

with $\dot{y}_i$ the solutions of

$$\begin{cases} -\operatorname{div}(A\nabla \dot{y}_i + H\nabla y_i) = 0 & \text{in } \mathcal{D}'(\Omega), \\ \dot{y}_i \in H^1_0(\Omega). \end{cases} \tag{3.7}$$

Now, for $\varepsilon > 0$, we consider $A_\varepsilon \in M(\Omega)$ in the conditions of (3.1). Then for $1 \le i \le k$, we define $y_{i,\varepsilon}$ as the solutions of

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla y_{i,\varepsilon}) = f_i & \text{in } \mathcal{D}'(\Omega), \\ y_{i,\varepsilon} \in H^1_0(\Omega). \end{cases} \tag{3.8}$$

Taking $y^*_{i,\varepsilon} - y_{i,\varepsilon}$ as test function in the difference of (3.5) and (3.8), and dividing by ε, we get

$$\begin{aligned} &\frac{1}{\varepsilon}\int_\Omega (A+\varepsilon H)\nabla(y^*_{i,\varepsilon} - y_{i,\varepsilon})\nabla(y^*_{i,\varepsilon} - y_{i,\varepsilon})\,dx \\ &\quad = \int_\Omega \frac{A_\varepsilon - (A+\varepsilon H)}{\varepsilon}\nabla(y_{i,\varepsilon} - y^*_{i,\varepsilon})\nabla(y^*_{i,\varepsilon} - y_{i,\varepsilon})\,dx \\ &\qquad + \int_\Omega \frac{A_\varepsilon - (A+\varepsilon H)}{\varepsilon}\nabla y^*_{i,\varepsilon}\nabla(y^*_{i,\varepsilon} - y_{i,\varepsilon})\,dx. \end{aligned} \tag{3.9}$$

By the ellipticity of $A+\varepsilon H$ (for ε small enough) and (3.1), we deduce from (3.9) the existence of $c > 0$ such that

$$\frac{1}{\varepsilon}\|y^*_{i,\varepsilon} - y_{i,\varepsilon}\|^2_{H^1_0(\Omega)} \le c\|y^*_{i,\varepsilon} - y_{i,\varepsilon}\|_{H^1_0(\Omega)}\|\mu_\varepsilon\|_{L^2(\Omega)},$$

with

$$\mu_\varepsilon = \frac{A_\varepsilon - (A+\varepsilon H)}{\varepsilon}\nabla y^*_{i,\varepsilon}.$$

From (3.1), (3.6), and the Lebesgue-dominated convergence theorem, we deduce that μ_ε converges strongly to zero in $L^2(\Omega)^N$. Thus,

$$\lim_{\varepsilon\to 0} \frac{y^*_{i,\varepsilon} - y_{i,\varepsilon}}{\varepsilon} = 0 \quad \text{in } H^1_0(\Omega),$$

which, by (3.6), implies

$$\lim_{\varepsilon\to 0} \frac{y_{i,\varepsilon} - y_i}{\varepsilon} = \dot{y}_i \quad \text{in } H^1_0(\Omega). \tag{3.10}$$

On the other hand, since A is a solution of (1.1) and $A_\varepsilon \in M(\Omega)$, we have

$$\frac{J(y_\varepsilon) - J(y)}{\varepsilon} \geq 0 \quad \forall\, \varepsilon > 0,$$

with $y_\varepsilon = (y_{1,\varepsilon}, \dots, y_{k,\varepsilon})$ and $y = (y_1, \dots, y_k)$. From (3.10) and the Fréchet derivability of J, we get

$$\sum_{i=1}^{k} \langle \partial_i J(y), \dot{y}_i \rangle = \lim_{\varepsilon \to 0} \frac{J(y_\varepsilon) - J(y)}{\varepsilon} \geq 0. \tag{3.11}$$

But taking $\dot{y}_i$ as the test function in (3.2) and p_i as the test function in (3.7), we have

$$\sum_{i=1}^{k} \langle \partial_i J(y), \dot{y}_i \rangle = \sum_{i=1}^{k} \int_\Omega A \nabla p_i \nabla \dot{y}_i \, dx = -\sum_{i=1}^{k} \int_\Omega H \nabla y_i \nabla p_i \, dx.$$

This proves

$$\int_\Omega H : C \, dx = \sum_{i=1}^{k} \int_\Omega H \nabla y_i \nabla p_i \, dx \leq 0 \quad \forall\, H \in K_A(M(\Omega)). \tag{3.12}$$

Now let H be in $\bar{K}_A(M(\Omega))$. For $\delta > 0$ we define

$$G_\delta = \left\{ M \in L^\infty(\Omega, \mathcal{M}_N^s) / \ \left| \int_\Omega C : (M - H) \, dx \right| < \delta \right\}.$$

Since G_δ is a neighborhood of H in the weak-$*$ topology of $L^\infty(\Omega, \mathcal{M}_N^s)$, there exists H_δ in $G_\delta \cap K_A(M(\Omega))$ and then, from (3.12), we get

$$\int_\Omega C : H \, dx = \int_\Omega C : H_\delta \, dx + \int_\Omega C : (H - H_\delta) \, dx < \delta$$

for every $\delta > 0$. This proves (3.4). □

Remark 3.3. The above theorem is still true if the elements of $M(\Omega)$ are not necessarily symmetric by changing A to A^t in the definition (3.2) of the functions p_i and taking

$$C = \sum_{i=1}^{k} \nabla p_i \otimes \nabla y_i.$$

Remark 3.4. If $M(\Omega)$ is convex, the condition (3.4) implies

$$\int_\Omega A : C \, dx = \max \left\{ \int_\Omega B : C \, dx \, / \ B \in M(\Omega) \right\}. \tag{3.13}$$

Remark 3.5. Theorem 3.2 still holds if we take $K_A(M(\Omega))$ as the cone of matrices $H \in L^\infty(\Omega, \mathcal{M}_N^s(\Omega))$ such that for every sequence $\Phi_\varepsilon^1, \dots, \Phi_\varepsilon^k$, which respectively converges in $L^2(\Omega)^N$ to $\Phi^1, \dots, \Phi^k$, there exists $A_\varepsilon \in M(\Omega)$ such that

$$\frac{A_\varepsilon - A}{\varepsilon} \Phi_\varepsilon^i \to H \Phi^i \qquad \text{in } L^2(\Omega)^N, \qquad i = 1, \dots, k.$$

The advantage of this definition is the following: If $M(\Omega)$ is the set we obtain by mixing r materials with proportions fixed, then for every $\xi_1, \dots, \xi_{N-1} \in \mathbf{R}^N$, the set

$$\{(B\xi_1, \dots, B\xi_{N-1})/B \in M(\Omega)\} \subset L^\infty(\Omega)^{N-1}$$

is convex (see [24], [25]). So, with this definition of $K_A(M(\Omega))$, the matrices of the form $B - A$ belong to $K_A(M(\Omega))$ if $k \leq N-1$. Thus, (3.13) still holds in this case. Later we will deduce this result (see Corollary 4.20) for more general choices of $M(\Omega)$, using simply Definition 3.1 of admissible directions.

4. Calculus of admissible directions. In the following, let us calculate explicitely some admissible directions by imposing additional hypotheses about $M(\Omega)$.

DEFINITION 4.1. *We say that $M(\Omega)$ is local if there exists a multivaluated application $F : x \in \Omega \to F(x) \subset \mathcal{M}_N^s$ such that*

$$M(\Omega) = \{B \in L^\infty(\Omega, \mathcal{M}_N^s)/B(x) \in F(x) \text{ a.e. } x \in \Omega\},$$

where F is measurable in the sense that

$$\{x \in \Omega/F(x) \cap G \neq \emptyset\} \text{ is measurable} \quad \forall\, G \subset \mathcal{M}_N^s \text{ open.}$$

As it is proved in [21], the local property is satisfied in several typical examples of $M(\Omega)$. A first consequence of assuming $M(\Omega)$ is local follows.

PROPOSITION 4.2. *We assume $M(\Omega)$ is local. We consider $A \in M(\Omega)$, $H_1, \dots, H_m \in K_A(M(\Omega))$, $\omega_1, \dots, \omega_m \subset \Omega$ measurable such that $|\omega_i \cap \omega_j| = 0$ if $i \neq j$. Then the matrix $H = \sum_{i=1}^m H_i\, \chi_{\omega_i}$ belongs to $K_A(M(\Omega))$.*

Proof. By Definition 3.1, for every $i \in \{1, \dots, m\}$ there exists $A_\varepsilon^i \in M(\Omega)$ and $c > 0$ (which can be taken independent of i) such that

$$\|A_\varepsilon^i - A\|_{L^\infty(\Omega, \mathcal{M}_N^s)} \leq c\varepsilon, \qquad \lim_{\varepsilon \to 0} \frac{A_\varepsilon^i - A}{\varepsilon} = H_i \quad \text{a.e. in } \Omega.$$

Taking then

$$A_\varepsilon = \sum_{i=1}^m A_\varepsilon^i\, \chi_{\omega_i} + A\, \chi_{\Omega \setminus \cup_{i=1}^m \omega_i},$$

which belongs to $M(\Omega)$ because $M(\Omega)$ is local, we have

$$\|A_\varepsilon - A\|_{L^\infty(\Omega, \mathcal{M}_N^s)} \leq c\varepsilon, \qquad \lim_{\varepsilon \to 0} \frac{A_\varepsilon - A}{\varepsilon} = H \quad \text{a.e. in } \Omega,$$

and then H belongs to $K_A(M(\Omega))$. □

Remark 4.3. It is not difficult to show that the above result remains true if we replace $K_A(M(\Omega))$ by $\bar{K}_A(M(\Omega))$.

Using Proposition 4.2, we get the following.

PROPOSITION 4.4. *In the assumptions of Theorem 3.2, if $M(\Omega)$ is local, we have*

$$H : C \leq 0 \quad \text{a.e. in } \Omega\ \forall\, H \in \bar{K}_A(M(\Omega)). \tag{4.1}$$

Proof. By Proposition 4.2, for every $H \in K_A(M(\Omega))$ and every $\omega \subset \Omega$ measurable, the matrix $H\, \chi_\omega$ belongs to $K_A(M(\Omega))$. So, using (3.4), we get

$$\int_\omega H : C\, dx \leq 0. \tag{4.2}$$

Since $\bar{K}_A(M(\Omega))$ is the closure of $K_A(M(\Omega))$ in the weak-$*$ topology of $L^\infty(\Omega, \mathcal{M}_N^s)$, we deduce that (4.2) holds, in fact, for every $H \in \bar{K}_A(M(\Omega))$ and every $\omega \subset \Omega$ measurable, which implies (4.1). □

Let us now see how assuming $M(\Omega)$ is local permits us to obtain admissible directions.

THEOREM 4.5. *We suppose that $M(\Omega)$ is local and closed for the H-convergence. We consider a linear subspace $W \subset \mathbf{R}^N$ of dimension ℓ and a measurable bounded subset $T \subset W$ such that $|T|_\ell$ is positive. Then, for every $A, B \in M(\Omega)$, the matrix $H \in L^\infty(\Omega, \mathcal{M}_N^s)$ defined by*

$$H(x)e_i = (B(x) - A(x)) \left(e_i + \frac{1}{|T|_\ell} \int_T \nabla_z^W \hat{w}_i(x,z)\, d_\ell(z) \right) \tag{4.3}$$

for $1 \le i \le N$ and a.e. $x \in \Omega$ belongs to $K_A(M(\Omega))$. In (4.3), the function $\hat{w}_i$ is defined by

$$\begin{cases} \hat{w}_i(x,.) \in H^1_{loc}(W), \quad \nabla_z^W \hat{w}_i(x,.) \in L^2(W,W), \\ \displaystyle\int_W (A(x)\ \chi_{W\setminus T} + B(x)\ \chi_T) \nabla_z^W \hat{w}_i(x,.) \nabla_z^W \hat{v}\, d_\ell(z) \\ \qquad \displaystyle = \int_T (A(x) - B(x)) e_i \nabla_z^W \hat{v}\, d_\ell(z), \\ \forall\, \hat{v} \in H^1_{loc}(W), \quad \nabla_z^W \hat{v} \in L^2(W,W) \quad \text{a.e. } x \in \Omega. \end{cases} \tag{4.4}$$

Proof. We consider A, B, W, and T as in the statement of the theorem. For an orthonormal basis $\{e'_1, \dots, e'_\ell\}$ of W, we denote

$$Y = \left\{ \sum_{i=1}^{\ell} \lambda_i e'_i \,/\, -\frac{1}{2} < \lambda_i < \frac{1}{2},\ 1 \le i \le \ell \right\} \subset W.$$

For $\varepsilon > 0$ small enough, we denote $T_\varepsilon = \varepsilon^{\frac{1}{\ell}} T \subset Y$, $\tilde{T}_\varepsilon = \bigcup_{k \in \mathbf{Z}^\ell} (T_\varepsilon + \sum_{i=1}^{\ell} k_i e'_i)$, and we define $\tilde{A}_\varepsilon : \Omega \times W \to \mathcal{M}_N^s$ by

$$\tilde{A}_\varepsilon(x,y) = A(x)(1 - \chi_{\tilde{T}_\varepsilon}(y)) + B(x)\ \chi_{\tilde{T}_\varepsilon}(y).$$

Since $M(\Omega)$ is local and closed for the H-convergence, the matrix A_ε obtained by taking, for ε fixed, the H-limit when δ tends to zero of the matrices $x \to \tilde{A}_\varepsilon(x, \frac{1}{\delta} P^W(x))$ belongs to $M(\Omega)$. Since the matrices $\tilde{A}_\varepsilon$ are a tensorial product of functions which only depend on x and functions which only depend on y, it is well known (see, e.g., [5], [2]) that A_ε is given by

$$A_\varepsilon(x) e_i = \int_Y \tilde{A}_\varepsilon(x,y)(\nabla_y^W w_{i,\varepsilon} + e_i)\, d_\ell(y), \tag{4.5}$$

where $w_{i,\varepsilon}$ is the unique solution of

$$\begin{cases} w_{i,\varepsilon} \in L^2(\Omega, H^1_\sharp(Y)/\mathbf{R}), \\ \displaystyle\int_Y \tilde{A}_\varepsilon(x,y)(\nabla_y^W w_{i,\varepsilon}(x,y) + e_i) \nabla_y^W v(y)\, d_\ell(y) = 0 \\ \forall\, v \in H^1_\sharp(Y)/\mathbf{R} \quad \text{a.e. } x \in \Omega. \end{cases} \tag{4.6}$$

Let us study the asymptotic behavior of A_ε. First, we remark that for every $v \in H^1_\sharp(Y)/\mathbf{R}$ and a.e. $x \in \Omega$, we have

$$\int_Y \tilde{A}_\varepsilon(x,y)e_i\nabla_y^W v(y)\,d_\ell(y) = \int_Y A(x)e_i\nabla_y^W v(y)\,d_\ell(y) + \int_{T_\varepsilon}(B(x)-A(x))e_i\nabla_y^W v(y)\,d_\ell(y),$$

but for a.e. $x \in \Omega$, the first term on the right-hand side vanishes. So, $w_{i,\varepsilon}$ satisfies

$$\text{(4.7)} \qquad \begin{cases} \displaystyle\int_Y \tilde{A}_\varepsilon(x,y)\nabla_y^W w_{i,\varepsilon}(x,y)\nabla_y^W v(y)\,d_\ell(y) \\ \qquad \displaystyle = \int_{T_\varepsilon}(A(x)-B(x))e_i\nabla_y^W v(y)\,d_\ell(y) \\ \forall\, v \in H^1_\sharp(Y)/\mathbf{R} \quad \text{a.e. } x \in \Omega. \end{cases}$$

For $1 \le i \le N$, we take $w_{i,\varepsilon}$ as the test function in (4.7). Then we get

$$\int_Y \tilde{A}_\varepsilon(x,y)\nabla_y^W w_{i,\varepsilon}(x,y)\nabla_y^W w_{i,\varepsilon}(x,y)\,d_\ell(y) = \int_{T_\varepsilon}(A(x)-B(x))e_i\nabla_y^W w_{i,\varepsilon}(x,y)\,d_\ell(y)$$

for a.e. $x \in \Omega$. Using then (2.1) and the Cauchy–Schwarz inequality, we deduce there exists $c > 0$ such that

$$\text{(4.8)} \qquad \int_Y |\nabla_y^W w_{i,\varepsilon}(x,y)|^2\,d_\ell(y) \le c|T_\varepsilon|_\ell = c\varepsilon$$

for $\varepsilon > 0$ small enough and a.e. $x \in \Omega$.

We define $\hat{w}_{i,\varepsilon} : \Omega \times (\varepsilon^{-\frac{1}{\ell}}Y) \to \mathbf{R}$ by

$$\hat{w}_{i,\varepsilon}(x,z) = \varepsilon^{-\frac{1}{\ell}} w_{i,\varepsilon}(x, \varepsilon^{\frac{1}{\ell}} z).$$

From (4.8), we deduce that $\nabla_z^W \hat{w}_{i,\varepsilon}\ \chi_{\varepsilon^{-\frac{1}{\ell}}Y}$ is bounded in $L^\infty(\Omega, L^2(W,W))$. So there exists a subsequence of ε, which we still denote by ε, which converges weak-$*$ in $L^\infty(\Omega, L^2(W,W))$. Since the curl of the limit is zero, it is the gradient of a function $\hat{w}_i \in L^\infty(\Omega, H^1_{loc}(W))$. Once we prove that $\hat{w}_i$ is the solution of (4.4), we conclude that the whole of the sequence converges.

We consider $\hat{v} \in \mathcal{D}(W)$ and $\varepsilon > 0$ small enough, such that $\varepsilon^{\frac{1}{\ell}}\,\mathrm{supp}(\hat{v}) \subset Y$, then we define $v_\varepsilon \in H^1_\sharp(Y)$ by

$$v_\varepsilon(y) = \varepsilon^{\frac{1}{\ell}}\hat{v}(\varepsilon^{-\frac{1}{\ell}}y) \quad \text{a.e. } y \in Y.$$

Taking v_ε as the test function in (4.7), using the change of variables $z = \varepsilon^{-\frac{1}{\ell}}y$, and integrating with respect to x in a measurable set ω, we get

$$\int_\omega \int_{\varepsilon^{-\frac{1}{\ell}}Y} (A(x)\ \chi_{W\setminus T}(z) + B(x)\ \chi_T(z))\nabla_z^W \hat{w}_{i,\varepsilon}(x,z)\nabla_z^W \hat{v}(z)\,d_\ell(z)\,dx = \int_\omega \int_T (A(x)-B(x))e_i\nabla_z^W \hat{v}(z)\,d_\ell(z)\,dx.$$

Passing to the limit in this equality and taking into account the arbitrariness of $\hat{v}$ and ω, and the density of $\mathcal{D}(\mathbf{R}^N)/\mathbf{R}$ in the factor space of functions with gradient in $L^2(\mathbf{R}^N)$ over $\mathbf{R}$ (see, e.g., [9]), we show that $\hat{w}_i$ is the solution of (4.4) for $1 \le i \le N$.

Let us now prove

$$\nabla_z^W \hat{w}_{i,\varepsilon}\ \chi_{\varepsilon^{-\frac{1}{\ell}}Y} \to \nabla_z^W \hat{w}_i \quad \text{in } L^2(\Omega, L^2(W,W)) \tag{4.9}$$

for $1 \le i \le N$. For this purpose, it is enough to take $w_{i,\varepsilon}$ as the test function in (4.7), to use the change of variables $z = \varepsilon^{-\frac{1}{\ell}} y$, to take $\hat{w}_i$ as the test function in (4.4), and to use that $\nabla_z^W \hat{w}_{i,\varepsilon}\ \chi_{\varepsilon^{-\frac{1}{\ell}}Y}$ converges to $\nabla_z^W \hat{w}_i$, weak-$*$ in $L^\infty(\Omega, L^2(W,W))$. This gives

$$\begin{aligned}
&\int_\Omega \int_{\varepsilon^{-\frac{1}{\ell}}Y} (A\ \chi_{W\setminus T} + B\ \chi_T)\nabla_z^W \hat{w}_{i,\varepsilon} \nabla_z^W \hat{w}_{i,\varepsilon}\, d_\ell(z)\, dx \\
&\quad = \tfrac{1}{\varepsilon} \int_\Omega \int_Y \tilde{A}_\varepsilon \nabla_y^W w_{i,\varepsilon} \nabla_y^W w_{i,\varepsilon}\, d_\ell(y)\, dx = \frac{1}{\varepsilon} \int_\Omega \int_{T_\varepsilon} (A-B) e_i \nabla_y^W w_{i,\varepsilon}\, d_\ell(y)\, dx \\
&\quad = \int_\Omega \int_T (A-B) e_i \nabla_z^W \hat{w}_{i,\varepsilon}\, d_\ell(z)\, dx \to \int_\Omega \int_T (A-B) e_i \nabla_z^W \hat{w}_i\, d_\ell(z)\, dx \\
&\quad = \int_\Omega \int_W (A\ \chi_{W\setminus T} + B\ \chi_T)\nabla_z^W \hat{w}_i \nabla_z^W \hat{w}_i\, d_\ell(z)\, dx.
\end{aligned}$$

This implies (4.9).

Now, for $i, j \in \{1, \dots, N\}$ and a.e. $x \in \Omega$, we write

$$\begin{aligned}
A_\varepsilon(x) e_i e_j &= \int_Y \tilde{A}_\varepsilon(x,y)[\nabla_y^W w_{i,\varepsilon} + e_i] e_j\, d_\ell(y) \\
&= \int_Y \tilde{A}_\varepsilon(x,y) \nabla_y^W w_{i,\varepsilon} e_j\, d_\ell(y) + A(x) e_i e_j + \varepsilon |T|_l (B(x) - A(x)) e_i e_j.
\end{aligned} \tag{4.10}$$

Taking $w_{i,\varepsilon}$ as the test function in the problem satisfied by $w_{j,\varepsilon}$ and using the change of variables $z = \varepsilon^{-\frac{1}{\ell}} y$, we have

$$\begin{aligned}
&\int_Y \tilde{A}_\varepsilon(x,y) \nabla_y^W w_{i,\varepsilon} e_j\, d_\ell(y) = -\int_Y \tilde{A}_\varepsilon(x,y) \nabla_y^W w_{j,\varepsilon} \nabla_y^W w_{i,\varepsilon}\, d_\ell(y) \\
&\quad = -\varepsilon \int_{\varepsilon^{-\frac{1}{\ell}}Y} (A(x)\ \chi_{W\setminus T}(z) + B(x)\ \chi_T(z)) \nabla_z^W \hat{w}_{j,\varepsilon} \nabla_z^W \hat{w}_{i,\varepsilon}\, d_\ell(z).
\end{aligned}$$

So, from (4.10), we get

$$\begin{aligned}
\left(\frac{A_\varepsilon - A}{\varepsilon}\right) e_i e_j &= |T|_\ell (B(x) - A(x)) e_i e_j \\
&\quad - \int_{\varepsilon^{-\frac{1}{\ell}}Y} (A(x)\ \chi_{W\setminus T}(z) + B(x)\ \chi_T(z)) \nabla_z^W \hat{w}_{j,\varepsilon} \nabla_z^W \hat{w}_{i,\varepsilon}\, d_\ell(z).
\end{aligned}$$

Since $\nabla_z^W \hat{w}_{j,\varepsilon}$, $\nabla_z^W \hat{w}_{i,\varepsilon}$ are bounded in $L^\infty(\Omega, L^2(W,W))$, we deduce that $\frac{A_\varepsilon - A}{\varepsilon}$ is bounded in $L^\infty(\Omega, \mathcal{M}_N^s)$. On the other hand, using (4.9), we have

$$\begin{aligned}
&\lim_{\varepsilon \to 0} \int_{\varepsilon^{-\frac{1}{\ell}}Y} (A(x)\ \chi_{W\setminus T}(z) + B(x)\ \chi_T(z)) \nabla_z^W \hat{w}_{j,\varepsilon} \nabla_z^W \hat{w}_{i,\varepsilon}\, d_\ell(z) \\
&\quad = \int_W (A(x)\ \chi_{W\setminus T}(z) + B(x)\ \chi_T(z)) \nabla_z^W \hat{w}_j \nabla_z^W \hat{w}_i\, d_\ell(z) \\
&\quad = -\int_T (B(x) - A(x)) \nabla_z^W \hat{w}_i e_j\, d_\ell(z),
\end{aligned}$$

where we have used the problem satisfied by $\hat{w}_j$ and where the limit is taken in $L^1(\Omega)$. So we have proved that the matrix

$$e_i \to |T|_\ell (B(x) - A(x)) \left(e_i + \frac{1}{|T|_\ell} \int_T \nabla_z^W \hat{w}_i(x,z)\, d_\ell(z) \right)$$

is in $K_A(M(\Omega))$. Since $K_A(M(\Omega))$ is a cone, we conclude the proof of the theorem. □

Remark 4.6. Theorem 4.5 applies, for example, to the case where $M(\Omega)$ is the set of materials which can be obtained by homogenization, mixing m materials with the prescribed volume. These sets usually appear in problems of optimal design. Some applications are given in the last section of the paper; see also, e.g., [3], [8], [13], [25], and references therein.

Remark 4.7. The method used in the proof of Theorem 4.5 to obtain admissible directions, which consists of putting an inclusion of a tensor B in a background of tensor A, is a variation of the classical Weierstrass test. Related ideas have been used, for example, by K. A. Lurie (see [8] and references therein).

Remark 4.8. In Theorem 4.5 the expression of H when T is contained in a subspace W of dimension l can be obtained from the corresponding result to $W = \mathbf{R}^N$ (and then the case $W = \mathbf{R}^N$ can be consider as the most interesting one). It is enough to consider the matrix H_ε corresponding to $W = \mathbf{R}^N$ and $T_\varepsilon \subset \mathbf{R}^N$ defined by

$$T_\varepsilon = \{x + \varepsilon y :\ x \in B(0,1) \cap W^\perp,\ y \in T\}$$

and then pass to the limit in ε. The proof of Theorem 4.5 given above has the advantage that we do not need to realize this second limit.

Remark 4.9. The expression (4.3) of the admissible direction H has the difficulty that the function $\hat{w}_i$ is not explicit. However, as we said in the introduction, we think that it can be used, for example, to obtain a steepest descent direction. Then the function $\hat{w}_i$ can be calculated numerically. For this purpose we recall that by (3.11) in the proof of Theorem 3.2 we have

$$J(y_\varepsilon) \sim J(y) + \varepsilon \sum_{i=1}^{k} \langle \partial_i J(y), \dot{y}_i \rangle = J(y) - \varepsilon \int_\Omega H : C\, dx$$

and then, since $M(\Omega)$ is local, an idea to obtain the steepest direction is to maximize the product $H : C$ in the closure of the matrices H given by (4.3). By Remark 4.8, it is enough to consider $W = \mathbf{R}^N$. We remark that the set of such H is bounded and it is not difficult to show that its closure is convex and thus is essentially a ball (for some norm).

In Theorem 4.5, the set T does not depend on $x \in \Omega$. Thanks to Proposition 4.2 we can, in fact, take T depending on x. A result in this sense, which we use later in Theorem 4.12, is the following.

LEMMA 4.10. *Assume $M(\Omega)$ is local and closed for the H-convergence. We consider a linear subspace $W \subset \mathbf{R}^N$ of dimension ℓ, a measurable bounded subset $T_0 \subset W$, and a matrix $E \in L^\infty(\Omega, \mathcal{L}(W,W))$ which is invertible a.e. in Ω and such that E^{-1} also belongs to $L^\infty(\Omega, \mathcal{L}(W,W))$. Then, taking $T(x) = E(x)T_0$, for a.e. $x \in \Omega$, the matrix H defined by (4.3) with $T = T(x)$ is in $\bar{K}_A(M(\Omega))$ for every $A, B \in M(\Omega)$.*

Proof. For E in the conditions of the lemma, it is well known that there exists a sequence $E_n = \sum_{j=1}^{m(n)} E_{j,n}\, \chi_{\omega_{j,n}}$ with $E_{j,n} \in \mathcal{L}(W,W)$, $\omega_{j,n} \subset \Omega$ measurable,

$\cup_{j=1}^{m(n)}\omega_{j,n} = \Omega$, $\omega_{j,n} \cap \omega_{l,n} = \emptyset$ if $l \neq j$, and such that E_n converges strongly to E in $L^\infty(\Omega, \mathcal{L}(W,W))$. If $n \in \mathbf{N}$ is large enough (denoting by $\|\cdots\|$ the norm in $L^\infty(\Omega, \mathcal{L}(W,W))$), we have $\|E_n - E\| < \|E^{-1}\|^{-1}$, and then E_n is also invertible and satisfies

$$\|E_n^{-1} - E^{-1}\| \leq \frac{\|E^{-1}\|\|E - En\|}{\|E^{-1}\|^{-1} - \|E - E_n\|}.$$

So E_n^{-1} also converges to E^{-1} in $L^\infty(\Omega, \mathcal{L}(W,W))$.

We now take $T_n(x) = E_n(x)T_0$ and we denote by $H_n \in L^\infty(\Omega, \mathcal{M}_N^s)$ the matrix defined by

$$H_n(x)e_i = (B(x) - A(x))\left(e_i + \frac{1}{|T_n|_\ell}\int_{T_n} \nabla_z^W \hat{w}_{i,n}(x,z)\, d_\ell(z)\right),$$

where, for $i \in \{1, \ldots, n\}$, $\hat{w}_{i,n}$ is the solution of

$$\text{(4.11)} \qquad \begin{cases} \hat{w}_{i,n}(x,.) \in H^1_{loc}(W), \quad \nabla_z^W \hat{w}_{i,n}(x,.) \in L^2(W,W), \\ \displaystyle\int_W (A\ \chi_{W\setminus T_n} + B\ \chi_{T_n})\nabla_z^W \hat{w}_{i,n} \nabla_z^W \hat{v}\, d_\ell(z) \\ \qquad \displaystyle = \int_{T_n} (A-B)e_i \nabla_z^W \hat{v}\, d_\ell(z) \\ \forall\, \hat{v} \in H^1_{loc}(W), \quad \nabla_z^W \hat{v} \in L^2(W,W) \quad \text{a.e. } x \in \Omega. \end{cases}$$

From Theorem 4.5 and Proposition 4.2, this function belongs to $K_A(M(\Omega))$. Using the change of variables $\tilde{w}_{i,n}(\tilde{z}) = w_{i,n}(E_n\tilde{z})$, we deduce that $\tilde{w}_{i,n}$ is the solution of

$$\begin{cases} \tilde{w}_{i,n}(x,.) \in H^1_{loc}(W), \quad \nabla_{\tilde{z}}^W \tilde{w}_{i,n}(x,.) \in L^2(W,W), \\ \displaystyle\int_W (E_n^{-1})^t (A\ \chi_{W\setminus T_0} + B\ \chi_{T_0}) E_n^{-1} \nabla_{\tilde{z}}^W \tilde{w}_{i,n} \nabla_{\tilde{z}}^W \tilde{v}\, d_\ell(z) \\ \qquad \displaystyle = \int_{T_0} (E_n^{-1})^t (A-B) e_i \nabla_{\tilde{z}}^W \tilde{v}\, d_\ell(z) \\ \forall\, \tilde{v} \in H^1_{loc}(W), \quad \nabla_{\tilde{z}}^W \tilde{v} \in L^2(W,W) \quad \text{a.e. } x \in \Omega. \end{cases}$$

From the uniform convergence of E_n^{-1} to E^{-1}, we easily deduce that $\nabla_{\tilde{z}}^W \tilde{w}_{i,n}$ converges a $\nabla_{\tilde{z}}^W \tilde{w}_i$ in $L^\infty(\Omega, L^2(W,W))$, with $\tilde{w}_i$ the solution of

$$\begin{cases} \tilde{w}_i(x,.) \in H^1_{loc}(W), \quad \nabla_{\tilde{z}}^W \tilde{w}_i(x,.) \in L^2(W,W), \\ \displaystyle\int_W (E^{-1})^t (A\ \chi_{W\setminus T_0} + B\ \chi_{T_0}) E^{-1} \nabla_{\tilde{z}}^W \tilde{w}_i \nabla_{\tilde{z}}^W \tilde{v}\, d_\ell(z) \\ \qquad \displaystyle = \int_{T_0} (E^{-1})^t (A-B) e_i \nabla_{\tilde{z}}^W \tilde{v}\, d_\ell(z) \\ \forall\, \tilde{v} \in H^1_{loc}(W), \quad \nabla_{\tilde{z}}^W \tilde{v} \in L^2(W,W) \quad \text{a.e. } x \in \Omega. \end{cases}$$

Returning to the old variables, we then deduce that $\nabla_z^W \hat{w}_{i,n}$ converges to $\nabla_z^W \hat{w}_i$ in $L^\infty(\Omega, L^2(W,W))$, with $\hat{w}_i$ the solution of (4.4). Thus, H_n converges strongly to H in $L^\infty(\Omega, \mathcal{M}_N^s)$. So H belongs to $\bar{K}_A(M(\Omega))$. □

Let us now obtain the solutions $\hat{w}_i$ of (4.4) for some particular choices of T and then use Proposition 4.4 to obtain explicit optimality conditions.

LEMMA 4.11. *If $M(\Omega)$ is local and closed for the H-convergence, then, for every linear subspace $W \subset \mathbf{R}^N$ of dimension ℓ and every $A, B \in M(\Omega)$, the matrix*

$$(B-A) - (B-A)(B+(\ell-1)A)_W^{-1} P^W (B-A)$$

is in $\bar{K}_A(M(\Omega))$.

Proof. Let W be a linear subspace of $\mathbf{R}^N$ of dimension l. Since A and A^{-1} are in $L^\infty(\Omega, \mathcal{M}_N^s)$, there exists $R \in L^\infty(\Omega, \mathcal{L}(W,W))$, with $R^{-1} \in L^\infty(\Omega, \mathcal{L}(W,W))$, such that $RA_W R^t = I_W$ a.e. in Ω. We define T_0 the unitary ball in W and for a.e. $x \in \Omega$ we take $T(x) = R(x)^{-1} T_0$ and $H(x)$ the matrix given by (4.3) with $T = T(x)$. From Lemma 4.10, H belongs to $\bar{K}_A(M(\Omega))$.

The problem is to calculate the solution $\hat{w}_i$, $1 \le i \le N$, of (4.4). For this purpose, the idea is to use the change of variables $z' = Rz$, which transforms (4.4) in a similar problem, where A and T are respectively replaced by the identity and the unitary ball in W. This problem can be solved by using in a suitable way the fundamental solution of the laplacian. Doing this and returning to the old variables we deduce that (up to a function which only depends on x) $\hat{w}_i$ is given by

$$\hat{w}_i(x,z) = \begin{cases} \mu_i(x) z & \text{in T}, \\ \dfrac{\mu_i(x) z}{|Rz|^\ell} & \text{in } W \setminus T, \end{cases} \qquad 1 \le i \le N,$$

where $\mu_i(x) = (B(x) + (\ell-1)A(x))_W^{-1} P^W (A(x) - B(x)) e_i$. Then by (4.3) we deduce

$$H(x) e_i = (B(x) - A(x))(e_i + \mu_i(x)).$$

Taking into account the expression of μ_i, we finish the proof of the theorem. ∎

Using Lemma 4.11 and condition (3.4), we deduce the following theorem.

THEOREM 4.12. *In the assumptions of Theorem 3.2, if $M(\Omega)$ is local and closed for the H-convergence, then, for every linear subspace $W \subset \mathbf{R}^N$ of dimension ℓ, we have*

(4.12)
$$C : \{(A-B) + (A-B)(B+(\ell-1)A)_W^{-1} P^W (A-B)\} \ge 0 \quad \text{a.e. in } \Omega \ \forall\, B \in M(\Omega).$$

Remark 4.13. The condition (4.12) can also be written

(4.13)
$$C : (A-B) + \min_{1 \le \ell \le N} \min_{\dim(W) = l} C : (A-B)(B+(\ell-1)A)_W^{-1} P^W (A-B) \ge 0.$$

Thus, the better choice for W is to consider just the subspace which gives the minimum in this expression. This can also be related to the choice of the steepest descent direction, mentioned in Remark 4.9. If we restrict ourselves to the set of matrices H of the form

$$H = B - A - (A-B)(B+(\ell-1)A)_W^{-1} P^W (A-B),$$

then to choose the matrix giving the maximum of $H : C$ is equivalent to solving the minimization problem which appears in (4.13).

Remark 4.14. The condition (4.12) must be compared with the usual one when $M(\Omega)$ is convex, which is

$$C : (A-B) \ge 0 \quad \forall\, B \in M(\Omega), \text{ a.e. in } \Omega. \tag{4.14}$$

In general we get a perturbation of this condition with a term of second growth in $(B-A)$. We will see in Corollary 4.18 how condition (4.12) implies (4.14), at least in a subset of Ω.

COROLLARY 4.15. *In the assumptions of Theorem* 3.2*, if $M(\Omega)$ is local and closed for the H-convergence, then, for every $B \in M(\Omega)$, we have*

$$C : (A-B) + \frac{(A-B)C(A-B)\xi\xi}{B\xi\xi} \geq 0 \quad \forall\, \xi \in \mathbf{R}^N \setminus \{0\} \text{ a.e. in } \Omega. \tag{4.15}$$

Proof. It is enough, for $\xi \in \mathbf{R}^N \setminus 0$, to take $W = \{\lambda\xi \,/\, \lambda \in \mathbf{R}\}$ in Theorem 4.12 and to use that in this case:

$$(A-B)B_W^{-1}P^W(A-B) = \frac{(A-B)\xi \otimes (A-B)\xi}{B\xi\xi}. \qquad \square$$

Remark 4.16. The condition (4.15) is equivalent to

$$C : (A-B) + \min_{B\xi\xi=1} B[B^{-1}(A-B)C(A-B)]\xi\xi \geq 0$$

or, equivalently (observe that $B^{-1}(A-B)C(A-B)$ is symmetric with respect to the scalar product given by $(\xi|\eta) = B\xi\eta$ for every $\xi, \eta \in \mathbf{R}^N$),

$$C : (A-B) + \min\{\lambda :\ \lambda \text{ eigenvalue of } B^{-1}(A-B)C(A-B)\} \geq 0. \tag{4.16}$$

Remark 4.17. The inequality (4.15) has been obtained taking $\ell = 1$ in (4.12); then it comes just from a lamination in the direction ξ. So Corollary 4.15 holds if we assume only $M(\Omega)$ is local and stable under lamination (and not necessarily by H-convergence). The sets of matrices stable under lamination have been characterized by Francfort and Milton in [10] and [15]. In particular, it has been shown that, under a suitable change of variables, the set $M(\Omega)$, assumed stable under lamination, is convex. Corollary 4.15 can also be obtained from this result. In fact, deriving the usual formula for the lamination of two matrices A and B in the direction ξ, it has been proved in [25] that the matrix

$$B - A - \frac{(A-B)\xi \otimes (A-B)\xi}{B\xi\xi}$$

is an admissible direction. However, this has not been applied in our knowledge to the obtaining of optimality conditions for problem (1.1). Most of the results we obtain in the following can be deduced using (4.15), and then one can conjecture that the choice $\ell = 1$ is the best one in (4.12) (or even using all the matrices H given by Theorem 4.5) or, equivalently (see Remark 4.13), that the minimum in (4.13) is attained for $\ell = 1$. An easy counterexample shows that this is not true in general; it is enough to consider $C = B = I$ and $A = mI$ with $m > 1$. Then the minimum of the eigenvalues of $B^{-1}(A-B)C(A-B)$ is $(m-1)^2$ while taking, for example, $W = \mathbf{R}^N$ we have

$$\begin{aligned} C : (A-B)(B+(N-1)A)_W^{-1}P^W(A-B) \\ = C : (A-B)(B+(N-1)A)^{-1}(A-B) \\ = \frac{N(m-1)^2}{1+(N-1)m} < (m-1)^2. \end{aligned}$$

COROLLARY 4.18. *In the assumptions of Theorem* 3.2, *with* $M(\Omega)$ *local and closed for the* H*-convergence, we define*

$$\Omega^- = \{x \in \Omega : \ \exists\, \lambda \leq 0 \text{ eigenvalue of } C(x)\}$$

and

$$\Omega_B = \{x \in \Omega : \ \mathrm{Ker}(A(x) - B(x)) \neq \{0\}\} \quad \forall\, B \in M(\Omega).$$

Then we have

$$C : (A - B) \geq 0 \quad \text{a.e. in } \Omega^- \cup \Omega_B \quad \forall\, B \in M(\Omega). \tag{4.17}$$

Proof. Let B be in $M(\Omega)$. For a.e. $x \in \Omega_B$, we choose $\xi(x) \in \mathrm{Ker}(A(x) - B(x)) \setminus \{0\}$, and for a.e. $x \in \Omega^- \setminus \Omega_B$, we take $e(x)$ as an eigenvector associated with a nonpositive eigenvalue of $C(x)$ and $\xi(x) = (A(x) - B(x))^{-1} e(x)$. Then, taking $\xi = \xi(x)$ in (4.15), we obtain (4.17). □

By the above result, it is interesting to learn how many nonpositive eigenvalues have the matrix C. In this sense, we give the following theorem.

THEOREM 4.19. *For* $\xi_1, \ldots, \xi_k, \eta_1, \ldots, \eta_k \in \mathbf{R}^N \setminus \{0\}$, *we define*

$$\phi_i^+ = \frac{\xi_i}{|\xi_i|} + \frac{\eta_i}{|\eta_i|}, \qquad \phi_i^- = \frac{\xi_i}{|\xi_i|} - \frac{\eta_i}{|\eta_i|},$$

$$m = \dim(\mathrm{Span}\{\xi_i, \eta_i \,/\, 1 \leq i \leq k\}) = \dim(\mathrm{Span}\{\phi_i^+, \phi_i^- \,/\, 1 \leq i \leq k\}),$$

$$m^+ = \dim(\mathrm{Span}\{\phi_i^+ \,/\, 1 \leq i \leq k\}), \qquad m^- = \dim(\mathrm{Span}\{\phi_i^- \,/\, 1 \leq i \leq k\}),$$

$$\tilde{C}_i = \frac{1}{2}(\xi_i \otimes \eta_i + \eta_i \otimes \xi_i), \quad 1 \leq i \leq k, \quad \tilde{C} = \sum_{i=1}^{k} \tilde{C}_i.$$

Then we have the following:

(i) *For* $1 \leq i \leq k$, *the matrix* $\tilde{C}_i$ *has as eigenvalues* $\frac{1}{2}(\xi_i\eta_i + |\xi_i||\eta_i|) \geq 0$, $\frac{1}{2}(\xi_i\eta_i - |\xi_i||\eta_i|) \leq 0$, *with respective eigenvectors* ϕ_i^+, ϕ_i^-. *The other eigenvalues of* $\tilde{C}_i$ *are zero.*

(ii) *If* k^+ *and* k^- *are, respectively, the number of positive and negative eigenvalues of* $\tilde{C}$, *we have*

$$m - m^- \leq k^+ \leq m^+, \qquad m - m^+ \leq k^- \leq m^-. \tag{4.18}$$

Proof. The proof of (i) is easy to verify. In order to prove (ii), we use the Courant–Fischer characterization of the eigenvalues:

$$\lambda_{i+1} = \max_{\dim E \leq i} \min_{\substack{\phi \in E^\perp \\ |\phi| = 1}} \tilde{C}\phi\phi,$$

where $\lambda_1 \leq \cdots \leq \lambda_N$ are the eigenvalues of $\tilde{C}$.

Taking $i = m^-$ and $E = \mathrm{Span}\{\phi_j^- \,/\, 1 \leq j \leq k\}$, statement (i) gives

$$\lambda_{m^-+1} \geq \min_{\substack{\phi \in E^\perp \\ |\phi| = 1}} \tilde{C}\phi\phi \geq 0. \tag{4.19}$$

So $k^- \leq m^-$.

Now, if $m-m^- = 0$, then clearly $k^+ \geq m-m^-$. In another case, we consider $E = \mathrm{Span}(\{\phi_j^- \,/\, 1 \leq j \leq k\} \cup \{\phi_j^-, \phi_j^+ \,/\, 1 \leq j \leq k\}^\perp)$, which has dimension $m^- + N - m$. If $\phi \in E^\perp$, then, as above, statement (i) gives $\tilde{C}_j \phi\phi \geq 0$, $1 \leq j \leq k$, and then $\tilde{C}\phi\phi \geq 0$. Moreover, $\tilde{C}\phi\phi = 0$ iff $\tilde{C}_j \phi\phi = 0$, $1 \leq j \leq k$, which, by statement (i), implies that ϕ is orthogonal to ϕ_j^-, ϕ_j^+, $1 \leq j \leq k$; i.e., ϕ is in E and so $\phi = 0$. Thus, taking $i = m^- + N - m$ in (4.19) and using the compactness of the unitary ball in $\mathbf{R}^N$, we get

$$\lambda_{m^-+N-m+1} \geq \min_{\substack{\phi \in E^\perp \\ |\phi|=1}} \tilde{C}\phi\phi > 0,$$

and thus $k^+ \geq m - m^-$.

The other inequalities in (4.18) follow analogously. □

As a consequence we get the following.

COROLLARY 4.20. *In the assumptions of Theorem* 3.2, *if* $M(\Omega)$ *is local and closed for the* H*-convergence and* $k \leq N-1$, *then condition* (4.14) *holds.*

Proof. We apply Theorem 4.19 to $\xi_i = \nabla y_i(x)$, $\eta_i = \nabla p_i(x)$, $1 \leq i \leq k$, a.e. $x \in \Omega$. In this case $\tilde{C} = C(x)$. Since, clearly, the number m^+ which appears in this result is less than or equal to $k \leq N-1$, we deduce that the number of positive eigenvalues of C is less than or equal to $N-1$, and then there exists at least a nonpositive eigenvalue of C a.e. in Ω. Corollary 4.18 gives then (4.14). □

5. Invariability by rotations. In the applications, it is a natural hypothesis to assume that $M(\Omega)$ is invariable by rotations. We show in this section that this assumption implies that the eigenvectors of A and C agree.

DEFINITION 5.1. *We say that* $M(\Omega)$ *is invariable by rotations if for every* $B \in M(\Omega)$ *and every* $Q \in L^\infty(\Omega, \mathcal{M}_N)$, *with* $Q \in \mathcal{O}_N$ *a.e. in* Ω, *the matrix* QBQ^t *belongs to* $M(\Omega)$.

We have the following result.

PROPOSITION 5.2. *In the assumptions of Theorem* 3.2, *if* $M(\Omega)$ *is invariable by rotations, then* A *and* C *are mutually diagonalizable a.e. in* Ω.

Proof. Let us first prove that given a skew-symmetric matrix R and a measurable set $\omega \subset \Omega$, the function $(RA + AR^t)\, \chi_\omega$ belongs to $K_A(M(\Omega))$. To this purpose we define $G : \mathcal{M}_N \to \mathcal{M}_N \times \mathbf{R}$ by $G(M) = (MM^t, \det(M))$. Since $\mathrm{Ker}(G'(I))$ coincides with the space of skew-symmetric matrices, it is known (see, e.g., [1]) that for $\varepsilon \in \mathbf{R}$ with $|\varepsilon|$ small enough, there exists $P_\varepsilon \in \mathcal{M}_N$ such that $G(P_\varepsilon) = G(I)$ or, equivalently, $P_\varepsilon \in \mathcal{O}_N$, and $(P_\varepsilon - I)/\varepsilon$ converges to R. Defining then

$$A_\varepsilon = P_\varepsilon A P_\varepsilon^t\, \chi_\omega + A\, \chi_{\Omega \setminus \omega}$$

and using that $M(\Omega)$ is invariable by rotations, we deduce that A_ε belongs to $M(\Omega)$ and $(A_\varepsilon - A)/\varepsilon$ converges to $(RA + AR^t)\, \chi_\omega$ in $L^\infty(\Omega, \mathcal{M}_N^s)$. Thus $(RA + AR^t)\, \chi_\omega$ belongs to $K_A(M(\Omega))$.

Using now that the set of skew-symmetric matrices is a vectorial space, condition (3.4), and the arbitrariness of ω, we deduce

$$2(RA) : C = (RA + AR^t) : C = 0 \quad \text{a.e. in } \Omega.$$

For $i, j \in \{1, \ldots, N\}$, $i \neq j$, we take in the above equation R as the matrix

defined by

$$R_{lk} = \begin{cases} 1 & \text{if } l = i,\ k = j, \\ -1 & \text{if } l = j,\ k = i, \\ 0 & \text{in another case.} \end{cases}$$

Then we get

$$(AC)_{ij} - (CA)_{ij} = 0 \quad \text{a.e. in } \Omega,$$

i.e., A and C commute, and then they are mutually diagonalizable. □

Remark 5.3. From Proposition 5.2, assuming that the matrix C is known and that their eigenvalues are all different, we must look for the optimal solution A on the set of matrices of $M(\Omega)$ which have the same eigenvectors as C a.e. So it can be interesting to write condition (4.13) assuming that B is also mutually diagonalizable with C. If we restrict ourselves to the spaces W which are generated by eigenvectors of C, we get the following result: In the conditions of Proposition 5.2, if c_i, a_i, $i \in \{1, \dots, N\}$, are, respectively, the eigenvalues of C and A, then for every $B \in M(\Omega)$ mutually diagonalizable with A and C, with eigenvalues $b_1, \dots, b_N$, we have

$$\sum_{i=1}^{N} c_i(a_i - b_i) + \min_{1 \le \ell \le N} \min_{1 \le i_1 < \cdots < i_\ell \le N} \sum_{j=1}^{\ell} \frac{c_{i_j}(a_{i_j} - b_{i_j})^2}{b_{i_j} + (\ell - 1)a_{i_j}} \ge 0. \tag{5.1}$$

We also note that by Remark 4.16, for A, B, C as above, the condition (5.1) implies in particular (4.15).

Assuming stronger hypotheses, we can improve Proposition 5.2. Proposition 5.4 below is related to a theorem due to Lewis [11], which applies to the optimization of a function $h : \mathcal{M}_N^s \to \mathbf{R}$ convex and invariable by rotations (see also [12], where there is a review of results corresponding to optimization problems on symmetric matrices).

PROPOSITION 5.4. *In the assumptions of Theorem* 3.2*, we assume* $M(\Omega)$ *invariable by rotations and at least one of the following hypotheses:*

(i) $M(\Omega)$ *is convex.*
(ii) $M(\Omega)$ *is* H*-closed and* $N \ge 3$.
(iii) $M(\Omega)$ *is* H*-closed,* $N = 2$ *and* $k = 1$.

Then there exists $Q \in L^\infty(\Omega, \mathcal{M}_N)$*, with* $Q \in \mathcal{O}_N$ *a.e. in* Ω*, such that*

$$\begin{aligned} QAQ^t &= \operatorname{diag}(a_1, \dots, a_N), \\ QCQ^t &= \operatorname{diag}(c_1, \dots, c_N), \end{aligned} \tag{5.2}$$

and $a_1 \le \cdots \le a_N$, $c_1 \le \cdots \le c_N$.

Proof. From Proposition 5.2 there exists $Q \in L^\infty(\Omega, \mathcal{M}_N)$, with $Q \in \mathcal{O}_N$ a.e. in Ω, such that (5.2) holds. Clearly, we can also assume $c_1 \le \cdots \le c_N$ a.e. in Ω. We consider $i, j \in \{1, \dots, N\}$, $i \ne j$, and we take $L \in \mathcal{O}_N$, defined by

$$Le_i = e_j, \quad Le_j = -e_i, \quad Le_l = e_l \quad \forall\, l \ne i, j.$$

Since $M(\Omega)$ is invariable by rotations, the matrix $B = (LQ)^t \operatorname{diag}(a_1, \dots, a_N) LQ$ belongs to $M(\Omega)$. Let us now see that if one of the hypotheses (i), (ii) or (iii) hold, then

$$C : (A - B) \ge 0 \quad \text{a.e. in } \Omega. \tag{5.3}$$

For this purpose, we define $\tilde{M}(\Omega)$ as the convex hull of

$$\{SAS^t \,/\, S \in L^\infty(\Omega, \mathcal{M}_N),\ S \in \mathcal{O}_N \text{ a.e. in } \Omega\},$$

if (i) holds and as the H-closure of this set in cases (ii) and (iii). Then $\tilde{M}(\Omega)$ is contained in $M(\Omega)$ and A belongs to $\tilde{M}(\Omega)$, so A is also a solution of (1.1) with $M(\Omega)$ replaced by $\tilde{M}(\Omega)$.

In case (i), $\tilde{M}(\Omega)$ is convex and local, so from (4.1) and $B - A$ in $K_A(M(\Omega))$, we deduce (5.3).

In cases (ii) and (iii), $\tilde{M}(\Omega)$ is local and H-closed. Moreover, if (ii) holds, then $\mathrm{Ker}(B - A) \neq 0$ a.e. in Ω, while if we have (iii) then, from Theorem 4.19, C has at least a nonpositive eigenvalue. So, in both situations, we deduce (5.3) from (4.17).

From (5.3) we get

$$0 \leq C : (A - B) = c_i(a_i - a_j) + c_j(a_j - a_i) = (c_i - c_j)(a_i - a_j) \quad \text{a.e. in } \Omega.$$

This finishes the proof of Proposition 5.4. □

Remark 5.5. As we have seen in the proof of Proposition 5.4, the order relation between the eigenvalues of C and A is a consequence of (5.3) with B defined as above. So if $N = 2$ and $M(\Omega)$ is H-closed, by Corollary 4.18, the thesis of Proposition 5.4 still holds on the set where C has a least a nonnegative eigenvalue. Where the two eigenvalues are positive, assuming $c_1 < c_2$, the condition (4.15) implies

$$a_1 < a_2 \quad \text{or} \quad a_1 > a_2 \quad \text{and} \quad c_2 a_2 \leq c_1 a_1. \tag{5.4}$$

Related to this inequality, we also remark that if in the conditions of Proposition 5.4 (ii) there are two eigenvalues of C, c_i, c_j such that $c_i \leq c_j \leq 0$, then besides $a_i \leq a_j$, we have $|c_j| a_j \leq |c_i| a_i$.

6. Applications. In this section let us show how the condition (3.4) and the consequences we have obtained from it can be used, in some cases, to obtain $A\nabla y_i$, $A\nabla p_i$, $1 \leq i \leq k$, as explicit functions of ∇y_i, ∇p_i and then, from (1.2) and (3.2), to reduce the optimality conditions given in Theorem 3.2 to a nonlinear system in ∇y_i, ∇p_i. The main difficulty in carrying out this idea is that obtaining the H-closure of a subset of $L^\infty(\Omega, \mathcal{M}_N)$ is a very difficult problem, which has only been solved in some particular cases (see [23], [14], [13], [8], [16], [25]). To simplify the exposition, we have chosen two simple problems where the H-closure is well known. The first consists of the mixture of two homogeneous isotropic materials in dimension two (the problem can also be studied analogously for higher dimensions). This problem has also been studied in [3] and [7]. In this case the set $M(\Omega)$ is convex. In the second problem we consider a nonconvex situation corresponding to a polycrystal in dimension two.

First problem. We start by recalling the following result which has been proved in [23] and [14].

THEOREM 6.1. *We assume $N = 2$. For $0 < \alpha \leq \beta$ and $\theta \in L^\infty(\Omega)$ with $0 \leq \theta \leq 1$ a.e. in Ω, the set $\mathcal{M}_\theta(\Omega)$ of the H-limits of the sequences $\alpha I\ \chi_{\omega_n} + \beta(1 - \chi_{\omega_n})I$, such that $\omega_n \subset \Omega$ are measurable sets and satisfy χ_{ω_n} converges weakly-$*$ in $L^\infty(\Omega)$ to θ, is characterized as follows:*

$\mathcal{M}_\theta \subset L^\infty(\Omega, \mathcal{M}_2^s)$ is the set of matrices such that their eigenvalues λ_1, λ_2 satisfy

the following inequalities a.e. in Ω*:*

$$\begin{cases} \lambda^- \le \lambda_1, \lambda_2 \le \lambda^+, \\ \displaystyle\sum_{i=1}^{2} \frac{1}{\lambda_i - \alpha} \le \frac{1}{\lambda^- - \alpha} + \frac{1}{\lambda^+ - \alpha}, \\ \displaystyle\sum_{i=1}^{2} \frac{1}{\beta - \lambda_i} \le \frac{1}{\beta - \lambda^-} + \frac{1}{\beta - \lambda^+}, \end{cases} \tag{6.1}$$

where λ^-*,* λ^+ *are given by*

$$\lambda^+ = \alpha\theta + \beta(1-\theta), \qquad \lambda^- = \left(\frac{\theta}{\alpha} + \frac{1-\theta}{\beta}\right)^{-1}. \tag{6.2}$$

Let us see how the results obtained in the previous sections permit us to obtain A from C, and then from ∇y_i, ∇p_i, $1 \le i \le k$, for some choices of $M(\Omega)$ related to $\mathcal{M}_\theta(\Omega)$.

In the following, we define

$$\check{\Omega} = \{x \in \Omega \,/\, (c_1(x), c_2(x)) \neq (0,0)\}, \tag{6.3}$$

where c_1, c_2 are the eigenvalues of the matrix C.

PROPOSITION 6.2. *In the assumptions of Theorem* 3.2*, if* $M(\Omega)$ *is the set of matrices defined in Theorem* 6.1 *for a fixed function* θ *and* c_1*,* c_2*,* $c_1 \le c_2$*, are the eigenvalues of* C*, then there exists an associated basis* $\{\mu_1, \mu_2\}$ *of eigenvectors of* C *such that a.e. in* Ω*, we have*

$$A\mu_i = a_i \mu_i, \quad i = 1, 2, \tag{6.4}$$

where a.e. in $\check{\Omega}$*, the functions* a_1*,* a_2 *are given by the following:*

$$\textit{If } c_2 < 0 \textit{ and } \sqrt{\frac{c_2}{c_1}} \ge \frac{\alpha}{\alpha + (\beta-\alpha)\theta} \Rightarrow \begin{cases} a_1 = \alpha + \dfrac{\alpha(\beta-\alpha)(1-\theta)}{2\alpha + (\beta-\alpha)\theta}\left(1 + \sqrt{\dfrac{c_2}{c_1}}\right), \\ a_2 = \alpha + \dfrac{\alpha(\beta-\alpha)(1-\theta)}{2\alpha + (\beta-\alpha)\theta}\left(1 + \sqrt{\dfrac{c_1}{c_2}}\right). \end{cases}$$

$$\textit{If } c_1 > 0 \textit{ and } \sqrt{\frac{c_2}{c_1}} \le \frac{\beta}{\alpha + (\beta-\alpha)\theta} \Rightarrow \begin{cases} a_1 = \beta - \dfrac{\beta(\beta-\alpha)\theta}{\alpha + \beta + (\beta-\alpha)\theta}\left(1 + \sqrt{\dfrac{c_2}{c_1}}\right), \\ a_2 = \beta - \dfrac{\beta(\beta-\alpha)\theta}{\alpha + \beta + (\beta-\alpha)\theta}\left(1 + \sqrt{\dfrac{c_1}{c_2}}\right). \end{cases}$$

In another case $a_1 = \lambda^-$*,* $a_2 = \lambda^+$*, where* λ^-*,* λ^+ *are defined by* (6.2)*.*

Proof. In this case, the set $M(\Omega)$ is local, convex, and invariable by rotations. So we can apply Proposition 5.4 and (4.14), which imply that a.e. in Ω, A satisfies (6.4) for a basis of eigenvectors of C, and $a_1c_1 + a_2c_2$ is the maximum of $\lambda_1c_1 + \lambda_2c_2$, with λ_1, λ_2 in the set defined by (6.1). Solving this maximum problem we get the expressions a_1 and a_2 given in Proposition 6.2. ∎

The above result assumes that θ is known, but clearly this is not a realistic situation. Next, we consider two examples where θ also varies. In the first one we impose the condition

$$\frac{1}{|\Omega|}\int_\Omega \theta \, dx = s, \tag{6.5}$$

with $s \in (0,1)$. This means we know the proportion of the materials defined by α and β but not its local distribution. This usually holds when one material is better than the other but it is also more expensive. In the second situation we consider the case where we do not have any restriction on θ.

PROPOSITION 6.3. *In the assumptions of Theorem* 3.2*, if for* $s \in (0,1)$ *given,* $M(\Omega)$ *is the set of matrices defined in Theorem* 6.1*, with* $\theta \in L^\infty(\Omega)$, $0 \le \theta \le 1$ *a.e. in* Ω *and such that* (6.5) *holds, then the matrix* A *satisfies the thesis of Proposition* 6.2*, where the corresponding function* θ *is such that defining* $F \in L^\infty(\check{\Omega})$ *by*

$$F(x) = \begin{cases} \dfrac{\alpha(\beta^2-\alpha^2)(\sqrt{-c_1}-\sqrt{-c_2})^2}{(2\alpha+(\beta-\alpha)\theta)^2} & \text{if } c_2 < 0 \text{ and } \sqrt{\dfrac{c_2}{c_1}} \ge \dfrac{\alpha}{\alpha+(\beta-\alpha)\theta}, \\ -\dfrac{\beta(\beta^2-\alpha^2)(\sqrt{c_1}+\sqrt{c_2})^2}{(\alpha+\beta+(\beta-\alpha)\theta)^2} & \text{if } c_1 > 0 \text{ and } \sqrt{\dfrac{c_2}{c_1}} \le \dfrac{\beta}{\alpha+(\beta-\alpha)\theta}, \\ -(\beta-\alpha)\left(\dfrac{\alpha\beta}{(\alpha+(\beta-\alpha)\theta)^2}c_1 + c_2\right) & \text{in another case,} \end{cases}$$

there exists $r \in \mathbf{R}$ *which satisfies*

$$\begin{cases} F(x) \le r & \text{a.e. in } \{\theta = 0\} \cap \check{\Omega}, \\ F(x) = r & \text{a.e. in } \{0 < \theta < 1\} \cap \check{\Omega}, \\ F(x) \ge r & \text{a.e. in } \{\theta = 1\} \cap \check{\Omega}. \end{cases} \tag{6.6}$$

Proof. We remark that if θ is the corresponding function associated with A in the definition of the elements of $M(\Omega)$ and we define $M_\theta(\Omega)$ as in the statement of Theorem 6.1, then A is also the solution of (1.1) with $M(\Omega)$ replaced by $M_\theta(\Omega)$. So Proposition 6.2 applies.

Let us now vary θ. For every $\theta^* \in L^\infty(\Omega)$ such that $0 \le \theta^* \le 1$ a.e. in Ω,

$$\frac{1}{|\Omega|}\int_\Omega \theta^*\,dx = s,$$

and $\theta^* = \theta$ a.e. in $\Omega \setminus \check{\Omega}$, we define $A_{\theta^*} \in M(\Omega)$ as the function given by Proposition 6.2 applied to θ^* a.e. in $\check{\Omega}$ and $A_{\theta^*} = A$ a.e. in $\Omega \setminus \check{\Omega}$. Deriving A_{θ^*} with respect to θ^*, we obtain an admissible direction. Then, using condition (3.4), we get

$$\int_\Omega F\vartheta\,dx \le 0 \tag{6.7}$$

for every $\vartheta \in L^\infty(\Omega)$ such that $\vartheta = 0$ a.e. in $\Omega \setminus \check{\Omega}$, $\vartheta \ge 0$ a.e. in $\{\theta = 0\} \cap \check{\Omega}$, $\vartheta \le 0$ a.e. in $\{\theta = 1\} \cap \check{\Omega}$, and

$$\int_\Omega \vartheta\,dx = 0.$$

It is easy to check that this implies the existence of $r \in \mathbf{R}$, which satisfies the statement of the proposition. □

Remark 6.4. The expression of F is strictly decreasing with respect to θ. Then, from (6.6), it is possible to obtain θ as a function of c_1, c_2 and r.

Remark 6.5. In Proposition 6.3, if $r \ge 0$, then $\theta = 0$ a.e. in the set

$$\{c_1 > 0\} \bigcap \left\{\sqrt{\frac{c_2}{c_1}} \le \frac{\beta}{\alpha}\right\} \cap \check{\Omega}.$$

Analogously, if $r \leq 0$, then $\theta = 1$ a.e. in the set

$$\{c_2 < 0\} \bigcap \left\{ \sqrt{\frac{c_2}{c_1}} \geq \frac{\alpha}{\beta} \right\} \cap \check{\Omega}.$$

We finish with the following result.

PROPOSITION 6.6. *In the assumptions of Theorem* 3.2*, let* $M(\Omega)$ *be the set of matrices defined in Theorem* 6.1*, where* θ *is any function in* $L^\infty(\Omega)$ *such that* $0 \leq \theta \leq 1$ *a.e. in* Ω*, and denote by* c_1*,* c_2*,* $c_1 \leq c_2$*, the eigenvalues of* C*. Then there exists an associated basis* $\{\mu_1, \mu_2\}$ *of eigenvectors of* C *such that a.e. in* Ω *we have* (6.4)*. Moreover, a.e. in* $\check{\Omega}$*, the functions* a_1*,* a_2 *are given by the following:*

If $0 \leq c_1 \leq c_2$*, then* $a_1 = a_2 = \beta$*.*

If $c_1 \leq c_2 \leq 0$*, then* $a_1 = a_2 = \alpha$*.*

If $c_1 < 0 < c_2$*, then*

$$\begin{cases} \dfrac{\alpha}{\beta} \geq -\dfrac{c_1}{c_2} \Rightarrow a_1 = a_2 = \beta, \\ \dfrac{\beta}{\alpha} > -\dfrac{c_1}{c_2} > \dfrac{\alpha}{\beta} \Rightarrow a_1 = \sqrt{-\dfrac{\alpha\beta c_2}{c_1}}, \quad a_2 = \alpha + \beta - \sqrt{-\dfrac{\alpha\beta c_1}{c_2}}, \\ -\dfrac{c_1}{c_2} \geq \dfrac{\beta}{\alpha} \Rightarrow a_1 = a_2 = \alpha. \end{cases}$$

Proof. We proceed similarly to Proposition 6.3, but now, in the condition (6.7), the function ϑ does not necessarily satisfy

$$\int_\Omega \vartheta \, dx = 0.$$

This implies that the function F given in Proposition 6.3 satisfies (6.6) with $r = 0$, which easily gives the result. ∎

Second problem. Given a diagonal matrix $\Lambda = \operatorname{diag}(\alpha, \beta)$ with $0 < \alpha < \beta$, let us now consider the optimization problem (1.1) when $M(\Omega)$ is the H-closure of the matrices of the form $R(x)\Lambda R(x)^t$, where R is measurable, and $R(x)$ belongs to $\mathcal{O}_2$ for a.e. $x \in \Omega$ (observe that to assume Λ diagonal is not a restriction). This set $M(\Omega)$ is known (see, e.g., [25], [16]) and agrees with the set of functions $B \in L^\infty(\Omega, \mathcal{M}_2^s)$ such that for a.e. $x \in \Omega$, the eigenvalues $b_1(x)$ and $b_2(x)$ of $B(x)$ satisfy $\alpha \leq b_1(x)$, $b_2(x) \leq \beta$, $b_1(x)b_2(x) = \alpha\beta$. For this choice of $M(\Omega)$, we have the following result.

PROPOSITION 6.7. *In the assumptions of Theorem* 3.2*, if* c_1 *and* c_2*, with* $c_1 \leq c_2$*, are the eigenvalues of* C*, then there exists an associated basis* $\{\mu_1, \mu_2\}$ *of eigenvectors of* C *such that a.e. in* Ω*, we have*

$$A\mu_i = a_i \mu_i, \quad i = 1, 2, \tag{6.8}$$

where a.e. in the set $\check{\Omega}$ *defined by* (6.3)*, the functions* a_1*,* a_2 *are given by*

$$c_2 < 0 \quad \textit{and} \quad \frac{c_2}{c_1} > \frac{\alpha}{\beta} \Rightarrow \begin{cases} a_1 = \sqrt{\alpha\beta}\sqrt{\dfrac{c_2}{c_1}}, \\ a_2 = \sqrt{\alpha\beta}\sqrt{\dfrac{c_1}{c_2}}. \end{cases}$$

If $c_1 > 0$ *and* $\frac{c_2}{c_1} \le \frac{\beta}{\alpha}$, *then there exist three possibilites:*

$$\begin{cases} a_1 = \alpha, \\ a_2 = \beta, \end{cases} \quad or \quad \begin{cases} a_1 = \beta, \\ a_2 = \alpha, \end{cases} \quad or \quad \begin{cases} a_1 = \sqrt{\alpha\beta}\sqrt{\dfrac{c_2}{c_1}}, \\ a_2 = \sqrt{\alpha\beta}\sqrt{\dfrac{c_1}{c_2}}. \end{cases}$$

In another case $a_1 = \alpha$, $a_2 = \beta$.

Proof. Since the set $M(\Omega)$ is invariable by rotations, we can apply Proposition 5.2 to deduce that for a.e. $x \in \Omega$, there exists a basis $\{\mu_1(x), \mu_2(x)\}$ of $\mathbf{R}^2$ such that

$$C(x)\mu_i(x) = c_i(x)\mu_i(x), \quad A(x)\mu_i(x) = a_i(x)\mu_i(x), \quad i = 1, 2,$$

with $c_1(x) \le c_2(x)$, $a_1(x), a_2(x) \in \mathbf{R}$. From the definition of $M(\Omega)$, we also have that for a.e. $x \in \Omega$, there exists $t^*(x) \in [1, \frac{\beta}{\alpha}]$ such that

$$a_1(x) = \alpha t^*(x), \qquad a_2(x) = \frac{\beta}{t^*(x)}.$$

Moreover, from (5.1), we deduce that t^* satisfies

$$\begin{aligned} &(t^*(x) - t)\left(\alpha c_1(x) - \frac{\beta c_2(x)}{t^*(x)t}\right) \\ &\quad + (t^*(x) - t)^2 \min\left\{\frac{\alpha c_1(x)}{t}, \frac{\beta c_2(x)}{(t^*(x))^2 t}, \frac{1}{t^*(x) + t}\left(\alpha c_1(x) + \frac{\beta c_2(x)}{t^*(x)t}\right)\right\} \ge 0 \end{aligned} \tag{6.9}$$

for every $t \in [1, \frac{\beta}{\alpha}]$ and a.e. $x \in \Omega$. In the set where $t^*(x) = 1$, we have $t^*(x) - t < 0$ for every $t \in (1, \frac{\beta}{\alpha}]$. So, dividing by $1 - t$ and taking t converging to 1 on the right, we deduce

$$\alpha c_1 - \beta c_2 \le 0 \quad \text{a.e. in } \{x \in \Omega : \ t^*(x) = 1\}. \tag{6.10}$$

Analogously, we deduce

$$\beta c_1 - \alpha c_2 \ge 0 \quad \text{a.e. in } \left\{x \in \Omega : \ t^*(x) = \frac{\beta}{\alpha}\right\}, \tag{6.11}$$

$$\alpha c_1 - \frac{\beta c_2}{(t^*(x))^2} = 0 \quad \text{a.e. in } \left\{x \in \Omega : \ 1 < t^*(x) < \frac{\beta}{\alpha}\right\}, \tag{6.12}$$

where the statement (6.12) implies

$$c_1 c_2 > 0, \quad t^* = \sqrt{\frac{\beta c_2}{\alpha c_1}} \quad \text{a.e. in } \left\{x \in \check{\Omega} : \ 1 < t^*(x) < \frac{\beta}{\alpha}\right\}. \tag{6.13}$$

Analyzing the different cases which appear depending on the sign of c_1 or c_2, we easily conclude from (6.10), (6.11), and (6.13) the proof of Proposition 6.7. □

Remark 6.8. We have deduced (6.10), (6.11), and (6.13) from inequality (6.9). One could conjecture that this inequality gives, in fact, more information. However, a simple calculus shows that the statements (6.10), (6.11), and (6.13) also imply (6.9).

Remark 6.9. Proposition 6.7 does not give the expressions of a_1 and a_2 in the set where $c_1 > 0$ and $\frac{c_2}{c_1} \le \frac{\beta}{\alpha}$; it gives three possibilities. The possibility $a_1 = \alpha$, $a_2 = \beta$

seems to be the most natural in order to stick continuously with the values of a_1 and a_2 in the other zones. We also note that if $M(\Omega)$ was convex then, using that $B - A$ is an admissible direction for every $B \in M(\Omega)$, we should obtain in place of (6.9) that

$$\alpha t^* + \frac{\beta}{t^*} = \max\left\{\alpha t + \frac{\beta}{t} : \ t \in \left[1, \frac{\beta}{\alpha}\right]\right\},$$

which implies that $t^* = 1$ (and then $a_1 = \alpha$, $a_2 = \beta$) a.e. on the set where $c_1 > 0$ and $\frac{c_2}{c_1} \leq \frac{\beta}{\alpha}$ as well as the expressions of a_1 and a_2 given in Proposition 6.7 in the other cases. However, since $M(\Omega)$ is not convex, this reasoning is not good and thus the only conclusion we obtain is that stated in Proposition 6.7.

REFERENCES

[1] V. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin, *Optimal Control*, Consultants Bureau, New York, 1987.

[2] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[3] G. Allaire, *Shape Optimization by the Homogenization Method*, Appl. Math. Sci. 146, Springer-Verlag, New York, 2002.

[4] H. Attouch, *Variational Convergence for Functions and Operators*, Applicable Mathematics Series, Pitman, London, 1984.

[5] A. Bensoussan, J. L. Lions, and G. Papanicolau, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North–Holland, Amsterdam, 1978.

[6] J. Casado Díaz, J. Couce Calvo, and J. D. Martín Gómez, *Sobre el control de la matriz de difusión*, in Actas encuentro de matemáticos andaluces (Sevilla, 13–17 November 2000), E. Briales-Morales, A. Carriazo-Rubio, T. Chacón-Rebollo, P. Real-Jurado, and A. Romero-Jiménez, eds., Publicaciones de la Universidad de Sevilla, 2001, pp. 237–244.

[7] J. Casado Díaz, J. Couce Calvo, and J. D. Martín Gómez, *Sobre la identificación de la matriz de difusión mediante varios experimentos*, in Actas XVII CEDYA, VII CMA (Salamanca, 24–28 September 2001), L. Ferragut and A. Santos, eds., 2001, CD-ROM.

[8] A. V. Cherkaev, *Variational Methods for Structural Optimization*, Appl. Math. Sci. 140, Springer-Verlag, New York, 2000.

[9] J. Deny and J. L. Lions, *Les espaces de Beppo Levi*, Ann. Inst. Fourier (Grenoble), 5 (1953–1954), pp. 305–370.

[10] G. Francfort and G. W. Milton, *Sets of conductivity and elasticity tensors stable under lamination*, Comm. Pure Appl. Math., 47 (1994), pp. 257–279.

[11] A. S. Lewis, *Convex analysis on the Hermitian matrices*, SIAM J. Optim., 6 (1996), pp. 164–177.

[12] A. S. Lewis and M. L. Overton, *Eigenvalue optimization*, in Acta Numerica, Acta Numer. 5, Cambridge University Press, Cambridge, UK, 1996, pp. 149–190.

[13] K. A. Lurie, *Applied Optimal Control Theory of Distributed Systems*, Plenum Press, New York, 1993.

[14] K. A. Lurie and A. V. Cherkaev, *Exact estimates of the conductivity of a binary mixture of isotropic materials*, Proc. Roy. Soc. Edinburgh Sect. A, 104 (1986), pp. 21–38.

[15] G. W. Milton, *A link between sets of tensors stable under lamination and quasiconvexity*, Comm. Pure Appl. Math., 47 (1994), pp. 959–1003.

[16] G. W. Milton, *The Theory of Composites*, Cambridge Monogr. Appl. Comput. Math., Cambridge University Press, Cambridge, UK, 2002.

[17] F. Murat, *Théorèmes de non-existence pour des problèmes de contrôle dans le coefficients*, C. R. Acad. Sci. Paris Sér. A-B, 274 (1972), pp. 395–398.

[18] F. Murat and L. Tartar, *On the control of coefficients in partial differential equations*, in Topics in the Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, A. Cherkaev and R. Kohn, eds., Birkhäuser Boston, Boston, 1997, pp. 1–8.

[19] F. Murat and L. Tartar, *H-convergence*, in Topics in the Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, A. Cherkaev and R. Kohn, eds., Birkhäuser Boston, Boston, 1997, pp. 21–43.

[20] F. Murat and L. Tartar, *Calculus of variations and homogenization*, in Topics in the Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, A. Cherkaev and R. Kohn, eds., Birkhäuser Boston, Boston, 1997, pp. 139–173.

[21] U. Raitums, *On the local representation of G-closure*, Arch. Ration. Mech. Anal., 158 (2001), pp. 213–234.

[22] S. Spagnolo, *Sulla convergenza di soluzioni di equazione paraboliche ed ellipticheé*, Ann. Scuola Norm. Sup. Pisa (3), 22 (1968), pp. 571–597.

[23] L. Tartar, *Estimations fines de coefficients homogénéisés*, in Ennio de Giorgi colloquium (Paris, 1983), Res. Notes Math. 125, P. Kree, ed., Pitman, London, 1985, pp. 168–187.

[24] L. Tartar, *Remarks on optimal design problems*, in Calculus of Variations, Homogenization and Continuum Mechanics, Ser. Adv. Math. Appl. Sci. 18, G. Bouchitté, G. Buttazzo, and P. Suquet, eds., World Scientific, Singapore, 1994, pp. 279–296.

[25] L. Tartar, *An Introduction to the homogenization method in optimal design*, in Optimal Shape Design (CIM/CIME, Summer School, Trôia, 1–6 June 1998), Lecture Notes in Math. 1740, A. Cellina and A. Ornelas, eds., Springer-Verlag, Berlin, 2000, pp. 47–156.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 240–255

OPTIMAL CONTROL OF BILATERAL OBSTACLE PROBLEMS*

MAÏTINE BERGOUNIOUX† AND SUZANNE LENHART‡

Abstract. We consider an optimal control problem where the state satisfies a bilateral elliptic variational inequality and the control functions are the upper and lower obstacles. We seek a state that is close to a desired profile and the H^2 norms of the obstacles are not too large. Existence results are given and an optimality system is derived. A particular case is studied that needs no compactness assumption, via a monotonicity method.

Key words. optimal control, obstacle problem, variational inequalities, semilinear elliptic equations

AMS subject classifications. 49J20, 49M25

DOI. 10.1137/S0363012902416912

1. Introduction.

We consider an optimal control problem where the state satisfies a bilateral elliptic variational inequality and the control functions are the upper and lower obstacles. We seek a state that is close to a desired profile and for which the H^2 norms of the obstacles are not too large.

This type of problem appears in shape optimization [11, 19]. It may concern, for example, the optimal shape for a dam. The obstacle gives the form to be designed such that the pressure of the fluid inside the dam is close to a desired value. This is equivalent in some sense to controlling the free boundary (see [14], for example).

There are few papers about the control of the obstacle in variational inequalities. In an early paper [10], Bucur, Buttazzo, and Trabeschi give an existence result for a problem of the following type:

$$\min\left\{F(g), g \in X_\psi(\Omega), \int_\Omega g(x)dx = c\right\},$$

where $X_\psi(\Omega) = \{g : \Omega \to \bar{\mathbb{R}},\ g \leq \psi\}$ and F is increasing γ lower semicontinuous. The monotonicity assumption on F allows the avoidance of the compactness assumption used in the present paper. In [3] the problem is studied with completely different methods (the γ-convergence is not used), using once again an implicit monotonicity assumption. Then [1, 2, 9, 17, 18] generalized that result to include source terms, semilinear and quasilinear elliptic operators, and parabolic operators, replacing the monotonicity hypothesis by a compactness one. One can refer also to the book by Chipot [13] and its references for background and estimates for bilateral variational inequalities. For recent work on control in lower order terms, see the works by Chen [12] on semilinear elliptic bilateral variational inequalities and by Bergounioux [8] on semilinear elliptic variational inequalities. The new feature in this paper is the control of the two obstacles in the bilateral case.

The motivation of our work is threefold. First, as mentioned above, many shape optimization problems can be modeled as the problem we describe here below. Sec-

*Received by the editors November 3, 2002; accepted for publication (in revised form) October 31, 2003; published electronically June 15, 2004.
http://www.siam.org/journals/sicon/43-1/41691.html
†Département de Mathématiques, Université d'Orléans, Laboratoire MAPMO, UFR Sciences, BP 6759, 45067 Orléans, France (Maitine.Bergounioux@labomath.univ-orleans.fr).
‡Mathematics Department, University of Tennessee, Knoxville, TN 37996-1300 (lenhart@math.utk.edu).

ondly, as usual, in optimal control theory, we are looking for a first order necessary optimality system that allows us to compute the solution exactly (often not the case) or numerically. The numerical strategy can be the direct resolution of the optimality system by a shooting method or any other fixed point method, or the setting of adapted algorithms whose convergence may be proved using Lagrange multipliers; see [5, 6, 7]. Therefore, we claim that the establishment of optimality conditions is the first step before any numerical analysis.

Thirdly, from the theoretical point of view, the problem is involved in a wider class of (open) problems, which can be (formally) described as follows:

$$\min\{J(u,\chi),\ u=\mathcal{T}(\chi),\ \chi\in U_{ad}\subset\mathcal{U}\},$$

where $\mathcal{T}$ is an operator which associates u to χ, where u is a (or the only) solution to

$$\forall v\in K(u,\chi),\quad \langle\mathcal{A}(u,\chi),u-v\rangle\geq 0,$$

where K is a *multiapplication* from $\mathcal{X}\times\mathcal{U}$ to $2^{\mathcal{X}}$, where $\mathcal{X}$ is a Banach space and $\mathcal{U}$ a Hilbert space. Let us give an example: let $\mathcal{Y}$ be a Banach space and A a differential operator (linear or not), parabolic or elliptic from $\mathcal{Y}$ to the dual space $\mathcal{Y}'$, and h an application from $\mathbb{R}\times\mathbb{R}\times\mathbb{R}$ to $\mathbb{R}$. The differential equation that relates the control χ to the state function u (i.e., the state "equation") is

$$\langle Au,v-u,\rangle_{\mathcal{Y},\mathcal{Y}'}+h(u,\chi,v)-h(u,\chi,u)\geq(\chi,v-u)\qquad\forall v\in\mathcal{Y},$$

where

1. $h(u,\chi,v)=h(v)$ gives the classical variational inequalities;
2. $h(u,\chi,v)=h(\chi,v)$ gives (for example) obstacle problems (where the obstacle is the control): this is the problem we investigate here;
3. $h(u,\chi,v)=h(u,v)$ leads to quasi-variational inequalities whose study is very delicate.

Let us specify the problem under consideration as follows.

Let Ω be an open bounded subset of $\mathbb{R}^n$ with a smooth boundary $\partial\Omega$. We consider the bilinear form $a(.,.)$ defined on $H_o^1(\Omega)\times H_o^1(\Omega)$ by

$$a(u,v)=\sum_{i,j=1}^{n}\int_\Omega a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,dx+\sum_{i=1}^{n}\int_\Omega b_i\frac{\partial u}{\partial x_i}v\,dx+\int_\Omega cu\,v\,dx,\tag{1.1}$$

where a_{ij},b_i,c belong to $L^\infty(\Omega)$. Moreover, we assume that a_{ij} belongs to $\mathcal{C}^{0,1}(\bar\Omega)$ (the space of Lipschitz continuous functions in Ω) and that c is nonnegative. The bilinear form $a(.,.)$ is continuous on $H_o^1(\Omega)\times H_o^1(\Omega)$,

$$\exists M>0,\ \forall(u,v)\in H_o^1(\Omega)\times H_o^1(\Omega),\qquad a(u,v)\leq M\,\|u\|_{H_o^1(\Omega)}\|v\|_{H_o^1(\Omega)},\tag{1.2}$$

and is coercive:

$$\exists\alpha>0,\ \forall u\in H_o^1(\Omega),\qquad a(u,u)\geq\alpha\|u\|^2_{H_o^1(\Omega)}.\tag{1.3}$$

We shall define $\|\cdot\|_V$, the norm in the Banach space V, and more precisely $\|\cdot\|_2$ the $L^2(\Omega)$-norm. In the same way, $\langle\cdot,\cdot\rangle$ denotes the duality product between $H^{-1}(\Omega)$ and $H_o^1(\Omega)$, and $(\cdot,\cdot)_2$ the $L^2(\Omega)$-inner product. We call $A\in\mathcal{L}(H_o^1(\Omega),H^{-1}(\Omega))$ the linear (elliptic) operator associated with a such that $\langle Au,v\rangle=a(u,v)$. Given $\varphi,\psi\in H_o^1(\Omega)$, we set

$$K(\varphi,\psi)=\{u\in H_o^1(\Omega)\mid\varphi\leq u\leq\psi\text{ a.e. in }\Omega\},\tag{1.4}$$

which is a nonempty, closed, convex subset of $H_o^1(\Omega)$. All inequalities as $u \leq \psi$ are understood in the almost everywhere sense. We choose $f \in L^2(\Omega)$ as a source term. For any $\varphi, \psi \in H_o^1(\Omega)$, it is well known (see [4], for example) that the variational inequality

$$\forall v \in K(\varphi,\psi), \qquad a(u, v-u) \geq (f, v-u),\ u \in K(\varphi,\psi), \tag{1.5}$$

has a unique solution u that belongs to $H_o^1(\Omega)$. In addition, if $\varphi, \psi \in H^2(\Omega)$, then u belongs to $H^2(\Omega) \cap H_o^1(\Omega)$ (see [4], for example). From now on, we consider H^2-obstacle functions so that we may define the operator $\mathcal{T}$ from $(H^2(\Omega) \cap H_o^1(\Omega)) \times (H^2(\Omega) \cap H_o^1(\Omega))$ to $H^2(\Omega) \cap H_o^1(\Omega)$ such that $\mathcal{T}(\varphi,\psi) = u$ is the unique solution to the variational inequality (1.5). It is known that this operator is not differentiable (and even not continuous if we define it on the whole space $H_o^1(\Omega)$).

Let U_{ad} be the set of admissible controls defined as follows:

$$U_{ad} = \{(\varphi,\psi) \in (H^2(\Omega) \cap H_o^1(\Omega)) \times (H^2(\Omega) \cap H_o^1(\Omega)) \ \mid \ \varphi \leq \psi \ \}.$$

Now, we consider the optimal control problem $(\mathcal{P})$ defined as follows:

$$\min\left\{ J(\varphi,\psi) \stackrel{def}{=} \frac{1}{2}\int_\Omega (\mathcal{T}(\varphi,\psi)-z)^2 dx + \frac{\nu}{2}\left(\int_\Omega \left((\Delta\varphi)^2 + (\Delta\psi)^2\right)\, dx\right), (\varphi,\psi) \in U_{ad} \right\},$$

where $z \in L^2(\Omega)$. In what follows we require continuity properties for $\mathcal{T}$; therefore we need H^2-a priori estimates. We could assume that U_{ad} is H^2-bounded, but this choice leads to technical difficulties in deriving an optimality system. An equivalent theoretical tool for getting such estimates is to involve the H^2-norm in the objective functional. That is why we have added the term $\int_\Omega \left((\Delta\varphi)^2 + (\Delta\psi)^2\right)\, dx$. The positive real number ν can be small but it is fixed: it ensures that the weight of the H^2-norm is not too large with respect to the least squared minimization to drive the state $u = \mathcal{T}(\varphi,\psi)$ as close as possible to a desired profile z. Of course, this is a regularizing term which gives compactness properties. It is not too unrealistic to look for smooth obstacles: the gradient is bounded, and thus we avoid oscillations, and the curvature is bounded as well.

Let us give an outline of the paper. The next section is devoted to the study of the state-inequality and some properties of the operator $\mathcal{T}$. Then we give an existence result for a solution of $(\mathcal{P})$. Section 3 is devoted to the optimality system. We consider an approximate optimality system and convergence results to pass to the limit in this system. In the last section, we present a particular case, where the H^2-boundedness assumption is weakened and replaced by an H^1-boundedness assumption. We use monotonicity tools to deal with the lack of compactness and obtain a "complete" optimality system.

2. Properties of the state operator $\mathcal{T}$. We consider the following technique (see [4, 3]) to approximate the variational inequality by a semilinear equation. More precisely, we define

$$\beta(r) = \begin{cases} 0 & \text{if } r \geq 0, \\ -r^2 & \text{if } r \in [-\frac{1}{2}, 0], \\ r + \frac{1}{4} & \text{if } r < -\frac{1}{2}. \end{cases} \tag{2.1}$$

Note that $\min\{0, r\} \le \beta(r) \le 0$ and $\beta \in \mathcal{C}^1(\mathbb{R})$ with

$$\beta'(r) = \begin{cases} 0 & \text{if } r \ge 0, \\ -2r & \text{if } r \in [-\frac{1}{2}, 0], \\ 1 & \text{if } r < -\frac{1}{2}. \end{cases} \tag{2.2}$$

We introduce the following semilinear elliptic equation:

$$Au + \frac{1}{\delta}\left(\beta(u-\varphi) - \beta(\psi - u)\right) = f \text{ in } \Omega,\ u = 0 \text{ on } \partial\Omega. \tag{2.3}$$

As $\beta(\cdot - \varphi) - \beta(\psi - \cdot)$ is nondecreasing, it is known that the above equation has a unique solution $u^\delta \in H^2(\Omega) \cap H_o^1(\Omega)$, and we set $u^\delta = \mathcal{T}^\delta(\varphi, \psi)$.

THEOREM 2.1. *Let $(\varphi^\delta, \psi^\delta) \in U_{ad}$ be a sequence strongly convergent in $H_o^1(\Omega)$ to some (φ, ψ) as δ tends to 0. Then the sequence $u^\delta = \mathcal{T}^\delta(\varphi^\delta, \psi^\delta)$ converges to $u = \mathcal{T}(\varphi, \psi)$ strongly in $H_o^1(\Omega)$.*

Proof. For every $(\varphi^\delta, \psi^\delta) \in U_{ad}$, we set $u^\delta = \mathcal{T}^\delta(\varphi^\delta, \psi^\delta)$. Equation (2.3) is equivalent to

$$\forall v \in H_o^1(\Omega), \qquad a(u^\delta, v) + \frac{1}{\delta}\left(\beta(u^\delta - \varphi^\delta) - \beta(\psi^\delta - u^\delta), v\right)_2 = (f, v)_2. \tag{2.4}$$

First, we choose $v = u^\delta - \varphi^\delta$. Equation (2.4) gives

$$a(u^\delta, u^\delta - \varphi^\delta) + \frac{1}{\delta}\int_\Omega \beta(u^\delta - \varphi^\delta)(u^\delta - \varphi^\delta)dx = \frac{1}{\delta}\int_\Omega \beta(\psi^\delta - u^\delta)(u^\delta - \varphi^\delta)dx + \int_\Omega f(u^\delta - \varphi^\delta)dx.$$

Note that if $u^\delta(x) - \varphi^\delta(x) \ge 0$, then $\beta(u^\delta(x) - \varphi^\delta(x)) = 0$; otherwise $\beta(u^\delta(x) - \varphi^\delta(x)) \le 0$. In any case we have

$$\beta(u^\delta - \varphi^\delta)(u^\delta - \varphi^\delta) \ge 0 \quad \text{a.e. in } \Omega.$$

Similarly, if $\psi^\delta(x) \le u^\delta(x)$, then with $\varphi^\delta(x) \le \psi^\delta(x)$ we get

$$\beta(\psi^\delta - u^\delta)(u^\delta - \varphi^\delta) \le 0 \quad \text{a.e. in } \Omega,$$

since $\beta(\psi^\delta(x) - u^\delta(x)) = 0$ if $\psi^\delta(x) > u^\delta(x)$. By the above, we obtain

$$a(u^\delta, u^\delta - \varphi^\delta) \le (f, u^\delta - \varphi^\delta)_2,$$

$$a(u^\delta, u^\delta) \le a(u^\delta, \varphi^\delta) + (f, u^\delta - \varphi^\delta)_2,$$

$$\alpha\|u^\delta\|^2_{H_o^1(\Omega)} \le M\|u^\delta\|_{H_o^1(\Omega)}\|\varphi^\delta\|_{H_o^1(\Omega)} + \|f\|_2\|u^\delta\|_2 + \|f\|_2\|\varphi^\delta\|_2,$$

$$\alpha\|u^\delta\|^2_{H_o^1(\Omega)} \le M\|u^\delta\|_{H_o^1(\Omega)}\|\varphi^\delta\|_{H_o^1(\Omega)} + \|f\|_2\|u^\delta\|_{H_o^1(\Omega)} + \|f\|_2\|\varphi^\delta\|_{H_o^1(\Omega)}.$$

This gives the existence of a constant C_1 depending only on f and a such that

$$\|u^\delta\|_{H_o^1(\Omega)} \le C_1\|\varphi^\delta\|_{H_o^1(\Omega)}. \tag{2.5}$$

Similarly, once again using $\varphi^\delta \le \psi^\delta$, we have the following estimate:

$$\|u^\delta\|_{H_o^1(\Omega)} \le C_2\|\psi^\delta\|_{H_o^1(\Omega)}, \tag{2.6}$$

where $C_2 \ge 0$ depends only on f and a.

• Next, we estimate $\beta(u^\delta-\varphi^\delta)-\beta(\psi^\delta-u^\delta)$. As $0\le\beta'\le 1$, $\beta(u^\delta-\varphi^\delta)-\beta(\psi^\delta-u^\delta)\in H^1_o(\Omega)$. Using (2.4) with $v=\beta(u^\delta-\varphi^\delta)-\beta(\psi^\delta-u^\delta)$ gives

$$\begin{aligned}
&\frac{1}{\delta}\|\beta(u^\delta-\varphi^\delta)-\beta(\psi^\delta-u^\delta)\|_2^2\\
&\quad= \ -a(u^\delta,\beta(u^\delta-\varphi^\delta)-\beta(\psi^\delta-u^\delta))+\big(f,\beta(u^\delta-\varphi^\delta)-\beta(\psi^\delta-u^\delta)\big)_2\\
&\quad\le M\|u^\delta\|_{H^1_o}\left(\|\beta(u^\delta-\varphi^\delta)\|_{H^1_o}+\|\beta(\psi^\delta-u^\delta)\|_{H^1_o}\right)\\
&\qquad+\|f\|_2\left(\|\beta(u^\delta-\varphi^\delta)\|_2+\|\beta(\psi^\delta-u^\delta)\|_2\right)\\
&\quad\le\left(M\|u^\delta\|_{H^1_o}+\|f\|_2\right)\left(\|\beta(u^\delta-\varphi^\delta)\|_{H^1_o}+\|\beta(\psi^\delta-u^\delta)\|_{H^1_o}\right).
\end{aligned}$$

As $\beta(0)=0$ and $0\le\beta'\le 1$, then $\|\beta(v)\|_2\le\|v\|_2$. In addition,

$$\frac{\partial\beta(v)}{\partial x_i}=\beta'(v)\frac{\partial v}{\partial x_i}$$

yields that $\|\nabla\beta(v)\|_2\le\|\nabla v\|_2$. Finally, $\|\beta(v)\|_{H^1_o(\Omega)}\le\|v\|_{H^1_o(\Omega)}$, and we get

$$\|\beta(u^\delta-\varphi^\delta)-\beta(\psi^\delta-u^\delta)\|_2^2\le\delta\left(M\|u^\delta\|_{H^1_o}+\|f\|_2\right)\left(\|u^\delta-\varphi^\delta\|_{H^1_o}+\|\ \psi^\delta-u^\delta\|_{H^1_o}\right).$$

With estimates (2.5) and (2.6) we obtain

$$\|\beta(u^\delta-\varphi^\delta)-\beta(\psi^\delta-u^\delta)\|_2^2\le\delta\left(C_3\|\varphi^\delta\|_{H^1_o}+\|f\|_2\right)\left(C_4\|\varphi^\delta\|_{H^1_o}+C_5\|\psi^\delta\|_{H^1_o}\right). \tag{2.7}$$

• Now, we consider a sequence $(\varphi^\delta,\psi^\delta)\in U_{ad}$ strongly convergent in $H^1_o(\Omega)$ to some (φ,ψ). As U_{ad} is H^1-closed, $(\varphi,\psi)\in U_{ad}$. In addition, $(\varphi^\delta,\psi^\delta)$ is bounded in $H^1_o(\Omega)$ uniformly with respect to δ. Therefore, estimates (2.5) and/or (2.6) imply the weak convergence of $u^\delta=\mathcal{T}(\varphi^\delta,\psi^\delta)$ to some $u\in H^1_o(\Omega)$.

Relation (2.7) proves that $\beta(u^\delta-\varphi^\delta)-\beta(\psi^\delta-u^\delta)$ strongly converges to 0 in $L^2(\Omega)$. With the strong convergence of $(\varphi^\delta,\psi^\delta,u^\delta)$ to (φ,ψ,u) in $L^2(\Omega)$ and $0\le\beta'\le 1$, we get $\beta(u-\varphi)-\beta(\psi-u)=0$ and

$$\beta(u-\varphi)=\beta(\psi-u)\qquad\text{a.e. in }\Omega;$$

if $u(x)<\varphi(x)\le\psi(x)$, then $\beta(u-\varphi)(x)<0$ and $\beta(\psi-u)(x)=0$: this is not possible. Similarly, if $\varphi(x)\le\psi(x)<u(x)$, then $\beta(u-\varphi)(x)=0$ and $\beta(\psi-u)(x)>0$. The only remaining possibility is $\varphi\le u\le\psi$, that is, $u\in K(\varphi,\psi)$.

• Let us prove that $u=\mathcal{T}(\varphi,\psi)$. As we already know that $u\in K(\varphi,\psi)$, it is sufficient to prove that

$$\forall v\in K(\varphi,\psi),\qquad a(u,v-u)\ge(f,v-u)_2.$$

We choose $v\in K(\varphi,\psi)$ and set $v^\delta=\inf\left(\sup\left(v,\varphi^\delta\right),\psi^\delta\right)$. Then $v^\delta\in K(\varphi^\delta,\psi^\delta)$ and v^δ strongly converges to v in $H^1_o(\Omega)$. Equation (2.4) with $v=v^\delta-u^\delta$ gives

$$a(u^\delta,v^\delta-u^\delta)=\frac{1}{\delta}\int_\Omega\left(\beta(\psi^\delta-u^\delta)-\beta(u^\delta-\varphi^\delta)\right)(v^\delta-u^\delta)\,dx+\int_\Omega f(v^\delta-u^\delta)\,dx.$$

- ▷ If $\varphi^\delta\le u^\delta\le\psi^\delta$, then $\beta(\psi^\delta-u^\delta)=\beta(u^\delta-\varphi^\delta)=0$.
- ▷ If $u^\delta<\varphi^\delta\le\psi^\delta$, then $\beta(\psi^\delta-u^\delta)=0$ and $v^\delta-u^\delta>0$. As $\beta(u^\delta-\varphi^\delta)\le 0$, we get $(\beta(\psi^\delta-u^\delta)-\beta(u^\delta-\varphi^\delta))(v^\delta-u^\delta)\ge 0$.
- ▷ If $\varphi^\delta\le\psi^\delta<u^\delta$, then $\beta(u^\delta-\varphi^\delta)=0$ and $v^\delta-u^\delta<0$. As $\beta(\psi^\delta-u^\delta)\le 0$, so that $(\beta(\psi^\delta-u^\delta)-\beta(u^\delta-\varphi^\delta))(v^\delta-u^\delta)\ge 0$.

Finally,

$$a(u^\delta, v^\delta - u^\delta) \geq (f, v^\delta - u^\delta)_2,$$
$$a(u^\delta, u^\delta) \leq a(u^\delta, v^\delta) - (f, v^\delta - u^\delta)_2.$$

We may pass to the limit and use the lower semicontinuity of a:

$$a(u,u) \leq \liminf_{\delta\to 0} a(u^\delta, u^\delta) \leq \lim_{\delta\to 0} a(u^\delta, v^\delta) - (f, v^\delta - u^\delta)_2 = a(u,v) - (f, v-u)_2.$$

This gives $a(u, v-u) \geq (f, v-u)_2$ for every $v \in K(\varphi,\psi)$ and $u = \mathcal{T}(\varphi,\psi)$.

• It remains to prove the strong convergence of u^δ to u in $H^1_o(\Omega)$. We again use $w^\delta = \inf(\sup(u, \varphi^\delta), \psi^\delta)$ that strongly converges to u in $H^1_o(\Omega)$. It is sufficient to prove that $w^\delta - u^\delta$ strongly converges to 0 in $H^1_o(\Omega)$. We use (2.4) once again:

$$a(w^\delta - u^\delta, w^\delta - u^\delta) = a(w^\delta, w^\delta - u^\delta) - a(u^\delta, w^\delta - u^\delta)$$
$$= a(w^\delta, w^\delta - u^\delta) + \frac{1}{\delta}\int_\Omega \big(\beta(u^\delta - \varphi^\delta) - \beta(\psi^\delta - u^\delta)\big)(w^\delta - u^\delta)\,dx - \int_\Omega f(w^\delta - u^\delta)\,dx\,.$$

A similar analysis as above shows that $(\beta(u^\delta - \varphi^\delta) - \beta(\psi^\delta - u^\delta))(w^\delta - u^\delta) \leq 0$ a.e. in Ω. Thus

$$a(w^\delta - u^\delta, w^\delta - u^\delta) \leq a(w^\delta, w^\delta - u^\delta) - \big(f, w^\delta - u^\delta\big)_2\,,$$
$$\alpha\|w^\delta - u^\delta\|^2_{H^1_o(\Omega)} \leq a(w^\delta, w^\delta - u^\delta) - \big(f, w^\delta - u^\delta\big)_2\,.$$

The right-hand side is convergent to 0 (with the strong convergence of w^δ to u in $H^1_o(\Omega)$ and the weak convergence of $w^\delta - u^\delta$ to 0 in $H^1_o(\Omega)$). Therefore $w^\delta - u^\delta \to 0$ strongly in $H^1_o(\Omega)$. □

Remark 2.1. Note that the sequence $(\varphi^\delta, \psi^\delta)$ belongs to $H^2(\Omega)\times H^2(\Omega)$, and it is sufficient to assume that $(\varphi^\delta, \psi^\delta)$ weakly converges to some (φ,ψ) in $H^2(\Omega)\times H^2(\Omega)$ to get the conclusion of Theorem 2.1.

COROLLARY 2.1. *For any $(\varphi,\psi) \in U_{ad}$, the sequence $u^\delta = \mathcal{T}^\delta(\varphi,\psi)$ strongly converges to $u = \mathcal{T}(\varphi,\psi)$ in $H^1_o(\Omega)$.*

COROLLARY 2.2. *There exists a constant C depending only on f and a such that, for any $(\varphi,\psi) \in U_{ad}$,*

$$\|\mathcal{T}(\varphi,\psi)\|_{H^1_o(\Omega)} \leq C \min\big(\|\varphi\|_{H^1_o(\Omega)}, \|\psi\|_{H^1_o(\Omega)}\big)\,. \tag{2.8}$$

Proof. We choose $\varphi^\delta = \varphi$, $\psi^\delta = \psi$, and $u^\delta = \mathcal{T}^\delta(\varphi,\psi)$; as u^δ strongly converges to $\mathcal{T}(\varphi,\psi)$ in $H^1_o(\Omega)$, we pass to the limit in (2.5) and (2.6). □

Let us conclude this section with a continuity result for the operator $\mathcal{T}$.

THEOREM 2.2. *$\mathcal{T}$ is continuous from U_{ad} endowed with the $H^2(\Omega)\times H^2(\Omega)$ sequential weak topology to $H^1_o(\Omega)$ endowed with the sequential weak topology.*

Proof. Assume that $(\varphi_k, \psi_k) \in U_{ad}$ is a sequence that weakly converges to (φ,ψ) in $H^2(\Omega)\times H^2(\Omega)$. Then (φ,ψ) belongs to U_{ad}, and (φ_k,ψ_k) strongly converges to (φ,ψ) in $H^1_o(\Omega)\times H^1_o(\Omega)$ as well. We set $u_k = \mathcal{T}(\varphi_k,\psi_k)$.

Let $v \in K(\varphi,\psi)$ and set (as previously) $v_k = \inf(\sup(v,\varphi_k),\psi_k) \in K(\varphi_k,\psi_k)$ that strongly converges to v in $H^1_o(\Omega)$. As $u_k = \mathcal{T}(\varphi_k,\psi_k)$, we get $a(u_k, v_k - u_k) \geq (f, v_k - u_k)_2$, that is,

$$a(u_k,u_k) \geq a(u_k, v_k) - (f, v_k - u_k)_2.$$

Using Corollary 2.2, u_k is bounded in $H_o^1(\Omega)$ and weakly converges to some u (up to a subsequence). Using the lower semicontinuity of a in the previous relation gives

$$a(u,u) \geq a(u,v) - (f, v-u)_2.$$

In addition, $\varphi_k \leq u_k \leq \psi_k$ implies $\varphi \leq u \leq \psi$. Therefore $u = \mathcal{T}(\varphi,\psi)$, and the whole sequence converges. ∎

Now, we turn back to the optimal control problem $(\mathcal{P})$. We first give an existence result for the solution.

THEOREM 2.3. *Problem $(\mathcal{P})$ has (at least) an optimal solution (φ^*, ψ^*).*

Proof. Let (φ_k, ψ_k) be a minimizing sequence. As $J(\varphi_k, \psi_k)$ is bounded, φ_k and ψ_k are H^2-bounded and converge to some φ^* and ψ^*, respectively, weakly in $H^2(\Omega)$ (and strongly in $H_o^1(\Omega)$ (up to a subsequence)). The weak cluster point belongs to U_{ad} and with Theorem 2.2 we know that $u_k = \mathcal{T}(\varphi_k, \psi_k)$ weakly converges to $u^* = \mathcal{T}(\varphi^*, \psi^*)$ in $H_o^1(\Omega)$. Then,

$$\begin{aligned} J(\varphi^*, \psi^*) &= \frac{1}{2}\int_\Omega (\mathcal{T}(\varphi^*, \psi^*) - z)^2 dx + \frac{\nu}{2}\int_\Omega \left((\Delta\varphi^*)^2 + (\Delta\psi^*)^2\right) dx \\ &\leq \liminf_{k\to\infty} J(\varphi_k, \psi_k) = \inf(\mathcal{P}). \end{aligned}$$

Thus (φ^*, ψ^*) is an optimal solution to $(\mathcal{P})$. ∎

3. Optimality system.

3.1. An approximate problem. We first consider an approximate problem and establish an optimality system for this problem. Let (φ^*, ψ^*) be an optimal solution to $(\mathcal{P})$ and $u^* = \mathcal{T}(\varphi^*, \psi^*)$. For any $\delta > 0$, we define

$$J_\delta(\varphi,\psi) \stackrel{def}{=} \frac{1}{2}\left[\int_\Omega \left(\mathcal{T}^\delta(\varphi,\psi) - z\right)^2 dx + \nu\int_\Omega \left((\Delta\varphi)^2 + (\Delta\psi)^2\right) dx + \|\varphi-\varphi^*\|_2^2 + \|\psi-\psi^*\|_2^2\right].$$

The last term in J_δ is an adapted penalization term that focuses on a chosen solution (φ^*, ψ^*). Let us define an approximate optimal control problem as follows:

$$(\mathcal{P}_\delta) \qquad \min\{J_\delta(\varphi,\psi),\ (\varphi,\psi) \in U_{ad}\}.$$

THEOREM 3.1. *Problem $(\mathcal{P}_\delta)$ has (at least) an optimal solution $(\varphi^\delta, \psi^\delta)$. Moreover, the sequence $(\varphi^\delta, \psi^\delta)$ weakly converges to (φ^*, ψ^*) in $H^2(\Omega)$, while $u^\delta = \mathcal{T}^\delta(\varphi^\delta, \psi^\delta)$ strongly converges to $u^* = \mathcal{T}(\varphi^*, \psi^*)$ in $H_o^1(\Omega)$.*

Proof. The functional J_δ is obviously lower semicontinuous and coercive. Therefore, Problem $(\mathcal{P}_\delta)$ has (at least) an optimal solution $(\varphi^\delta, \psi^\delta)$. We call $u^\delta = \mathcal{T}^\delta(\varphi^\delta, \psi^\delta)$ and note that, for any $\delta > 0$,

$$J_\delta(\varphi^\delta, \psi^\delta) \leq J_\delta(\varphi^*, \psi^*) = \frac{1}{2}\left[\int_\Omega \left(\mathcal{T}^\delta(\varphi^*, \psi^*) - z\right)^2 dx + \nu\int_\Omega \left((\Delta\varphi^*)^2 + (\Delta\psi^*)^2\right) dx\right]; \tag{3.1}$$

using Corollary 2.1, $\mathcal{T}^\delta(\varphi^*, \psi^*) \to u^* = \mathcal{T}(\varphi^*, \psi^*)$ strongly in $H_o^1(\Omega)$, and $J_\delta(\varphi^*, \psi^*) \to J(\varphi^*, \psi^*)$. Therefore, there exist $\delta_o > 0$ and a constant j^* such that

$$\forall \delta \leq \delta_o, \qquad J_\delta(\varphi^\delta, \psi^\delta) \leq j^* < +\infty.$$

Therefore φ^δ and ψ^δ are H^2-bounded uniformly with respect to $\delta \leq \delta_o$. We apply Theorem 2.1: up to subsequences, we get

$$\varphi^\delta \to \tilde{\varphi} \text{ and } \psi^\delta \to \tilde{\psi} \text{ weakly in } H^2(\Omega) \text{ and strongly in } H_o^1(\Omega) \text{ and}$$
$$u^\delta \to \tilde{u} = \mathcal{T}(\tilde{\varphi}, \tilde{\psi}) \text{ strongly in } H_o^1(\Omega).$$

As U_{ad} is weakly closed, then $(\tilde{\varphi}, \tilde{\psi}) \in U_{ad}$. Using the lower semicontinuity of J_δ and (3.1), we obtain

$$\begin{aligned} J(\tilde{\varphi}, \tilde{\psi}) + \frac{1}{2}\|\tilde{\varphi} - \varphi^*\|_2^2 + \frac{1}{2}\|\tilde{\psi} - \psi^*\|_2^2 &\leq \liminf_{\delta \to 0} J_\delta(\varphi^\delta, \psi^\delta) \\ &\leq \limsup_{\delta \to 0} J_\delta(\varphi^\delta, \psi^\delta) \\ &\leq \lim_{\delta \to 0} J_\delta(\varphi^*, \psi^*) = J(\varphi^*, \psi^*) \\ &\leq J(\tilde{\varphi}, \tilde{\psi}). \end{aligned}$$

Here, we used that $(\tilde{\varphi}, \tilde{\psi})$ is an admissible pair for problem $(\mathcal{P})$. This yields that $\|\tilde{\varphi} - \varphi^*\|_2^2 + \|\tilde{\psi} - \psi^*\|_2^2 \leq 0$; thus $\tilde{\varphi} = \varphi^*$, $\tilde{\psi} = \psi^*$, and the whole sequence is convergent. In addition,

$$\lim_{\delta \to 0} J_\delta(\varphi^\delta, \psi^\delta) = J(\varphi^*, \psi^*). \qquad \square$$

3.2. An approximate optimality system. We first establish a (necessary) optimality system for $(\mathcal{P}_\delta)$, using the following result on the Gâteaux-derivative of the operator $\mathcal{T}^\delta$.

LEMMA 3.1. *The mapping $\mathcal{T}^\delta$ is Gâteaux-differentiable at any $(\varphi, \psi) \in U_{ad}$:*

$$\forall (\xi, \eta) \in H_o^1(\Omega) \times H_o^1(\Omega), \quad \frac{\mathcal{T}^\delta(\varphi + t\xi, \psi + t\,\eta) - \mathcal{T}^\delta(\varphi, \psi)}{t} \overset{w}{\rightharpoonup} v^\delta \text{ in } H_o^1(\Omega), \text{ as } t \to 0,$$

where v^δ is the solution of the sensitivity equation

$$\begin{aligned} &Av^\delta + \frac{1}{\delta}\left(\beta'(w^\delta - \varphi) + \beta'(\psi - w^\delta)\right) v^\delta = \frac{1}{\delta}\left(\beta'(w^\delta - \varphi)\,\xi + \beta'(\psi - w^\delta)\,\eta\right) \text{ in } \Omega, \\ &v^\delta = 0 \text{ on } \partial\Omega, \end{aligned}$$

where $w^\delta = \mathcal{T}^\delta(\varphi, \psi)$.

Proof. The proof is similar to [3, Lemma 5.1]. $\square$

In what follows, we keep the notation of subsection 3.1. As usual we define the (approximate) adjoint state of the problem as the solution $p^\delta \in H_o^1(\Omega)$ of

$$A^* p^\delta + \frac{1}{\delta}\left(\beta'(u^\delta - \varphi^\delta) + \beta'(\psi^\delta - u^\delta)\right) p^\delta = u^\delta - z \text{ in } \Omega, \ p^\delta = 0 \text{ on } \partial\Omega,$$

where A^* denotes the adjoint operator of A. As $(\varphi^\delta, \psi^\delta)$ is a solution to $(\mathcal{P}_\delta)$, we have

$$\forall (\varphi, \psi) \in U_{ad}, \qquad \frac{d}{dt} J_\delta\left(\varphi^\delta + t(\varphi - \varphi^\delta), \psi^\delta + t(\psi - \psi^\delta)\right)_{|t=0} \geq 0.$$

This gives

$$\begin{aligned} \int_\Omega \left(\chi^\delta(u^\delta - z) + \nu\Delta\varphi^\delta\Delta(\varphi - \varphi^\delta) + \nu\Delta\psi^\delta\Delta(\psi - \psi^\delta)\right) dx \\ + \int_\Omega \left((\varphi^\delta - \varphi^*)(\varphi - \varphi^\delta) + (\psi^\delta - \psi^*)(\psi - \psi^\delta)\right) dx \geq 0, \end{aligned}$$

where $\chi^\delta \in H_o^1(\Omega)$ satisfies

$$\begin{aligned} &A\chi^\delta + \frac{1}{\delta}\left(\beta'(u^\delta - \varphi^\delta) + \beta'(\psi^\delta - u^\delta)\right)\chi^\delta \\ &\quad = \frac{1}{\delta}\left(\beta'(u^\delta - \varphi^\delta)(\varphi - \varphi^\delta) + \beta'(\psi^\delta - u^\delta)(\psi - \psi^\delta)\right) \text{ in } \Omega. \end{aligned}$$

Using the definition of p^δ, we obtain

$$a^*(p^\delta,\chi^\delta)+\int_\Omega \frac{\beta'(u^\delta-\varphi^\delta)+\beta'(\psi^\delta-u^\delta)}{\delta}p^\delta\chi^\delta dx + \nu\int_\Omega \left(\Delta\varphi^\delta\Delta(\varphi-\varphi^\delta)+\Delta\psi^\delta\Delta(\psi-\psi^\delta)\right)dx + \int_\Omega \left((\varphi^\delta-\varphi^*)(\varphi-\varphi^\delta)+(\psi^\delta-\psi^*)(\psi-\psi^\delta)\right)\,dx\geq 0.$$

Here a^* denotes the adjoint form of a (associated with the adjoint operator A^*). Then,

$$a(\chi^\delta,p^\delta)+\int_\Omega \frac{\beta'(u^\delta-\varphi^\delta)+\beta'(\psi^\delta-u^\delta)}{\delta}\chi^\delta p^\delta dx + \nu\int_\Omega \left(\Delta\varphi^\delta\Delta(\varphi-\varphi^\delta)+\Delta\psi^\delta\Delta(\psi-\psi^\delta)\right)dx + \int_\Omega \left((\varphi^\delta-\varphi^*)(\varphi-\varphi^\delta)+(\psi^\delta-\psi^*)(\psi-\psi^\delta)\right)\,dx\geq 0;$$

we obtain

$$\int_\Omega \frac{\beta'(u^\delta-\varphi^\delta)}{\delta}p^\delta(\varphi-\varphi^\delta)+\frac{\beta'(\psi^\delta-u^\delta)}{\delta}p^\delta(\psi-\psi^\delta)\,dx + \nu\int_\Omega \left(\Delta\varphi^\delta\Delta(\varphi-\varphi^\delta)+\Delta\psi^\delta\Delta(\psi-\psi^\delta)\right)dx + \int_\Omega \left((\varphi^\delta-\varphi^*)(\varphi-\varphi^\delta)+(\psi^\delta-\psi^*)(\psi-\psi^\delta)\right)\,dx\geq 0.$$

In what follows we set

$$\mu_1^\delta=\frac{\beta'(u^\delta-\varphi^\delta)}{\delta}p^\delta,\quad \mu_2^\delta=\frac{\beta'(\psi^\delta-u^\delta)}{\delta}p^\delta,\quad \text{and } \mu^\delta=\mu_1^\delta+\mu_2^\delta\,(\in L^2(\Omega)). \tag{3.2}$$

Finally, we obtain the following result.

THEOREM 3.2. *Assume that $(\varphi^\delta,\psi^\delta)$ is an optimal solution to $(\mathcal{P}_\delta)$ and $u^\delta=\mathcal{T}^\delta(\varphi^\delta,\psi^\delta)$. Then there exist $p^\delta\in H_o^1(\Omega)\cap H^2(\Omega)$ and $\mu_1^\delta,\mu_2^\delta\in L^2(\Omega)$ such that the following optimality system is satisfied:*

$$Au^\delta+\frac{1}{\delta}\left(\beta(u^\delta-\varphi^\delta)-\beta(\psi^\delta-u^\delta)\right)=f \text{ in } \Omega,\ u^\delta=0 \text{ on } \partial\Omega, \tag{3.3a}$$

$$A^*p^\delta+\mu_1^\delta+\mu_2^\delta=u^\delta-z \text{ in } \Omega,\ p^\delta=0 \text{ on } \partial\Omega, \tag{3.3b}$$

$$\forall(\varphi,\psi)\in U_{ad},\quad \left(\mu_1^\delta+\varphi^\delta-\varphi^*,\varphi-\varphi^\delta\right)_2+\left(\mu_2^\delta+\psi^\delta-\psi^*,\psi-\psi^\delta\right)_2 + \nu\left(\Delta\varphi^\delta,\Delta(\varphi-\varphi^\delta)\right)_2+\nu\left(\Delta\psi^\delta,\Delta(\psi-\psi^\delta)\right)_2\geq 0. \tag{3.3c}$$

Let us give additional properties of μ_1^δ and μ_2^δ, as follows.

PROPOSITION 3.1. *The supports of μ_1^δ and μ_2^δ are disjoint so that*

$$\left(\mu_1^\delta,\mu_2^\delta\right)_2=0 \text{ and } \|\mu^\delta\|_2^2=\|\mu_1^\delta\|_2^2+\|\mu_2^\delta\|_2^2.$$

It follows that

$$\mu^\delta=\begin{cases}\mu_1^\delta & \text{on } \{\ x\in\Omega|\ u^\delta(x)<\varphi^\delta(x)\ \}\overset{def}{=}\omega_\delta^1,\\ \mu_2^\delta & \text{on } \{\ x\in\Omega|\ u^\delta(x)>\psi^\delta(x)\ \}\overset{def}{=}\omega_\delta^2,\\ 0 & \text{elsewhere.}\end{cases} \tag{3.4}$$

Moreover, $\mu_i^\delta\in H_o^1(\Omega)$ for $i=1,2$.

Proof. The first assertion is clear because of the definition of μ_i^δ. Note that ω_δ^i are open subsets of Ω since u^δ, ϕ^δ, and ψ^δ are continuous ($n \le 3$).

Let us compute the derivative of μ_1^δ in the distribution sense (the same proof holds for μ_2^δ); we recall that $\mu_1^\delta = \frac{\eta(u^\delta - \varphi^\delta)}{\delta}\, p^\delta$, where

$$\eta(r) = \begin{cases} 0 & \text{if } r \ge 0, \\ -2r & \text{if } r \in [-\frac{1}{2}, 0], \\ 1 & \text{if } r \le -\frac{1}{2}. \end{cases}$$

The function η is derivable as a distribution, and its derivative is

$$\eta'(r) = \begin{cases} 0 & \text{if } r \in \,]-\infty - \frac{1}{2}[\cup]0, +\infty[, \\ -2 & \text{if } r \in [-\frac{1}{2}, 0], \end{cases}$$

because η is continuous and piecewise $\mathcal{C}^1$. So μ_1^δ is differentiable in the distribution sense and

$$\frac{\partial \mu_1^\delta}{\partial x_i} = \frac{1}{\delta}\,\eta'(u^\delta - \varphi^\delta)\,\frac{\partial (u^\delta - \varphi^\delta)}{\partial x_i}\,p^\delta + \frac{1}{\delta}\,\eta(u^\delta - \varphi^\delta)\frac{\partial p^\delta}{\partial x_i}\ .$$

As $\eta(u^\delta - \varphi^\delta), \eta'(u^\delta - \varphi^\delta) \in L^\infty(\Omega)$, $p^\delta \in H^2(\Omega) \cap H_o^1(\Omega) \subset L^\infty(\Omega)$ (since $n \le 3$), and $\frac{\partial (u^\delta - \varphi^\delta)}{\partial x_i}, \frac{\partial p^\delta}{\partial x_i} \in L^2(\Omega)$, then $\frac{\partial \mu_1^\delta}{\partial x_i} \in L^2(\Omega)$. Therefore $\mu_1^\delta \in H^1(\Omega)$. As $\mu_1^\delta = 0$ on $\partial\Omega$ (since $p^\delta \in H_o^1(\Omega)$), then $\mu_1^\delta \in H_o^1(\Omega)$. ∎

3.3. Optimality system for ($\mathcal{P}$). We would like to pass to the limit in the previous system. We already know the weak convergence of $(\varphi^\delta, \psi^\delta)$ in H^2 and the strong convergence of u^δ in $H_o^1(\Omega)$. We have to estimate p^δ and μ_i^δ, $i = 1, 2$.

THEOREM 3.3. *When $\delta \to 0$, p^δ weakly converges in $H_o^1(\Omega)$ to some p^* (taking a subsequence). The sequence μ^δ is bounded in $H^{-1}(\Omega)$ and weakly converges to some $\mu^* \in H^{-1}(\Omega)$.*

Proof. We use (3.3b) to obtain

$$a^*(p^\delta, p^\delta) + \int_\Omega \frac{\beta'(u^\delta - \varphi^\delta) + \beta'(\psi^\delta - u^\delta)}{\delta}(p^\delta)^2\,dx = \left(u^\delta - z, p^\delta\right)_2. \tag{3.5}$$

As $\beta' \ge 0$, this implies

$$\alpha \|p^\delta\|^2_{H_o^1(\Omega)} \le \|u^\delta - z\|_2 \|p^\delta\|_2,$$

$$\|p^\delta\|_{H_o^1(\Omega)} \le \frac{1}{\alpha}\|u^\delta - z\|_2. \tag{3.6}$$

This implies that p^δ weakly converges to some p^* in $H_o^1(\Omega)$. Therefore $A^* p^\delta$ is bounded in $H^{-1}(\Omega)$ uniformly with respect to δ and $\mu^\delta = -A^* p^\delta + u^\delta - z$ as well. Thus, there exists $\mu^* \in H^{-1}(\Omega)$ such that μ^δ weakly converges to μ^* in $H^{-1}(\Omega)$. Therefore we may pass to the limit in (3.3b); this gives

$$A^* p^* + \mu^* = u^* - z \text{ in } \Omega,\ p^* = 0 \text{ on } \partial\Omega\ . \tag{3.7}$$ ∎

Let $\chi \in H^2(\Omega) \cap H_o^1(\Omega)$ and choose $\varphi = \varphi^\delta + \chi$, $\psi = \psi^\delta + \chi$. Obviously, $(\varphi, \psi) \in U_{ad}$, and we use relation (3.3c) to obtain

$$\left(\mu^\delta + \varphi^\delta - \varphi^* + \psi^\delta - \psi^*, \chi\right)_2 + \nu\left(\Delta\varphi^\delta + \Delta\psi^\delta, \Delta\chi\right)_2 \ge 0;$$

that is (since we have chosen any $\chi \in H^2(\Omega) \cap H_o^1(\Omega)$), $\forall \chi \in H^2(\Omega) \cap H_o^1(\Omega)$,

$$\left(\mu^\delta + \varphi^\delta - \varphi^* + \psi^\delta - \psi^*, \chi\right)_2 + \nu \left(\Delta\varphi^\delta + \Delta\psi^\delta, \Delta\chi\right)_2 = 0. \tag{3.8}$$

Let us set $h^\delta = (\Delta\varphi^\delta + \Delta\psi^\delta) \in L^2(\Omega)$ so that relation (3.8) reads

$$\forall \chi \in H^2(\Omega) \cap H_o^1(\Omega), \qquad \left(\mu^\delta + \varphi^\delta - \varphi^* + \psi^\delta - \psi^*, \chi\right)_2 + \nu \left(h^\delta, \Delta\chi\right)_2 = 0.$$

Using the previous relation with $\chi \in \mathcal{D}(\Omega)$ gives

$$-\nu\, \Delta h^\delta = \mu^\delta + \varphi^\delta - \varphi^* + \psi^\delta - \psi^* \tag{3.9}$$

in the sense of distributions, and relation (3.8) is equivalent to

$$\forall \chi \in H^2(\Omega) \cap H_o^1(\Omega), \qquad \left(-\Delta h^\delta, \chi\right)_2 + \left(h^\delta, \Delta\chi\right)_2 = 0. \tag{3.10}$$

As $\mu^\delta + \varphi^\delta - \varphi^* + \psi^\delta - \psi^* \in L^2(\Omega)$, we have

$$h^\delta \in \mathcal{V} \stackrel{def}{=} \{\ h \in L^2(\Omega)\ \mid\ \Delta h \in L^2(\Omega)\ \}.$$

Therefore, the traces $h^\delta_{|\partial\Omega}$ and $\frac{\partial h^\delta}{\partial n}{}_{|\partial\Omega}$ can be defined in $H^{-1/2}(\partial\Omega)$ and $H^{-3/2}(\partial\Omega)$, respectively (see Lions [15, p. 229], for example). Using a generalized Greens' formula gives

$$\left(h^\delta, \Delta\chi\right)_2 = \left(\Delta h^\delta, \chi\right)_2 + \int_{\partial\Omega} h^\delta \frac{\partial\chi}{\partial n}\, d\sigma$$

for any $\chi \in H^2(\Omega) \cap H_o^1(\Omega)$. Then with (3.10) we obtain

$$\forall \chi \in H^2(\Omega) \cap H_o^1(\Omega), \qquad \int_{\partial\Omega} h^\delta \frac{\partial\chi}{\partial n}\, d\sigma = 0,$$

that is, with the surjectivity of the trace application (see Lions and Magenes [16, p. 47]),

$$\forall \zeta \in H^{\frac{1}{2}}(\partial\Omega), \qquad \int_{\partial\Omega} h^\delta \zeta\, d\sigma = 0.$$

This implies $h^\delta_{|\partial\Omega} = 0$. Therefore, h^δ is the unique solution of

$$-\nu\Delta h^\delta = \mu^\delta + \varphi^\delta - \varphi^* + \psi^\delta - \psi^* \in L^2(\Omega), \qquad h^\delta = 0 \text{ on } \partial\Omega,$$

and belongs to $H^2(\Omega) \cap H_o^1(\Omega)$. Moreover, we know that μ^δ is weakly convergent to μ^* in $H^{-1}(\Omega)$ and that $(\varphi^\delta, \psi^\delta)$ is strongly convergent to (φ^*, ψ^*) in $H_o^1(\Omega)$. Therefore h^δ is weakly convergent to h^* weakly in $H_o^1(\Omega)$ and $-\nu\Delta h^* = \mu^*$. Uniqueness of the limit implies that $h^* = \Delta(\varphi^* + \psi^*)$.

As $0 \le \beta' \le 1$, we get with (3.5)

$$a^*(p^\delta, p^\delta) \le \left(u^\delta - z, p^\delta\right)_2.$$

Using the lower semicontinuity of a^* gives

$$a^*(p^*, p^*) \le \liminf_{\delta \to +\infty} \left(u^\delta - z, p^\delta\right)_2 = \left(u^* - z, p^*\right)_2.$$

Thus, with (3.7) we get

$$\langle A^*p^*, p^*\rangle = (u^* - z, p^*)_2 - \langle \mu^*, p^*\rangle \le (u^* - z, p^*)_2\,,$$

where $\langle\cdot,\cdot\rangle$ denotes the (H^{-1}, H^1_o) duality pairing. Thus $\langle \mu^*, p^*\rangle \ge 0$. Therefore, we obtain the following optimality system.

THEOREM 3.4. *Let (φ^*, ψ^*) be an optimal solution to $(\mathcal{P})$. Then $\Delta(\varphi^* + \psi^*) \in H^1_o(\Omega)$ and there exist $p^* \in H^1_o(\Omega)$ and $\lambda^* \ge 0$ in $H^{-1}(\Omega)$ such that the following optimality system is satisfied:*

$$u^* = \mathcal{T}(\varphi^*, \psi^*), \tag{3.11a}$$

$$A^*p^* = u^* - z^* - \mu^* \text{ in } \Omega,\ p^* = 0 \text{ on } \partial\Omega, \tag{3.11b}$$

$$\langle p^*, \mu^*\rangle \ge 0, \tag{3.11c}$$

$$\mu^* = \mu_1^* + \mu_2^*, \text{ with } \mu_1^* = -\lambda^* - \nu\Delta^2\varphi^* \text{ and } \mu_2^* = \lambda^* - \nu\Delta^2\psi^*, \tag{3.11d}$$

$$\langle \lambda^*, \varphi^* - u^*\rangle = 0 \text{ and } \langle \lambda^*, u^* - \psi^*\rangle = 0. \tag{3.11e}$$

Proof. We have already proved (3.11a)–(3.11c). It remains to show (3.11d) and (3.11e).

Setting $\varphi = \varphi^\delta - \chi$ and $\psi = \psi^\delta$ in relation (3.3c) with $\chi \in H^2(\Omega) \cap H^1_o(\Omega)$, $\chi \ge 0$, we get

$$\forall \chi \ge 0, \qquad \left(\mu_1^\delta + \varphi^\delta - \varphi^*, \chi\right)_2 + \nu\left(\Delta\varphi^\delta, \Delta\chi\right)_2 \le 0.$$

The application $\chi \mapsto (\mu_1^\delta + \varphi^\delta - \varphi^*, \chi)_2 + \nu(\Delta\varphi^\delta, \Delta\chi)_2$ from $H^2(\Omega) \cap H^1_o(\Omega)$ to $\mathbb{R}$ is clearly linear and continuous. Therefore, there exists a measure λ_1^δ (in the dual of $H^2(\Omega) \cap H^1_o(\Omega)$) such that

$$\left(\mu_1^\delta + \varphi^\delta - \varphi^*, \chi\right)_2 + \nu\left(\Delta\varphi^\delta, \Delta\chi\right)_2 = \left\langle \lambda_1^\delta, \chi\right\rangle.$$

As $\langle \lambda_1^\delta, \chi\rangle \le 0$, then λ_1^δ is a nonpositive measure. If we choose $\chi \in \mathcal{D}(\Omega)$, we get

$$\left(\mu_1^\delta + \varphi^\delta - \varphi^*, \chi\right)_2 + \nu\left(\Delta^2\varphi^\delta, \chi\right)_2 = \left\langle \lambda_1^\delta, \chi\right\rangle$$

in $\mathcal{D}'(\Omega)$, so that the measure is

$$\lambda_1^\delta = \mu_1^\delta + \varphi^\delta - \varphi^* + \nu\Delta^2\varphi^\delta.$$

Similarly, there exists a nonnegative measure: $\lambda_2^\delta = \mu_2^\delta + \psi^\delta - \psi^* + \nu\Delta^2\psi^\delta$.
In addition, we have seen that

$$\mu_1^\delta + \mu_2^\delta = \mu^\delta = -\nu\Delta h^\delta - (\varphi^\delta - \varphi^* + \psi^\delta - \psi^*),$$

where $h^\delta = (\Delta\varphi^\delta + \Delta\psi^\delta) \in H^2(\Omega) \cap H^1_o(\Omega)$. Thus

$$\lambda_1^\delta + \lambda_2^\delta = 0.$$

Finally we get the existence of a nonnegative measure λ^δ such that

$$\mu_1^\delta = -(\lambda^\delta + \varphi^\delta - \varphi^* + \nu\Delta^2\varphi^\delta) \text{ and } \mu_2^\delta = \lambda^\delta - (\psi^\delta - \psi^* + \nu\Delta^2\psi^\delta). \tag{3.12}$$

Now we claim that μ_i^δ, $i = 1, 2$, are bounded as measures by a constant independent of δ. Indeed, we use the following result: *let ω be an open subset of Ω and $\mu \in L^2(\Omega)$ $(\subset H^{-1}(\Omega))$. Then $\mu_{|\omega} \in H^{-1}(\omega)$ and $\|\mu_{|\omega}\|_{H^{-1}(\omega)} \le \|\mu\|_{H^{-1}(\Omega)}$.*

We have noticed with relation (3.4) that $\mu_{i|_{\omega_i^\delta}}^\delta = \mu_i^\delta$ for $i = 1, 2$. Moreover, the support of μ_i^δ is included in ω_i^δ and $\mu_i^\delta \in L^2(\Omega)$; thus

$$\|\mu_i^\delta\|_{H^{-1}(\Omega)} = \sup_{\substack{\varphi \in H_o^1(\Omega) \\ \|\varphi\|_{H_o^1} \leq 1}} \left\langle \mu_i^\delta, \varphi \right\rangle_{H^{-1}, H_o^1} = \sup_{\substack{\varphi \in H_o^1(\Omega) \\ \|\varphi\|_{H_o^1} \leq 1}} \left(\mu_i^\delta, \varphi\right)_{L^2(\omega_i^\delta)} = \sup_{\substack{\varphi \in H_o^1(\omega_i^\delta) \\ \|\varphi\|_{H_o^1} \leq 1}} \left\langle \mu_i^\delta, \varphi \right\rangle_{H^{-1}, H_o^1}$$

$$= \|\mu_i^\delta\|_{H^{-1}(\omega_i^\delta)} = \|\mu_{|_{\omega_i^\delta}}^\delta\|_{H^{-1}(\omega_i^\delta)} \leq \|\mu^\delta\|_{H^{-1}(\Omega)}.$$

As μ^δ is bounded in $H^{-1}(\Omega)$, then μ_i^δ is bounded in $H^{-1}(\Omega)$ as well by a constant independent of δ. Therefore μ_i^δ converges to μ_i^* (as a measure) and λ^δ converges to a nonnegative measure λ^*, as well, with

$$\mu_1^* = -\lambda^* - \nu\Delta^2\varphi^* \text{ and } \mu_2^* = \lambda^* - \nu\Delta^2\psi^* \ .$$

Let us show the complementarity relation (3.11e). Inserting (3.12) into relation (3.3c) with $\varphi = \psi = u^\delta$ gives

$$\forall \delta, \qquad \left(\lambda^\delta, \varphi^\delta - u^\delta\right) + \left(\lambda^\delta, u^\delta - \psi^\delta\right) \geq 0;$$

passing to the limit with respect to δ yields

$$\langle \lambda^*, \varphi^* - u^* \rangle + \langle \lambda^*, u^* - \psi^* \rangle \geq 0.$$

As $\lambda^* \geq 0$ (as a measure) and $\varphi^* - u^* \leq 0$, $u^* - \psi^* \leq 0$, we get

$$\langle \lambda^*, \varphi^* - u^* \rangle + \langle \lambda^*, u^* - \psi^* \rangle = 0.$$

Note that we have the sum of two terms that are separately nonpositive. Therefore

$$\langle \lambda^*, \varphi^* - u^* \rangle = \langle \lambda^*, u^* - \psi^* \rangle = 0.$$

This achieves the proof. □

4. A particular case. In some particular cases we may obtain a more complete optimality system using monotonicity methods instead of compactness methods. Let us consider the special case where the obstacle H^2-norm that occurs in the cost functional is replaced by the H^1-norm given by the way of the bilinear form a, with

$$a(u, v) = (\nabla u, \nabla v)_2 \text{ and } f \equiv 0.$$

More precisely, we define the following cost functional:

$$\hat{J}(\varphi, \psi) \stackrel{def}{=} \frac{1}{2} \int_\Omega (\mathcal{T}(\varphi, \psi) - z)^2 \, dx + \frac{\nu}{2} \left(\|\nabla\varphi\|_2^2 + \|\nabla\psi\|_2^2 \right).$$

In this case, obstacle functions are not necessarily H^2-regular, and the admissible control set is defined as follows:

$$\hat{U}_{ad} = \{ \ (\varphi, \psi) \in H_o^1(\Omega) \times H_o^1(\Omega) \ \mid \ \varphi \leq \psi \ \}.$$

We consider the corresponding optimal control problem:

$$(\hat{\mathcal{P}}) \qquad \min \{ \ \hat{J}(\varphi, \psi) \, , \, (\varphi, \psi) \in \hat{U}_{ad} \ \}.$$

THEOREM 4.1. *Problem* $(\hat{\mathcal{P}})$ *has (at least) an optimal control solution* $(\hat{\varphi}, \hat{\psi})$ *with* $\hat{\varphi} = \hat{\psi} = \hat{u} = \mathcal{T}(\hat{u}, \hat{u})$ *for some* $\hat{u} \in H_o^1(\Omega)$.

Proof. Let $(\varphi_k, \psi_k) \in \hat{U}_{ad}$ be a minimizing sequence:

$$\lim_{k \to +\infty} \hat{J}(\varphi_k, \psi_k) = \inf(\hat{\mathcal{P}}).$$

Therefore, (φ_k, ψ_k) is bounded in $H_o^1(\Omega)$ and weakly converges to some $(\hat{\varphi}, \hat{\psi})$ in $H_o^1(\Omega)$; moreover, $(\hat{\varphi}, \hat{\psi}) \in \hat{U}_{ad}$. Relation (2.8) of Corollary 2.2 yields that $u_k = \mathcal{T}(\varphi_k, \psi_k)$ is bounded in $H_o^1(\Omega)$ as well and weakly converges to some $\hat{u}$ in $H_o^1(\Omega)$. Note that $\hat{u} \in K(\hat{u}, \hat{u})$ and $\forall v \in K(\hat{u}, \hat{u})$, $a(\hat{u}, v - \hat{u}) = 0 = (f, v - \hat{u})_2$. Thus $\hat{u} = \mathcal{T}(\hat{u}, \hat{u})$. As $\hat{J}$ is lower semicontinuous, we have

$$\begin{aligned} \hat{J}(\hat{u}, \hat{u}) &= \frac{1}{2} \int_\Omega (\hat{u} - z)^2 \, dx + \frac{\nu}{2} \left(\|\nabla \hat{u}\|_2^2 + \|\nabla \hat{u}\|_2^2 \right) \\ &\leq \liminf_{k \to +\infty} \frac{1}{2} \int_\Omega (u_k - z)^2 \, dx + \frac{\nu}{2} \left(\|\nabla u_k\|_2^2 + \|\nabla u_k\|_2^2 \right). \end{aligned}$$

Note that if $u = \mathcal{T}(\varphi, \psi)$, we have

$$\forall v \in K(\varphi, \psi), \qquad (\nabla u, \nabla u)_2 \leq (\nabla u, \nabla v)_2 \ .$$

This gives

$$\forall v \in K(\varphi, \psi), \qquad \|\nabla u\|_2 \leq \|\nabla v\|_2,$$

and with $v = \varphi$ and $v = \psi$,

$$\|\nabla \mathcal{T}(\varphi, \psi)\|_2 \leq \|\nabla \varphi\|_2 \quad \text{and} \quad \|\nabla \mathcal{T}(\varphi, \psi)\|_2 \leq \|\nabla \psi\|_2. \tag{4.1}$$

These inequalities will replace the compactness assumption. Indeed,

$$\begin{aligned} \hat{J}(\hat{u}, \hat{u}) &\leq \liminf_{k \to +\infty} \frac{1}{2} \int_\Omega (u_k - z)^2 \, dx + \frac{\nu}{2} \left(\|\nabla \varphi_k\|_2^2 + \|\nabla \psi_k\|_2^2 \right) \\ &= \liminf_{k \to +\infty} \hat{J}(\varphi_k, \psi_k) \\ &= \inf(\hat{\mathcal{P}}). \end{aligned}$$

Thus $(\hat{u}, \hat{u})$ is a solution. □

In this very case, we may give the generic form for the optimal solution, as follows.

THEOREM 4.2. *Any solution* $(\tilde{\varphi}, \tilde{\psi})$ *to* $(\hat{\mathcal{P}})$ *satisfies*

$$\tilde{\varphi} = \tilde{\psi} = \mathcal{T}(\tilde{\varphi}, \tilde{\psi}).$$

Proof. We have just found a solution in this form. Let $(\tilde{\varphi}, \tilde{\psi})$ be another optimal solution, and set $\tilde{u} = \mathcal{T}(\tilde{\varphi}, \tilde{\psi})$. Then

$$\begin{aligned} \hat{J}(\tilde{u}, \tilde{u}) &= \frac{1}{2} \int_\Omega (\tilde{u} - z)^2 \, dx + \frac{\nu}{2} \left(\|\nabla \tilde{u}\|_2^2 + \|\nabla \tilde{u}\|_2^2 \right) \\ &\leq \frac{1}{2} \int_\Omega (\tilde{u} - z)^2 \, dx + \frac{\nu}{2} \left(\|\nabla \tilde{\varphi}\|_2^2 + \|\nabla \tilde{\psi}\|_2^2 \right) \text{ with (4.1)} \\ &= \hat{J}(\tilde{\varphi}, \tilde{\psi}) = \inf(\hat{\mathcal{P}}). \end{aligned}$$

Thus $\hat{J}(\tilde{u},\tilde{u}) = \hat{J}(\tilde{\varphi},\tilde{\psi})$ and

$$2\,\|\nabla\tilde{u}\|_2^2 = \|\nabla\tilde{\varphi}\|_2^2 + \|\nabla\tilde{\psi}\|_2^2\ .$$

With (4.1), we get

$$\|\nabla\tilde{\varphi}\|_2^2 + \|\nabla\tilde{\psi}\|_2^2 = \|\nabla\tilde{u}\|_2^2 + \|\nabla\tilde{u}\|_2^2 \le \|\nabla\tilde{u}\|_2^2 + \|\nabla\tilde{\psi}\|_2^2.$$

Therefore $\|\nabla\tilde{\varphi}\|_2^2 \le \|\nabla\tilde{u}\|_2^2$, and with (4.1) we obtain $\|\nabla\tilde{\varphi}\|_2 = \|\nabla\tilde{u}\|_2$ (and $\|\nabla\tilde{\psi}\|_2 = \|\nabla\tilde{u}\|_2$ in a similar way). Moreover, $(\nabla\tilde{u},\nabla\tilde{\varphi})_2 \le \|\nabla\tilde{u}\|_2\|\nabla\tilde{\varphi}\|_2 = \|\nabla\tilde{u}\|_2^2 = \|\nabla\tilde{\varphi}\|_2^2$; as $\tilde{u} = \mathcal{T}(\tilde{\varphi},\tilde{\psi})$, we already have $\|\nabla\tilde{u}\|_2^2 \le (\nabla\tilde{u},\nabla\tilde{\varphi})_2$, so that

$$(\nabla\tilde{u},\nabla\tilde{\varphi})_2 = \|\nabla\tilde{u}\|_2^2 = \|\nabla\tilde{\varphi}\|_2^2\ .$$

Similarly $(\nabla\tilde{u},\nabla\tilde{\psi})_2 = \|\nabla\tilde{u}\|_2^2$. Finally,

$$\|\nabla(\tilde{u}-\tilde{\varphi})\|_2^2 = \|\nabla\tilde{u}\|_2^2 + \|\nabla\tilde{\varphi}\|_2^2 - 2\,(\nabla\tilde{u},\nabla\tilde{\varphi})_2 = 0.$$

Thus $\nabla\tilde{u} = \nabla\tilde{\varphi}$ (and similarly $\nabla\tilde{u} = \nabla\tilde{\psi}$). As $\tilde{u}$, $\tilde{\varphi}$, and $\tilde{\psi}$ belong to $H_o^1(\Omega)$, this gives $\tilde{u} = \tilde{\varphi} = \tilde{\psi}$. □

Let us define $\hat{J}^*$ on $H_o^1(\Omega)$ as follows:

$$\hat{J}^*(w) = \hat{J}(w,w) = \frac{1}{2}\int_\Omega (w-z)^2\,dx + \nu\|\nabla w\|_2^2\ .$$

We have proved that any optimal pair (u,u) satisfies

$$\hat{J}^*(u) = \inf(\hat{\mathcal{P}}).$$

We consider the following optimal control problem:

$$(\hat{\mathcal{P}}^*) \qquad \min\{\ \hat{J}^*(v),\ v \in H_o^1(\Omega)\ \}.$$

It is clear that $(\hat{\mathcal{P}}^*)$ has a unique solution $\hat{u}^*$ since $\hat{J}^*$ is continuous, coercive, and strictly convex on $H_o^1(\Omega)$. Problems $(\hat{\mathcal{P}})$ and $(\hat{\mathcal{P}}^*)$ are equivalent in the following way.

THEOREM 4.3. *$(\hat{\mathcal{P}})$ has a unique optimal solution $(\hat{u},\hat{u})$, and $\hat{u}^* = \hat{u}$ is also the unique optimal solution to $(\hat{\mathcal{P}}^*)$. Conversely, if $\hat{u}^*$ is the solution to $(\hat{\mathcal{P}}^*)$, $(\hat{u}^*,\hat{u}^*)$ is the optimal solution to $(\hat{\mathcal{P}})$.*

Proof. Let us choose an optimal solution to $(\hat{\mathcal{P}})$: $(\bar{u},\bar{u})$. Then

$$\forall(\varphi,\psi) \in U_{ad}, \qquad \hat{J}^*(\bar{u}) = \hat{J}(\bar{u},\bar{u}) \le \hat{J}(\varphi,\psi);$$

in particular, we choose $\varphi = \psi$. Then,

$$\forall\varphi \in H_o^1(\Omega), \qquad \hat{J}^*(\bar{u}) \le \hat{J}(\varphi,\varphi) = \hat{J}^*(\varphi).$$

Therefore $\bar{u}$ is the optimal solution to $(\hat{\mathcal{P}}^*)$. As the solution to $(\hat{\mathcal{P}}^*)$ is unique, the solution to $(\hat{\mathcal{P}})$ is unique as well.

Conversely, let $\hat{u}^*$ be the solution to $(\hat{\mathcal{P}}^*)$, and $(\hat{u},\hat{u})$ the solution to $(\hat{\mathcal{P}})$. We get

$$\hat{J}(\hat{u}^*,\hat{u}^*) = \hat{J}^*(\hat{u}^*) \le \hat{J}^*(\hat{u}) = \hat{J}(\hat{u},\hat{u}) = \inf(\hat{\mathcal{P}}).$$

As the solution to $(\hat{\mathcal{P}})$ is unique, then $\hat{u}^* = \hat{u}$. □

It is easy now to derive optimality conditions, since $(\hat{\mathcal{P}}^*)$ is an unconstrained problem.

COROLLARY 4.1. *The optimal solution to $(\hat{\mathcal{P}})$ is characterized by*

$$u^* = \mathcal{T}(u^*,u^*) \text{ and } -2\nu\Delta u^* + u^* = z \text{ in } \Omega,\ u^* = 0 \text{ on } \partial\Omega.$$

REFERENCES

[1] D. R. Adams and S. Lenhart, *An obstacle control problem with a source term*, Appl. Math. Optim., 47 (2002), pp. 79–95.

[2] D. R. Adams and S. Lenhart, *Optimal control of the obstacle for a parabolic variational inequality*, J. Math. Anal. Appl., 268 (2002), pp. 602–614.

[3] D. R. Adams, S. Lenhart, and J. Yong, *Optimal control of the obstacle for an elliptic variational inequality*, Appl. Math. Optim., 38 (1998), pp. 121–140.

[4] V. Barbu, *Analysis and Control of Non Linear Infinite Dimensional Systems*, Math. Sci. Engrg. 190, Academic Press, San Diego, 1993.

[5] M. Bergounioux and K. Kunisch, *Augmented Lagrangian techniques for elliptic state constrained optimal control problems*, SIAM J. Control Optim., 35 (1997), pp. 1524–1543.

[6] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[7] M. Bergounioux and K. Kunisch, *Primal-dual strategy for state-constrained optimal control problems*, Comput. Optim. Appl., 22 (2002), pp. 193–224.

[8] M. Bergounioux, *Optimal control of semilinear elliptic obstacle problems*, J. Nonlinear Convex Anal., 3 (2002), pp. 25–39.

[9] M. Bergounioux and S. Lenhart, *Optimal control of the obstacle in semilinear variational inequalities*, Positivity, to appear.

[10] D. Bucur, G. Buttazzo, and P. Trabeschi, *An existence result for optimal obstacles*, J. Funct. Anal., 162 (1999), pp. 96–119.

[11] G. Buttazzo and A. Wagner, *On the optimal shape of a rigid body supported by an elastic membrane*, Nonlinear Anal., 39 (2000), pp. 47–63.

[12] Q. Chen, *Optimal control for semilinear evolutionary variational bilateral problems*, J. Math. Anal. Appl., 277 (2003), pp. 303–323.

[13] M. Chipot, *Variational Inequalities and Flow in Porous Media*, Springer-Verlag, New York, 1984.

[14] A. Friedman, *Variational Principles and Free-Boundary Problems*, Wiley, New York, 1982.

[15] J. L. Lions, *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Dunod-Gauthier-Villars, Paris, 1968.

[16] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Vol. 1, Dunod, Paris, 1968.

[17] H. Lou, *On the regularity of an obstacle control problem*, J. Math. Anal. Appl., 258 (2001), pp. 32–51.

[18] H. Lou, *An Optimal Control Problem Governed by Quasilinear Variational Inequalities*, preprint.

[19] J. Sokolowski and J.-P. Zolesio, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, Springer-Verlag, New York, 1992.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 256–276

SEPARATION PRINCIPLES FOR INPUT-OUTPUT AND INTEGRAL-INPUT-TO-STATE STABILITY*

D. ANGELI†, B. INGALLS‡, E. D. SONTAG§, AND Y. WANG¶

Abstract. We present new characterizations of input-output-to-state stability. This is a notion of detectability formulated in the ISS (input-to-state stability) framework. Equivalent properties are presented in terms of asymptotic estimates of the state trajectories based on the magnitudes of the external input and output signals. These results provide a set of *separation principles* for input-output-to-state stability—characterizations of the property as conjunctions of weaker stability notions. When applied to the notion of integral ISS, these characterizations yield analogous results.

Key words. nonlinear systems, input-to-output stability, detectability, Lyapunov method

AMS subject classifications. 93D20, 93D05, 93D09

DOI. 10.1137/S0363012902419047

1. Introduction.

Detectability is a central notion in control theory. It plays a major role both in static state-feedback design (LaSalle's invariance principle and Jurdjevic–Quinn control) as well as in stabilization by means of dynamic output feedback or observer design. Several possibilities are available when formulating such a notion in the context of nonlinear control. According to the specific problem under consideration, they capture some or most of the useful features of its linear counterpart. One approach that has proved to be especially powerful for systems subject to exogenous disturbances is to define zero-detectability in terms of estimates involving (possibly nonlinear) gains with respect to input and output norms. This leads to the so-called input-output-to-state stability (IOSS) property. Such a notion not only allows one to extend LaSalle-type stability results to the case of nonautonomous systems [2], but it also provides a machinery, fully compatible with the formalism of the input-to-state stability (ISS) property [8, 9, 13, 14, 15, 17, 18, 19, 23, 24, 25, 30, 31], that helps one understand relevant issues such as minimum-phase behavior or certainty equivalence [20, 11].

Although general nonlinear systems may often exhibit an overwhelming variety of behaviors, it turns out that many of the "reasonable" formulations of the detectability property (meaning at least compatible with the linear notion of detectability) are equivalent to each other. In this paper, we discuss characterizations of IOSS in terms of the asymptotic behavior of system solutions. This leads to several useful decompositions of the IOSS property in terms of weaker notions. These *separation principles* are in direct analogy to those previously provided for ISS [28] and input-to-output sta-

*Received by the editors November 28, 2002; accepted for publication (in revised form) October 30, 2003; published electronically June 15, 2004.
http://www.siam.org/journals/sicon/43-1/41904.html

†Dipartimento Sistemi e Informatica, Università di Firenze, Via di Santa Marta 3, 50139 Firenze, Italy (angeli@dsi.unifi.it).

‡Department of Applied Mathematics, University of Waterloo, Waterloo, ON, N2L3G1, Canada (bingalls@math.uwaterloo.ca).

§Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (sontag@control.rutgers.edu). The research of this author was supported in part by US Air Force grant F49620-01-1-0063 and by NSF grant CCR-0206789.

¶Department of Mathematics, Florida Atlantic University, Boca Raton, FL 33431 (ywang@math.fau.edu). The research of this author was supported in part by NSF grant DMS-0072620 and by Chinese Natural Science Foundation grant 60228003.

bility [12]. These results all generalize from differential equations the fact that global asymptotic stability can be characterized by the combination of (neutral) stability and attractivity.

As an application of our results, we will also discuss several ways of reformulating the notion of integral input-to-state stability (iISS) (cf. [5, 26]) in terms of asymptotic gains. This is accomplished by treating the iISS property as the IOSS property for suitable auxiliary systems. Similar results were also obtained for the so-called derivative-ISS property (DISS) in the recent work [7].

As mentioned, the main results are (far from obvious) generalizations to systems with outputs of the analogous separation principles which appeared in [28] dealing with the ISS notion. Actually, the ISS case is a special case of IOSS when the output map is identically zero. However, it takes much more effort to handle the general case of nonzero output maps. It can be seen from later sections that IOSS amounts to the requirement of convergence to 0 (or to balls whose radii are proportional to the norms of the input signals) *only* for those trajectories which evolve in a certain set constrained by the output signals. In the ISS case, by comparison, the output map is identically zero, so the constraints become trivial, and the restricted set becomes the whole state space.

2. Basic definitions. Consider systems in the following general form:

$$\dot{x}(t) = f(x(t), u(t)), \quad y(t) = h(x(t)), \tag{1}$$

where, for each $t \geq 0$, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{U}$, a subset of $\mathbb{R}^m$. We assume that the maps $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}^p$ are locally Lipschitz continuous, with $f(0,0) = 0$ and $h(0) = 0$. The symbol $|\cdot|$ denotes the usual Euclidean norms. The open ball in $\mathbb{R}^l$ centered at the origin with radius r will be denoted by $\mathcal{B}_l(r)$.

By an *input* we mean a measurable and locally essentially bounded function $u : \mathcal{I} \to \mathbb{U}$, where $\mathcal{I}$ is a subinterval of $\mathbb{R}$ which contains the origin. Whenever the domain $\mathcal{I}$ of an input u is not specified, it will be understood that $\mathcal{I} = \mathbb{R}_{\geq 0}$. Given a system with input-value set $\mathbb{U}$, we will also often consider the same system with inputs restricted to some subset $\mathcal{O} \subseteq \mathbb{U}$. We use $\mathcal{M}_{\mathcal{O}}$ to denote the set of all such inputs.

Given any input u and any $\xi \in \mathbb{R}^n$, the unique maximal solution of the initial value problem $\dot{x} = f(x,u)$, $x(0) = \xi$ (defined on some maximal open subinterval of $\mathcal{I}$) is denoted by $x(\cdot, \xi, u)$. When $\mathcal{I} = \mathbb{R}_{\geq 0}$, this maximal subinterval has the form $[0, T_{\xi,u})$. The system is said to be *forward complete* if for every initial state ξ and for every input u defined on $\mathbb{R}_{\geq 0}$, $T_{\xi,u} = +\infty$. The corresponding output is denoted by $y(\cdot, \xi, u)$, that is, $y(t, \xi, u) = h(x(t, \xi, u))$ on the domain of definition of the solution.

The L_∞-norm (possibly infinite) of a function v defined on $\mathcal{I}$ is denoted by $\|v\|$, i.e.,

$$\|v\| = (\text{ess})\sup\{|v(t)|\,, t \in \mathcal{I}\}.$$

In particular, for a maximal trajectory $x(\cdot, \xi, u)$ and the corresponding output function $y(\cdot, \xi, u)$ of (1) defined on $[0, T_{\xi,u})$, $\|u\|$, $\|x\|$, and $\|y\|$ denote the L_∞-norm of $u(\cdot)$, $x(\cdot, \xi, u)$, and $y(\cdot, \xi, u)$, respectively, on $[0, T_{\xi,u})$. We will make a slight abuse of notation and use sup and lim sup to mean the essential supremum where appropriate. For a function v defined on an interval $\mathcal{I}$, if $\mathcal{I}_1 \subseteq \mathcal{I}$, we use $v_{\mathcal{I}_1}$ to denote the restriction of v to $\mathcal{I}_1$, i.e., $v_{\mathcal{I}_1}(t) = v(t)$ if $t \in \mathcal{I}_1$, and $v_{\mathcal{I}_1}(t) = 0$ otherwise. Notice the following fact: for a measurable function v defined on some interval $[0, T)$ for some $T \leq \infty$,

$$\limsup_{t \to T} |v(t)| = \lim_{t \to T} \|v_{[t,T)}\|. \tag{2}$$

We use standard terminology (cf. [10]): $\mathcal{N}$ is the class of continuous, increasing functions from $[0,\infty)$ to $[0,\infty)$; $\mathcal{K}$ is the set of $\mathcal{N}$ functions γ that are strictly increasing and satisfy $\gamma(0)=0$; $\mathcal{K}_\infty$ is the set of $\mathcal{K}$ functions that are unbounded; $\mathcal{L}$ is the set of functions $[0,+\infty)\to[0,+\infty)$ which are continuous, decreasing, and converge to 0 as their argument tends to $+\infty$; $\mathcal{KL}$ is the class of functions $[0,\infty)^2\to[0,\infty)$ which are class $\mathcal{K}$ on the first argument and class $\mathcal{L}$ on the second one. A positive definite function $\gamma:[0,\infty)\to[0,\infty)$ is one such that $\gamma(0)=0$ and $\gamma(s)>0$ for all $s>0$.

The following notions were introduced in [22] (see also [4]) and [16, 29], respectively.

DEFINITION 2.1. *The system* (1) *satisfies the* unboundedness observability *(UO) property if, for each state ξ and control u such that $T_{\xi,u}<\infty$, it holds that $\limsup_{t\to T_{\xi,u}}|y(t,\xi,u)|=+\infty$, that is, for each state ξ and control u,*

$$T_{\xi,u}<\infty \quad\Rightarrow\quad \|y\|=+\infty.$$

DEFINITION 2.2. *The system* (1) *is* input-output-to-state stable *(IOSS) if there exist some $\beta\in\mathcal{KL}$, $\gamma_u\in\mathcal{K}$, and $\gamma_y\in\mathcal{K}$ such that*

$$|x(t,\xi,u)|\le\beta(|\xi|,t)+\gamma_u(\|u\|)+\gamma_y\left(\|y_{[0,t]}\|\right) \tag{3}$$

for all $t\in[0,T_{\xi,u})$, all $\xi\in\mathbb{R}^n$, and all $u(\cdot)$. □

Clearly, the IOSS property implies the UO property. The following local version of it will also be used in the proof of our main result.

DEFINITION 2.3. *The system* (1) *is* locally IOSS *if there exist $\delta>0$ and functions $\beta\in\mathcal{KL}$, $\gamma_u,\gamma_y\in\mathcal{K}$ so that for any $\xi\in\mathbb{R}^n$, any $u(\cdot)$,*

$$\max\{|\xi|,\|u\|,\|y\|\}\le\delta\ \Rightarrow\ |x(t,\xi,u)|\le\beta(|\xi|,t)+\gamma_u(\|u\|)+\gamma_y(\|y_{[0,t]}\|) \tag{4}$$

for all $t\in[0,T_{\xi,u})$. □

Below we discuss several other properties for systems as in (1) regarding estimates of the state variables on the basis of external information provided by past input and output signals.

2.1. A catalog of properties.

DEFINITION 2.4. *Consider system* (1). *We say that*

- *the* input-output limit property *(IO-LIM) holds if for some $\gamma_u,\gamma_y\in\mathcal{K}$,*

$$\inf_{t\in[0,T_{\xi,u})}|x(t,\xi,u)|\le\max\{\gamma_u(\|u\|),\gamma_y(\|y\|)\}\quad\forall\,\xi\in\mathbb{R}^n,\ \forall\,u(\cdot); \tag{5}$$

- *the* input-output asymptotic gain property *(IO-AG) holds if solutions are ultimately bounded by some nonlinear gain function of $\|u\|$ and $\|y\|$, that is, for some $\gamma_u,\gamma_y\in\mathcal{K}$,*

$$\limsup_{t\to T_{\xi,u}}|x(t,\xi,u)|\le\max\{\gamma_u(\|u\|),\gamma_y(\|y\|)\}\quad\forall\,\xi\in\mathbb{R}^n,\ \forall\,u(\cdot); \tag{6}$$

- *the* input-output-to-state boundedness property *(IO-BND) holds if for some $\sigma_0,\sigma_u,\sigma_y\in\mathcal{N}$, it holds that*

$$|x(t,\xi,u)|\le\max\{\sigma_0(|\xi|),\sigma_u(\|u\|),\sigma_y(\|y_{[0,t]}\|)\} \tag{7}$$

for all $\xi\in\mathbb{R}^n$, all $u(\cdot)$, and all $t\in[0,T_{\xi,u})$;

- *the* input-output global stability property *(IO-GS) holds if the functions* $\sigma_0, \sigma_u, \sigma_y$ *in* (7) *can be taken to be of class* $\mathcal{K}$.

Thinking of these detectability properties as "stability modulo inputs and outputs," we can identify IOSS with asymptotic stability, IO-GS with (neutral) stability, and IO-AG with attractivity. In this context it seems perfectly natural that IOSS should be equivalent to the combination of IO-GS and IO-AG, and indeed that is one of the decompositions which appears in our main result. Related results follow by considering other "basic" stability-like notions, such as IO-LIM.

It is not hard to see that each of the IO-AG, the IO-GS, and the IO-BND properties implies the UO condition. This follows from the fact that, for any ξ, u, if $T_{\xi,u} < \infty$, then $|x(t,\xi,u)| \to \infty$ as $t \to T_{\xi,u}$. That the IO-LIM property also implies the UO condition follows from the next remark.

Remark 2.5. The IO-LIM property can be defined equivalently by replacing the "inf" in (5) by "lim inf", that is,

$$\liminf_{t \to T_{\xi,u}} |x(t,\xi,u)| \le \max\{\gamma_u(\|u\|), \gamma_y(\|y\|)\}. \tag{8}$$

It is straightforward that (8) implies (5). To see that (5) implies (8), take any $T \in [0, T_{\xi,u})$. Applying (5) to $x(t, \xi_T, u_T)$ with $\xi_T = x(T,\xi,u)$ and $u_T(t) = u(t+T)$, one gets

$$\inf_{t \in [T, T_{\xi,u})} |x(t,\xi,u)| \le \max\{\gamma_u(\|u_T\|), \gamma_y(\|y_T\|)\} \le \max\{\gamma_u(\|u\|), \gamma_y(\|y\|)\}, \tag{9}$$

where $y_T = h(x(\cdot, \xi_T, u_T))$. Since T can be arbitrary, one obtains (8). □

Another characterization of IO-LIM is as follows. This statement is proved (along with some related characterizations of the IO-LIM property) in Appendix A.

LEMMA 2.6. *System* (1) *satisfies the IO-LIM property if and only if there exist* γ_u, $\gamma_y \in \mathcal{K}$ *such that*

$$\inf_{t \in [0, T_{\xi,u})} \Big\{ |x(t,\xi,u)| - \max\{\gamma_u(\|u_{[0,t]}\|), \gamma_y(\|y_{[0,t]}\|)\} \Big\} \le 0 \qquad \forall\, \xi \in \mathbb{R}^n,\ \forall\, u(\cdot). \tag{10}$$

The following implication will be needed.

LEMMA 2.7. *If* (1) *satisfies the IO-LIM property, then it satisfies the IO-BND property.*

Proof. Consider a system as in (1). For each subset $\mathcal{O}$ of the input space $\mathbb{U}$, each subset C of $\mathbb{R}^n$, and each $\mathcal{Y} \subset \mathbb{R}^p$ we denote

$$\mathcal{R}_{\mathcal{O}/\mathcal{Y}}(C) := \Big\{ x(t,\xi,u) :\ \xi \in C, u \in \mathcal{M}_{\mathcal{O}}, t \in [0, T_{\xi,u}) \\ \text{and } h(x(\lambda,\xi,u)) \in \mathcal{Y}\ \ \forall \lambda \in [0,t] \Big\}.$$

Then $\mathcal{R}_{\mathcal{O}/\mathcal{Y}}(C)$ is the reachable set with initial conditions in C, controls in $\mathcal{O}$, and subject to an output constraint.

Suppose a system as in (1) satisfies the IO-LIM property and, consequently, the UO property as well. By Lemma 2.6 one sees that (10) holds for some $\gamma_u, \gamma_y \in \mathcal{K}$. Pick an arbitrary $s > 0$. We let

$$\Omega = \mathcal{B}_n(2s),\ C = \mathrm{cl}(\Omega),\ K = \mathrm{cl}\left(\mathcal{B}_n\left(\frac{3s}{2}\right)\right),\ \mathcal{Y} = \mathrm{cl}\left(\mathcal{B}_p\left(\gamma_y^{-1}\frac{(s)}{2}\right)\right)$$

and

$$\mathcal{Y}_o = \mathcal{B}_p(\gamma_y^{-1}(s)),\ \mathcal{O} = \mathrm{cl}(\mathcal{B}_m(\gamma_u^{-1}(s))).$$

Finally we define $\sigma_0(s) := \sup\{|p| :\ p \in \mathcal{R}_{\mathcal{O}/\mathcal{Y}}(C)\}$.

Claim. $\sigma_0(s) < \infty$.

Proof. For any $\xi \in \mathbb{R}^n$ and any $u \in \mathcal{M}_{\mathcal{O}}$, by (10), there exists $\tau \in [0, T_{\xi,u})$ so that

$$|x(\tau, \xi, u)| \leq \frac{3}{2} \max\left\{s, \gamma_y(\|y_{[0,\tau]}\|)\right\}.$$

Considering separately the cases $\|y_{[0,\tau]}\| \leq \gamma_y^{-1}(s)$ and $\|y_{[0,\tau]}\| > \gamma_y^{-1}(s)$, we obtain that either

$$|x(\tau, \xi, u)| \leq \frac{3s}{2}$$

or there is some $t \in [0, \tau]$ so that

$$|h(x(t, \xi, u))| > \gamma_y^{-1}(s).$$

In other words, there exists $\tau \in [0, T_{\xi,u})$ such that either $x(\tau, \xi, u) \in K$ or for some $t \in [0, \tau]$, $h(x(t, \xi, u)) \notin \mathcal{Y}_o$. We then apply Lemma B.1 to conclude that $\mathcal{R}_{\mathcal{O}/\mathcal{Y}}(C)$ is bounded, and thus $\sigma_0(s) < \infty$.

Let $\sigma_u(r) = \sigma_0(\gamma_u(r))$ and $\sigma_y(r) = \sigma_0(\gamma_y(r))$. Now pick $\xi \in \mathbb{R}^n$, an input u, and $t \in [0, T_{\xi,u})$. Let $s = \max\{|\xi|, \gamma_u(\|u\|), \gamma_y(\|y_{[0,t]}\|)\}$. If $s = 0$, then $|x(t, 0, 0)| = 0$. If $s > 0$, by definition of σ_0 we have

$$|x(t, \xi, u)| \leq \sigma_0(s) \leq \max\{\sigma_0(|\xi|), \sigma_u(\|u\|), \sigma_y(\|y_{[0,t]}\|)\}. \tag{11}$$

This completes the proof of IO-LIM $\Rightarrow$ IO-BND. ∎

Remark 2.8. Note that in the above proof, if the function γ_u can be chosen to be the zero function, that is, if (10) can be strengthened to

$$\inf_{t \in [0, T_{\xi,u})} \left[|x(t, \xi, u)| - \gamma_y(\|y_{[0,t]}\|)\right] \leq 0,$$

then the function σ_u in (11) can be chosen to be the zero function. ∎

We next comment on some straightforward characterizations of the IO-AG property.

Remark 2.9. It is immediate from the definition that the IO-AG property (6) is equivalent to the UO property in combination with the following:

$$\limsup_{t \to \infty} |x(t, \xi, u)| \leq \max\left\{\gamma_u(\|u\|),\ \gamma_y\left(\|y\|\right)\right\} \tag{12}$$

for all ξ, u for which $T_{\xi,u} = \infty$. ∎

Combining this remark with (2), the following can be easily shown (using an argument as in Remark 2.5).

LEMMA 2.10. *A system as in* (1) *satisfies the IO-AG property if and only if it is UO and, for some* $\gamma_u, \gamma_y \in \mathcal{K}$, *the following holds for all* ξ *and* u *for which* $T_{\xi,u} = \infty$:

$$\limsup_{t \to \infty} |x(t, \xi, u)| \leq \max\left\{\gamma_u\left(\limsup_{t \to \infty} |u(t)|\right),\ \gamma_y\left(\limsup_{t \to \infty} |y(t)|\right)\right\}. \tag{13}$$

2.2. IOSS and OSS properties. Consider a system

$$\dot{x} = f(x),\ y = h(x) \tag{14}$$

without input. This can be considered as a system as in (1) with $\mathbb{U}$ consisting of a single point. We use $x(\cdot, \xi)$ to denote the solution of (14) with the initial state ξ defined on a maximal interval $[0, T_\xi)$, and we let $y(t, \xi) = h(x(t, \xi))$.

DEFINITION 2.11. *We say that the system* (14) *is*

- locally stable modulo outputs *(O-LS) if for any* $\varepsilon > 0$ *there exists* $\delta_\varepsilon > 0$ *such that for all* ξ *and all* $t \in [0, T_\xi)$ *it holds that*

$$\max\{|\xi|, \|y_{[0,t]}\|\} \le \delta_\varepsilon \ \Rightarrow\ |x(t, \xi)| \le \varepsilon; \tag{15}$$

- output-to-state stable *(OSS) (see* [29]*) if there exist some* $\beta \in \mathcal{KL}$ *and some* $\gamma \in \mathcal{K}$ *such that, for any trajectory* $x(\cdot)$ *of the system, it holds that*

$$|x(t, \xi)| \le \beta(|\xi|, t) + \gamma(\|y_{[0,t]}\|) \qquad \forall\, t \in [0, T_\xi). \tag{16}$$

For a system as in (1), we say that the system is *zero-input OSS* (zero-OSS) or *zero-input O-LS* (zero-O-LS) if the zero input system $\dot{x} = f(x, 0),\ y = h(x)$ is OSS or O-LS, respectively. The following technical lemma will be needed in the proof of our main result.

LEMMA 2.12. *The zero-OSS property implies the local IOSS property.*

Proof. Suppose the system (1) is zero-OSS. Then, by Theorem 3 in [29] there exists a smooth function $V : \mathbb{R}^n \to \mathbb{R}_{\ge 0}$ and class $\mathcal{K}_\infty$ functions α_1, α_2, α, and ρ such that $\alpha_1(|\xi|) \le V(\xi) \le \alpha_2(|\xi|)$ and

$$\frac{\partial V(\xi)}{\partial \xi} f(\xi, 0) \le -2\alpha(|\xi|) + \rho(|h(\xi)|) \qquad \forall\, \xi \in \mathbb{R}^n. \tag{17}$$

Let $\sigma : \mathbb{R}_{\ge 0} \times \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be defined as

$$\sigma(s, r) = sr + \max_{|\xi| \le s, |\mu| \le r} \left\{ \frac{\partial V(\xi)}{\partial \xi} [f(\xi, \mu) - f(\xi, 0)] \right\}. \tag{18}$$

Note that $\sigma(s, 0) = 0$ for all $s \ge 0$. Since V is smooth and $\frac{\partial V}{\partial \xi}(0) = 0$, we also have $\sigma(0, r) = 0$ for all $r \ge 0$. Furthermore, $\sigma(\cdot, r) \in \mathcal{K}$ for each $r \ge 0$ and $\sigma(s, \cdot) \in \mathcal{K}$ for each $s \ge 0$. By Corollary IV.5 in [5], there exist σ_1, $\sigma_2 \in \mathcal{K}_\infty$ so that $\sigma(s, r) \le \sigma_1(s)\sigma_2(r)$. Hence, combining (17) and (18), we get

$$\begin{aligned} \frac{\partial V(\xi)}{\partial \xi} f(\xi, \mu) &= \frac{\partial V(\xi)}{\partial \xi} f(\xi, 0) + \frac{\partial V(\xi)}{\partial \xi} [f(\xi, \mu) - f(\xi, 0)] \\ &\le -2\alpha(|\xi|) + \rho(|h(\xi)|) + \sigma(|\xi|, |\mu|) \\ &\le -2\alpha(|\xi|) + \rho(|h(\xi)|) + \sigma_1(|\xi|)\sigma_2(|\mu|). \end{aligned}$$

Therefore

$$\alpha^{-1}(\sigma_2(|\mu|) + \rho(|h(\xi)|)) \le |\xi| \le \sigma_1^{-1}(1) \quad \Rightarrow \quad \frac{\partial V(\xi)}{\partial \xi} f(\xi, \mu) \le -\alpha(|\xi|),$$

from which it follows that for some $\mathcal{K}_\infty$-functions χ_1, χ_2, and $\tilde{\alpha}$ and some $c > 0$, it holds that

$$\max\{\chi_1(|\mu|), \chi_2(|h(\xi)|)\} \le V(\xi) \le c \quad \Rightarrow \quad \frac{\partial V(\xi)}{\partial \xi} f(\xi, \mu) \le -\tilde{\alpha}(V(\xi)).$$

Pick any initial state ξ and input u. Denote the trajectory $x(t,\xi,u)$ by $x(t)$, and the output $y(t,\xi,u)$ by $y(t)$. Pick any $T \in [0, T_{\xi,\mu})$, and let

$$v^* = \max\{\chi_1(\|u\|), \chi_2(\|y_{[0,T]}\|)\}.$$

For almost all $t \in [0,T]$, we have

$$v^* \leq V(x(t)) \leq c \;\Rightarrow\; \frac{d}{dt}V(x(t)) \leq -\tilde{\alpha}(V(x(t))).$$

By Lemma 13 in [29], one sees that there exists some $\beta_0 \in \mathcal{KL}$ which depends only on $\tilde{\alpha}$ so that

$$\max_{t\in[0,T]} V(x(t)) \leq c \quad \Rightarrow\; V(x(t)) \leq \max\{\beta_0(V(x(0)),t), v^*\} \quad \forall\, t \in [0,T];$$

that is, if $V(x(t)) \leq c$ for all $t \in [0,T]$,

$$V(x(t)) \leq \max\{\beta_0(V(x(0)),t), \chi_1(\|u\|),\ \chi_2(\|y_{[0,T]}\|)\} \quad \forall\, t \in [0,T]. \tag{19}$$

Let $\hat{\beta}_0(r) = \beta_0(r,0)$. Without loss of generality, we assume that $\hat{\beta}_0(r) > r$, and thus, $\hat{\beta}_0 \in \mathcal{K}_\infty$. Let

$$\delta = \min\left\{\chi_1^{-1}\left(\frac{c}{2}\right),\ \chi_2^{-1}\left(\frac{c}{2}\right),\ \alpha_2^{-1}\left(\hat{\beta}_0^{-1}\left(\frac{c}{2}\right)\right)\right\}.$$

Claim. If $\max\{|x(0)|,\ \|u\|,\ \|y\|\} \leq \delta$, then $V(x(t)) \leq c$ for all $t \in [0, T_{\xi,u})$.

Suppose the claim fails. This means

$$\max\{\tilde{\beta}_0(\alpha_2(|x(0)|)),\ \chi_1(\|u\|),\ \chi_1(\|y\|)\} \leq \frac{c}{2},$$

but for some $t \in [0, T_{\xi,u})$, $V(x(t)) \geq c$. Let

$$t_0 = \inf\left\{t \geq 0 : V(x(t)) \geq \frac{c}{2}\right\}.$$

Then $t_0 < T_{\xi,u}$. By the continuity property of $x(\cdot)$, there is some $0 < \varepsilon < T_{\xi,u} - t_0$ such that on $[0, t_0+\varepsilon)$, $V(x(t)) < c$. By (19), we have

$$V(x(t)) \leq \max\{\hat{\beta}_0(V(x(0))),\ \chi_1(\|u\|),\ \chi_2(\|y\|)\} \leq \frac{c}{2}$$

for all $t \in [0, t_0+\varepsilon]$. This contradicts the definition of t_0.

Finally, applying (19) together with the proved claim, one sees that if

$$\max\{|x(0)|, \|u\|, \|y\|\} < \delta,$$

then

$$|x(t)| \leq \max\{\beta(|x(0)|,t),\ \gamma_1(\|u\|),\ \gamma_2(\|y_{[0,t]}\|)\}\ \forall\, t \in [0, T_{\xi,u}),$$

where $\beta(s,r) = \alpha_1^{-1}(\beta_0(\alpha_2(s),r)), \gamma_i(r) = \alpha_1^{-1}(\chi_i(r))$ for $i = 1,2$. □

3. Equivalent characterizations of IOSS. The following is our main result in the context of IOSS.

THEOREM 1. *Consider a system as in* (1) *with* $\mathbb{U} = \mathbb{R}^m$. *The following properties are equivalent:*

1. *(IOSS),*
2. *(IO-AG) & (IO-GS),*
3. *(IO-AG) & (zero-OSS),*
4. *(IO-AG) & (local IOSS),*
5. *(IO-AG) & (zero-O-LS),*
6. *(IO-LIM) & (IO-GS),*
7. *(IO-LIM) & (zero-OSS),*
8. *(IO-LIM) & (local IOSS),*
9. *(IO-LIM) & (zero-O-LS).*

Among the properties listed in Theorem 1, it is easy to see that the IOSS property implies every other one, and the zero-O-LS property is implied by any one of the local IOSS, the zero-OSS, and the IO-GS properties. Thus, to prove Theorem 1, it is enough to show the following implication:

$$\text{(IO-LIM) \& (zero-O-LS)} \Rightarrow \text{(IOSS)}.$$

We will proceed by the following technical lemmas.

LEMMA 3.1. *(IO-LIM) & (zero-O-LS)* $\Rightarrow$ *(zero-OSS).*

LEMMA 3.2. *(IO-LIM) & (local IOSS)* $\Rightarrow$ *(IO-LIM) & (IO-GS).*

LEMMA 3.3. *(IO-LIM) & (IO-GS)* $\Rightarrow$ *(IO-AG) & (IO-GS).*

Observe that Lemmas 3.1–3.3 together with Lemma 2.12 provide the following chain:

$$\begin{aligned}\text{(IO-LIM) \& (zero-O-LS)} &\Rightarrow \text{(IO-LIM) \& (zero-OSS)} \Rightarrow \text{(IO-LIM) \& (local IOSS)} \\ &\Rightarrow \text{(IO-LIM) \& (IO-GS)} \Rightarrow \text{(IO-AG) \& (IO-GS)}.\end{aligned}$$

To complete the proof of Theorem 1, we will need the following result.

PROPOSITION 3.4. *(IO-AG) & (IO-GS)* $\Rightarrow$ *(IOSS).*

The proofs of the lemmas and the proposition will be given in section 6.

Remark 3.5. Theorem 1 is a satisfying theoretical result in that it unifies a number of properties and provides a generalization of the separation principle for asymptotically stable differential equations. Moreover, the theorem is a useful tool for recognizing IOSS systems. The definition of IOSS rarely lends itself to direct verification. More often, this property is shown using the Lyapunov characterization provided in [16]. In cases where construction of an appropriate Lyapunov function proves difficult, Theorem 1 provides a number of alternative conditions which may be tested.

As an example, consider the following family of systems without inputs, with state $(x, z) \in \mathbb{R}^n \times \mathbb{R}$:

$$\begin{aligned}\dot{x} &= f(x), \\ \dot{z} &= |x|, \\ y &= z.\end{aligned} \tag{20}$$

Suppose that the system is forward complete. Provided that the function f is locally Lipschitz, this system satisfies the IO-LIM property, which can be shown as

follows. For each initial condition ξ, consider two cases. If $\int_0^\infty |x(s,\xi)|\, ds = \infty$, then $\lim_{t\to\infty} y(t,\xi) = \infty$, and so the IO-LIM bound (8) holds trivially for any $\gamma_y \in \mathcal{K}_\infty$. Otherwise, from the fact that $\int_0^\infty |x(s,\xi)|\, ds < \infty$, we have $\liminf_{t\to\infty} |x(t,\xi)| = 0 \leq \|y\|$. Thus, the IO-LIM estimate (8) holds in both cases.

Hence system (20) is known to satisfy the IOSS property (OSS in this case), provided that it satisfies one of the stability properties as in Theorem 1. For instance, if the system $\dot{x} = f(x)$ is locally stable, then the system (20) is (zero)-O-LS (since the z component of the state is trivially bounded by the output), and so the system enjoys the OSS property.

On the other hand, the Lyapunov approach would not be well suited to exploring the IOSS property in this situation, since little is assumed about the dynamics. ∎

4. On iISS. In this section, we indicate how the equivalences shown in Theorem 1 can be used to derive asymptotic characterizations of the iISS property.

DEFINITION 4.1 (see [26]). *A system as in* (1) *is* integral input-to-state stable *(iISS) if there exist functions $\beta \in \mathcal{KL}$, $\sigma \in \mathcal{K}$, and $\gamma \in \mathcal{K}$ such that, for all $\xi \in \mathbb{R}^n$ and all u, the solution $x(t,\xi,u)$ is defined for all $t \geq 0$, and*

$$|x(t,\xi,u)| \leq \beta(|\xi|, t) + \gamma\left(\int_0^t \sigma(|u(s)|)\, ds\right) \tag{21}$$

for all $t \geq 0$.

To make use of our main result, we reformulate the iISS property in terms of the IOSS property.

LEMMA 4.2. *System* (1) *is integral input-to-state stable with an estimate as in* (21) *if and only if the augmented system*

$$\dot{x} = f(x,u), \quad \dot{e} = \sigma(|u|), \quad y = e \tag{22}$$

is IOSS.

Proof. Let the augmented system be IOSS; then, for $e(0) = 0$ we have

$$\begin{aligned} |x(t,\xi,u)| &\leq |x(t,\xi,u)| + |e(t)| \leq \beta(|\xi|,t) + \gamma_1(\|u\|) + \gamma_2(\|y_{[0,t]}\|) \\ &= \beta(|\xi|,t) + \gamma_1(\|u\|) + \gamma_2\left(\int_0^t \sigma(|u(s)|)\, ds\right) \end{aligned} \tag{23}$$

for all $t \in [0, T_{\xi,u})$. By causality, (23) can be rewritten as

$$|x(t,\xi,u)| \leq \beta(|\xi|,t) + \gamma_1(\|u_{[0,t]}\|) + \gamma_2\left(\int_0^t \sigma(|u(s)|)\, ds\right). \tag{24}$$

Since on any finite interval, the integral term in (24) is finite, it follows that $T_{\xi,u} = \infty$. In turn, this implies that (24) holds on $[0,\infty)$. This estimate implies iISS for (1), by virtue of Theorem 1 in [6].

To see the converse, clearly, $|e(t)| = |y(t)| \leq \|y_{[0,t]}\|$. Also observe that the iISS property implies that the augmented system is forward complete. Thus, it is enough to show that a suitable estimate holds on $[0,\infty)$ for the x component of the state. By (21), we have

$$\begin{aligned} |x(t,\xi,u)| &\leq \beta(|\xi|,t) + \gamma\left(\int_0^t \sigma(|u(s)|)\, ds\right) \\ &= \beta(|\xi|,t) + \gamma(e(t) - e(0)) \\ &\leq \beta(|\xi|,t) + \gamma(|y(t)| + |y(0)|) \leq \beta(|\xi|,t) + \gamma(2\|y_{[0,t]}\|) \end{aligned} \tag{25}$$

for all $t \geq 0$. This completes the proof. ∎

4.1. Lyapunov characterizations of iISS. This IOSS formulation of the iISS property allows us to exploit known results on IOSS to develop new characterizations of iISS. For instance, the following new Lyapunov characterization for iISS follows directly from the Lyapunov characterization for IOSS presented in [16]. We prove the next two results under the assumption that the gain σ in (21) is locally Lipschitz. Remark 4.7 indicates how this can always be achieved through a simple modification.

THEOREM 2. *System* (1) *is iISS if and only if there exist functions* $\alpha_1, \alpha_2, \alpha$, γ_1, γ_2, σ *of class* $\mathcal{K}_\infty$ *and a smooth function* $V : \mathbb{R}^{n+1} \to \mathbb{R}_{\geq 0}$, *with*

$$\alpha_1(|\xi| + |\eta|) \leq V(\xi, \eta) \leq \alpha_2(|\xi| + |\eta|) \qquad \forall \xi \in \mathbb{R}^n, \eta \in \mathbb{R}, \tag{26}$$

such that the following dissipation inequality is satisfied:

$$\frac{\partial V(\xi,\eta)}{\partial \xi} f(\xi, \mu) + \frac{\partial V(\xi,\eta)}{\partial \eta} \sigma(|\mu|) \leq -\alpha(|\xi| + |\eta|) + \gamma_1(|\eta|) + \gamma_2(|\mu|) \tag{27}$$

for all $\xi \in \mathbb{R}^n,\ \eta \in \mathbb{R},\ \mu \in \mathbb{U}$. □

Remark 4.3. It is easy to see that estimates (26) and (27) imply

$$\alpha_1(|\xi|) \leq V(\xi, \eta) \leq \alpha_2(|\xi| + |\eta|) \qquad \forall \xi \in \mathbb{R}^n, \eta \in \mathbb{R}, \tag{28}$$

and for all $\xi \in \mathbb{R}^n,\ \eta \in \mathbb{R},\ \mu \in \mathbb{U}$,

$$\frac{\partial V(\xi,\eta)}{\partial \xi} f(\xi, \mu) + \frac{\partial V(\xi,\eta)}{\partial \eta} \sigma(|\mu|) \leq -\alpha(|\xi|) + \gamma_1(|\eta|) + \gamma_2(|\mu|). \tag{29}$$

On the other hand, suppose that for a given V, equations (28) and (29) hold for some $\alpha_1, \alpha_2, \alpha,\ \gamma_1,\ \gamma_2, \sigma$ of class $\mathcal{K}_\infty$; then one can again show that the following type of estimate holds for the x-component of (22):

$$|x(t, \xi, u)| \leq \beta(|\xi|, t) + \rho_1(\|e\|_{[0,t]}) + \rho_2(\|u\|) \qquad \forall\, t \geq 0,$$

where $\rho_1,\ \rho_2 \in \mathcal{K}$. Combining this with the fact that $|e(t)| \leq \|e\|_{[0,t]}$, one sees that the augmented system (22) is IOSS. Hence, the corresponding system as in (1) is iISS.

Thus, a system as in (1) is iISS if and only if there exists some smooth function $V : \mathbb{R}^{n+1} \to \mathbb{R}_{\geq 0}$ for which (28) and (29) hold for some $\mathcal{K}_\infty$-functions $\alpha_1, \alpha_2, \alpha, \gamma_1, \gamma_2$ and σ.

In [5], an equivalent Lyapunov characterization for iISS was formulated as in the following: for some $\alpha_1, \alpha_2, \gamma \in \mathcal{K}_\infty$ and some *continuous positive definite* function α

$$\alpha_1(|\xi|) \leq V(\xi) \leq \alpha_2(|\xi|) \qquad \forall \xi \in \mathbb{R}^n, \tag{30}$$

$$\frac{\partial V(\xi)}{\partial \xi} f(\xi, \mu) \leq -\alpha(|\xi|) + \gamma(|\mu|). \tag{31}$$

The significance of the new Lyapunov characterization by (28)–(29) (or (26)–(27)) is that in (27) or (29) one can require the function α be of class $\mathcal{K}_\infty$. This may lead to some interesting applications in feedback design. For instance, if V is an iISS-Lyapunov function defined by (30)–(31), then, given a $\mathcal{K}_\infty$-function, $\rho \circ V$ may fail to be an iISS-Lyapunov function. However, if V is an iISS-Lyapunov function satisfying (28)–(29), then, for any $\rho \in \mathcal{K}_\infty$, $\rho \circ V$ is again an iISS-Lyapunov function satisfying the same type of estimates as in (28)–(29) (cf. [27, 2] regarding changing "supply rates"). □

4.2. Asymptotic characterizations of iISS.

DEFINITION 4.4. *A system as in* (1) *satisfies the* bounded energy weakly converging state *(BEWCS) property if for some σ of class $\mathcal{K}_\infty$ the following holds:*

$$\int_0^{+\infty} \sigma(|u(s)|)\,ds < +\infty \Rightarrow \liminf_{t\to+\infty} |x(t,\xi,u)| = 0. \tag{32}$$

To be more precise, (32) means that for any ξ and any u for which

$$\int_0^{\infty} \sigma(|u(s)|)\,ds < \infty,$$

it holds that $T_{\xi,u} = \infty$, and $\liminf_{t\to+\infty} |x(t,\xi,u)| = 0$.

DEFINITION 4.5. *A system as in* (1) *satisfies the* bounded energy frequently bounded state *(BEFBS) property if for some σ of class $\mathcal{K}_\infty$ the following holds:*

$$\int_0^{+\infty} \sigma(|u(s)|)\,ds < +\infty \;\Rightarrow\; \liminf_{t\to+\infty} |x(t,\xi,u)| < +\infty. \tag{33}$$

To be more precise, (33) *means that for any ξ and any u for which $\int_0^{\infty} \sigma(|u(s)|)\,ds < \infty$, it holds that $T_{\xi,u} = \infty$ and $\liminf_{t\to+\infty} |x(t,\xi,u)| < \infty$.*

Remark 4.6. Note that, given any input function u, any $T < \infty$, and any $\mathcal{K}_\infty$-function σ, one has $\int_0^T \sigma(|u(s)|) < \infty$. Hence, together with the causality of the trajectories, the BEFBS property implies the forward completeness property. That is, if a system is BEFBS, then the system is forward complete. Since the BEWCS property implies the BEFBS property, the BEWCS property also implies the forward completeness property. □

We say that a system as in (1) is *zero-GAS* if the corresponding zero-input system $\dot{x} = f(x,0)$ is globally asymptotically stable, and is *zero-LS* if the zero-input system is locally (neutrally) stable.

THEOREM 3. *The following properties are equivalent for system* (1) *with* $\mathbb{U} = \mathbb{R}^m$:

1. *iISS,*
2. *BEWCS and zero-LS,*
3. *BEFBS and zero-GAS.*

Proof. Implication 1 $\Rightarrow$ 3 follows immediately from the definition of iISS. We show next 3 $\Rightarrow$ 2 and 2 $\Rightarrow$ 1.

[3 $\Rightarrow$ 2]. Assume that the system (1) is zero-GAS and satisfies (33) for some $\sigma \in \mathcal{K}_\infty$. By the Lyapunov characterization of zero-GAS in [5, Lemma IV.10], there exists a smooth Lyapunov function $U(\xi)$ such that for some $\tilde{\alpha}_1, \tilde{\alpha}_2 \in \mathcal{K}_\infty$,

$$\tilde{\alpha}_1(|\xi|) \le U(\xi) \le \tilde{\alpha}_2(|\xi|) \qquad \forall \xi \in \mathbb{R}^n,$$

and for some $\tilde{\alpha}, \gamma, \sigma_0 \in \mathcal{K}_\infty$, it holds that

$$\frac{\partial U(\xi)}{\partial \xi} f(\xi,\mu) \le -\tilde{\alpha}(|\xi|) + \gamma(|\xi|)\sigma_0(|\mu|) \qquad \forall \xi \in \mathbb{R}^n, \mu \in \mathbb{U}. \tag{34}$$

By Proposition II.5 in [5], there exists some smooth $\mathcal{K}$-function χ such that for the function V defined by $V = \chi \circ U$ it holds that

$$\frac{\partial V(\xi)}{\partial \xi} f(\xi,\mu) \le -\rho(|\xi|) + \sigma_1(|\mu|) \qquad \forall \xi \in \mathbb{R}^n, \mu \in \mathbb{U}, \tag{35}$$

for some $\mathcal{K}_\infty$ function σ_1 and some positive definite function ρ. Let $\tilde{\sigma} = \max\{\sigma_1, \sigma\}$, where σ is as in (33). Pick $\xi \in \mathbb{R}^n$ and u with $\int_0^\infty \tilde{\sigma}(|u(s)|)\, ds < +\infty$. By the BEFBS assumption,

$$m := \liminf_{t \to +\infty} |x(t, \xi, u)| < +\infty.$$

We want to show $m = 0$. For the sake of contradiction, assume $m > 0$. For any $r > 0$, let

$$w(r) := \max_{|\xi| \le r} V(\xi).$$

Claim. $w(3m) - w(2m) > 0$.

Proof. Suppose $w(3m) - w(2m) = 0$. Then there exists some ξ_0 with $|\xi_0| \le 2m$ at which V takes the maximum value $w(3m)$. Since ξ_0 is an interior point of the open ball centered at 0 with radius $3m$, it follows that $\frac{\partial}{\partial \xi} V(\xi_0) = 0$. This contradicts (35) applied with $\mu = 0$.

We let T be such that

$$\int_T^{+\infty} \tilde{\sigma}(|u(s)|)\, ds < w(3m) - w(2m).$$

By the definition of m, there exists $\tau \ge T$ such that $|x(\tau, \xi, u)| < 2m$. By virtue of (35), for all $t \ge \tau$

$$\begin{aligned} V(x(t, \xi, u)) - V(x(\tau, \xi, u)) &\le \int_\tau^t \sigma_1(|u(s)|)\, ds \\ &< \int_T^{+\infty} \tilde{\sigma}(|u(s)|)\, ds < w(3m) - w(2m). \end{aligned} \tag{36}$$

Hence $V(x(t, \xi, u)) < w(3m)$ for all $t \ge \tau$. This implies that

$$U(x(t, \xi, u)) \le \chi^{-1}(w(3m)) \quad \forall\, t \ge \tau$$

(note that $w(3m)$ is in the range of the $\mathcal{K}$-function χ). Hence, $x(t, \xi, u)$ stays bounded on $[\tau, \infty)$, and consequently, $x(t, \xi, u)$ is bounded on $[0, \infty)$. Let $M > 0$ be such that $|x(t, \xi, u)| < M$ for all t. By (34), one sees that

$$\frac{d}{dt} U(x(t)) \le -\tilde{\alpha}(|x(t)|) + \gamma(M)\sigma_1(|u(t)|),$$

where $x(t)$ denotes the considered trajectory $x(t, \xi, u)$. This is enough to conclude that for some $\beta \in \mathcal{KL}$ and some $\alpha \in \mathcal{K}_\infty$ it holds that

$$\alpha(|x(t, \xi, u)|) \le \beta(|\xi|, t) + \int_0^t \tilde{\sigma}(|u(s)|)\, ds,$$

and therefore, as shown in [26], $|x(t, \xi, u)| \to 0$. This implies $m = 0$, which is clearly a contradiction. □

[2 ⇒ 1]. Suppose that (1) satisfies the BEWCS property with an estimate as in (32). Consider the auxiliary system (22), where σ is the energy supply function as in (32). Note that this system is forward complete (cf. Remark 4.6). By (32), one sees that for any $\gamma_y \in \mathcal{K}_\infty$ it holds that

$$\liminf_{t \to \infty} |x(t, \xi, u)| \le \gamma_y(\|y\|).$$

Therefore, for any choice of γ_u and $\gamma_y \in \mathcal{K}_\infty$, the following asymptotic property is true:

$$\liminf_{t\to+\infty} |x(t,\xi,u)| \le \max\{\gamma_u(\|u\|), \gamma_y(\|y\|)\}. \tag{37}$$

Since $|e(t)| = |y(t)| \le \|y\|$ for all $t \ge 0$, system (22) satisfies the IO-LIM property. Also, it follows from the zero-LS and the BEWCS properties that the system (1) is zero-GAS. Hence, the corresponding augmented system (22) is zero-OSS. By applying the main result in section 3, we find that (22) is IOSS, and therefore, by virtue of Lemma 4.2, system (1) is iISS. ∎

Remark 4.7. Notice that, in the construction of the augmented system (22), the only requirement on the function σ is that it must be class $\mathcal{K}_\infty$. In order to apply the results for IOSS systems, though, the local Lipschitz condition of the dynamics is needed. This issue can be dealt with as follows. First of all, one may always choose a function $\hat\sigma \in \mathcal{K}_\infty$ that majorizes σ so that $\hat\sigma^{-1}$ is locally Lipschitz. Then the iISS, the BEWCS, or the BEFBS estimates as in (21), (32), or (33), respectively, are not violated by replacing σ by $\hat\sigma$. Consider the change of input variables by $u \doteq \frac{w}{|w|}\hat\sigma^{-1}(|w|)$. Clearly $\hat\sigma(|u|) = |w|$, and therefore, iISS, BEWCS, or BEFBS of (1) with $\hat\sigma$ as the input energy supply function is easily seen to be equivalent to iISS, BEWCS, or BEFBS, respectively, for

$$\dot x = f(x, w\sigma^{-1}(|w|)/|w|) \tag{38}$$

with the new supply function $(s) = |s|$. The augmented system corresponding to (38) will therefore be

$$\dot x = f(x, w\sigma^{-1}(|w|)/|w|), \quad \dot e = |w|, \quad y = e. \tag{39}$$

The proofs of the previously derived results, when applied to system (39) and (38), provide proofs of the results in Theorem 3 for the original system (1). Therefore the Lipschitz condition on σ can be dropped. ∎

5. Uniform detectability. In this section we develop some machinery which will be needed to prove Theorem 1. We will derive a separation principle for the property we call *uniform OSS*. Here we consider systems with inputs (unlike in the definition of OSS), but we think of those inputs not as additive disturbances but as multiplicative time-varying uncertainties. We restrict the possible values of these inputs by considering systems as in (1) with $u \in \mathcal{M}_\mathcal{O}$ for some $\mathcal{O} \subset \mathbb{U}$. Throughout this section, we assume that $\mathcal{O}$ is compact.

DEFINITION 5.1. *For a system as in* (1)*, the* global detectability property *holds if the following hold:*

- *there exist* $\bar\sigma_1$, $\bar\sigma_2 \in \mathcal{K}$ *so that*

$$|x(t,\xi,u)| \le \max\{\bar\sigma_1(|\xi|), \bar\sigma_2(\|y_{[0,t]}\|)\} \quad \forall\, \xi \in \mathbb{R}^n,\ \forall\, u,\ \forall\, t \in [0, T_{\xi,u}); \tag{40}$$

- *there exists* $\gamma \in \mathcal{K}_\infty$ *so that*

$$\limsup_{t\to T_{\xi,u}} |x(t,\xi,u)| \le \gamma(\|y\|) \tag{41}$$

for all $\xi \in \mathbb{R}^n$*, all* $u \in \mathcal{M}_\mathcal{O}$. ∎

Note that the last condition (41) is just the IO-AG condition as in (6) with $\gamma_u = 0$. It can be seen that if a system is globally detectable, then the system satisfies the UO property.

DEFINITION 5.2. *We say that the* uniform output-to-state stability property *holds for* (1) *with* $u \in \mathcal{M}_{\mathcal{O}}$ *if for some* $\tilde{\gamma} \in \mathcal{K}_\infty$ *and some* $\beta \in \mathcal{KL}$ *the system satisfies*

$$|x(t,\xi,u)| \leq \max\{\beta(|\xi|,t), \tilde{\gamma}(\|y_{[0,t]}\|)\} \tag{42}$$

for all $\xi \in \mathbb{R}^n$, *all* $u \in \mathcal{M}_{\mathcal{O}}$, *and all* $t \in [0, T_{\xi,u})$. □

Clearly, the uniform output-to-state stability property implies the global detectability property. The main result of this section says that the converse is true as well when $\mathcal{O}$ is compact.

THEOREM 4. *Consider a system as in* (1) *with* $\mathcal{O}$ *compact. The system is* globally detectable *if and only if it is* uniformly output-to-state stable.

To prove Theorem 4, we will need the following result.

LEMMA 5.3. *Consider system* (1) *with* $u \in \mathcal{M}_{\mathcal{O}}$ *for some compact set* $\mathcal{O}$. *Assume that the system* (1) *satisfies the global detectability property. Then there exists some* $\gamma_1 \in \mathcal{K}$ *such that for all* $\varepsilon > 0$ *and all* $r > 0$ *there exists* $T_{\varepsilon,r}$ *so that for any* $\xi \in \mathbb{R}^n$ *with* $|\xi| \leq r$ *and for any* $u \in \mathcal{M}_{\mathcal{O}}$, *if* $T_{\varepsilon,r} < T_{\xi,u}$, *then*

$$|x(t,\xi,u)| \leq \max\left\{\varepsilon,\ \gamma_1\left(\|y_{[0,t]}\|\right)\right\}$$

for all $t \in [T_{\varepsilon,r}\,,\ T_{\xi,u})$.

Proof. Suppose a system satisfies the global detectability property for $u \in \mathcal{M}_{\mathcal{O}}$ as in (40) and (41). Let $\rho(r) := \max\{\bar{\sigma}_2(r), \gamma(r)\}$. It can be seen that, with $z = x$, $w = y$, and $\hat{z}(t) = \max\{|z(t)| - \rho(|w(t)|),\ 0\}$, the system is globally error-detectable as defined in [3] (see also Appendix B of this paper). Hence, by Lemma B.6, the system is uniformly globally error-detectable. Combining this with Remark B.5, we have proved Lemma 5.3. □

Modifying $T_{\varepsilon,r}$ if necessary, we may restate Lemma 5.3 as in the following.

COROLLARY 5.4. *Consider system* (1) *with* $u \in \mathcal{M}_{\mathcal{O}}$ *for some compact set* $\mathcal{O}$. *Assume that the system* (1) *satisfies the global detectability property. Then there exists a continuous map* $T : \mathbb{R}_{>0} \times \mathbb{R}_{\geq 0}$ *with the properties that for each* $r > 0$, $T(\cdot, r)$ *is decreasing, and for each* $\varepsilon > 0$, $T(\varepsilon, \cdot)$ *is increasing, so that for any* $\varepsilon > 0$, $|\xi| \leq r$, *and any* $u \in \mathcal{M}_{\mathcal{O}}$, *the following holds:*

$$T(\varepsilon,r) < T_{\xi,u} \ \Rightarrow\ \Big\{|x(t,\xi,u)| \leq \max\left\{\varepsilon,\ \gamma_1\left(\|y_{[0,t]}\|\right)\right\} \quad \forall\, t \in [T(\varepsilon,r)\,,\ T_{\xi,u})\Big\}. \tag{43}$$

The following lemma on $\mathcal{KL}$ functions will also be needed. This fact is proved in [21] and is stated as Lemma 4.1 in [1]. (That reference requires φ to take nonnegative values, but this can always be assumed without loss of generality, simply replacing φ by $\max\{\varphi, 0\}$.)

PROPOSITION 5.5. *If a function* $\varphi : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ *satisfies*

- *for all* $r > 0$, $\varepsilon > 0$, *there exists some* $T = T(\varepsilon, r) > 0$ *so that* $\varphi(s,t) < \varepsilon$ *for all* $s \leq r$ *and* $t \geq T$;
- *for all* $\varepsilon > 0$, *there exists* $\delta > 0$ *so that* $\varphi(s,t) \leq \varepsilon$ *for all* $s \leq \delta$ *and all* $t \geq 0$,

then there exists some $\beta \in \mathcal{KL}$ *so that* $\varphi(s,t) \leq \beta(s,t)$ *for all* $s \geq 0$, $t \geq 0$. □

We are now ready to prove Theorem 4.

Proof of Theorem 4. Suppose that the system (1) with $u \in \mathcal{M}_{\mathcal{O}}$ is globally detectable. Then estimate (40) holds for some $\bar{\sigma}_1, \bar{\sigma}_2 \in \mathcal{K}$, and there exist some

$\gamma_1 \in \mathcal{K}$ and some $T(\cdot,\cdot)$ as in Corollary 5.4 such that (43) holds. Without loss of generality, we assume that $\gamma_1 = \bar{\sigma}_2$. For each $s \geq 0$, $t \geq 0$, set

$$\varphi(s,t) := \sup\left\{|x(t,\xi,u)| - \gamma_1(\|y_{[0,t]}\|) :\ |\xi| \leq s,\ u \in \mathcal{M}_{\mathcal{O}},\ t < T_{\xi,u}\right\}$$

(with the convention that $\sup \emptyset = -\infty$).

It follows from definitions of $T(\cdot,\cdot)$ and φ that for any $r > 0$ and any $\varepsilon > 0$, $\varphi(s,t) < \varepsilon$ for all $s \leq r$ and $t \geq T(\varepsilon, r)$. By (40), $\varphi(s,t) \leq \bar{\sigma}_1(|\xi|)$ for all $s \geq 0$ and $t \geq 0$. Hence, φ satisfies both conditions as in Proposition 5.5, and thus, there exists some $\beta \in \mathcal{KL}$ so that $\varphi(s,t) \leq \beta(s,t)$ for all $s \geq 0$, $t \geq 0$. Combining this with the definition of φ, we get

$$|x(t,\xi,u)| \leq \beta(|\xi|, t) + \gamma_1(\|y_{[0,t]}\|)$$

for all $\xi \in \mathbb{R}^n$, all $u \in \mathcal{M}_{\mathcal{O}}$, and all $t < T_{\xi,u}$. ∎

6. Proofs of the technical lemmas. To prove Lemmas 3.1–3.3, we need the following result.

DEFINITION 6.1. *A system as in* (1) *satisfies the* input-output local stability property *(IO-LS) if there exist $\delta > 0$ and $\mathcal{K}$ functions $\alpha_0, \alpha_u, \alpha_y \in \mathcal{K}$ so that for all ξ, all $u(\cdot)$, and all $t \in [0, T_{\xi,u})$ it holds that*

$$\max\{|\xi|, \|u\|, \|y_{[0,t]}\|\} \leq \delta \Rightarrow |x(t,\xi,u)| \leq \max\{\alpha_0(|\xi|), \alpha_u(\|u\|), \alpha_y(\|y_{[0,t]}\|)\}. \tag{44}$$

LEMMA 6.2. *(IO-LS) & (IO-BND) ⇔ (IO-GS).*

Proof. The implication (IO-GS) ⇒ (IO-LS) & (IO-BND) is obvious; therefore, we only show the converse. Let δ, α_0, α_u, α_y, σ_0, σ_u, σ_y be given as in (44) and (7). Pick a constant $c > 0$ and three class-$\mathcal{K}$ functions $\beta_\star$ such that for each $\star \in \{0, u, y\}$, $\sigma_\star(s) \leq \beta_\star(s) + c$ for all $s > 0$. Pick class-$\mathcal{K}$ functions $\gamma_\star$ so that for each $\star \in \{0, u, y\}$ it holds that

$$\gamma_\star(s) \geq \begin{cases} \max\{\alpha_\star(s), 3\beta_\star(s)\} & \forall\ 0 \leq s \leq \delta, \\ 3[\beta_\star(s) + 3c] & \forall s \geq \delta. \end{cases} \tag{45}$$

Consider any ξ and $u(\cdot)$. Then (IO-GS) holds with gains γ_0, γ_u, and γ_y. Indeed, if $|\xi| \leq \delta$, $\|u\| \leq \delta$, and $\|y_{[0,t]}\| \leq \delta$, this follows by (IO-LS). If instead $|\xi| > \delta$, the definition of (IO-BND) implies

$$\begin{aligned} |x(t,\xi,u)| &\leq \sigma_0(|\xi|) + \sigma_u(\|u\|) + \sigma_y(\|y_{[0,t]}\|) \\ &\leq \beta_0(|\xi|) + \beta_u(\|u\|) + \beta_y(\|y_{[0,t]}\|) + 3c \\ &\leq [\beta_0(|\xi|) + 3c] + [\gamma_u(\|u\|) + \gamma_y(\|y_{[0,t]}\|)]/3 \\ &\leq (1/3)[\gamma_0(|\xi|) + \gamma_u(\|u\|) + \gamma_y(\|y_{[0,t]}\|)] \\ &\leq \max\{\gamma_0(|\xi|), \gamma_u(\|u\|), \gamma_y(\|y_{[0,t]}\|)\}. \end{aligned} \tag{46}$$

The case $\|u\| > \delta$ can be treated in a similar way. Finally, we are left to deal with the case $\delta < \|y_{[0,t]}\| < +\infty$. (The case $\|y\| = +\infty$ is trivial.) By (UO), $x(t,\xi,u)$ is well defined, and therefore the argument as in (46) can be repeated. ∎

Remark 6.3. Observe that in the proof of Lemma 6.2, if the functions α_u and σ_u as in (44) and (7) can be chosen zero, and if the $\|u\|$ term is not presented in the $\max\{\cdots\} \leq \delta$ phrase in (44), then the function γ_u in (46) can also be chosen zero. That is, one has the following.

If for some $\alpha_0,\ \alpha_y \in \mathcal{K}$, $\sigma_0,\ \sigma_y \in \mathcal{N}$ it holds that

$$\max\{|\xi|\,, \|y_{[0,t]}\|\} \le \delta \Rightarrow |x(t,\xi,u)| \le \max\{\alpha_0(|\xi|),\ \alpha_y(\|y_{[0,t]}\|)\}, \tag{47}$$

and

$$|x(t,\xi,u)| \le \max\{\sigma_0(|\xi|),\ \sigma_y(\|y_{[0,t]}\|)\} \qquad \forall\, \xi \in \mathbb{R}^n,\ \forall\, u,\ \forall\, t \in [0, T_{\xi,u}), \tag{48}$$

then for some $\tilde{\sigma}_0, \tilde{\sigma}_y \in \mathcal{K}$ it holds that

$$|x(t,\xi,u)| \le \max\{\tilde{\sigma}_0(|\xi|),\ \tilde{\sigma}_y(\|y_{[0,t]}\|)\} \qquad \forall\, \xi \in \mathbb{R}^n,\ \forall\, u,\ \forall\, t \in [0, T_{\xi,u}). \tag{49}$$

6.1. Proof of Lemma 3.3. Let $\gamma_u, \gamma_y, \sigma_0, \sigma_u, \sigma_y \in \mathcal{K}$ be as in (5) and (7). Pick any ξ, u and any $\varepsilon > 0$. By (5), there is some $T \in [0, T_{\xi,u})$ such that

$$|x(T,\xi,u)| \le \max\{\gamma_u(\|u\|),\ \gamma_y(\|y\|)\} + \varepsilon.$$

Applying (7) to the initial state $\xi_T := x(T,\xi,u)$ and the control $u_T(t) := u(t+T)$ (whose corresponding output function is $y_T(t,\xi_T,u_T) = y(t+T,\xi,u)$), we conclude that

$$\begin{aligned}\sup_{t\ge T} |x(t,\xi,u)| &= \sup_{0\le t< T_{\xi_T,u_T}} |x(t,\xi_T,u_T)| \le \max\{\sigma_0(|\xi_T|), \sigma_u(\|u_T\|), \sigma_y(\|y\|)\} \\ &\le \max\left\{\sigma_0(\gamma_u(\|u\|)+\varepsilon), \sigma_0(\gamma_y(\|y\|)+\varepsilon), \sigma_u(\|u\|), \sigma_y(\|y\|)\right\}.\end{aligned}$$

Letting $\varepsilon \to 0$, we get

$$\limsup_{t\to T_{\xi,u}} |x(t,\xi,u)| \le \max\{\hat{\gamma}_u(\|u\|),\ \hat{\gamma}_y(\|y\|)\}, \tag{50}$$

where $\hat{\gamma}_u(s) = \max\{\sigma_0(\gamma_u(s)), \sigma_u(s)\}$, $\hat{\gamma}_y(s) = \max\{\sigma_0(\gamma_y(s)), \sigma_y(s)\}$. □

Remark 6.4. We remark that if γ_u and σ_u as in (5) and (7) can be chosen to be the zero function, then the gain function $\hat{\gamma}_u$ in (50) can also be chosen to be the zero function. □

6.2. Proof of Lemma 3.1. Consider the zero-input system

$$\dot{x} = f(x,0), \quad y = h(x). \tag{51}$$

Denote the trajectory corresponding to an initial state ξ by $x(t,\xi)$. This trajectory is defined on a maximal interval $[0, T_\xi)$. Applying Lemma 2.7 and Remark 2.8 to the zero-input system, we know that there exists some $\sigma_0,\ \sigma_y \in \mathcal{N}$ such that for every trajectory of the system the following holds:

$$|x(t,\xi)| \le \max\{\sigma_0(|\xi|),\ \sigma_y(\|y_{[0,t]}\|)\} \qquad \forall\, t \in [0, T_\xi). \tag{52}$$

By Remark 6.3, one sees that, together with the zero-O-LS property (as defined in (15)), (52) implies that the system is locally detectable, that is, for some $\tilde{\sigma}_0, \tilde{\sigma}_y \in \mathcal{K}$, (49) holds.

Applying Remark 6.4 to (51) with the IO-LIM estimate (5) with $\gamma_u = 0$ and (49), one sees that the following holds for every trajectory $x(t,\xi)$ of (51):

$$\limsup_{t\to T_\xi} |x(t,\xi)| \le \gamma\left(\|y_{[0,T_\xi)}\|\right). \tag{53}$$

The combination of (52) and (53) means that the system (51) is globally detectable. By Theorem 4, the system (51) is OSS. □

6.3. Proof of Lemma 3.2. Clearly (local-IOSS) $\Rightarrow$ (IO-LS). Moreover, by Lemma 2.7, (IO-LIM) $\Rightarrow$ (IO-BND). Combining this with Lemma 6.2, we get

$$\text{(local IOSS) \& (IO-LIM)} \Rightarrow \text{(IO-LS) \& (IO-BND)} \Leftrightarrow \text{(IO-GS)}, \tag{54}$$

which completes the proof. □

6.4. Proof of Proposition 3.4. In order to complete the proof, the following result will be needed.

LEMMA 6.5. *Suppose that system* (1) *satisfies the IO-AG and IO-GS properties. Then there exists some locally Lipschitz* $\varphi \in \mathcal{K}_\infty$ *such that the system*

$$\dot{x}(t) = f(x(t), d(t)\varphi(|x(t)|)), \qquad y(t) = h(x(t)), \tag{55}$$

where d *denotes the disturbance functions taking values in the closed unit ball of* $\mathbb{R}^m$*, is globally detectable.*

Proof. Let σ_0, σ_u, $\sigma_y \in \mathcal{K}$ be such that the IO-GS estimate (7) holds, and let γ_u, $\gamma_y \in \mathcal{K}$ be such that the IO-AG estimate (6) holds for the system. Without loss of generality, we may assume that $\sigma_u = \gamma_u$ and $\sigma_y = \gamma_y$. Pick any locally Lipschitz $\mathcal{K}_\infty$-function φ such that $\sigma_u(\varphi(s)) \le s/2$ for all $s \ge 0$. Below we show that with the function φ, the corresponding system (55) is globally detectable. For any $\xi \in \mathbb{R}^n$ and any d, we let $x_\varphi(t, \xi, d)$ denote the corresponding solution for (55), $y_\varphi = h(x_\varphi)$, and let $[0, \tau_{\xi,d})$ denote the maximal interval of the solution. Observe that for any ξ, d it holds that

$$x_\varphi(t, \xi, d) = x(t, \xi, u_d) \qquad \forall\, 0 \le t < \tau_{\xi,d},$$

where $u_d(t) = d(t)\varphi(|x_\varphi(t, \xi, d)|)$. Consequently,

$$\begin{aligned} |x_\varphi(t, \xi, d)| &\le \max\left\{\sigma_0(|\xi|),\ \sigma_u\left(\varphi\left(\|x_\varphi(\cdot, \xi, d)_{[0,t)}\|\right)\right),\ \sigma_y(\|(y_\varphi)_{[0,t)}\|)\right\} \\ &\le \max\left\{\sigma_0(|\xi|),\ \frac{\|x_\varphi(\cdot, \xi, d)_{[0,t)}\|}{2},\ \sigma_y(\|(y_\varphi)_{[0,t)}\|)\right\} \qquad \forall\, 0 \le t < \tau_{\xi,d}\,. \end{aligned}$$

Thus, for any $0 \le t < \tau_{\xi,d}$,

$$\|x_\varphi(\cdot, \xi, d)_{[0,t)}\| \le \max\left\{\sigma_0(|\xi|),\ \frac{\|x_\varphi(\cdot, \xi, d)_{[0,t)}\|}{2},\ \sigma_y(\|(y_\varphi)_{[0,t)}\|)\right\}.$$

Consequently, for any $0 \le t < \tau_{\xi,d}$,

$$\|x_\varphi(\cdot, \xi, d)_{[0,t)}\| \le \max\left\{\sigma_0(|\xi|),\ \sigma_y(\|(y_\varphi)_{[0,t)}\|)\right\},$$

and in particular,

$$|x_\varphi(t, \xi, d)| \le \max\left\{\sigma_0(|\xi|),\ \sigma_y\left(\|(y_\varphi)_{[0,t)}\|\right)\right\} \tag{56}$$

for all $0 \le t < \tau_{\xi,d}$. This shows that property (40) holds. Below we show that the attractivity property as in (41) holds for the system (55). By (56), the system (55) satisfies the UO property. Pick any ξ, d. Suppose $\tau_{\xi,d} = \infty$ and $\|y\| < \infty$. Then, again, by (56),

$$\limsup_{t\to\infty} |x_\varphi(t, \xi, d)| < \infty. \tag{57}$$

By the IO-AG property (6), with Lemma 2.10,

$$\begin{aligned}\limsup_{t\to\infty}|x_\varphi(t,\xi,d)| &\le \max\left\{\gamma_u\left(\limsup_{t\to\infty}|u_d(t)|\right),\ \gamma_y\left(\limsup_{t\to\infty}|y_\varphi(t)|\right)\right\}\\ &\le \max\left\{\gamma_u\left(\varphi\left(\limsup_{t\to\infty}|x_\varphi(t,\xi,d)|\right)\right),\ \gamma_y\left(\limsup_{t\to\infty}|y_\varphi(t)|\right)\right\}\\ &\le \max\left\{\frac{1}{2}\limsup_{t\to\infty}|x_\varphi(t,\xi,d)|,\ \gamma_y\left(\limsup_{t\to\infty}|y_\varphi(t)|\right)\right\}.\end{aligned}$$

Combining this with (57), we get

$$\limsup_{t\to\infty}|x_\varphi(t,\xi,d)| \le \gamma_y\left(\limsup_{t\to\infty}|y_\varphi(t)|\right).$$

Again, with the UO property, we see that the attractivity property (41) holds for the system (55). Thus, the system is globally detectable. □

Proof of Proposition 3.4. Suppose that a system as in (1) satisfies the IO-AG and the IO-GS properties. By Lemma 6.5, there exists some locally Lipschitz $\varphi \in \mathcal{K}_\infty$ such that the corresponding system (55) is globally detectable. By Theorem 4, the system (55) is uniformly output-to-state stable. Applying Theorem 2.16 together with Corollary 3.6 of [16], one sees that the system (1) is IOSS. (Following the work in [16], if a system (1) admits a locally Lipschitz $\varphi \in \mathcal{K}_\infty$ such that the corresponding system (55) is uniformly OSS, then the system (1) is called robustly OSS.) □

Appendix A. Characterizations of IO-LIM. Consider the following variations of the IO-LIM property.

DEFINITION A.1. *For system* (1)*, we say that*

- *the asymptotic IO-LIM property holds if for some* $\gamma_u, \gamma_y \in \mathcal{K}$*,*

$$\liminf_{t\to T_{\xi,u}}|x(t,\xi,u)| \le \limsup_{t\to T_{\xi,u}}\max\{\gamma_u(|u(t)|),\gamma_y(|y(t)|)\} \tag{58}$$

  for all $\xi \in \mathbb{R}^n$*, all* $u(\cdot)$*;*
- *the causal IO-LIM property holds if for some* $\gamma_u, \gamma_y \in \mathcal{K}$*,*

$$\inf_{t\in[0,T_{\xi,u})}\Big\{|x(t,\xi,u)| - \max\{\gamma_u(\|u_{[0,t]}\|),\gamma_y(\|y_{[0,t]}\|)\}\Big\} \le 0 \tag{59}$$

  for all $\xi \in \mathbb{R}^n$*, all* $u(\cdot)$*;*
- *the asymptotic causal IO-LIM property holds if*

$$\liminf_{t\to T_{\xi,u}}\Big\{|x(t,\xi,u)| - \max\{\gamma_u(\|u_{[0,t]}\|),\gamma_y(\|y_{[0,t]}\|)\}\Big\} \le 0 \tag{60}$$

  for all $\xi \in \mathbb{R}^n$*, all* $u(\cdot)$*.* □

PROPOSITION A.2. *For a system as in* (1)*, the following properties are equivalent:*

1. *IO-LIM,*
2. *asymptotic IO-LIM,*
3. *causal IO-LIM,*
4. *asymptotic causal IO-LIM.* □

Proof. [1 ⇒ 2]: Suppose the IO-LIM property holds for system (1) with $\gamma_u, \gamma_y \in \mathcal{K}_\infty$ as in (5). Fix ξ, u. Pick any $T \in [0, T_{\xi,u})$. Let $\xi_T = x(T,\xi,u)$ and $u_T(t) = u(t+T)$. Applying (5) to ξ_T with u_T, we get

$$\inf_{t\ge T}|x(t,\xi,u)| = \inf_{t\ge 0}|x(t,\xi_T,u_T)| \le \max\left\{\gamma_u(\|u_{[T,T_{\xi,u})}\|),\ \gamma_y(\|y_{[T,T_{\xi,u})}\|)\right\}.$$

This implies that

$$\liminf_{t\to T_{\xi,u}} |x(t,\xi,u)| \le \lim_{T\to T_{\xi,u}} \left\{\max\left\{\gamma_u(\|u_{[T,T_{\xi,u})}\|),\ \gamma_y(\|y_{[T,T_{\xi,u})}\|)\right\}\right\}.$$

With (2), we get (58).

[$2 \Rightarrow 3$]: Assume that (58) holds for some $\gamma_u, \gamma_y \in \mathcal{K}$. Below we show that this will imply (59) with the same γ_u, γ_y functions. Suppose that this fails for some trajectory $x(t,\xi,u)$. It then follows that

$$|x(t,\xi,u)| \ge \max\{\gamma_u(\|u_{[0,t]}\|),\ \gamma_y(\|y_{[0,t]}\|)\} \quad \forall\, t \in [0, T_{\xi,u}).$$

Taking the lim inf on both sides of the inequality, one gets

$$\begin{aligned}\liminf_{t\to T_{\xi,u}} |x(t,\xi,u)| &\ge \liminf_{t\to T_{\xi,u}} \max\{\gamma_u(\|u_{[0,t]}\|),\ \gamma_y(\|y_{[0,t]}\|)\}\\ &= \max\{\gamma_u(\|u_{[0,T_{\xi,u})}\|),\ \gamma_y(\|y_{[0,T_{\xi,u})}\|)\},\end{aligned}$$

which contradicts (58).

[$3 \Rightarrow 4$]: The proof of this portion is trivial.

[$4 \Rightarrow 1$]: An estimate as in (60) in particular implies that

$$\liminf_{t\to T_{\xi,u}} \Big[|x(t,\xi,u)| - \max\{\gamma_u(\|u\|),\ \gamma_y(\|y\|)\}\Big] \le 0,$$

which is the same as (8). By Remark 2.5, IO-LIM holds. □

Appendix B. The error detectability properties. In some of the proofs in this paper, we have used some results from the recent work [3]. They are stated here, in the interest of making this paper self-contained. The first statement is a special case of Corollary 4.3 in [3].

LEMMA B.1. *For a given unboundedness observable system of the form* (1)*, assume*

- *a compact subset $\mathcal{O}$ of the input set $\mathbb{U}$;*
- *subsets K, C, and Ω of the state space $\mathbb{R}^n$ such that K and C are compact, Ω is open, and $K \subset \Omega \subset C$;*
- *subsets $\mathcal{Y}$ and $\mathcal{Y}_0$ of the output space $\mathbb{R}^p$ such that $\mathcal{Y}$ is compact, $\mathcal{Y}_0$ is open, and $\mathcal{Y} \subset \mathcal{Y}_0$*

such that, for all $\xi \in C$ and all inputs $u \in \mathcal{M}_{\mathcal{O}}$, there exists $t \in [0, T_{\xi,u})$ for which

$$x(t,\xi,u) \in K \quad \text{or} \quad h(x(t,\xi,u)) \notin \mathcal{Y}_0. \tag{61}$$

Then $\mathcal{R}_{\mathcal{O}/\mathcal{Y}}(C)$ is bounded. □

Considered a system as in the following:

$$\dot{x}(t) = f(x(t),u(t)),\ z(t) = \varphi(x(t)),\ w(t) = k(x(t)), \tag{62}$$

where $\varphi(\cdot), k(\cdot)$ are continuous maps from $\mathbb{R}^n$ to $\mathbb{R}^k$ and $\mathbb{R}^l$, respectively, for some k and l, and where f satisfies the same assumptions as were made for system (1). Following the work in [3], we call z the output signal and w the measurement signal. The inputs are functions in $\mathcal{M}_{\mathcal{O}}$ for some $\mathcal{O} \subset \mathbb{U}$.

DEFINITION B.2. *The system* (62) *is said to satisfy the* unboundedness observability property through w *(UO through w) if, for each ξ and u such that $T_{\xi,u} < \infty$, it holds that*

$$\limsup_{t\to T_{\xi,u}} |w(t)| = \infty.$$

The following definitions were proposed in [3].

DEFINITION B.3. *We say that system* (62) *is* globally error-detectable *if there exists some* $\gamma \in \mathcal{K}$ *such that for the map* $\hat{z}(\cdot)$ *defined by*

$$\hat{z}(t) = \max\{|z(t)| - \gamma(|w(t)|),\ 0\} \tag{63}$$

the following hold:

- *(Local uniform output-stability modulo measurements): for some* $\sigma_1, \sigma_2 \in \mathcal{K}$ *and some* $\delta > 0$*, it holds that*

$$|\hat{z}(0)| < \delta \Longrightarrow |\hat{z}(t)| \leq \sigma_1(|\hat{z}(0)|) + \sigma\left(\|w\|_{[0,t]}\right) \qquad \forall\, 0 \leq t < T_{\xi,u} \tag{64}$$

  for all ξ *and* $u \in \mathcal{M}_{\mathcal{O}}$.
- *(Asymptotic-detectablity):* $\inf_{0 \leq t < T_{\xi,u}} |\hat{z}(t)| = 0$ *for all* ξ *and* $u \in \mathcal{M}_{\mathcal{O}}$. ∎

DEFINITION B.4. *We say that system* (1) *is* uniformly globally error-detectable *if there exists some* $\gamma \in \mathcal{K}$ *such that for the map* $\hat{z}(\cdot)$ *defined by* (63) *the following hold:*

- *the local uniform output-stability modulo measurements property as in Definition* B.3 *holds, and*
- *(uniform asy-detectability): for any* $\varepsilon > 0$ *and any* $\kappa > 0$*, there exists* $T_{\varepsilon,\kappa}$ *so that for any* $\xi \in \mathbb{R}^n$ *with* $|\xi| \leq \kappa$ *and for any* u *there exists some* $\tau < \min\{T_{\varepsilon,\kappa},\ T_{\xi,u}\}$ *such that*

$$|\hat{z}(\tau)| \leq \varepsilon. \tag{65}$$

Remark B.5. The uniformly global error-detectability implies the following property: there exists some $\gamma_1 \in \mathcal{K}$ such that for all $\varepsilon > 0$ and all $\kappa > 0$ there exists $T_{\varepsilon,\kappa}$ so that for any $\xi \in \mathbb{R}^n$ with $|\xi| \leq \kappa$ and for any u, if $T_{\varepsilon,\kappa} < T_{\xi,u}$, then

$$|z(t)| \leq \varepsilon + \gamma_1\left(\|w\|_{[0,t)}\right)$$

for all $t \in [T_{\varepsilon,\kappa}\,,\ T_{\xi,u})$. ∎

A main result in [3] is the following.

LEMMA B.6. *Let* $\mathcal{O}$ *be compact. Assume that system* (62) *satisfies the UO property through the measurement* w*. Then the system is globally error-detectable in* $u \in \mathcal{M}_{\mathcal{O}}$ *if and only if it is uniformly globally error-detectable in* $u \in \mathcal{M}_{\mathcal{O}}$. ∎

REFERENCES

[1] F. ALBERTINI AND E. D. SONTAG, *Continuous control-Lyapunov functions for asymptotically controllable time-varying systems*, Internat. J. Control, 72 (1999), pp. 1630–1641.

[2] D. ANGELI, *Input-to-state stability of PD-controlled robotic systems*, Automatica, 35 (1999), pp. 1285–1290.

[3] D. ANGELI, B. INGALLS, E. SONTAG, AND Y. WANG, *Uniform global asymptotic stability of differential inclusions*, J. Dynam. Control Systems, to appear.

[4] D. ANGELI AND E. D. SONTAG, *Forward completeness, unboundedness observability, and their Lyapunov characterizations*, Systems Control Lett., 38 (1999), pp. 209–217.

[5] D. ANGELI, E. D. SONTAG, AND Y. WANG, *A characterization of integral input to state stability*, IEEE Trans. Automat. Control, 45 (2000), pp. 1082–1097.

[6] D. ANGELI, E. D. SONTAG, AND Y. WANG, *Further equivalences and semiglobal versions of integral input to state stability*, Dynam. Control, 10 (2000), pp. 127–149.

[7] D. ANGELI, E. D. SONTAG, AND Y. WANG, *Input-to-state stability with respect to inputs and their derivatives*, Internat. J. Robust Nonlinear Control, 13 (2003), pp. 1035–1056.

[8] P. D. CHRISTOFIDES AND A. TEEL, *Singular perturbations and input-to-state stability*, IEEE Trans. Automat. Control, 41 (1996), pp. 1645–1650.

[9] R. A. Freeman and P. V. Kokotović, *Robust Nonlinear Control Design, State-Space and Lyapunov Techniques*, Birkhäuser Boston, Cambridge, MA, 1996.
[10] W. Hahn, *Stability of Motion*, Springer-Verlag, Berlin, 1967.
[11] J. Hespanha and A. Morse, *Certainty equivalence implies detectability*, Systems Control Lett., 36 (1999), pp. 1–13.
[12] B. Ingalls, E. Sontag, and Y. Wang, *Generalizations of asymptotic gain characterizations of ISS to input-to-output stability*, in Proceedings of the 2001 American Control Conference, Arlington, VA, IEEE publishing, Los Alamitos, CA, pp. 2279–2284.
[13] A. Isidori, *Nonlinear Control Systems* II, Springer-Verlag, London, 1999.
[14] Z.-P. Jiang, A. Teel, and L. Praly, *Small-gain theorem for ISS systems and applications*, Math. Control Signals Systems, 7 (1994), pp. 95–120.
[15] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.
[16] M. Krichman, E. D. Sontag, and Y. Wang, *Input-output-to-state stability*, SIAM J. Control Optim., 39 (2001), pp. 1874–1928.
[17] M. Krstić and H. Deng, *Stabilization of Uncertain Nonlinear Systems*, Springer-Verlag, London, 1998.
[18] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear and Adaptive Control Design*, John Wiley & Sons, New York, 1995.
[19] M. Krstić and Z. H. Li, *Inverse optimal design of input-to-state stabilizing nonlinear controllers*, IEEE Trans. Automat. Control, 43 (1998), pp. 336–350.
[20] D. Liberzon, A. Morse, and E. D. Sontag, *Output-input stability and minimum-phase nonlinear systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 422–436.
[21] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.
[22] F. Mazenc, L. Praly, and W. Dayawansa, *Global stabilization by output feedback: Examples and counterexamples*, Systems Control Lett., 23 (1994), pp. 119–125.
[23] D. Nešić and A. R. Teel, *Input-to-state stability for nonlinear time-varying systems via averaging*, Math. Control Signals Systems, 14 (2001), pp. 257–280.
[24] R. Sepulchre, M. Jankovic, and P. V. Kokotović, *Constructive Nonlinear Control*, Springer, London, 1997.
[25] E. D. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
[26] E. D. Sontag, *Comments on integral variants of input-to-state stability*, Systems Control Lett., 34 (1998), pp. 93–100.
[27] E. D. Sontag and A. Teel, *Changing supply function in input/state stable systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1476–1478.
[28] E. D. Sontag and Y. Wang, *New characterizations of the input to state stability property*, IEEE Trans. Automat. Control, 41 (1996), pp. 1283–1294.
[29] E. D. Sontag and Y. Wang, *Output-to-state stability and detectability of nonlinear systems*, Systems Control Lett., 29 (1997), pp. 279–290.
[30] A. R. Teel, L. Moreau, and D. Nešić, *A note on the robustness of input-to-state stability*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando (2001), pp. 875–880.
[31] J. Tsinias, *Stochastic Input-to-State stability and applications to global feedback stabilization*, Internat. J. Control, 71 (1998), pp. 907–930.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 277–300

REDUCTION OF CONTROLLED LAGRANGIAN AND HAMILTONIAN SYSTEMS WITH SYMMETRY*

DONG EUI CHANG† AND JERROLD E. MARSDEN‡

Abstract. We develop reduction theory for controlled Lagrangian and controlled Hamiltonian systems with symmetry. Reduction theory for these systems is needed in a variety of examples, such as a spacecraft with rotors, a heavy top with rotors, and underwater vehicle dynamics. One of our main results shows the equivalence of the method of reduced controlled Lagrangian systems and that of reduced controlled Hamiltonian systems in the case of simple mechanical systems with symmetry.

Key words. controlled Lagrangian systems, controlled Hamiltonian systems, reduction, equivalence

AMS subject classifications. 70Q05, 93C15, 34H05

DOI. 10.1137/S0363012902412951

1. Introduction.

We develop the theory of symmetry reduction for the class of controlled Lagrangian (CL) and controlled Hamiltonian (CH) systems with symmetry and show the equivalence of the method of reduced CL systems and that of reduced CH systems for simple mechanical systems. The phrases "controlled Lagrangian" and "controlled Hamiltonian" systems were coined in [20], and their technical definition is recalled below. The concept of a CH system used here is the same as that of a "port-controlled Hamiltonian system" in [32]. We prefer to use the term "controlled Hamiltonian" in a way that is parallel to "controlled Lagrangian." The notion of an "implicit Hamiltonian system" is related to that of a CH system but is based on the more general notion of Dirac structures rather than symplectic and Poisson structures, and a type of reduction theory (different from what is given here) in that context was developed in [6]. The notion of an "implicit Lagrangian system" is developed, also using Dirac structures, in [35], and in future work we hope to explore the reduction theory of controlled implicit Hamiltonian and Lagrangian systems and the relation between them.

CL systems. We will begin by describing the Lagrangian side. To help design stabilizing controllers for mechanical systems whose equations are written in Euler–Lagrange form with control forces, the CL method was developed by a number of authors, including [3], [4], [7], [10], [11], [12], [13], [14], [15], [19], [20], [22], [23], [33]. In the study of CL systems, it is important to take into account external forces and control forces, including gyroscopic forces.

The main idea of the CL method is to proceed in the following steps:

1. Define a (feedback-transformation) equivalence relation among CL systems.
2. Given a CL system, find an equivalent CL system such that the energy of the second CL system has a minimum at the equilibrium of interest and the system is forced or controlled by a dissipative (plus gyroscopic) force.

*Received by the editors August 10, 2002; accepted for publication (in revised form) October 23, 2003; published electronically June 15, 2004. This research was partially supported by Caltech and by ONR contract N00014-02-1-0826.

http://www.siam.org/journals/sicon/43-1/41295.html

†Centre Automatique et Systèmes, École Nationale Supérieure des Mines de Paris, 60, bd. Saint-Michel, 75272 Paris CEDEX 06, France (dchang@cas.ensmp.fr).

‡Control and Dynamical Systems 107-81, California Institute of Technology, Pasadena, CA 91125 (marsden@cds.caltech.edu).

3. Using the energy as a Lyapunov function, obtain asymptotic stability of the equilibrium in the second CL system.
4. Since the two systems are equivalent, the equivalence relation gives an asymptotically stabilizing controller for the first system.

CH systems. A similar approach has been taken on the Hamiltonian side, too, where the main governing equations are in Hamiltonian form (see [5], [9], [32], [34] and references therein).

Equivalence. There have been some recent studies concerning the equivalence of both methods. First of all, [5] showed that the CH method is more general than the CL method for simple mechanical systems and used a limited definition of CL systems. Then [20] extended the theory of CL systems and showed the full equivalence of the method of CL systems with the method of CH systems for simple mechanical systems.

Symmetry. The main goal of the present paper is to include symmetry in the discussion. We do so in a systematic and general way. In previous works, symmetry was considered in certain cases (see, for instance, [15], [23], and [26]), but a comprehensive theory has been lacking.

In this paper, we define the notion of G-invariant CL and CH systems and develop a reduction theory for them. In addition, we trace through the equivalence of the CL and CH systems under the reduction process.

Reduction theory. Reduction theory for mechanical systems has been studied both on the Lagrangian side and on the Hamiltonian side. It has proven to be a powerful tool for studying the dynamics, stability, and control of many mechanical systems, such as rigid body and spacecraft systems, underwater vehicles, heavy tops, and fluid systems.

On the Lagrangian side, one reduces variational principles, an idea due to [29] and [30]. A comprehensive account of this theory is given in [17]. On the other hand, on the Hamiltonian side, one reduces symplectic and Poisson structures. See, for instance, [1], [27], and [28] and references therein. We shall merge these ideas with those of CL and CH systems to develop the notions and basic properties of reduced CL and CH systems.

Equivalence and reduction. Once we have developed the reduction theory, we will be in a position to define an equivalence relation among G-invariant CL (resp., CH) systems; in this context, we show that two reduced CL (resp., CH) systems are equivalent if and only if their unreduced G-invariant CL (resp., CH) systems are equivalent. This is given in Theorems 2.10 and 3.9.

We also show that this reduced equivalence relation is a feedback-transform equivalence relation (see Theorems 2.11 and 3.10). Based on these results and the equivalence of the methods of CL and CH systems for simple mechanical systems proven in [20], we prove that the method of reduced CL systems and that of reduced CH systems are equivalent for reduced simple mechanical systems (this is given in Theorem 4.3).

Example. In the final section, we review the example of a satellite controlled via a spinning rotor and show that the Euler–Poincaré matching conditions derived in [8] and [15] directly in the reduced context can be obtained in a natural way through the reduction process. We study the stabilization of the heavy top with rotors, using the reduced CL method as an illustration.

2. Reduction of CL systems with symmetry. We begin this section with some general notation, then we introduce the notion of a CL system with symmetry, and finally study the theory of reduction for such systems.

We first summarize some general notation that will be used. We refer to [2], [17], [25], and [28] for more details and background information.

Manifolds and bundles. Let Q be the configuration manifold, and let $\tau_Q : TQ \to Q$ and $\pi_Q : T^*Q \to Q$ be the tangent bundle projection and the cotangent bundle projection, respectively. The second order tangent bundle $\tau_Q^{(2)} : T^{(2)}Q \to Q$ is defined as follows. For $\bar{q} \in Q$, elements of $T_q^{(2)}Q$ are equivalence classes of curves in Q, the equivalence relation being defined as follows: two curves $q_i(t)$, $i = 1, 2$, with $q_1(\bar{t}) = q_2(\bar{t}) = \bar{q}$, are equivalent, by definition, if and only if in any local chart we have $q_1^{(l)}(\bar{t}) = q_2^{(l)}(\bar{t})$ for $l = 1, 2$, where $q^{(l)}(t)$ denotes the derivative of order l. The second order bundle has local coordinates given by $(q^i, \dot{q}^i, \ddot{q}^i)$ and may be thought of as a subbundle of the second tangent bundle TTQ via the embedding $(q^i, \dot{q}^i, \ddot{q}^i) \mapsto (q^i, \dot{q}^i, \dot{q}^i, \ddot{q}^i)$. As is explained in [28], the second order tangent bundle is the basic space on which the Euler–Lagrange operator of mechanics is defined; in fact, given a Lagrangian L, the associated Euler–Lagrange operator $\mathcal{EL}$ induces a bundle map $\mathcal{EL}(L) : T^{(2)}Q \to T^*Q$.

For a manifold M, $\mathcal{F}(M)$ denotes the set of smooth real-valued functions on M.

Symmetry groups. Let G be a Lie group acting (on the left) on Q freely and properly so that $\pi_G(Q) : Q \to Q/G$ becomes a principal bundle. The tangent (resp., cotangent) lift of the action of G on Q defines an action of G on TQ (resp., T^*Q), which is automatically free and proper as well, so that the maps $\tau_{/G} : TQ \to TQ/G$ (resp., $\pi_{/G} : T^*Q \to T^*Q/G$) also define principal bundles. When M is a manifold on which G acts, we let $[m]_G$ denote an equivalence class of $m \in M$ in the quotient space M/G. Even though we do not explicitly denote the manifold M in this notation, it will be clear which manifold is meant from the context. The space TQ/G becomes a vector bundle with base Q/G by inheriting the vector bundle structure of TQ as follows:

$$[u_q]_G + \lambda[v_q]_G = [u_q + \lambda v_q]_G,$$

where $\lambda \in \mathbb{R}$; $u_q, v_q \in T_qQ$; and $[u_q]_G, [v_q]_G$ are their equivalence classes in the quotient space TQ/G. The fiber $(TQ/G)_x$ is isomorphic, as a vector space, to T_qQ for each $x = [q]_G \in Q/G$, $q \in Q$ (see Lemma 2.4.1 in [17]). In the same manner, the space T^*Q/G becomes a vector bundle with base Q/G.

Vertical lifts. Let V be a vector bundle over a manifold Q. The vertical lift of a vector $w_q \in V_q$ along the vector $v_q \in V_q$ is the vector, $\mathrm{vlift}_{v_q}(w_q) \in T_{v_q}V$, defined by

$$\mathrm{vlift}_{v_q}(w_q) = \left.\frac{d}{dt}\right|_{t=0} (v_q + tw_q).$$

The vertical lift of a fiber-preserving map $F : V \to V$ is a section, $\mathrm{vlift}(F) : V \to TV$, defined by

$$\mathrm{vlift}(F)(v_q) = \mathrm{vlift}_{v_q}(F(v_q)). \tag{2.1}$$

The vertical lift of a subbundle W of V is the subbundle of TV defined by

$$\mathrm{vlift}(W) = \{\mathrm{vlift}_{v_q}(w_q) \mid v_q \in V_q, w_q \in W_q, q \in Q\}. \tag{2.2}$$

CL systems. We next recall the definition of a CL system from [20].

Definition 2.1. *A* controlled Lagrangian (CL) system *is a triple* (L, F, W), *where the function* $L : TQ \to \mathbb{R}$ *is the* Lagrangian, *the fiber-preserving map* F :

$TQ \to T^*Q$ is the (external) force map*, and the subbundle W of T^*Q is called the* control subbundle*. When one chooses a feedback control law $u : TQ \to W$, the triple (L, F, u) will denote the* closed-loop CL system*.*

Euler–Lagrange operator. The Euler–Lagrange operator $\mathcal{EL}$ assigns to a Lagrangian $L : TQ \to \mathbb{R}$ a bundle map $\mathcal{EL}(L) : T^{(2)}Q \to T^*Q$, which may be written in local coordinates (and using the index summation convention) as

$$\mathcal{EL}(L)(q, \dot{q}, \ddot{q}) = \left(\frac{d}{dt} \frac{\partial L}{\partial \dot{q}^i}(q, \dot{q}) - \frac{\partial L}{\partial q^i}(q, \dot{q}) \right) \mathbf{d}q^i,$$

in which it is understood that one regards the first term on the right-hand side as a function on the second order tangent bundle $T^{(2)}Q$ by formally applying the chain rule and replacing dq/dt by $\dot{q}$ everywhere. The equations of motion of a CL system (L, F, W) with a choice of feedback control $u : TQ \to W$ are written as

$$\mathcal{EL}(L)(q, \dot{q}, \ddot{q}) = F(q, \dot{q}) + u(q, \dot{q}), \tag{2.3}$$

which may be derived from the Lagrange–d'Alembert principle.

CL systems with symmetry. We next define the notion of a G-invariant CL system on TQ and the associated notion of a reduced CL system on TQ/G, where G is a Lie group acting on Q.

Definition 2.2. *Let G be a Lie group action on Q. A* G-invariant controlled Lagrangian (G-CL) system *is a CL system, (L, F, W), where L is a G-invariant Lagrangian, F is a G-equivariant force map, and W is a G-invariant subbundle of T^*Q.*

Reduction of CL systems with symmetry. Now we can define the notion of a reduced CL system.

Definition 2.3. *A* reduced controlled Lagrangian (RCL) system *is a triple (l, f, U), where the function $l : TQ/G \to \mathbb{R}$ is called the* reduced Lagrangian*, the fiber-preserving map $f : TQ/G \to T^*Q/G$ is called the* reduced force map*, and the subbundle U of T^*Q/G is called the* reduced control subbundle*. A* feedback control *for an RCL system is a fiber-preserving map of TQ/G into U.*

Suppose that we are given a G-CL system (L, F, W). The G-invariance of L induces the reduced Lagrangian l on TQ/G satisfying

$$l \circ \tau_{/G} = L. \tag{2.4}$$

The equivariance of F induces the reduced force map $[F]_G : TQ/G \to T^*Q/G$ satisfying

$$[F]_G \circ \tau_{/G} = \pi_{/G} \circ F. \tag{2.5}$$

Similarly, a G-invariant control subbundle W induces a reduced control subbundle W/G in a natural way; namely, we have $W = \pi_{/G}^{-1}(W/G)$.

These considerations lead to the following definition.

Definition 2.4. *The RCL system of a G-CL system (L, F, W) is the triple $(l, [F]_G, W/G)$, where l is the reduced Lagrangian satisfying* (2.4) *and $[F]_G$ is the reduced force satisfying* (2.5)*.*

One may ask whether there exists a G-CL system on TQ when one is given a RCL system on TQ/G. The following proposition proves its existence and uniqueness.

Proposition 2.5. *Given a RCL system (l, f, U), there is a unique G-CL system (L, F, W) whose RCL system is (l, f, U).*

Proof. Define L by (2.4). Define a force map F on TQ as follows: for $v_q, w_q \in T_qQ$,

$$\langle F(v_q),\, w_q\rangle = \langle f \circ \tau_{/G}(v_q),\, \tau_{/G}(w_q)\rangle. \tag{2.6}$$

One can check the G-equivariance of F. One can also check that relation (2.6) defines the unique *fiber-preserving* map F of TQ to T^*Q. Let $W := \pi_{/G}^{-1}(U)$. By construction, (L, F, W) is the unique G-CL system whose RCL system is (l, f, U). □

By Proposition 2.5, we can write an arbitrary RCL in the form of the RCL of a G-CL without loss of generality. Additionally, the proof of Proposition 2.5 establishes the following assertion: given a fiber-preserving map $f : TQ/G \to T^*Q/G$, there exists the unique fiber-preserving map $F : TQ \to T^*Q$ satisfying

$$f \circ \tau_{/G} = \pi_{/G} \circ F.$$

Reduced equations of motion. Given a G-CL system (L, F, W), the G-invariance of L implies the G-equivariance of the bundle map $\mathcal{EL}(L) : T^{(2)}Q \to T^*Q$, which induces a quotient map

$$\mathcal{REL}(l) := [\mathcal{EL}(L)]_G : T^{(2)}Q/G \to T^*Q/G,$$

which depends only on the reduced Lagrangian l on TQ/G induced from L. The operator $\mathcal{REL}$ is called the reduced Euler–Lagrange operator. The equations of motion of an RCL $(l, [F]_G, W/G)$ with a choice of control $[u]_G : TQ/G \to W/G$ are given by

$$\mathcal{REL}(l)([q, \dot q, \ddot q]_G) = [F]_G([q, \dot q]_G) + [u]_G([q, \dot q]_G).$$

To write computable equations of $\mathcal{REL}$, one normally chooses a principal connection on the principal bundle $Q \to Q/G$ to identify the quotient bundles,

$$TQ/G \text{ with } T(Q/G) \oplus \tilde{\mathfrak{g}},$$
$$T^{(2)}Q/G \text{ with } T^{(2)}(Q/G) \times_{Q/G} 2\tilde{\mathfrak{g}},$$

and

$$T^*Q/G \text{ with } T^*(Q/G) \oplus \tilde{\mathfrak{g}}^*,$$

where $\tilde{\mathfrak{g}} := Q \times_G \mathfrak{g}$ is the associated adjoint bundle, $\tilde{\mathfrak{g}}^* := Q \times_G \mathfrak{g}^*$ is the associated coadjoint bundle, $T^{(2)}(Q/G) \times_{Q/G} 2\tilde{\mathfrak{g}}$ is the product bundle over Q/G, $2\tilde{\mathfrak{g}} := \tilde{\mathfrak{g}} \oplus \tilde{\mathfrak{g}}$, and $\oplus$ is the Whitney sum (see Lemmas 2.4.2 and 3.2.2 in [17]). For example, a principal connection $A : TQ \to \mathfrak{g}$ on the principal bundle $\pi : Q \to Q/G$ induces the bundle isomorphism $\alpha_A : TQ/G \to T(Q/G) \oplus \tilde{\mathfrak{g}}$ as follows:

$$\alpha_A([q, \dot q]_G) = T\pi(q, \dot q) \oplus [q, A(q, \dot q)]_G. \tag{2.7}$$

With these identifications, $\mathcal{REL}$ induces the Lagrange–Poincaré operator $\mathcal{LP}$ as follows: for a reduced Lagrangian l,

$$\mathcal{LP}(l) : T^{(2)}(Q/G) \times_{Q/G} 2\tilde{\mathfrak{g}} \to T^*(Q/G) \oplus \tilde{\mathfrak{g}}^*. \tag{2.8}$$

Hence, the reduced Euler–Lagrange operator $\mathcal{REL}$ may be replaced by the Lagrange–Poincaré operator $\mathcal{LP}$ in this paper as far as one chooses a connection on $Q \to Q/G$. Further details may be found in [17].

Let us give the local coordinate expression of the Lagrange–Poincaré operator $\mathcal{LP}$ induced from a connection A on the principal bundle $Q \to Q/G$. This will be used in section 5. We choose a local trivialization of the bundle $Q \to Q/G$ to be $X \times G \to X$, where X is an open subset of $\mathbb{R}^r$ with $r = \dim(Q/G)$. Then, at any tangent vector $(x, g, \dot{x}, \dot{g}) \in T_{(x,g)}(X \times G)$, we have

$$A(x, g, \dot{x}, \dot{g}) = \mathrm{Ad}_g(A_e(x) \cdot \dot{x} + \xi),$$

where A_e is the $\mathfrak{g}$-valued 1-form on X defined by $A_e(x)\cdot\dot{x} = A(x, e, \dot{x}, 0)$ and $\xi = g^{-1}\dot{g}$. The bundle isomorphism α_A in this case becomes

$$\alpha_A([x, g, \dot{x}, \dot{g}]_G) = (x, \dot{x}) \oplus (x, \Omega),$$

where $\Omega = A_e(x) \cdot \dot{x} + \xi$ and we choose a local trivialization of the associated bundle $\tilde{\mathfrak{g}}$ to be $X \times \mathfrak{g} \to X$. The Lagrange–Poincaré operator $\mathcal{LP}$ for a reduced Lagrangian l on $TQ/G = T(Q/G) \oplus \tilde{\mathfrak{g}}$ gives

$$\mathcal{LP}(l) = \begin{pmatrix} \dfrac{d}{dt}\dfrac{\partial l}{\partial \dot{x}^\alpha} - \dfrac{\partial l}{\partial x^\alpha} + \dfrac{\partial l}{\partial \Omega^a}(B^a_{\beta\alpha}\dot{x}^\beta + C^a_{db}\Omega^d A^b_\alpha) \\ \dfrac{d}{dt}\dfrac{\partial l}{\partial \Omega^b} - \dfrac{\partial l}{\partial \Omega^a}(C^a_{db}\Omega^d - C^a_{db}A^d_\alpha \dot{x}^\alpha) \end{pmatrix}, \tag{2.9}$$

where $B^a_{\beta\alpha}$ is the curvature of the connection $A_e = (A^a_\alpha)$, C^a_{bd} are the structure constants of the Lie algebra $\mathfrak{g}$, and $\Omega = (\Omega^a)$ and $x = (x^\alpha)$ with $a = 1, \ldots, \dim \mathfrak{g}$ and $\alpha = 1, \ldots, \dim(Q/G)$. More details of the derivation of (2.9) may be found in sections 3.3 and 4.2 of [17]. In particular, if one chooses a local trivial connection, i.e., $A_e = 0$, then the Lagrange–Poincaré equation in (2.9) is given by

$$\mathcal{LP}(l) = \begin{pmatrix} \dfrac{d}{dt}\dfrac{\partial l}{\partial \dot{x}^\alpha} - \dfrac{\partial l}{\partial x^\alpha} \\ \dfrac{d}{dt}\dfrac{\partial l}{\partial \xi^b} - C^a_{db}\xi^d \dfrac{\partial l}{\partial \xi^a} \end{pmatrix}. \tag{2.10}$$

We briefly mention the relation between trajectories of G-CL systems and trajectories of RCL systems. Let (L, F, W) be a G-CL system and $(l, [F]_G, W/G)$ its RCL system. Choose an arbitrary *G-equivariant* feedback control law $u : TQ \to W$ for (L, F, W). The control u induces a reduced map $[u]_G : TQ/G \to W/G$. Then, if $(q(t), \dot{q}(t)) \in TQ$ is a trajectory of the closed-loop system (L, F, u), then $\tau_{/G}(q(t), \dot{q}(t)) \in TQ/G$ is the trajectory of the closed-loop system $(l, [F]_G, [u]_G)$.

Simple CL systems. We define simple CL systems, which include most mechanical systems in engineering applications.

Definition 2.6. *A CL system (L, F, W) on TQ is called* simple *if its Lagrangian $L : TQ \to \mathbb{R}$ is of the form kinetic minus potential energy as follows:*

$$L(q, \dot{q}) = \frac{1}{2} m_q(\dot{q}, \dot{q}) - V(q), \tag{2.11}$$

where m is a (generalized) mass tensor, i.e., a nondegenerate symmetric $(0, 2)$-tensor. A reduced CL system $(l, [F]_G, W/G)$ is called simple *if the reduced Lagrangian l is induced by a G-invariant simple Lagrangian L on TQ. The acronym (R)SCL will denote "(reduced) simple controlled Lagrangian."*

When a simple G-invariant Lagrangian L is given by (2.11), its RSCL $l : TQ/G \to \mathbb{R}$ is given by

$$l([q,\dot{q}]_G) = \frac{1}{2}[m]_G([q,\dot{q}]_G,[q,\dot{q}]_G) - [V]_G([q]_G),$$

where $[m]_G \in \Gamma(Q/G, T^*Q/G \otimes T^*Q/G)$ is the reduced mass tensor induced from the G-invariance of the mass tensor $m \in \Gamma(Q, T^*Q \otimes T^*Q)$ and $[V]_G : Q/G \to \mathbb{R}$ is the reduced potential energy.

SCL-equivalence. We now recall the fundamental definition of CL-equivalence from [20] as follows.

DEFINITION 2.7. *Two SCL systems* (L_1, F_1, W_1) *and* (L_2, F_2, W_2) *are said to be* CL-equivalent*, or simply,* $(L_1, F_1, W_1) \overset{L}{\sim} (L_2, F_2, W_2)$*, if the following* Euler–Lagrange matching conditions *hold:*

ELM-1: $W_1 = m_1 m_2^{-1}(W_2)$,

ELM-2: $\mathrm{Im}[\mathcal{EL}(L_1) - F_1 - m_1 m_2^{-1}(\mathcal{EL}(L_2) - F_2)] \subset W_1$,

where m_i *is the mass tensor of* L_i *and* Im *means the pointwise image of the map in brackets.*

The following proposition from [20] explains the significance of the CL-equivalence property. It shows that in a very natural sense the two control systems can be made to correspond by using an appropriate choice of control.

PROPOSITION 2.8. *Suppose that two SCL systems* (L_i, F_i, W_i)*,* $i = 1, 2$*, are CL-equivalent. Then, for an arbitrary control law for one system, there exists a control law for the other system such that the two closed-loop systems produce the same equations of motion. The explicit relation between the two control laws* u_i*,* $i = 1, 2$*, is given by*

$$u_1 = \mathcal{EL}(L_1) - F_1 - m_1 m_2^{-1}(\mathcal{EL}(L_2) - F_2) + m_1 m_2^{-1} u_2, \tag{2.12}$$

where m_i *is the mass tensor of* L_i*,* $i = 1, 2$*.*

Proof. Recall that the Euler–Lagrange operator is given by

$$\mathcal{EL}(L)(q,\dot{q},\ddot{q})_j = m_{ij}\ddot{q}^i + \frac{\partial m_{ij}}{\partial q^k}\dot{q}^i\dot{q}^k - \frac{1}{2}\frac{\partial m_{ik}}{\partial q^j}\dot{q}^i\dot{q}^k + \frac{\partial V}{\partial q^j},$$

where the Lagrangian L is given by

$$L = \frac{1}{2}m_{ij}\dot{q}^i\dot{q}^j - V(q).$$

One can solve (2.3) for $\ddot{q}$. Denote by $\ddot{q}_{L_i}$ the expression of the acceleration $\ddot{q}$ obtained from the closed-loop SCL system (L_i, F_i, u_i), $i = 1, 2$. Then,

$$m_1(\ddot{q}_{L_1} - \ddot{q}_{L_2}) = u_1 - m_1 m_2^{-1} u_2 - \left[(\mathcal{EL}(L_1) - F_1) - m_1 m_2^{-1}(\mathcal{EL}(L_2) - F_2)\right].$$

The conditions ELM-1 and ELM-2 imply that (2.12) holds if and only if $\ddot{q}_{L_1} = \ddot{q}_{L_2}$ if and only if they produce the same equations of motion. Notice that the term

$$\mathcal{EL}(L_1) - m_1 m_2^{-1}\mathcal{EL}(L_2)$$

in (2.12) can be regarded as a map defined on TQ because the acceleration $\ddot{q}$ cancels out. □

RSCL-equivalence. We now define an equivalence relation among RSCL systems on TQ/G.

DEFINITION 2.9. *Two RSCL systems* $(l_i, [F_i]_G, W_i/G)$, $i = 1, 2$, *are said to be* reduced-CL-equivalent (RCL-equivalent), *or simply*

$$(l_1, [F_1]_G, W_1/G) \overset{L}{\sim} (l_2, [F_2]_G, W_2/G),$$

if the following reduced Euler–Lagrange matching conditions *hold:*

RELM-1: $W_1/G = [m_1]_G [m_2]_G^{-1}(W_2/G)$,

RELM-2: $\operatorname{Im}[\mathcal{REL}(l_1) - [F_1]_G - [m_1]_G [m_2]_G^{-1}(\mathcal{REL}(l_2) - [F_2]_G)] \subset W_1/G$,

where $[m_i]_G$ *is the reduced mass tensor of* l_i, $i = 1, 2$, *and* Im *means the pointwise image of the map in brackets.*

Equivalence commutes with reduction. The following theorem explains the relationship between the CL-equivalence relation among G-SCL systems and the RCL-equivalence relation among RSCL systems.

THEOREM 2.10. *Two G-SCL systems are CL-equivalent if and only if their associated RSCL systems are RCL-equivalent.*

Proof. Let (L, F, W) be a G-SCL system, and $(l, [F]_G, W/G)$ its associated RSCL system. Then, the theorem follows from the G-invariance of W and the following relations:

$$\mathcal{REL}(l) \circ \tau_{/G}^{(2)} = \pi_{/G} \circ \mathcal{EL}(L), \qquad [F]_G \circ \tau_{/G} = \pi_{/G} \circ F,$$

where $\tau_{/G}^{(2)} : T^{(2)}Q \to T^{(2)}Q/G$ is the G-quotient map. □

Hence, one can check the RCL-equivalence of two RSCL systems in two ways; one is to directly check it, and the other is to check the CL-equivalence of their associated *unreduced* G-SCL systems. In practice, it is more convenient to check it directly at the reduced level; we shall see an example of this in section 5.

The following theorem explains the property of the RCL-equivalence relation.

THEOREM 2.11. *Suppose that two RSCL systems* $(l_i, [F_i]_G, W_i/G)$, $i = 1, 2$, *are RCL-equivalent. Then, for an arbitrary control law for one system, there exists a control law for the other system such that the two closed-loop systems produce the same equations of motion. The explicit relation between the two control laws* $[u_i]_G$, $i = 1, 2$, *is given by*

$$\begin{aligned}[u_1]_G = {} & \mathcal{REL}(l_1) - [F_1]_G - [m_1]_G [m_2]_G^{-1}(\mathcal{REL}(l_2) - [F_2]_G) \\ & + [m_1]_G [m_2]_G^{-1} [u_2]_G, \end{aligned} \tag{2.13}$$

where $[m_i]_G$ *is the reduced mass tensor of* l_i, $i = 1, 2$.

Proof. Let $[u_i]_G$ be a control for $(l_i, [F_i]_G, W_i/G)$, $i = 1, 2$. Let (L_i, F_i, W_i) be the unreduced G-CL system of $(l_i, [F_i]_G, W_i/G)$, $i = 1, 2$. By Theorem 2.10, the two G-SCL systems are CL-equivalent. By Proposition 2.8, the two closed-loop G-SCL systems (L_i, F_i, u_i), $i = 1, 2$, produce the same equations of motion when u_1 and u_2 satisfy (2.12). Hence, the two closed-loop RSCL systems $(l_i, [F_i]_G, [u_i]_G)$, $i = 1, 2$, produce the same equations of motion when $[u_1]_G$ and $[u_2]_G$ satisfy (2.13), because each term in (2.12) is G-equivariant. Also notice that, for any choice of $[u_i]_G$, one can choose the other $[u_j]_G$ such that (2.13) holds. □

One can prove Theorem 2.11 by comparing the expressions of "accelerations" of both equations using (2.9), as in the proof of Proposition 2.8. For this purpose, one needs to choose a connection on $Q \to Q/G$ because one has to split the variations

to write down the equations of motion in coordinates, because the Euler–Lagrange equations come from the variational principles (see (2.9) of this paper and Chapter 3 of [17] for more detail). In the current proof of Theorem 2.11, we were able to bypass this route by Theorem 2.10.

3. Reduction of CH systems with symmetry. There is a Hamiltonian counterpart to CL systems called *CH systems*. We study the reduction of CH systems with symmetry in this section.

CH systems. We first recall the definition of CH systems from [20].

DEFINITION 3.1. *A* CH system *is a quadruple (H, B, F, W), where the function $H : T^*Q \to \mathbb{R}$ is called the Hamiltonian, $B \in \Gamma(\wedge^2 TT^*Q)$ is called an* almost Poisson tensor *(the main point being that it need not satisfy the Jacobi identity) on T^*Q, the fiber-preserving map $F : T^*Q \to T^*Q$ is called the* (external) force map*, and the subbundle W of T^*Q is called the* control subbundle*. When a feedback control law $u : T^*Q \to W$ is chosen, the quadruple (H, B, F, u) denotes the closed-loop CH system.*

We remark that we choose to use almost Poisson tensors rather than almost symplectic forms, as it is more general and, moreover, is convenient in performing Poisson reduction. The vector field $X_{(H,B,F,u)}$ of a CH system (H, B, F, W) with a control law u is given by

$$X_{(H,B,F,u)} = B^\sharp \mathbf{d}H + \mathrm{vlift}(F) + \mathrm{vlift}(u),$$

where $\mathrm{vlift}(F)$ and $\mathrm{vlift}(u)$ are the vertical lifts defined in (2.1).

Reduction of G-invariant CH systems. We define G-invariant CH systems on T^*Q and reduced CH systems on T^*Q/Q as follows.

DEFINITION 3.2. *Let G be a Lie group action on Q. A* G-invariant CH (G-CH) system *is a CH system (H, B, F, W), where H, B, F, and W are all G-invariant.*

DEFINITION 3.3. *A* reduced CH (RCH) system *is a quadruple (h, b, f, U), where the function $h : T^*Q/G \to \mathbb{R}$ is called the* reduced Hamiltonian*, $b \in \Gamma(\wedge^2 T(T^*Q/G))$ is called the* reduced almost Poisson tensor*, the fiber-preserving map $f : T^*Q/G \to T^*Q/G$ is called the* reduced force map*, and the subbundle U of T^*Q/G is called the* reduced control subbundle.

Suppose that we are given a G-CH system (H, B, F, W) on T^*Q. The G-invariant Hamiltonian $H : T^*Q \to \mathbb{R}$ induces the reduced Hamiltonian $h : T^*Q/G \to \mathbb{R}$ as follows:

$$H = h \circ \pi_{/G}. \tag{3.1}$$

The G-invariance of the almost Poisson tensor $B \in \Gamma(\wedge^2 TT^*Q)$ induces a reduced almost Poisson tensor $[B]_G \in \Gamma(\wedge^2 T(T^*Q/G))$ as follows: for $f_1, f_2 \in \mathcal{F}(T^*Q/G)$,

$$[B]_{G\,[q,p]_G}(\mathbf{d}f_1, \mathbf{d}f_2) = B_{(q,p)}(\pi^*_{/G}\mathbf{d}f_1, \pi^*_{/G}\mathbf{d}f_2). \tag{3.2}$$

This is well defined since

$$\begin{aligned} B_{g(q,p)}(\mathbf{d}(f_1 \circ \pi_{/G}), \mathbf{d}(f_2 \circ \pi_{/G})) &= B_{(q,p)}(g^*\mathbf{d}(f_1 \circ \pi_{/G}), g^*\mathbf{d}(f_2 \circ \pi_{/G})) \\ &= B_{(q,p)}(\mathbf{d}(f_1 \circ \pi_{/G} \circ g), \mathbf{d}(f_2 \circ \pi_{/G} \circ g)) \\ &= B_{(q,p)}(\mathbf{d}(f_1 \circ \pi_{/G}), \mathbf{d}(f_2 \circ \pi_{/G})) \end{aligned}$$

for any $g \in G$, where we used the G-invariance of B in the first equality. One can easily check that $[B]_G$ is skew-symmetric. The G-invariance of F induces the reduced

force $[F]_G : T^*Q/G \to T^*Q/G$ satisfying

$$[F]_G \circ \pi_{/G} = \pi_{/G} \circ F. \tag{3.3}$$

This discussion motivates the following definition.

DEFINITION 3.4. *The* RCH system of a G-CH system (H, B, F, W) *is a quadruple* $(h, [B]_G, [F]_G, W/G)$, *where* h *is the reduced Hamiltonian defined in* (3.1), $[B]_G$ *is the reduced almost Poisson tensor defined in* (3.2), *and* $[F]_G$ *is the reduced force defined in* (3.3).

Similarly to Proposition 2.5, the following proposition explains the relations between G-CH systems and RCH systems.

PROPOSITION 3.5. *Given a RCH system* (h, b, f, U), *there is a (not necessarily unique)* G*-CH system* (H, B, F, W) *whose RCH system is* (h, b, f, U).

Proof. Define H by $H = h \circ \pi_{/G}$. Define a force map F on T^*Q as follows: for $\alpha_q \in T_q^*Q$, $v_q \in T_qQ$

$$\langle F(\alpha_q), v_q \rangle = \langle f \circ \pi_{/G}(\alpha_q), \tau_{/G}(v_q) \rangle.$$

Choose a connection on the principal bundle $T^*Q \to T^*Q/G$. (See Chapter 2, Theorem 2.1 in [25] for the proof of the existence.) Then we can split TT^*Q into the vertical space V and the horizontal space H as $TT^*Q = \mathrm{V} \oplus \mathrm{H}$. This induces the decomposition of T^*T^*Q as $T^*T^*Q = \mathrm{H}^\circ \oplus \mathrm{V}^\circ$, where H° and V° are the annihilators of H and V, respectively. Let $\mathrm{hor} : T(T^*Q/G) \to \mathrm{H}$ be the horizontal lift. Then its dual map $\mathrm{hor}^* : \mathrm{V}^\circ \to T^*(T^*Q/G)$ is an isomorphism. For simplicity, we use H° (resp., V°) as the projection of T^*T^*Q onto H° (resp., V°). Define an almost Poisson tensor B on T^*Q as follows: for $\alpha, \beta \in T_p^*T^*Q$

$$B(\alpha, \beta) := b(\mathrm{hor}^* \, \mathrm{V}^\circ \, \alpha, \, \mathrm{hor}^* \, \mathrm{V}^\circ \, \beta).$$

One can check that this almost Poisson tensor is G-invariant. We now show that $\pi_{/G} : T^*Q \to T^*Q/G$ is the Poisson map; i.e., $b = [B]_G$. Let h_1, h_2 be two functions on T^*Q/G. Then $\mathbf{d}(h_i \circ \pi_{/G}) \in V^\circ$, $i = 1, 2$. Thus, $\mathrm{hor}^* \, \mathbf{d}(h_i \circ \pi_{/G}) = \mathbf{d}h_i$, $i = 1, 2$. Hence,

$$\begin{aligned} [B]_G(\mathbf{d}h_1, \mathbf{d}h_2) &= B(\mathbf{d}(h_1 \circ \pi_{/G}), \mathbf{d}(h_2 \circ \pi_{/G})) \\ &= b(\mathbf{d}h_1, \mathbf{d}h_2). \end{aligned}$$

It follows that $[B]_G = b$. Let $W = \pi_{/G}^{-1}(U)$. Then one can see that (H, B, F, W) is a G-CH system and that its RCH system coincides with (h, b, f, U). This completes the proof. □

Notice in Proposition 3.5 that there can be more than one B satisfying $[B]_G = b$, which is the source of the nonuniqueness of the G-CH systems (H, B, F, W) in Proposition 3.5. By Proposition 3.5, we can write an arbitrary RCH system in the form of the RCH system of a G-CH system without loss of generality.

Reduced CH dynamics. Given a G-CH system (H, B, F, W), let $(h, [B]_G, [F]_G, W/G)$ be its RCH system. The (reduced) Hamiltonian vector field $X_{(h,[B]_G,[F]_G,[u]_G)}$ of $(h, [B]_G, [F]_G, W/G)$ with a control $[u]_G \in W/G$ is given by

$$X_{(h,[B]_G,[F]_G,[u]_G)} = [B]_G^\sharp \mathbf{d}h + \mathrm{vlift}([F]_G) + \mathrm{vlift}([u]_G), \tag{3.4}$$

where $\mathrm{vlift}([F]_G)$ and $\mathrm{vlift}([u]_G)$ are the vertical lifts defined in (2.1). Let $X_{(H,B,F,u)}$ be the vector field of (H, B, F, W) with control $u \in W$. Then we have

$$X_{(h,[B]_G,[F]_G,[u]_G)} \circ \pi_{/G} = T\pi_{/G} \cdot X_{(H,B,F,u)}. \tag{3.5}$$

The CH-equivalence relation. We shall first recall the CH-equivalence relation among CH systems on T^*Q from [20].

DEFINITION 3.6. *Two CH systems* (H_i, B_i, F_i, W_i), $i = 1, 2$, *are said to be* CH-equivalent*, or simply,* $(H_1, B_1, F_1, W_1) \overset{H}{\sim} (H_2, B_2, F_2, W_2)$, *if the following* Hamiltonian matching conditions *hold:*

HM-1: $W_1 = W_2$,

HM-2: $\mathrm{Im}[B_1^\sharp \mathbf{d}H_1 + \mathrm{vlift}(F_1) - B_2^\sharp \mathbf{d}H_2 - \mathrm{vlift}(F_2)] \subset \mathrm{vlift}(W_1)$,

where $\mathrm{vlift}(W_1)$ *is the vertical lift of* W_1 *defined in* (2.2).

The following proposition explains the significance of the CH-equivalence relation.

PROPOSITION 3.7. *Suppose that two CH systems* (H_i, B_i, F_i, W_i), $i = 1, 2$, *are CH-equivalent. Then, for an arbitrary control law for one system, there exists a control law for the other system such that the two closed-loop systems produce the same equations of motion. The explicit relation between the two control laws* u_i, $i = 1, 2$, *is given by*

$$\mathrm{vlift}(u_1) = -B_1^\sharp \mathbf{d}H_1 - \mathrm{vlift}(F_1) + B_2^\sharp \mathbf{d}H_2 + \mathrm{vlift}(F_2) + \mathrm{vlift}(u_2).$$

Proof. Just compare $X_{(H_1,B_1,F_1,u_1)}$ and $X_{(H_2,B_2,F_2,u_2)}$ with controls $u_1 : T^*Q \to W_1$ and $u_2 : T^*Q \to W_2$. □

RCH-equivalence. We now introduce an equivalence relation among RCH systems on T^*Q/G as follows.

DEFINITION 3.8. *Two RCH systems,* $(h_i, [B_i]_G, [F_i]_G, W_i/G)$, $i = 1, 2$, *are said to be* reduced-CH-equivalent (RCH-equivalent)*, or simply*

$$(h_1, [B_1]_G, [F_1]_G, W_1/G) \overset{H}{\sim} (h_2, [B_2]_G, [F_2]_G, W_2/G),$$

if the following reduced Hamiltonian matching conditions *hold:*

RHM-1: $W_1/G = W_2/G$,

RHM-2: $\mathrm{Im}[[B_1]_G^\sharp \mathbf{d}h_1 + \mathrm{vlift}([F_1]_G) - [B_2]_G^\sharp \mathbf{d}h_2 - \mathrm{vlift}([F_2]_G)] \subset \mathrm{vlift}(W_1/G)$,

where $\mathrm{vlift}(W_1/G)$ *is the vertical lift of the subbundle* W_1/G *defined in* (2.2).

Reduction commutes with equivalence. The following theorem explains the relation between the RCH-equivalence relation among RCH systems on T^*Q/G and the CH-equivalence relation among G-CH systems on T^*Q.

THEOREM 3.9. *Two G-CH systems are CH-equivalent if and only if their associated RCH systems are RCH-equivalent.*

Proof. Use Definitions 3.6 and 3.8 as well as the relation (3.5). □

THEOREM 3.10. *Suppose that two RCH systems* $(h_i, [B_i]_G, [F_i]_G, W_i/G)$, $i = 1, 2$, *are RCH-equivalent. Then, for an arbitrary control law for one system, there exists a control law for the other system such that the two closed-loop RCH systems produce the same equations of motion. The explicit relation between the two control laws* $[u_i]_G$, $i = 1, 2$, *is given by*

$$\mathrm{vlift}([u_1]_G) = -[B_1]_G^\sharp \mathbf{d}h_1 - \mathrm{vlift}([F_1]_G) + [B_2]_G^\sharp \mathbf{d}h_2 + \mathrm{vlift}([F_2]_G) + \mathrm{vlift}([u_2]_G).$$

Proof. The proof follows from a straightforward computation using (3.4). □

Simple CH systems. Let us review the definition of simple CH systems from [20].

DEFINITION 3.11. *A CH system* (H, B, F, W) *on* T^*Q *is called* simple *if the Hamiltonian* H *has the form kinetic plus potential energy,*

$$H(q, p) = \frac{1}{2}\langle p,\, m_q^{-1} p\rangle + V(q),$$

and its almost Poisson tensor B is nondegenerate and is of the form

$$B(q,p) = \begin{bmatrix} 0 & K(q)^T \\ -K(q) & J(q,p) \end{bmatrix} \tag{3.6}$$

in cotangent coordinates (q,p) on T^*Q, where $K(q), J(q,p)$ are $n \times n$ matrices with $n = \dim Q$. We call H the simple Hamiltonian and B the simple almost Poisson tensor. The acronym, SCH, denotes "simple controlled Hamiltonian."

One can check that the statement that B has the form (3.6) is independent of the choice of cotangent bundle coordinates on T^*Q.

Let B be a simple almost Poisson tensor. Then the relation

$$\text{vlift}(\psi_B) = B \circ \Theta \tag{3.7}$$

defines a unique $\psi_B \in \Gamma(\text{Aut}(T^*Q))$, where Θ is the canonical 1-form[1] on T^*Q and B is regarded as a bundle map $B : T^*T^*Q \to TT^*Q$. In other words, there exists a unique $\psi_B \in \Gamma(\text{Aut}(T^*Q))$ such that the following diagram commutes:

$$\begin{array}{ccc} T^*T^*Q & \xrightarrow{B} & TT^*Q \\ \uparrow{\scriptstyle \Theta} & & \uparrow{\scriptstyle \text{vlift}} \\ T^*Q & \xrightarrow{\psi_B} & T^*Q \end{array}$$

See [20] for the proof of the existence of ψ_B. When B is given in coordinates as in (3.6), the map ψ_B is given in coordinates by

$$\psi_B(q,p) = (q, K(q)p).$$

Now, we make the following definition of the reduced simple CH system.

DEFINITION 3.12. *An RCH system $(h, [B]_G, [F]_G, W/G)$ is called a reduced simple CH system (or* RSCH *system) if it is the reduced CH system of a G-invariant simple CH system.*

Simple almost Poisson tensors and their reductions. Recall that in Definition 3.11 we defined simple almost Poisson tensors on T^*Q using local coordinates. Here, we characterize the reduced simple almost Poisson tensors using local coordinates, so we may assume that $Q = G \times X$, where G is a Lie group acting on the manifold X trivially. Recall the following identifications by left translation of G:

$$T^*G = G \times \mathfrak{g}^*, \quad TT^*G = (G \times \mathfrak{g}^*) \times (\mathfrak{g} \times \mathfrak{g}^*), \quad T^*T^*G = (G \times \mathfrak{g}^*) \times (\mathfrak{g}^* \times \mathfrak{g}).$$

We use (g_a, μ_a, x_i, p_i) as local coordinates for $T^*Q = G \times \mathfrak{g}^* \times T^*X$, and (μ_a, x_i, p_i) for $T^*Q/G = \mathfrak{g}^* \times T^*X$, where $a = 1, \ldots, \dim G$ and $i = 1, \ldots, \dim X$. We will use $\{e_a\}$ as a basis for $\mathfrak{g}$, and $\{e_a^*\}$ as its dual basis. Let $B \in \Gamma(\wedge^2 TT^*Q)$ be a G-invariant simple almost Poisson tensor. Then it is of the following form:

$$\begin{aligned} B(g,\mu,x,p) = {} & A_{ab}(x)(e_a \otimes e_b^* - e_b^* \otimes e_a) + E_{ia}(x)(\partial_{x_i} \otimes e_b^* - e_b^* \otimes \partial_{x_i}) \\ & + C_{ai}(x)(e_a \otimes \partial_{p_i} - \partial_{p_i} \otimes e_a) + D_{ij}(x)(\partial_{x_i} \otimes \partial_{p_j} - \partial_{p_j} \otimes \partial_{x_i}) \\ & + R_{ab}(\mu,x,p) e_a^* \otimes e_b^* + S_{ai}(\mu,x,p)(e_a^* \otimes \partial_{p_i} - \partial_{p_i} \otimes e_a^*) \\ & + U_{ij}(\mu,x,p)\partial_{p_i} \otimes \partial_{p_j}. \end{aligned} \tag{3.8}$$

[1] The canonical 1-form Θ on T^*Q is given by $\Theta = p_i \mathbf{d}q^i$ in cotangent bundle coordinates $(q^i, p_j) \in T^*Q$.

In the matrix form, B is given by

$$B = \begin{bmatrix} O & O & A(x) & C(x) \\ O & O & E(x) & D(x) \\ -A(x)^T & -E(x)^T & R(\mu,x,p) & S(\mu,x,p) \\ -C(x)^T & -D(x)^T & -S(\mu,x,p)^T & U(\mu,x,p) \end{bmatrix},$$

where we used the basis for T_zT^*Q in the following order: $e_a,\ \partial_{x_i},\ e_a^*,\ \partial_{p_i}$. The nondegeneracy condition for B is given by

$$\operatorname{rank}\begin{bmatrix} A & C \\ E & D \end{bmatrix} = \dim Q.$$

The reduced simple Poisson tensor $[B]_G$ is given by

$$\begin{aligned}[B]_G(\mu,x,p) &= E_{ia}(x)(\partial_{x_i}\otimes e_b^* - e_b^*\otimes\partial_{x_i}) + D_{ij}(x)(\partial_{x_i}\otimes\partial_{p_j} - \partial_{p_j}\otimes\partial_{x_i}) \\ &\quad + R_{ab}(\mu,x,p)e_a^*\otimes e_b^* + S_{ai}(\mu,x,p)(e_a^*\otimes\partial_{p_i} - \partial_{p_i}\otimes e_a^*) \\ &\quad + U_{ij}(\mu,x,p)\partial_{p_i}\otimes\partial_{p_j}. \end{aligned} \tag{3.9}$$

In a matrix form,

$$[B]_G = \begin{bmatrix} R(\mu,x,p) & -E(x)^T & S(\mu,x,p) \\ E(x) & O & D(x) \\ -S(\mu,x,p)^T & -D(x)^T & U(\mu,x,p) \end{bmatrix},$$

where we used the basis for $T_{[z]}(T^*Q/G)$ in the following order: $e_a^*,\ \partial_{x_i},\ \partial_{p_i}$. The nondegeneracy condition for B induces the following rank condition for $[B]_G$:

$$\operatorname{rank}[E\ \ D] = \dim X. \tag{3.10}$$

Remark. One could argue that it might be better if we could characterize all the tensors $b \in \Gamma(\wedge^2 T(T^*Q/G))$ for which there exists a G-invariant simple almost Poisson tensor B such that $b = [B]_G$. Then we could define RSCH systems without reference to G-invariant SCH systems. This point has to be studied more, and we think that the use of connections is crucial; see [31] and [17].

4. Equivalence between reduced simple Lagrangian and Hamiltonian systems. In this section we show the equivalence between the method of RSCL systems on TQ/G and that of RSCH systems on T^*Q/G.

The CL-CH equivalence theorem. In [19] and [20], the equivalence between the method of SCL systems on TQ and that of SCH systems on T^*Q was shown. It is summarized in the following theorem.[2]

THEOREM 4.1. *The method of CL systems is equivalent to that of CH systems for simple mechanical systems in the following sense:*

1. *For any two simple CL systems (L_i, F_i^L, W_i^L), $i = 1, 2$, there exist two associated simple CH systems*[3] *(H_i, B_i, F_i^H, W_i^H), $i = 1, 2$, such that*

$$\begin{aligned}&(L_1, F_1^L, W_1^L) \overset{L}{\sim} (L_2, F_2^L, W_2^L) \\ &\quad\Longleftrightarrow\ (H_1, B_1, F_1^H, W_1^H) \overset{H}{\sim} (H_2, B_2, F_2^H, W_2^H).\end{aligned} \tag{4.1}$$

[2]The statement of Theorem 4.1 is slightly different from that of Corollary 4.1 in [20] in that there appears an additional term, $\psi_{B_2} \circ \psi_{B_1}^{-1} = m_{H_2}(m_{H_1})^{-1}$, in statement 2. However, one can still prove Theorem 4.1 of this paper using section 4 of [20]. For example, see [19] for the proof.

[3]Refer to section 4 of [20] to learn how to find the associated CH systems.

2. *For any two simple CH systems* (H_i, B_i, F_i^H, W_i^H), $i = 1, 2$, *there exist two associated simple CL systems* (L_i, F_i^L, W_i^L), $i = 1, 2$, *such that*

$$(H_1, B_1, F_1^H, W_1^H) \overset{H}{\sim} (H_2, B_2, F_2^H, W_2^H)$$
$$\iff (L_1, F_1^L, W_1^L) \overset{L}{\sim} (L_2, F_2^L, W_2^L) \quad \textit{and} \quad \psi_{B_2} \circ \psi_{B_1}^{-1} = m_{H_2}(m_{H_1})^{-1},$$

with m_{H_i} *the mass tensor of* H_i *and* ψ_{B_i} *defined in* (3.7) *for* $i = 1, 2$.

In statement 1, *one does not need the condition* $\psi_{B_2} \circ \psi_{B_1}^{-1} = m_{H_2}(m_{H_1})^{-1}$ *along with the CL equivalence condition for the two CL systems, because it automatically holds for the associated CH systems.*

This result implies the following: suppose that we want to find all SCL systems which are equivalent to a given SCL system. One can directly search for them on the Lagrangian side using the CL equivalence relation. Alternatively, one can first find an SCH system which is associated with the given SCL system, secondly search for all SCH systems which are CH-equivalent to this SCH system, and finally transform those SCH systems to SCL systems. Those SCL systems are all SCL systems CL-equivalent to the original SCL system. In the similar way, one can find all the SCH systems which are CH-equivalent to a given SCH system, directly or with a CL-equivalence relation. (Refer to [20] and [19] for more detail.) Hence, one can describe a given simple mechanical system as an SCL system or as an SCH system and then apply the CL method or the CH method, correspondingly. Both procedures are equivalent.

We now restrict Theorem 4.1 to G-invariant systems.

THEOREM 4.2. *The method of* G*-invariant SCL systems and that of* G*-invariant SCH systems are equivalent. In other words, Theorem* 4.1 *restricted to* G*-invariant simple CL and CH systems holds.*

Proof. For the proof, one just needs to keep track of the G-invariance in the proof of Theorem 4.1 in [20] and the proof of Theorem 3.2.1 in [19]. □

The following theorem explains the equivalence between the method of RSCL systems on TQ/G and that of RSCH systems on T^*Q/G.

THEOREM 4.3. *The method of RSCL systems is equivalent to that of RSCH systems in the following sense:*

1. *For two given RSCL systems* $(l_i, [F_i^L]_G, W_i^L/G)$, $i = 1, 2$, *there exist two associated RSCH systems* $(h_i, [B_i]_G, [F_i^H]_G, W_i^H/G)$, $i = 1, 2$, *such that*

$$(l_1, [F_1^L]_G, W_1^L/G) \overset{L}{\sim} (l_2, [F_2^L]_G, W_2^L/G)$$
$$\iff (h_1, [B_1]_G, [F_1^H]_G, W_1^H/G) \overset{H}{\sim} (h_2, [B_2]_G, [F_2^H]_G, W_2^H/G). \tag{4.2}$$

2. *For two given RSCH systems* $(h_i, [B_i]_G, [F_i^H]_G, W_i^H/G)$, $i = 1, 2$, *there exist two associated RSCL systems* $(l_i, [F_i^L]_G, W_i^L/G)$, $i = 1, 2$, *such that*

$$(h_1, [B_1]_G, [F_1^H]_G, W_1^H/G) \overset{H}{\sim} (h_2, [B_2]_G, [F_2^H]_G, W_2^H/G)$$
$$\iff (l_1, [F_1^L]_G, W_1^L/G) \overset{L}{\sim} (l_2, [F_2^L]_G, W_2^L/G) \ \textit{and} \ [\psi_{B_2} \circ \psi_{B_1}^{-1}]_G = [m_{H_2}]_G [m_{H_1}]_G^{-1},$$

where $[m_{H_i}]_G$ *is the reduced mass tensor of* h_i.

Proof. Let us prove statement 1. For given two RSCL systems $(l_i, [F_i^L]_G, W_i^L/G)$, $i = 1, 2$, consider their unreduced G-SCL systems (L_i, F_i^L, W_i^L), $i = 1, 2$, with $L_i = l_i \circ \tau_{/G}$ (see Proposition 2.5). Theorem 4.2 implies that there are two G-SCH systems

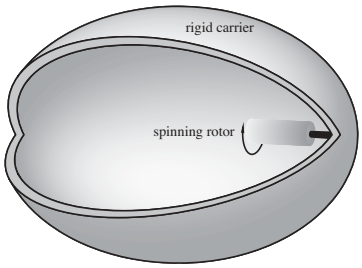


FIG. 5.1. *A satellite with a rotor along the third body axis.*

(H_i, B_i, F_i^H, W_i^H), $i = 1, 2$, such that (4.1) holds. Let $(h_i, [B_i]_G, [F_i^H]_G, W_i^H/G)$ be the RSCH system of (H_i, B_i, F_i^H, W_i^H) for $i = 1, 2$. Then (4.2) follows from Theorems 2.10 and 3.9 and (4.1). Now we prove statement 2. In this case, use Proposition 3.5 instead of Proposition 2.5, and then proceed in a similar manner. ∎

Hence, one can describe a given simple mechanical system as an SCL/SCH system, apply the CL/CH reduction, and then apply the reduced CL/CH method, correspondingly. Both procedures are equivalent.

Remark. Notice in (4.2) that $\psi_{B_2} \circ \psi_{B_1}^{-1}$ is G-equivariant even though each of ψ_{B_i} may not be. This equivariance is a consequence of the following commutative diagram:

$$\begin{array}{ccccc} TT^*Q & \xleftarrow{B_1} & T^*Q^*Q & \xrightarrow{B_2} & TT^*Q \\ \uparrow \text{vlift} & & \uparrow \Theta & & \uparrow \text{vlift} \\ T^*Q & \xleftarrow{\psi_{B_1}} & T^*Q & \xrightarrow{\psi_{B_2}} & T^*Q \end{array}$$

It follows that for $\alpha \in T^*Q$, $B_2 B_1^{-1}(\text{vlift}(\alpha)) = \text{vlift}(\psi_{B_2} \circ \psi_{B_1}^{-1}(\alpha))$. One can easily check that vlift is G-equivariant, i.e., $\text{vlift}(g\alpha) = g\,\text{vlift}(\alpha)$ for $g \in G$. Then the G-equivariance of $\psi_{B_2} \circ \psi_{B_1}^{-1}$ follows from the G-equivariance of B_1, B_2, and vlift and from the injectivity of vlift.

5. Examples. We review the example of a satellite with a rotor and the Euler–Poincaré matching conditions presented in [15] in the framework of the current paper. We then apply the RCL method to the stabilization of the heavy top with rotors. This will show the application of CL systems only. For an excellent application of CH systems to the problem of underwater vehicle stabilization by internal rotors, refer to [34] and [15].

5.1. Satellite with a rotor and Euler–Poincaré matching.

Satellite with a rotor (see [15]). We consider the example of a satellite with a rotor aligned along the third principal axis of the body (see Figure 5.1). The configuration space is $Q = G \times X = \text{SO}(3) \times S^1$, with the first factor being the satellite attitude and the second factor being the rotor angle. The Lie group $G = \text{SO}(3)$ acts on the first factor of Q only. We take a trivial (flat) connection on Q such that $TQ/G \simeq \mathfrak{g} \times TX$. Use $((\Omega_1, \Omega_2, \Omega_3), (\phi, \dot{\phi}))$ as coordinates for $\mathfrak{so}(3) \times TS^1 \simeq \mathbb{R}^3 \times TS^1$. This system is described by the RSCL system $(l_1, [F_1]_G = 0, W_1/G)$ given by

$$l_1(\Omega, \dot{\phi}) = \frac{1}{2}\big(\lambda_1 \Omega_1^2 + \lambda_2 \Omega_2^2 + (I_3 + J_3)\Omega_3^2 + 2J_3\Omega_3\dot{\phi} + J_3\dot{\phi}^2\big)$$

and $W_1/G = \text{span}\{\mathbf{d}\phi\}$, where $\lambda_1 > \lambda_2 > \lambda_3 := J_3 + I_3$. Notice that l_1 does not depend on ϕ. Recall that the reduced Euler–Lagrange operator $\mathcal{REL}$ induces the Lagrange–Poincaré operator $\mathcal{LP}$ in (2.8) with respect to the trivial connection on $Q \to X$. By (2.10), this Lagrange–Poincaré operator $\mathcal{LP}(l_1)$ is given by

$$\mathcal{LP}(l_1) = \begin{bmatrix} \dfrac{d}{dt}\dfrac{\partial l_1}{\partial \Omega} - \dfrac{\partial l_1}{\partial \Omega} \times \Omega \\ \dfrac{d}{dt}\dfrac{\partial l_1}{\partial \dot{\phi}} - \dfrac{\partial l_1}{\partial \phi} \end{bmatrix}, \tag{5.1}$$

where we switched the first and second components of (2.10).

Consider another RSCL system $(l_2, 0, W_2/G)$ with $W_2 = W_1$ and

$$l_2(\Omega, \dot{\phi}) = \frac{1}{2}\big(\lambda_1\Omega_1^2 + \lambda_2\Omega_2^2 + (I_3 + J_3)\Omega_3^3 + 2J_3\Omega_3\dot{\phi} + \rho J_3\dot{\phi}^2\big)$$

with $\rho \in \mathbb{R}$. Here we allow only one free parameter ρ as in [15], but one can consider a more general form of CL system. One can check that RELM-1 and RELM-2 in Definition 2.9 are satisfied. Since these two CL systems are equivalent, one has only to design a controller for the second system, which will give an asymptotically stabilizing controller for the first system by Theorem 2.11. See [8] for the discussion on asymptotic stabilization of the rotation about the middle axis in the body-fixed frame. There, it is shown how to choose ρ and the dissipative input for stability of the equivalent system $(l_2, 0, W_2/G)$. This leads to an asymptotically stabilizing controller for the original system $(l_1, 0, W_1/G)$.

Euler–Poincaré matching. Here we briefly sketch the proof that the set of Euler–Poincaré matching conditions in [12] and [15] is a special case of the reduced Euler–Lagrange matching conditions in this paper. This set of matching conditions can handle such examples as a satellite with a rotor and underwater vehicles with internal rotors. Let $Q = G \times X$ be the configuration space, where G is a Lie group acting trivially on the manifold X. We choose the trivial (flat) connection on $Q \to X$ to write down the Lagrange–Poincaré equation on $TQ/G \simeq \mathfrak{g} \times TX$ with the Lie algebra $\mathfrak{g}$ of the Lie group G. We use $\eta = (\eta^\alpha)$ as coordinates for $\mathfrak{g}$, and $(\theta, \dot{\theta}) = (\theta^a, \dot{\theta}^a)$ as coordinates for TX. By (2.10), the Lagrange–Poincaré operator $\mathcal{LP}$ with respect to the *trivial* connection is given by

$$\mathcal{LP}(l) = \begin{pmatrix} \dfrac{d}{dt}\dfrac{\partial l}{\partial \eta^\alpha} - C^\beta_{\gamma\alpha}\eta^\gamma \dfrac{\partial l}{\partial \eta^\beta} \\ \dfrac{d}{dt}\dfrac{\partial l}{\partial \dot{\theta}^a} - \dfrac{\partial l}{\partial \theta^a} \end{pmatrix} \tag{5.2}$$

for any reduced Lagrangian $l = l(\eta^\alpha, \dot{\theta}^a, \theta)$, where $C^\beta_{\gamma\alpha}$ are the structure constants of the Lie algebra $\mathfrak{g}$. In (5.2) we wrote the vertical part of $\mathcal{LP}(l)$ first, while in (2.10) the vertical part was written in the second component.

Let $(l, 0, T^*X)$ be the given RSCL system with the reduced Lagrangian

$$l(\eta^\alpha, \dot{\theta}^a) = \frac{1}{2}g_{\alpha\beta}\eta^\alpha\eta^\beta + g_{\alpha a}\eta^\alpha\dot{\theta}^a + \frac{1}{2}g_{ab}\dot{\theta}^a\dot{\theta}^b,$$

where $g_{\alpha\beta}, g_{\alpha a}, g_{ab}$ are constant functions on TQ/G. Notice that this Lagrangian is cyclic in the variables θ^a and that the controls act only on the cyclic variables. Let

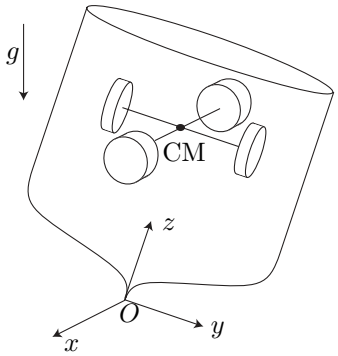


FIG. 5.2. *Heavy top with two rotors, each consisting of two rigidly coupled disks. The center of mass is at CM.*

$(l_{\tau,\sigma,\rho}, 0, T^*X)$ be another RSCL system with the reduced Lagrangian of the following form:

$$l_{\tau,\sigma,\rho} = l(\eta^\alpha, \dot\theta^a + \tau^a_\alpha \eta^\alpha) + \frac{1}{2}\sigma_{ab}\tau^a_\alpha \tau^b_\beta \eta^\alpha \eta^\beta + \frac{1}{2}(\rho_{ab} - g_{ab})(\dot\theta^a + g^{ac}g_{c\alpha}\eta^\alpha + \tau^a_\alpha \eta^\alpha)(\dot\theta^b + g^{bc}g_{c\beta}\eta^\beta + \tau^b_\beta \eta^\beta), \tag{5.3}$$

which is exactly the equation (11) in [15]. See also [14] for the motivation of this choice of the form in (5.3). [15] proposes the following so-called Euler–Poincaré matching conditions:

EP-1: $\tau^a_\alpha = -\sigma^{ab}g_{b\alpha}$,
EP-2: $\sigma^{ab} + \rho^{ab} = g^{ab}$.

Then one can show that the two assumptions of EP-1 and EP-2 imply the RCL-equivalence of the two RSCL systems $(l, 0, T^*X)$ and $(l_{\tau,\sigma,\rho}, 0, T^*X)$. Hence, one can equivalently work with the second system to design controllers. Refer to [15] for the method of constructing a Lyapunov function using the energy and Casimir functions.

5.2. Heavy top with rotors. It is well known that an upright spinning top is unstable if the angular velocity is small. The motion of a heavy top and the stability of the Lagrange top are well studied in [28] and [24]. We use the CL method to asymptotically stabilize the upright spinning motion of a heavy top with small vertical angular velocity, including zero velocity. See Figure 5.2 for the heavy top system. One can notice that the system dynamics are not SO(3)-invariant because the gravitational force breaks the SO(3) symmetry, and thus we cannot perform the usual reduction of the system by the SO(3) group. However, there is a way of doing the SO(3)-reduction of this system by considering this system as one depending on a parameter in $\mathbb{R}^3$.

In this section, we will first review the general theory of reduction of systems depending on a parameter and then apply this reduction theory to the design of a controller for the heavy top system. We will not develop the whole theory of *CL systems depending on a parameter* and the reduction theory for those systems with symmetry because it is a straightforward modification of the theory in section 2 of this paper. Moreover, the complete theory of reduction for (uncontrolled) systems depending on a parameter is in [17].

Systems depending on a parameter. We here review the reduction theory for systems depending on an advected parameter, as presented in [17]. Consider a Lagrangian

$$L : T(G \times X) \times V^* \to \mathbb{R},$$

where G is a Lie group, X is a manifold, and V^* is the dual space of the vector space V. The value of L at the point $(g, x, \dot{g}, \dot{x}, a_0) \in T(G \times X) \times V^*$ will be denoted by $L(g, x, \dot{g}, \dot{x}, a_0)$, as usual, and we will think of a_0 as a parameter that remains fixed along the evolution of the system. Assume that there is an action of G on V, so there is an induced action on V^* such that $\langle ga_0, gb_0 \rangle = \langle a_0, b_0 \rangle$ for all $a_0 \in V^*$, $b_0 \in V$, $g \in G$. Assume that G acts trivially on X and that L is G-invariant, i.e.,

$$L(hg, x, h\dot{g}, \dot{x}, ha_0) = L(g, x, \dot{g}, \dot{x}, a_0)$$

for $h \in G$, $(g, x, \dot{g}, \dot{x}, a_0) \in T(G \times X) \times V^*$. Define the reduced Lagrangian $l : \mathfrak{g} \times TX \times V^* \to \mathbb{R}$ by

$$l(\xi, x, \dot{x}, a) = L(e, x, \xi, \dot{x}, a) \tag{5.4}$$

for $\xi \in \mathfrak{g}$, $(x, \dot{x}) \in TX$, $a \in V^*$, where $\xi = g^{-1}\dot{g}$ and $a = g^{-1}a_0$.

Fix $a_0 \in V^*$. Let $L_{a_0} : T(G \times X) \to \mathbb{R}$ be the restriction of L to $T(G \times X) \times \{a_0\}$. Then, the Euler–Lagrange operator $\mathcal{EL}(L_{a_0})$ induces the reduced Euler–Lagrange operator

$$\mathcal{REL}(l) : 2\tilde{\mathfrak{g}} \oplus T^{(2)}X \times TV^* \to \mathfrak{g}^* \times T^*X$$

and the equation in (5.6) as follows:

$$\mathcal{REL}(l) = \begin{pmatrix} \dfrac{d}{dt}\dfrac{\partial l}{\partial \xi} - \mathrm{ad}^*_\xi \dfrac{\partial l}{\partial \xi} - \dfrac{\partial l}{\partial a} \diamond a \\ \dfrac{d}{dt}\dfrac{\partial l}{\partial \dot{x}} - \dfrac{\partial l}{\partial x} \end{pmatrix} \tag{5.5}$$

and

$$\dot{a} = -\xi a, \tag{5.6}$$

where the map $\diamond : V \times V^* \to \mathfrak{g}^*$ is defined by $\langle b \diamond a, \eta \rangle = -\langle \eta a, b \rangle$ for $\eta \in \mathfrak{g}$, $a \in V^*$, $b \in V$. One may find the derivation of (5.5) and (5.6) in section 7.4 of [17].

The equations of motion of the reduced Lagrangian l with a $(\mathfrak{g}^* \times T^*X)$-valued (reduced) force f, are given by

$$\mathcal{REL}(l) = f \quad \text{and} \quad \dot{a} = -\xi a, \tag{5.7}$$

where f includes external forces and control forces.

Heavy top with two pairs of rotors. We first describe a heavy top with two pairs of rotors. We mount two pairs of rotors within the top so that each pair's rotation axis is parallel to the first and the second principal axes of the top; see Figure 5.2. Let I_1, I_2, I_3 be the moments of inertia of the top in the body-fixed frame. Let J_1, J_2 be the moments of inertia of the rotors around their rotation axes. Let J_{i1}, J_{i2}, J_{i3} be the moments of inertia of the ith rotor with $i = 1, 2$ around the first, the second, and the third principal axis, respectively. Let $\bar{I}_1 = I_1 + J_{11} + J_{21}$, $\bar{I}_2 = I_2 + J_{12} + J_{22}$,

and $\bar{I}_3 = I_3 + J_{13} + J_{23}$. Let $\lambda_1 = \bar{I}_1 + J_1$ and $\lambda_2 = \bar{I}_2 + J_2$. Let M be the total mass of the system, g the magnitude of the gravitational acceleration, and h the distance from the origin O to the center of mass of the system.

Let $G = \mathrm{SO}(3)$, $X = S^1 \times S^1$, $V^* = \mathbb{R}^3$. We use the following notation for coordinates:

$$\mathbf{R} \in \mathrm{SO}(3), \quad \theta = (\theta_1, \theta_2) \in S^1 \times S^1, \quad a \in \mathbb{R}^3.$$

We will use $\Omega = (\Omega_1, \Omega_2, \Omega_3) \in \mathbb{R}^3$ as coordinates for the Lie algebra $\mathfrak{so}(3)$ under the Lie algebra isomorphism, $\vee : (\mathfrak{so}(3), [,]) \to (\mathbb{R}^3, \times)$,

$$\begin{pmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{pmatrix}^{\vee} = (x, y, z).$$

Let $L : T(G \times X) \times V^* \to \mathbb{R}$ be the Lagrangian defined by

$$L(\mathbf{R}, \theta, \dot{\mathbf{R}}, \dot{\theta}, a) = \frac{1}{2} \begin{pmatrix} \Omega_1 \\ \Omega_2 \\ \Omega_3 \\ \dot{\theta}_1 \\ \theta_2 \end{pmatrix}^T \begin{pmatrix} \lambda_1 & 0 & 0 & J_1 & 0 \\ 0 & \lambda_2 & 0 & 0 & J_2 \\ 0 & 0 & \bar{I}_3 & 0 & 0 \\ J_1 & 0 & 0 & J_1 & 0 \\ 0 & J_2 & 0 & 0 & J_2 \end{pmatrix} \begin{pmatrix} \Omega_1 \\ \Omega_2 \\ \Omega_3 \\ \dot{\theta}_1 \\ \theta_2 \end{pmatrix} - Mgh\mathbf{R}^{-1}a \cdot \chi,$$

where $\Omega = (\Omega_1, \Omega_2, \Omega_3) = (\mathbf{R}^{-1}\dot{\mathbf{R}})^{\vee}$ is the body-fixed angular momentum and χ is the body-fixed unit vector on the line segment connecting the origin O with the body's center of mass, i.e., $\chi = (0, 0, 1)$ in the body-fixed frame.

Fix $\mathbf{k} := (0, 0, 1) \in \mathbb{R}^3$. Let $L_{\mathbf{k}}$ be the restriction of L to $T(G \times X) \times \{\mathbf{k}\}$ as follows:

$$L_{\mathbf{k}}(\mathbf{R}, \theta, \dot{\mathbf{R}}, \dot{\theta}) = L(\mathbf{R}, \theta, \dot{\mathbf{R}}, \dot{\theta}, \mathbf{k}).$$

Then one can check that $L_{\mathbf{k}}$ is the Lagrangian of the heavy top system in Figure 5.2. The actuation is exerted on each pair of rotors, so the control bundle U is given by $U = T^*X$.

By (5.4), the reduced Lagrangian $l : \mathfrak{so}(3) \times TX \times \mathbb{R}^3 \to \mathbb{R}$ is given by

$$\begin{aligned} l(\Omega, \theta, \dot{\theta}, \Gamma) = {} & \frac{1}{2}(\lambda_1\Omega_1^2 + \lambda_2\Omega_2^2 + \bar{I}_3\Omega_3^2 + 2J_1\Omega_1\dot{\theta}_1 + 2J_2\Omega_2\dot{\theta}_2) \\ & + \frac{1}{2}(J_1\dot{\theta}_1^2 + J_2\dot{\theta}_2^2) - Mgh\Gamma \cdot \chi, \end{aligned}$$

where $\Gamma = (\Gamma_1, \Gamma_2, \Gamma_3) = \mathbf{R}^{-1}\mathbf{k}$. Physically, the vector Γ represents the motion of the *unit vector* with the *opposite direction* of gravity as seen from the body. Recall that the reduced equations of motion are derived from (5.7) and (5.5).

Let us consider a new reduced Lagrangian $\tilde{l}$ defined by

$$\begin{aligned} \tilde{l}(\Omega, \theta, \dot{\theta}, \Gamma) = {} & \frac{1}{2}(\lambda_1\Omega_1^2 + \lambda_2\Omega_2^2 + \bar{I}_3\Omega_3^2 + 2J_1\Omega_1\dot{\theta}_1 + 2J_2\Omega_2\dot{\theta}_2) \\ & + \frac{1}{2}(\rho_1 J_1\dot{\theta}_1^2 + \rho_2 J_2\dot{\theta}_2^2) - Mgh\Gamma \cdot \chi, \end{aligned} \tag{5.8}$$

where $\rho_1, \rho_2 \in \mathbb{R}$ are free parameters to be chosen later. See [15] and [21] for the motivation of this choice of the form in (5.8).

Even though we have not developed the general theory of CL systems depending on a parameter and the associated reduction theory, one can check that Definition 2.9 and Theorem 2.11 hold for reduced CL systems depending on a parameter where one should use the reduced Euler–Lagrange operator in (5.5) instead of (2.9) or (2.10). Notice that the equation in (5.6) is common for all reduced CL systems depending on a parameter. With these modifications in mind, one can check that

$$(l, 0, T^*X) \overset{L}{\sim} (\tilde{l}, 0, T^*X).$$

By Theorem 2.11, we can work with the RCL system $(\tilde{l}, 0, T^*X)$ to design a controller.

Let us define the angular momentum, $\Pi = (\Pi_1, \Pi_2, \Pi_2)$, and the control momentum, $\tilde{J} = (\tilde{J}_1, \tilde{J}_2)$, as follows:

$$\Pi = \frac{\partial \tilde{l}}{\partial \Omega} = (\lambda_1\Omega_1 + J_1\dot{\theta}_1,\ \lambda_2\Omega_2 + J_2\dot{\theta}_2,\ \bar{I}_3\Omega_3), \tag{5.9}$$

$$\tilde{J} = \frac{\partial \tilde{l}}{\partial \dot{\theta}} = (J_1\Omega_1 + J_1\rho_1\dot{\theta}_1, J_2\Omega_2 + J_2\rho_2\dot{\theta}_2). \tag{5.10}$$

By (5.7) and (5.5), the equations of motion of the system $(\tilde{l}, 0, T^*X)$ with a choice of control $v = (v_1, v_2)$ are given as follows:

$$\dot{\Pi} = \Pi \times \Omega + Mgh\Gamma \times \chi, \tag{5.11}$$

$$\dot{\tilde{J}} = v, \tag{5.12}$$

$$\dot{\Gamma} = \Gamma \times \Omega. \tag{5.13}$$

These dynamics have two constants of motion,

$$\Pi \cdot \Gamma \quad \text{and} \quad \|\Gamma\| = 1,$$

where $\Pi \cdot \Gamma$ is the vertical component of the space-fixed angular momentum, and $\|\Gamma\| = \|\mathbf{R}^{-1}\mathbf{k}\| = \|\mathbf{k}\| = 1$.

Since the reduced Lagrangian $\tilde{l}$ (or l) does not depend on the rotor angle θ, and we are not interested in the angle of rotors but in the angular velocity of rotors, we will remove X from the *phase space* for the sake of simplicity. Hence, we will regard $\mathfrak{so}(3) \times \mathbb{R}^2 \times \mathbb{R}^3$ as a new phase space, where $\mathbb{R}^2$ is the velocity component of $TX = \mathbb{R}^2 \times \mathbb{R}^2$.

Let $\Omega(0)$, $\dot{\theta}(0)$, and $\Gamma(0)$ with $||\Gamma(0)||^2 = 1$ be an initial condition with

$$\Omega_3^\circ := \Pi(0) \cdot \frac{\Gamma(0)}{\bar{I}_3} < \sqrt{\frac{Mgh}{\bar{I}_3}}. \tag{5.14}$$

We are interested in the equilibrium $e = (\Omega_e, \dot{\theta}_e, \Gamma_e)$,

$$\Omega_e = (0, 0, \Omega_3^\circ),\ \ \dot{\theta_e} = (0, 0),\ \ \Gamma_e = (0, 0, 1) \tag{5.15}$$

or

$$\Omega_e = (0, 0, \Omega_3^\circ),\ \ \tilde{J}_e = (0, 0),\ \ \Gamma_e = (0, 0, 1),$$

which corresponds to the upright spinning top with the rotors at rest. Notice that this equilibrium lies in the same level set of $(\Pi \cdot \Gamma, ||\Gamma||^2)$ as the initial condition.

We construct a Lyapunov function using the energy-Casimir method (see [8] for more detail of this method). Set

$$E_{\tilde{\Phi}} = K + U + \Phi(\Pi \cdot \Gamma, ||\Gamma||^2) + \Psi(\tilde{J}_1, \tilde{J}_2), \tag{5.16}$$

where the potential energy U is given by $U(\Gamma) = Mgh\Gamma \cdot \chi = Mgh\Gamma_3$, and the kinetic energy K is given by

$$K = \frac{1}{2}\left(\lambda_1 - \frac{J_1}{\rho_1}\right)\Omega_1^2 + \frac{1}{2}\left(\lambda_2 - \frac{J_2}{\rho_2}\right)\Omega_2^2 + \frac{1}{2}\bar{I}_3\Omega_3^2 + \frac{\tilde{J}_1^2}{2J_1\rho_1} + \frac{\tilde{J}_2^2}{2J_2\rho_2},$$

in the new coordinates $(\Omega, \tilde{J}, \Gamma)$. Choose the function Ψ as follows:

$$\Psi(\tilde{J}_1, \tilde{J}_2) = \frac{\tilde{J}_1^2}{2\epsilon_1 J_1} + \frac{\tilde{J}_2^2}{2\epsilon_2 J_2}, \tag{5.17}$$

where coefficients ϵ_i will be determined later. Choose the function Φ of the form

$$\begin{aligned}\Phi(x, y) = &-\Omega_3^\circ(x - \bar{I}_3\Omega_3^\circ) + \frac{1}{2}\left(\bar{I}_3(\Omega_3^\circ)^2 - Mgh\right)(y-1)\\ &+ \frac{1}{2}a_1(x - \bar{I}_3\Omega_3^\circ)^2 + \frac{1}{2}a_2(y-1)^2,\end{aligned}$$

where the constants a_1 and a_2 are chosen such that

$$a_1 < \frac{-1}{\bar{I}_3}$$

and

$$4a_2 + a_1(\bar{I}_3\Omega_3^\circ)^2 + \bar{I}_3(\Omega_3^\circ)^2 - Mgh < \frac{\bar{I}_3(a_1\bar{I}_3\Omega_3^\circ - \Omega_3^\circ)^2}{1 + a_1\bar{I}_3}.$$

One can check that the equilibrium e is a critical point of $E_{\tilde{\Phi}}$. We now find conditions under which this critical point is a local *maximum*. First, choose ρ_i satisfying

$$\frac{\bar{I}_3(\Omega_3^\circ)^2 - Mgh}{(\Omega_3^\circ)^2} < \lambda_i - \frac{J_i}{\rho_i} < 0 \tag{5.18}$$

for $i = 1, 2$, and then we can choose ϵ_1 and ϵ_2 such that the second derivative of $E_{\tilde{\Phi}}$ becomes negative definite at e, which implies that $E_{\tilde{\Phi}}$ has a local maximum at e. For later use, we impose an additional condition on ρ_i and ϵ_i as follows:

$$J_i(\Omega_3^\circ)^2 + (\epsilon_i + \rho_i)\left((\Omega_3^\circ)^2(\bar{I}_3 - \lambda_i) - Mgh\right) \neq 0. \tag{5.19}$$

With (5.19), it is still possible to find ρ_i and ϵ_i to ensure negative definiteness of the second derivative of $E_{\tilde{\Phi}}$ at e.

The following choice of $v = (v_1, v_2)$,

$$v_i = c_i\left(\dot{\theta}_i + \frac{\tilde{J}_i}{\epsilon_i J_i}\right), \tag{5.20}$$

with $c_i > 0$ for $i = 1, 2$, implies

$$\frac{d}{dt}E_{\tilde{\Phi}} = \sum_{i=1}^{2} c_i\left(\dot{\theta}_i + \frac{\tilde{J}_i}{\epsilon_i J_i}\right)^2 \geq 0, \tag{5.21}$$

which proves the Lyapunov stability of the equilibrium e in the closed-loop system. The complete control law u for the original system $(l, 0, T^*X)$ can be obtained from Theorem 2.11.

Asymptotic stabilization will now be shown by using LaSalle's theorem. Since $E_{\tilde{\Phi}}$ has a local maximum at e, it is nondecreasing in time, and $\Pi \cdot \Gamma$ and $||\Gamma||^2$ are conserved, there is a number c such that the set

$$S = \{x \in \mathfrak{so}(3) \times \mathbb{R}^2 \times \mathbb{R}^3 \mid E_{\tilde{\Phi}} \geq c,\ \Pi \cdot \Gamma = \Pi_e \cdot \Gamma_e,\ ||\Gamma||^2 = 1\}$$

is nonempty, compact, and positively invariant. Define

$$\mathcal{E} = \{x \in S \mid \dot{E}_{\tilde{\Phi}} = 0\} = \{x \in S \mid v = 0\}.$$

Let $\mathcal{M}$ be the largest invariant subset of $\mathcal{E}$. One can show $\mathcal{M} = \{e\}$ by (5.19) after shrinking the set S if necessary. Thus, by LaSalle's theorem, e is asymptotically stable.

Here are the main points in the proof that $\mathcal{M} = \{e\}$. Let $(\Omega(t), \dot{\theta}(t), \Gamma(t))$ be a trajectory in $\mathcal{M}$. The condition $v = 0$ and (5.12) imply that $\tilde{J}(t)$ is constant. Hence, $\theta_i(t)$ and $\Omega_i(t)$ are constant for $i = 1, 2$. By (5.9), $\Pi_i(t)$, $i = 1, 2$ are constant. Then the third component of (5.11) becomes $\lambda_3 \dot{\Omega}_3(t) =$ constant. By the Lyapunov stability of the equilibrium, it follows $\dot{\Omega}_3(t) \equiv 0$. Hence, $\Omega_3(t)$ is constant. The first and second component of (5.11) imply that $\Gamma_1(t)$ and $\Gamma_2(t)$ are constant. Then the third component of (5.13) implies that $\Gamma_3(t)$ is constant. So far we have shown that the trajectory $(\Omega(t), \dot{\theta}(t), \Gamma(t))$, or $(\Pi(t), \dot{\theta}(t), \Gamma(t))$, is constant for all $t \geq 0$. Consider the map $f : \mathbb{R}^8 \to \mathbb{R}^{10}$ defined by

$$f(\Omega, \dot{\theta}, \Gamma) = \begin{pmatrix} \Pi \times \Omega + Mgh\Gamma \times \chi \\ \Gamma \times \Omega \\ J_1\Omega_1 + (\epsilon_1 + J_1\rho_1)\dot{\theta}_1 \\ J_2\Omega_2 + (\epsilon_2 + J_2\rho_2)\dot{\theta}_2 \\ \Pi \cdot \Gamma - \Pi_e \cdot \Gamma_e \\ \|\Gamma\|^2 - 1 \end{pmatrix} = \begin{pmatrix} \dot{\Pi} \\ \dot{\Gamma} \\ \epsilon_1 v_1 \\ \epsilon_2 v_2 \\ \Pi \cdot \Gamma - \Pi_e \cdot \Gamma_e \\ \|\Gamma\|^2 - 1 \end{pmatrix},$$

where Π is expressed in terms of $(\Omega, \dot{\theta})$ as in (5.9). Then one can see that all the trajectories lying in $\mathcal{M}$ are contained in the set $f^{-1}(O)$. In particular, the equilibrium $(\Omega_e, \dot{\theta}_e, \Gamma_e)$ in (5.15) is also contained in $f^{-1}(O)$. One can check that the rank of the Jacobian matrix Df at the equilibrium is the full rank 8 by (5.19). Thus, f is locally one-to-one around the equilibrium by Theorem 4.12 in [16]. Therefore, the only possible trajectory totally lying in $\mathcal{M}$ is the equilibrium point itself. It follows from LaSalle's theorem that the equilibrium is asymptotically stable.

Remarks. 1. The above procedure shows that the choice of control gains depends on the initial condition. This is unavoidable because we need to know the value of the constant of motion $\Pi \cdot \Gamma$, which the internal actuation cannot change; however, our suggested controller is robust to small errors in the measurement of the initial condition. Let $\tilde{e}$ be the equilibrium of the form (5.15), with $\tilde{\Omega}_3^\circ$ instead of Ω_3°. Suppose the Ω_3° used in constructing the control law is very close to the value $\tilde{\Omega}_3^\circ$. Let $\tilde{E}_{\tilde{\Phi}}$ be the function of the form (5.16), with Ω_3° replaced by $\tilde{\Omega}_3^\circ$. Then $\tilde{e}$ is a critical point of $\tilde{E}_{\tilde{\Phi}}$. By continuity, the second derivative of $\tilde{E}_{\tilde{\Phi}}$ at $\tilde{e}$ will remain negative definite, proving Lyapunov stability of $\tilde{e}$.

2. The same form of controller works for the asymptotic stabilization of the upright spinning top with $\Omega_3^\circ > \sqrt{Mgh/\bar{I}_3}$, which is the opposite of (5.14). All

that needs to be done is to choose ρ_i and ϵ_i to make $E_{\tilde{\Phi}}$ have a local minimum at the equilibrium and to choose negative c_i such that $E_{\tilde{\Phi}}$ decreases in time. LaSalle's theorem argument guarantees asymptotic stability.

6. Conclusions. In this paper we have studied the reduction of controlled Lagrangian (CL) and controlled Hamiltonian (CH) systems with symmetry. We have shown that the notion of equivalence of controlled systems is preserved by the reduction procedure. This leads to a natural derivation of the Bloch–Leonard–Marsden Euler–Poincaré matching conditions and shows in a precise sense how they are related to the unreduced Euler–Lagrange matching conditions. The theory also shows how to do the equivalent matching on the Hamiltonian side. We studied the examples of a rigid body with rotors (a spacecraft) as well as a heavy top with rotors to illustrate the theory.

In the future we will study more examples and also see to what extent this theory applies to controlled nonholonomic systems with or without symmetry, following [18] and [36], and to degenerate and implicit controlled Lagrangian and Hamiltonian systems, following [6] and [35].

Acknowledgments. The authors thank Hernán Cendra for helpful discussions and the anonymous referees for constructive comments.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed., Addison-Wesley, Reading, MA, 1978.

[2] R. Abraham, J. E. Marsden, and T. S. Ratiu, *Manifolds, Tensor Analysis, and Applications*, Appl. Math. 75, 2nd ed., Springer-Verlag, New York, 1988.

[3] D. Auckly, L. Kapitanski, and W. White, *Control of nonlinear underactuated systems*, Comm. Pure Appl. Math., 53 (2000), pp. 354–369.

[4] D. Auckly and L. Kapitanski, *On the λ-equations for matching control laws*, SIAM J. Control Optim., 41 (2002), pp. 1372–1388.

[5] G. Blankenstein, R. Ortega, and A. J. van Der Schaft, *The matching conditions of controlled Lagrangians and interconnection and damping assignment passivity based control*, Internat. J. Control, 75 (2002), pp. 645–665.

[6] G. Blankenstein and A. J. van der Schaft, *Symmetry and reduction in implicit generalized Hamiltonian systems*, Rep. Math. Phys., 47 (2001), pp. 57–100.

[7] A. M. Bloch, D. E. Chang, N. E. Leonard, and J. E. Marsden, *Controlled Lagrangians and the stabilization of mechanical systems* II: *Potential shaping*, IEEE Trans. Automat. Control, 46 (2001), pp. 1556–1571.

[8] A. M. Bloch, D. E. Chang, N. E. Leonard, J. E. Marsden, and C. Woolsey, *Asymptotic stabilization of Euler-Poincaré mechanical systems*, in Lagrangian and Hamiltonian Methods for Nonlinear Control: A Proceedings Volume from the IFAC Workshop, N. E. Leonard and R. Ortega, eds., Pergamon, Elmsford, NY, 2000, pp. 51–56.

[9] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and G. Sánchez De Alvarez, *Stabilization of rigid body dynamics by internal and external torques*, Automatica, 28 (1992), pp. 745–756.

[10] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and T. Ratiu, *The Euler–Poincaré Equations and double bracket dissipation*, Comm. Math. Phys., 175 (1996), pp. 1–42.

[11] A. M. Bloch, N. E. Leonard, and J. E. Marsden, *Stabilization of mechanical systems using controlled Lagrangians*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 2356–2361.

[12] A. M. Bloch, N. E. Leonard, and J. E. Marsden, *Matching and stabilization by the method of controlled Lagrangians*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 1446–1451.

[13] A. M. Bloch, N. E. Leonard, and J. E. Marsden, *Stabilization of the pendulum on a rotor arm by the method of controlled Lagrangians*, in Proceedings of the International Conference on Robotics and Automation, Detroit, MI, IEEE, Piscataway, NJ, 1999, pp. 500–505.

[14] A. M. Bloch, N. E. Leonard, and J. E. Marsden, *Controlled Lagrangians and the stabilization of mechanical systems* I: *The first matching theorem*, IEEE Trans. Automat. Control, 45 (2000), pp. 2253–2270.

[15] A. M. Bloch, N. E. Leonard, and J. E. Marsden, *Controlled Lagrangians and the stabilization of Euler–Poincaré mechanical systems*, Internat. J. Robust Nonlinear Control, 11 (2001), pp. 191–214.

[16] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1986.

[17] H. Cendra, J. E. Marsden, and T. S. Ratiu, *Lagrangian reduction by stages*, Mem. Amer. Math. Soc., 152 (2001), number 722.

[18] H. Cendra, J. E. Marsden, and T. S. Ratiu, *Geometric mechanics, Lagrangian reduction and nonholonomic systems*, in Mathematics Unlimited—2001 and Beyond, B. Enquist and W. Schmid, eds., Springer-Verlag, New York, 2001, pp. 221–273.

[19] D. E. Chang, *Controlled Lagrangian and Hamiltonian Systems*, Ph.D. Thesis, Control and Dynamical Systems Department, California Institute of Technology, Pasadena, CA, 2002.

[20] D. E. Chang, A. M. Bloch, N. E. Leonard, J. E. Marsden, and C. Woolsey, *Equivalence of controlled Lagrangian and controlled Hamiltonian systems*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 393–422.

[21] D. E. Chang, and J. E. Marsden, *Asymptotic stabilization of the heavy top using controlled Lagrangians*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 269–273.

[22] J. Hamberg, *General matching conditions in the theory of controlled Lagrangians*, in Proceedings of the 39th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 2519–2523.

[23] J. Hamberg, *Controlled Lagrangians, symmetries and conditions for strong matching*, in Lagrangian and Hamiltonian Methods for Nonlinear Control: A Proceedings Volume from the IFAC Workshop, N. E. Leonard and R. Ortega, eds., Pergamon, Elmsford, NY, 2000, pp. 62–67.

[24] D. D. Holm, J. E. Marsden, and T. S. Ratiu, *The Euler–Poincaré equations and semidirect products with applications to continuum theories*, Adv. Math., 137 (1998), pp. 1–81.

[25] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, John Wiley & Sons, New York, 1963.

[26] P. S. Krishnaprasad, *Lie–Poisson structures, dual-spin spacecraft and asymptotic stability*, Nonlinear Anal., 9 (1985), pp. 1011–1035.

[27] J. E. Marsden, *Lectures on Mechanics*, London Math. Soc. Lecture Note Ser. 174, Cambridge University Press, Cambridge, UK, 1992.

[28] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, 2nd ed., Texts in Appl. Math. 17, Springer-Verlag, New York, 1999.

[29] J. E. Marsden and J. Scheurle, *Lagrangian reduction and the double spherical pendulum*, Z. Angew. Math. Phys., 44 (1993), pp. 17–43.

[30] J. E. Marsden and J. Scheurle, *The reduced Euler–Lagrange equations*, Fields Inst. Commun., 1 (1993), pp. 139–164.

[31] R. Montgomery, J. E. Marsden, and T. S. Ratiu, *Gauged Lie-Poisson structures*, Contemp. Math., 28 (1984), pp. 101–114.

[32] A. J. van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*, Communication & Control Engineering Series, Springer-Verlag, New York, 2000.

[33] L. S. Wang and P. S. Krishnaprasad, *Gyroscopic control and stabilization*, J. Nonlinear Sci., 2 (1992), pp. 367–415.

[34] C. Woolsey and N. E. Leonard, *Underwater vehicle stabilization by internal rotors*, in Proceedings of the American Control Conference, San Diego, CA, 1999.

[35] H. Yoshimura and J. E. Marsden, *Variational Principles, Dirac Structure and Implicit Lagrangian Systems*, preprint.

[36] D. V. Zenkov, A. M. Bloch, and J. E. Marsden, *The Lyapunov–Malkin theorem and stabilization of the unicycle with rider*, Systems Control Lett., 45 (2002), pp. 293–302.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 301–324

HIGH-ORDER VARIATIONS FOR FAMILIES OF VECTOR FIELDS*

RONALD HIRSCHORN† AND ANDREW D. LEWIS†

Abstract. Sufficient conditions involving Lie brackets of arbitrarily high order are obtained for local controllability of families of vector fields. After providing a general framework for the generation of high-order control variations, we propose a specific method for generating such variations. The theory is applied to a number of nontrivial examples.

Key words. local controllability, nonlinear systems, higher-order conditions

AMS subject classification. 70H03

DOI. 10.1137/S0363012902400622

1. Introduction. In this paper we present a technique for generating high-order variations of families of vector fields. Our approach is motivated by the early work of Sussmann on local controllability [10]. As in [10], we consider S a set of analytic vector fields on $\Omega \subset \mathbb{R}^n$ and an *S-trajectory* to be a continuous curve which is a finite concatenation of integral curves of vector fields in S. A point q is *S-reachable from p* if there exists an S-trajectory $t \mapsto \gamma(t)$ such that $\gamma(0) = p$, $q = \gamma(t)$ for some $t \geq 0$, and *S-reachable from p in time $\leq T$* if $t \leq T$. We say S is *locally controllable* (hereafter abbreviated *l.c.*) if, for every $T > 0$, the set of points S-reachable from p in time $\leq T$ contains p in its interior. In [10] Sussmann defines the set S^1_p of Lie brackets of order two of vector fields in S. His main result is that S is l.c. at p if $0 \in \operatorname{int}(\operatorname{co}(S(p) \cup S^1_p(p)))$, where co stands for convex hull. The main contribution in this paper is the construction of sets of vector fields $S^2_p, S^3_p, \ldots$ of higher-order Lie brackets of vector fields in S. In Theorem 4.5 we summarize our results concerning the generation of these high-order variations. This method of generating variations leads to a controllability result, Theorem 3.7, which states that S is l.c. at p if

$$0 \in \operatorname{int}(\operatorname{co}(S(p) \cup S^1_p(p) \cup \cdots \cup S^m_p(p)))$$

for some $m \geq 1$. Of course, the problem of local controllability, especially for control affine systems, has been studied in detail. We refer particularly to [1, 2, 3, 5, 6, 7, 11, 12].

The paper is organized as follows. In section 2 we review Sussmann's results on local controllability and consider an example. In section 3 we introduce our high-order condition for local controllability, Theorem 3.7. In section 4 we introduce a concrete class of higher-order variations which allow us to apply Theorem 3.7. In section 5 we give some examples illustrating our results.

2. First-order conditions. Suppose that S is a set of vector fields on an open set $\Omega \subset \mathbb{R}^n$ and $0 \in \operatorname{co}(S(p))$ for some $p \in \Omega$, where $\operatorname{co}(S(p))$ is the convex hull in $\mathbb{R}^n \simeq T_p\Omega$ of the set of vectors $S(p) = \{X(p) \mid X \in S\}$. Then, as in [10], we let $L^0(S,p) \subset \mathbb{R}^n$ denote the unique linear subspace of maximal dimension such that

$$0 \in \operatorname{int}_{L^0(S,p)}(\operatorname{co}(S(p)) \cap L^0(S,p))$$

*Received by the editors January 7, 2002; accepted for publication (in revised form) November 4, 2003; published electronically June 25, 2004. This research was supported by the Natural Sciences and Engineering Research Council of Canada.
http://www.siam.org/journals/sicon/43-1/40062.html
†Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (ron@mast.queensu.ca, andrew@mast.queensu.ca).

and define

$$Z_p^0 = \{X \in S \mid X(p) \in L^0(S,p)\}.$$

Let $\tilde{S}_p^1$ denote the set of second-order Lie brackets $\tilde{S}_p^1 = \{[X,Y] \mid X, Y \in Z_p^0\}$, where $[X,Y](p) = dY_pX(p) - dX_pY(p)$. The following sufficient condition was established by Sussmann.

THEOREM 2.1 (see [10]). *Suppose that S is a finite set of vector fields such that $0 \in \operatorname{int}(\operatorname{co}(S(p) \cup \tilde{S}_p^1(p)))$. Then S is l.c. at p.*

Remark 2.2. A natural extension of this result would involve $\tilde{S}_p^2$, the set of all triple brackets of elements of Z_p^0. Sussmann points out that the corresponding second-order theorem, that S is l.c. at p if

$$0 \in \operatorname{int}(\operatorname{co}(S(p) \cup \tilde{S}_p^1(p) \cup \tilde{S}_p^2(p))), \tag{2.1}$$

is false. One consequence of our results is that this theorem does hold if $\tilde{S}_p^2$ is the restricted set of triple brackets of elements of Z_p^0 of the form $[X,[X,Y]]$. For example, if in $\mathbb{R}^3$ we take the vector fields

$$W = (1,z,0), \quad X = (-1,0,x^2), \quad Y = (0,1,0), \quad Z = (0,-1,0),$$

then (2.1) holds at $p = (0,0,0)$, but clearly the family is not l.c. at this point as one can never reach states with negative z coordinate.

3. Higher-order Lie brackets. In this section we develop our methodology for the generation of control variations involving arbitrarily high-order brackets of vector fields in S. Our method for doing so begins with some constructions involving what we call complementary sets of vector fields. After these considerations have been discussed in section 3.1, in section 3.2 we produce explicit S-trajectories, which give us control variations involving certain high-order Lie brackets of vector fields in S. In section 3.3 we apply these constructions to give a theorem on local controllability of S.

If X is a vector field, we denote its flow by $t \mapsto X_t(p)$. If X, Y are vector fields, we let $\operatorname{ad}_X Y$ denote the Lie bracket $[X,Y](p) = dY_pX(p) - dX_pY(p)$ (see [14]). We shall consider iterated brackets of vector fields from a family of vector fields, and so need the notion of degree of a bracket. For us, this will refer to the number of vector fields involved in the bracket. Thus a plain vector field has degree 1, and $[X,Y]$ is a bracket of degree 2. Of course, to be perfectly clear about this, one should use free Lie algebras [8]. However, the loss of rigor in what we do here does not merit the introduction of the additional terminology.

3.1. Complementary vector fields. A finite subset $\mathcal{X}_p \subset Z_p^0$ is said to be *complementary at p* if

$$0 \in \operatorname{int}_{\operatorname{aff}(\mathcal{X}_p(p))}(\operatorname{co}(\mathcal{X}_p(p))),$$

where aff denotes the affine hull. Equivalently, $\mathcal{X}_p$ is complementary if 0 can be written as a linear combination of the $X(p)$, $X \in \mathcal{X}_p$, with strictly positive coefficients. Clearly Z_p^0 is complementary at p. If Z_p^0 is convex, then there are many complementary sets. We note that Z_p^0 is convex if S is. Furthermore it is known that S is l.c. if and only if $\operatorname{co}(S)$ is l.c. While our results do not depend on S being convex, to simplify notation *we will assume that S is convex* for the rest of this paper. We will also assume that the family of vector fields has the property that $S(p) \subset T_p\Omega$ is compact.

PROPOSITION 3.1. *Suppose Z_p^0 is convex. Then for every $X \in Z_p^0$ there exists a vector field $Y \in Z_p^0$ such that $\{X, Y\}$ is complementary at p.*

Proof. Let $X \in Z_p^0$. From the definition of Z_p^0 there exist $\lambda_i > 0$ and $Y_i \in Z_p^0$ such that $\sum_{i=0}^k \lambda_i = 1$ and $\lambda_0 X(p) + \lambda_1 Y_1(p) + \cdots + \lambda_k Y_k(p) = 0$. Set $\lambda_* = \sum_{i=1}^k \lambda_i$ and $Y = \sum_{i=1}^k (\lambda_i/\lambda_*) Y_i$. Because Z_p^0 is convex, $Y \in Z_p^0$. This, together with the fact that $(\lambda_0 X + \lambda_* Y)(p) = 0$, completes the proof. □

Suppose that $\mathcal{X}_p = \{X^1, \ldots, X^k\} \subset Z_p^0$ is complementary at p. Then $\mathcal{X}_p$ gives rise to vector fields which vanish at p, namely those which can be expressed as $Z = \lambda_1 X^1 + \cdots + \lambda_k X^k$ for appropriate $\lambda_i > 0$. We define $\mathcal{Z}_p$ as the collection of all such vector fields Z. Since we assume that S is convex, we know that $Z \in S$ and thus

$$\mathcal{Z}_p = \{Z \mid \ Z \in S,\ Z(p) = 0\}.$$

Part of our approach will be to systematically consider rather general classes of S-trajectories. To this end, let π be a permutation of $\{1, \ldots, k\}$. We denote by $\mathcal{X}_t^\pi(p)$ the composition of integral curves of the vector fields in $\mathcal{X}_p$ with time rescaled, namely

$$\mathcal{X}_t^\pi(p) = X^{\pi(k)}_{\lambda_{\pi(k)} t} \circ \cdots \circ X^{\pi(1)}_{\lambda_{\pi(1)} t}(p),$$

where $\lambda_i > 0$ and $\sum_{i=0}^k \lambda_i X^i(p) = 0$. Note that $\mathcal{X}_t^\pi(p)$ is reachable in time $(\sum_i \lambda_i)t$. In spite of a rescaling of time, $\mathcal{X}_t^\pi(p)$ is an S-trajectory in the sense that all points of the form $\mathcal{X}_t^\pi(p)$ for t sufficiently small are the image of a proper S-trajectory. Let P_k denote the set of sequences of permutations of $\{1, 2, \ldots, k\}$. If $\eta \in P_k$, then $\eta = (\pi_\ell, \pi_{\ell-1}, \ldots, \pi_1)$ for some $\ell \in \mathbb{N}$, and we define a *$\mathcal{X}_p^1$-trajectory* $\mathcal{X}_t^\eta(p)$ to be the S-trajectory which is the composition of the curves $\mathcal{X}_t^{\pi_i}(p)$. Then the Campbell–Baker–Hausdorff formula [13] asserts that, for t sufficiently small, there exist vector fields $X^{\eta,i}$ and $X^{\pi,i}$ such that

$$\begin{aligned} \mathcal{X}_t^\pi(p) &= \left(\sum_{i=1}^k \lambda_{\pi(i)} X^{\pi(i)} + X^{\pi,1} t + X^{\pi,2} t^2 + o(t^2) \right)_t (p), \\ \mathcal{X}_t^\eta(p) &= \mathcal{X}_t^{\pi_s} \circ \cdots \circ \mathcal{X}_t^{\pi_1}(p) \\ &= (X^{\eta,1} + X^{\eta,2} t + o(t))_t(p), \end{aligned} \tag{3.1}$$

where $X^{\eta,1}$ is a multiple of $\sum_{i=1}^k \lambda_i X^i$ and hence vanishes at p. Note that the Campbell–Baker–Hausdorff formula also provides explicit expressions for these terms in the series. In any event, this leaves as dominant the second-order term $X^{\eta,2}(p)$. Sussmann [10] generates a richer class of S-trajectories, which allows him to prove his theorem on local controllability (stated as Theorem 2.1 here). However, the local controllability result can be proved using the smaller class of S-trajectories that we consider here. We also point out that, as with $\mathcal{X}_t^\pi(p)$ above, the point $\mathcal{X}_t^\eta(p)$ is reached by an S-trajectory after some time σt, $\sigma > 0$, has elapsed.

Remark 3.2. In (3.1) we have expressed $\mathcal{X}_t^\eta(p)$ as an integral curve for a "time-dependent" vector field $X(t) = X^{\eta,1} + X^{\eta,2} t + o(t)$. To make this more precise we fix $\tau > 0$ and let $\alpha_\tau(t)$ denote the integral curve of the vector field $X(\tau)$ through p, that is, $\alpha_\tau(t) = (X(\tau))_t(p)$, where $\frac{d}{dt}\alpha_\tau(t) = X(\tau)(\alpha_\tau(t))$ and $\alpha_\tau(0) = p$. Then $\mathcal{X}_t^\eta(p)$ denotes the point $\alpha_\tau(t)|_{\tau=t}$.

Remark 3.3. While our definition for S_p^1 differs slightly from Sussmann's $\tilde{S}_p^1$, we do have $\mathrm{co}(\tilde{S}_p^1) \subset \mathrm{co}(S_p^1)$. To show this we can utilize the limited set of permutations

used by Sussmann in his proof of his sufficiency condition for local controllability (Theorem 3 of [10]).

Before we define S^k_p, we motivate the notion of S-trajectories which approximate integral curves to orders higher than one. Let $X, Y \in S^1_p$. From the definition of S^1_p there exist S-trajectories $\mathcal{X}^\eta_t(p) = (X^1 + tX + o(t))_t(p)$ and $\mathcal{Y}^\eta_t(p) = (Y^1 + tY + o(t))_t(p)$ such that X^1 and Y^1 are linear combinations of vector fields in S that vanish at p. Now suppose that $(\lambda_1 X + \lambda_2 Y + \lambda_3 Z)(p) = 0$ for some $Z \in S$ and $\lambda_1, \lambda_2, \lambda_3 > 0$. Proceeding as above, while rescaling time to ensure compatibility between the vector fields in S and S^1_p, we construct the S-trajectory

$$\mathcal{X}_t(p) = \mathcal{X}^\eta_{\sqrt{\lambda_1}t} \circ \mathcal{Y}^\eta_{\sqrt{\lambda_2}t} \circ Z_{\lambda_3 t^2}(p).$$

From the Campbell–Baker–Hausdorff formula we obtain

$$\begin{aligned}
\mathcal{X}_t(p) &= \big(X^1 + X\sqrt{\lambda_1}t + o(t)\big)_{\sqrt{\lambda_1}t} \circ \big(Y^1 + Y\sqrt{\lambda_2}t + o(t)\big)_{\sqrt{\lambda_2}t} \circ Z_{\lambda_3 t^2}(p) \\
&= \Big(t(\sqrt{\lambda_1}X^1 + \sqrt{\lambda_2}Y^1) + t^2(\lambda_1 X + \lambda_2 Y + \lambda_3 Z + (1/2)\sqrt{\lambda_1\lambda_2}[Y^1, X^1]) \\
&\qquad + t^3 W + o(t^3)\Big)_1(p) \\
&= (tX^{\eta,1} + t^2 X^{\eta,2} + t^3 X^{\eta,3})_1(p)
\end{aligned}$$

for vector fields $X^{\eta,1}, X^{\eta,2}, X^{\eta,3}$ with the following properties:

1. $X^{\eta,1}$ is a linear combination of vector fields from S;
2. $X^{\eta,2}$ is a linear combination of degree 1 and 2 brackets of vector fields evaluated at p in S;
3. $X^{\eta,3}$ is a linear combination of degree 2 and 3 brackets of vector fields evaluated at p in S;
4. $X^{\eta,1}$ and $X^{\eta,2}$ vanish at p.

Since the coefficients of t and t^2 vanish at p, we have produced an S-trajectory which approximates, to the third order in t, the integral curve of $X^{\eta,3}$. We let $S^2_p(\mathcal{X}_p)$ denote the set of all such terms $X^{\eta,3}$, and S^2_p the union of the sets $S^2_p(\mathcal{X}_p)$ over all subsets $\mathcal{X}_p$ complementary at p.

3.2. Higher-order variations. We now define S^i_p for $i > 1$ inductively. Suppose that we have defined sets of vector fields $S^1_p, \dots, S^m_p$ with the following property: for any $X \in S^j_p$ there exists an S-trajectory of the form $\mathcal{X}^\eta_t(p) = (X^\eta(t))_t(p) = (tX^\eta(t))_1(p)$, with $X^\eta(t)$ a time-varying vector field so that for t sufficiently small $tX^\eta(t)$ can be represented by the convergent power series

$$tX^\eta(t) = tX^{\eta,1} + t^2 X^{\eta,2} + \cdots + t^{\sigma_j - 1} X^{\eta,\sigma_j - 1} + t^{\sigma_j} X^{\eta,\sigma_j} + o(t^{\sigma_j}),$$

where the vector fields $X^{\eta,1}, \dots, X^{\eta,\sigma_j - 1}$ vanish at p, $X = X^{\eta,\sigma_j}$, and σ_j is defined inductively by $\sigma_1 = 2$ and

$$\sigma_{k+1} = \frac{(\sigma_k + 1)\mathrm{lcm}\{\sigma_1, \dots, \sigma_k\}}{\sigma_k}, \tag{3.2}$$

where lcm denotes least common multiple. We note that one consequence of the above definition is that $\sigma_{m+1} > \sigma_m > \cdots > \sigma_1$. The reason for this definition becomes apparent in the proof of Lemma 3.4. Let $L^m(S, p)$ denote the unique linear subspace of maximal dimension such that

$$0 \in \mathrm{int}_{L^m(S,p)}\big(\mathrm{co}(S(p) \cup S^1_p(p) \cup \cdots \cup S^m_p(p)) \cap L^m(S, p)\big),$$

and set

$$Z_p^m = \{X \in S \cup S_p^1 \cup \cdots \cup S_p^m \mid X(p) \in L^m(S,p)\}.$$

A finite subset $\mathcal{X}_p \subset Z_p^m$ is said to be *complementary at* p if

$$0 \in \operatorname{int}_{\operatorname{aff}(\mathcal{X}_p(p))}\big(\operatorname{co}(\mathcal{X}_p(p))\big)$$

or, equivalently, if 0 can be written as a linear combination of the vectors $X(p), X \in \mathcal{X}_p$, with strictly positive coefficients. Suppose that $\mathcal{X}_p = \{X^1, \ldots, X^k\}$ is a subset of Z_p^m complementary at p, so that $\sum_{i=0}^k \lambda_i X^i(p) = 0$ for some $\lambda_i > 0$. Let π be a permutation of $\{1, 2, \ldots, k\}$. Then $X^i \in S$ or $X^i \in S_p^{m_i}$, where $m_i \in \{1, \ldots, m\}$. If $X^i \in S_p^{m_i}$, then, by our induction hypothesis, there exists an S-trajectory of the form $\mathcal{X}_t^{\eta_i}(p) = (X^{\eta_i}(t))_t(p)$, where the time-varying vector field $X^{\eta_i}(t)$ has the power series expansion

$$tX^{\eta_i}(t) = tX^{\eta_i,1} + \cdots + t^{\sigma_{m_i}-1}X^{\eta_i,\sigma_{m_i}-1} + t^{\sigma_{m_i}}X + o(t^{\sigma_{m_i}}) \tag{3.3}$$

and such that $X^{\eta_i,1}, \ldots, X^{\eta_i,\sigma_{m_i}-1}$ vanish at p. We rescale time by $t \mapsto \alpha_i t^{\gamma_i}$, where $\alpha_i = \lambda_i^{1/\sigma_{m_i}}$, and $\gamma_i = \operatorname{lcm}\{\sigma_1, \ldots, \sigma_m\}/\sigma_{m_i}$. If $X^i \in S$, we rescale time by $t \mapsto \alpha_i t^{\gamma_i}$, where $\gamma_i = \operatorname{lcm}\{\sigma_1, \ldots, \sigma_m\}$—in effect, we define $\sigma_0 = 1$. We denote by $\mathcal{X}_t^\pi(p)$ the S-trajectory

$$\mathcal{X}_t^\pi(p) = \mathcal{X}^{\eta_{\pi(k)}}_{\alpha_{\pi(k)}t^{\gamma_{\pi(k)}}} \circ \cdots \circ \mathcal{X}^{\eta_{\pi(1)}}_{\alpha_{\pi(1)}t^{\gamma_{\pi(1)}}}(p). \tag{3.4}$$

This rescaling is needed because, if $X \in S_p^k$, then, using a suitable control variation, we can generate an S-trajectory which achieves motion in the X direction to order σ_k in t. Finally, if $\eta \in P_k$, so that $\eta = (\pi_{s-1}, \pi_{k_{s-1}}, \ldots, \pi_1)$, we define $\mathcal{X}_t^\eta(p)$ to be the composition of the curves $\mathcal{X}_t^{\pi_i}(p)$ and say that $\mathcal{X}_t^\eta(p)$ is an $\mathcal{X}_p^{m+1}$*-trajectory*. Then

$$\mathcal{X}_t^\eta(p) = \mathcal{X}_t^{\pi_s} \circ \cdots \circ \mathcal{X}_t^{\pi_1}(p) = (X^\eta(t))_t(p).$$

For t sufficiently small the Campbell–Baker–Hausdorff formula yields

$$tX^\eta(t) = tX^{\eta,1} + \cdots + t^{\sigma_{m+1}-1}X^{\eta,\sigma_{m+1}-1} + t^{\sigma_{m+1}}X^{\eta,\sigma_{m+1}} + o(t^{\sigma_{m+1}}) \tag{3.5}$$

for vector fields $X^{\eta,i}$.

The following lemma makes clear why the inductive definitions of the σ_k's are as in (3.2). The idea essentially is that one needs to define time rescalings along vector fields in an S-trajectory to ensure that the desired term is the first nonzero term in the series expansion. This makes sense of our inductive definition of $S_p^m(p)$.

LEMMA 3.4. *The vector fields* $X^{\eta,1}, \ldots, X^{\eta,\sigma_{m+1}-1}$ *that appear in* (3.5) *vanish at* p.

Proof. Let $X \in S_p^i, Y \in S_p^j$, where $i, j \in \{0, \ldots, m\}$ and $S_p^0 = S$. By our induction hypotheses there exist time-varying vector fields $X^{\eta_i}(t), X^{\eta_j}(t)$, where

$$\begin{aligned} tX^{\eta_i}(t) &= (tX^{\eta_i,1} + \cdots + t^{\sigma_i-1}X^{\eta_i,\sigma_i-1} + t^{\sigma_i}X + o(t^{\sigma_i})), \\ tX^{\eta_j}(t) &= (tX^{\eta_j,1} + \cdots + t^{\sigma_j-1}X^{\eta_j,\sigma_j-1} + t^{\sigma_j}Y + o(t^{\sigma_j})), \end{aligned}$$

with $X^{\eta_j,k}, X^{\eta_i,\ell}$ vanishing at p as in (3.3) above. The corresponding S-trajectories are $\mathcal{X}_t^{\eta_j}(p) = (X^{\eta_j}(t))_t(p) = (tX^{\eta_j}(t))_1(p)$ and $\mathcal{X}_t^{\eta_i}(p) = (X^{\eta_i}(t))_t(p) = (tX^{\eta_i}(t))_1(p)$.

Rescaling time as above and concatenating these curves yields the S-trajectory

$$\begin{aligned}\beta(t) &= \mathcal{X}^{\eta_j}_{\alpha_j t^{\gamma_j}} \circ \mathcal{X}^{\eta_i}_{\alpha_i t^{\gamma_i}} \\ &= (X^{\eta_j}(\alpha_j t^{\gamma_j}))_{\alpha_j t^{\gamma_j}} \circ (X^{\eta_i}(\alpha_i t^{\gamma_i}))_{\alpha_i t^{\gamma_i}}(p) \\ &= (\alpha_j t^{\gamma_j} X^{\eta_j}(\alpha_j t^{\gamma_j}))_1 \circ (\alpha_i t^{\gamma_i} X^{\eta_i}(\alpha_i t^{\gamma_i}))_1(p) \\ &= \left(\sum_{k=1}^{\sigma_j+1} (\alpha_j t^{\gamma_j})^k X^{\eta_j,k} + o(t^{\gamma_j(\sigma_j+1)}) \right)_1 \circ \left(\sum_{\ell=1}^{\sigma_i+1} (\alpha_i t^{\gamma_i})^\ell X^{\eta_i,\ell} + o(t^{\gamma_i(\sigma_i+1)}) \right)_1 (p),\end{aligned}$$

where $X = X^{\eta_i,\sigma_i}$, $Y = X^{\eta_j,\sigma_j}$, and $X^{\eta_i,\ell}$ and $X^{\eta_j,k}$ vanish at p for $k < \sigma_j, \ell < \sigma_i$. For t sufficiently small, the Campbell–Baker–Hausdorff formula gives the coefficients of t in the power series expansion for $\beta(t)$. In particular, $\beta(t)$ can be written as a convergent power series whose terms are expressible as linear combinations of Lie brackets of the vector fields $X^{\eta_i,\ell}$ and $X^{\eta_j,k}$ and Lie brackets of these vector fields of all orders. Our induction hypothesis implies that $X^{\eta_i,\ell}$ and $X^{\eta_j,k}$ vanish at p if $k < \sigma_j$ and $\ell < \sigma_i$. Hence Lie brackets of these vector fields also vanish at p. Thus the lowest-order term with respect to t in the power series expansion for $\beta(t)$ which does not necessarily vanish at p will be

$$(\alpha_j t^{\gamma_j})^{\sigma_j} X^{\eta_j,\sigma_j} + (\alpha_i t^{\gamma_i})^{\sigma_i} X^{\eta_i,\sigma_i}.$$

From the above definitions

$$(\alpha_j t^{\gamma_j})^{\sigma_j} X^{\eta_j,\sigma_j} = ((\lambda_j^{1/\sigma_j})^{\sigma_j}) t^{\gamma_j \sigma_j} X^{\eta_j,\sigma_j}$$

and

$$(\alpha_i t^{\gamma_i})^{\sigma_i} X^{\eta_i,\sigma_i} = ((\lambda_i^{1/\sigma_i})^{\sigma_i}) t^{\gamma_i \sigma_i} X^{\eta_i,\sigma_i}.$$

Thus

$$(\alpha_j t^{\gamma_j})^{\sigma_j} X^{\eta_j,\sigma_j} + (\alpha_i t^{\gamma_i})^{\sigma_i} X^{\eta_i,\sigma_i} = t^{\operatorname{lcm}\{\sigma_1,\ldots,\sigma_m\}} (\lambda_j X^{\eta_j,\sigma_j} + \lambda_i X^{\eta_i,\sigma_i}).$$

The next (higher) power of t which appears in the power series for $\beta(t)$ is t^r, which has as coefficient the linear combination of vector fields

$$\begin{aligned}(\alpha_j t^{\gamma_j})^{\sigma_j+1} X^{\eta_j,\sigma_j+1} + (\alpha_i t^{\gamma_i})^{\sigma_i+1} X^{\eta_i,\sigma_i+1} \\ = \alpha_j^{\sigma_j+1} t^{\gamma_j(\sigma_j+1)} X^{\eta_j,\sigma_j+1} + \alpha_i^{\sigma_i+1} t^{\gamma_i(\sigma_i+1)} X^{\eta_i,\sigma_i+1}.\end{aligned}$$

We now show that $r \geq \sigma_{m+1}$. Since $\gamma_j(\sigma_j+1) = (\frac{\sigma_j+1}{\sigma_j})\operatorname{lcm}\{\sigma_1,\ldots,\sigma_m\}$; the sequence $\{\sigma_j\}$ is, by definition, monotone increasing; and $\sigma_{m+1} = (\frac{\sigma_m+1}{\sigma_m})\operatorname{lcm}\{\sigma_1,\ldots,\sigma_m\}$, we see that $\gamma_j(\sigma_j+1) > \sigma_{m+1}$ for $j < m$ and $\gamma_j(\sigma_j+1) = \sigma_{m+1}$ if $j = m$. Among the Lie brackets of order 2 in the power series expansion of $\beta(t)$ which do not vanish at p, the terms with the lowest power of t will have the form

$$[\alpha_j t^{\gamma_j} X^{\eta_j,1}, (\alpha_i t^{\gamma_i})^{\sigma_i} X^{\eta_i,\sigma_i}] = \alpha_j {\alpha_i}^{\sigma_i} t^{\gamma_j+\gamma_i\sigma_i} [X^{\eta_j,1}, X^{\eta_i,\sigma_i}].$$

Here we have t to the power $\gamma_j + \gamma_i\sigma_i$ and

$$\begin{aligned}\gamma_j + \gamma_i\sigma_i &= \frac{\operatorname{lcm}\{\sigma_1,\ldots,\sigma_m\}}{\sigma_j} + \operatorname{lcm}\{\sigma_1,\ldots,\sigma_m\} \\ &= \Big(\frac{\sigma_j+1}{\sigma_j}\Big)\operatorname{lcm}\{\sigma_1,\ldots,\sigma_m\} \\ &\geq \sigma_{m+1},\end{aligned}$$

with equality holding if and only if $j = m$. Lie brackets of order greater than 2 which are coefficients of t^s with $s \le \sigma_{m+1}$ clearly must vanish at p. Thus if $\ell = \operatorname{lcm}\{\sigma_1, \dots, \sigma_m\}$, then the power series expansion for $\mathcal{X}_t^\pi(p)$ defined by (3.4) is of the form $(tZ^1 + \cdots + t^{\ell-1}Z^{\ell-1} + t^\ell Z^\ell + t^r Z^r + o(t^r))_1(p)$, where $Z^1, \dots, Z^{\ell-1}$ vanish at p, $Z^\ell = \sum_{i=1}^m \lambda_i X^i$; hence by our choice of the λ_i's we have $Z^\ell(p) = 0$, and $r \ge \sigma_{m+1}$ with $r = \sigma_{m+1}$ if and only if one of the vector fields $X^i \in S_p^m$. Extending this argument to $\mathcal{X}_t^\eta(p)$ completes the proof. □

This lemma implies that $\mathcal{X}_t^\eta(p)$ is an S-trajectory which approximates, to order $t^{\sigma_{m+1}}$, the integral curve of $X^{\eta,\sigma_{m+1}}$ with time rescaled to $t^{\sigma_{m+1}}$. We let $S_p^{m+1}(\mathcal{X}_p)$ denote the set of all such terms $X^{\eta,\sigma_{m+1}}$, indexed over all $\mathcal{X}_p^{m+1}$-trajectories $\mathcal{X}_t^\eta(p)$.

DEFINITION 3.5. *S_p^{m+1} is defined to be the union of the sets $S_p^{m+1}(\mathcal{X}_p)$ over all subsets $\mathcal{X}_p$ complementary at p.*

We note that vector fields in S_p^m will be linear combinations of brackets of degree at most $m+1$ of vector fields in S. The following is a consequence of the above discussion.

PROPOSITION 3.6. *Suppose that $X \in S_p^m$. Then*

1. *for t sufficiently small, there exists an S-trajectory $\mathcal{X}_t^\eta(p)$ of the form*

$$\mathcal{X}_t^\eta(p) = (X^{\eta,1} + \cdots + t^{\sigma_m - 1}X^{\eta,\sigma_m} + t^{\sigma_m}X^{\eta,\sigma_m+1} + o(t^{\sigma_m}))_t(p), \tag{3.6}$$

   where $X = X^{\eta,\sigma_m}$ and the vector fields $X^{\eta,k}$ vanish at p for $k = 1, \dots, \sigma_m - 1$;
2. *if $X(p) = 0$, then X^{η,σ_m+1} in (3.6) belongs to S_p^{m+1};*
3. *the S-trajectory (3.6) has the form*

$$\mathcal{X}_t^\eta(p) = p + t^{\sigma_m}X(p) + o(t^{\sigma_m}),$$

   where X is a linear combination of brackets of vector fields in S of degrees up to and including $m+1$.

Proof. Assertion 1 follows from our definition of S_p^m. In particular, the fact that, for t sufficiently small, there exists an S-trajectory $\mathcal{X}_t^\eta(p)$ of the form

$$\mathcal{X}_t^\eta(p) = (X^{\eta,1} + \cdots + t^{\sigma_m - 1}X^{\eta,\sigma_m} + t^{\sigma_m}X^{\eta,\sigma_m+1} + o(t^{\sigma_m}))_t(p),$$

where $X = X^{\eta,\sigma_m}$ and the vector fields $X^{\eta,k}$ vanish at p for $k = 1, \dots, \sigma_m - 1$, follows from the definition of S_p^m and Lemma 3.4. For point 2, suppose that X also vanishes at p. Then, by definition, $\mathcal{X}_p = \{X\} \subset Z_p^m$ is a set of vector fields complementary at p, and hence the $\mathcal{X}_p^m$-trajectory $\mathcal{X}_t^\eta(p)$ is also a $\mathcal{X}_p^{m+1}$-trajectory, and then $X^{\eta,\sigma_m+1} \in S_p^{m+1}$ by definition.

For assertion 3 we write (3.6) in exponential form:

$$\begin{aligned}\mathcal{X}_t^\eta(p) &= \exp(tX^{\eta,1} + \cdots + t^{\sigma_m-1}X^{\eta,\sigma_m-1} + t^{\sigma_m}X + o(t^{\sigma_m}))(p)\\ &= p + t^{\sigma_m}X(p) + o(t^{\sigma_m}),\end{aligned}$$

since $X^{\eta,1}(p) = \cdots = X^{\eta,\sigma_m-1}(p) = 0$. □

3.3. A theorem on local controllability. The main result in this section is the following high-order sufficient condition for local controllability.

THEOREM 3.7. *Suppose that S is a set of vector fields on $\Omega \subset \mathbb{R}^n$ such that*

$$0 \in \operatorname{int}\big(\operatorname{co}(S(p) \cup S_p^1(p) \cup \cdots \cup S_p^m(p))\big)$$

for some $m \ge 1$. Then S is l.c. at p.

Proof. By assumption there exist vector fields $X^i_1, \dots, X^i_{k_i} \in S^i_p$ for $0 \le i \le m$ such that 0 is contained in the absolute interior of the convex hull of $\{X^i_j(p) \mid 0 \le i \le m,\ 1 \le j \le k_i\}$. Here we set $S^0_p = S$. In light of Proposition 3.6(3) we can find corresponding S-trajectories

$$\mathcal{X}^{\eta_{i,j}}_t(p) = p + \frac{t^{\sigma_i}}{\sigma_i} X^i_j(p) + o(t^{\sigma_i}).$$

Rescaling time by $t^{\sigma_i} = \sigma_i s_{i,j}$ for $s_{i,j} > 0$, we have $\tilde{\mathcal{X}}^{\eta_{i,j}}_t(p) = p + s_{i,j} X^i_j(p) + o(s_{i,j})$. The composition of such S-trajectories yields

$$\alpha(s_{1,1}, s_{1,2}, \dots, s_{m,k_m}) = p + \sum_{i=0}^{m} \sum_{j=1}^{k_i} s_{i,j} X^i_j(p) + o(s_{1,1} + s_{1,2} + \cdots + s_{m,k_m}).$$

This is the form of the S-trajectories used in the proof of Theorem 3 in [10]. We can then apply Lemma 4 of [10] to conclude that S is l.c. at p. ∎

Remark 3.8. The result may also be proved using techniques of Frankowska [4].

Remark 3.9. Suppose that $X \in S^m_p$ and that $Z^1, \dots, Z^\ell \in \mathcal{Z}_p$. Then the directions spanned by $\pm \mathrm{ad}_{Z^1} \circ \mathrm{ad}_{Z^2} \circ \cdots \circ \mathrm{ad}_{Z^\ell} X(p)$ can be considered as available directions for the purposes of local controllability, provided that there exists $Y \in S^m_p$ so that $X(p) + Y(p) = 0$. This may be argued by slightly generalizing Theorem 2.4 in [2].

4. A concrete class of higher-order variations. While Theorem 3.7 is interesting, it is not so easy to apply, as we have not been very concrete about describing tangent vectors in $S^m_p(p)$. In this section we provide a description of some such tangent vectors. Our description arises from developing S-trajectories associated with sequences of permutations. One of the consequences of our development is the identification of terms in the series expansion for the S-trajectories that are independent of permutation. These are obstructions to local controllability in our setup. In the parlance of Sussmann [12], these are fixed points of a group action in a free Lie algebra.

4.1. Variations associated with sequences of permutations. Suppose that $X, Y \in S$. Then, for t sufficiently small,

$$Y_t \circ X_t(p) = \big(A^0(X,Y) + A^1(X,Y)t + A^2(X,Y)t^2 + A^3(X,Y)t^3 + \cdots\big)_t(p),$$

where, from the Campbell–Baker–Hausdorff formula,

$$\begin{aligned}
A^0(X,Y) &= X + Y,\\
A^1(X,Y) &= \frac{1}{2}\mathrm{ad}_X Y,\\
A^2(X,Y) &= \frac{1}{12}(\mathrm{ad}^2_Y X + \mathrm{ad}^2_X Y),\\
A^3(X,Y) &= -\frac{1}{24}\mathrm{ad}_Y \mathrm{ad}^2_X Y,\\
A^4(X,Y) &= -\frac{1}{180}\mathrm{ad}_Y \mathrm{ad}^3_X Y - \frac{1}{120}[\mathrm{ad}_X Y, \mathrm{ad}^2_X Y] + \frac{1}{180}\mathrm{ad}^2_Y \mathrm{ad}^2_X Y\\
&\quad + \frac{1}{360}[\mathrm{ad}_X Y, \mathrm{ad}^2_Y X] - \frac{1}{720}\mathrm{ad}^4_X Y - \frac{1}{720}\mathrm{ad}^4_Y X,
\end{aligned} \tag{4.1}$$

and $A^k(X,Y)$ can, in principal, be expressed explicitly as functions of X,Y for all $k>0$. Let $\mathbb{N}$ denote the positive integers. If $X^i,Y^i\in S$, $\boldsymbol{s}=(s_1,\dots,s_k)\in\mathbb{N}^k$, then for $\pi\in P_k^0$, the group of permutations of $\{1,2,\dots,k\}$, we form the S-trajectory

$$\begin{aligned}\mathcal{X}_t^\pi(p)&=Y^{\pi(1)}_{t^{s_{\pi(1)}}}\circ X^{\pi(1)}_{t^{s_{\pi(1)}}}\circ Y^{\pi(2)}_{t^{s_{\pi(2)}}}\circ X^{\pi(2)}_{t^{s_{\pi(2)}}}\circ\cdots\circ Y^{\pi(k)}_{t^{s_{\pi(k)}}}\circ X^{\pi(k)}_{t^{s_{\pi(k)}}}(p)\\&=(Q_\pi^0+Q_\pi^1t+Q_\pi^2t^2+\cdots)_t(p),\end{aligned}\tag{4.2}$$

where the vector fields $Q_\pi^\ell=Q_\pi^\ell(X^1,Y^1,\dots,X^k,Y^k,\boldsymbol{s})$ are linear combinations of the vector fields $A^j(X^i,Y^i)$ and their Lie brackets. For example, for $\boldsymbol{s}=(1,\dots,1)$ we have

$$\begin{aligned}Q_\pi^0&=A^0(X^1,Y^1)+\cdots+A^0(X^k,Y^k),\\Q_\pi^1&=\sum_{i=1}^k A^1(X^i,Y^i)+\frac{1}{2}\sum_{1\le i<j\le k}[A^0(X^{\pi(i)},Y^{\pi(i)}),A^0(X^{\pi(j)},Y^{\pi(j)})].\end{aligned}$$

The order of the group P_k^0 is $k!$, and we define P_k^1 to be the elements of the $k!$-fold direct product of P_k^0 with itself, $\Pi_{i=1}^{k!}P_k^0$, of the form $\pi=(\pi_1,\dots,\pi_{k!})$, where $\pi_i\in P_k^0$ are *distinct*. We note that P_k^1 is a set with $\Gamma=k!!$ elements. If $\pi=(\pi_1,\dots,\pi_{k!})\in P_k^1$, we define a corresponding S-trajectory

$$\mathcal{X}_t^\pi(p)=\mathcal{X}_t^{\pi_1}\circ\cdots\circ\mathcal{X}_t^{\pi_{k!}}(p)=(Q_\pi^0+Q_\pi^1t+Q_\pi^2t^2+\cdots)_t(p),$$

where, as above, $Q_\pi^\ell=Q_\pi^\ell(X^1,Y^1,\dots,X^k,Y^k,\boldsymbol{s})$ is a linear combination of the vector fields $A^j(X^i,Y^i)$ and their Lie brackets. Similarly $\pi\in P_k^2$ if $\pi=(\pi_1,\dots,\pi_\gamma)$, where $\pi_i\in P_k^1$ and

$$\mathcal{X}_t^\pi(p)=\mathcal{X}_t^{\pi_1}\circ\cdots\circ\mathcal{X}_t^{\pi_\Gamma}(p)=(Q_\pi^0+Q_\pi^1t+Q_\pi^2t^2+\cdots)_t(p).$$

In this way we can inductively define subsets of permutations P_k^ℓ. It will be convenient to use the notation $\Gamma(k,\ell)$ to denote the cardinality of P_k^ℓ. Thus $\Gamma(k,0)=k!$ and $\Gamma(k,\ell+1)=\Gamma(k,\ell)!$. Note that if $\pi=(\pi_1,\dots,\pi_{\Gamma(k,\ell)})\in P_k^{\ell+1}$, where $\pi_i\in P_k^\ell$, then $\mathcal{X}_t^\pi(p)$ denotes the S-trajectory

$$\begin{aligned}\mathcal{X}_t^\pi(p)&=\mathcal{X}_t^{\pi_1}\circ\cdots\circ\mathcal{X}_t^{\pi_{\Gamma(k,\ell)}}(p)\\&=(Q_\pi^0+Q_\pi^1t+\cdots)_t(p).\end{aligned}\tag{4.3}$$

For example, $P_2^0=\{\pi_1,\pi_2\}$ with

$$\pi_1=\begin{pmatrix}1&2\\1&2\end{pmatrix},\quad \pi_2=\begin{pmatrix}1&2\\2&1\end{pmatrix}.$$

Thus

$$\begin{aligned}P_2^1&=\{(\pi_1,\pi_2),\ \ (\pi_2,\pi_1)\},\\P_2^2&=\{((\pi_1,\pi_2),(\pi_2,\pi_1)),\ \ ((\pi_2,\pi_1),(\pi_1,\pi_2))\}.\end{aligned}$$

If $\pi=(\pi_1,\pi_2)\in P_2^1$, then

$$\mathcal{X}_t^\pi(p)=Y^1_{t^{s_1}}\circ X^1_{t^{s_1}}\circ Y^2_{t^{s_2}}\circ X^2_{t^{s_2}}\circ Y^2_{t^{s_2}}\circ X^2_{t^{s_2}}\circ Y^1_{t^{s_1}}\circ X^1_{t^{s_1}},$$

and if $\pi=(\pi_2,\pi_1)\in P_2^1$, then

$$\mathcal{X}_t^\pi(p)=Y^2_{t^{s_2}}\circ X^2_{t^{s_2}}\circ Y^1_{t^{s_1}}\circ X^1_{t^{s_1}}\circ Y^1_{t^{s_1}}\circ X^1_{t^{s_2}}\circ Y^2_{t^{s_2}}\circ X^2_{t^{s_2}}.$$

Similar expressions then hold for the elements of P_2^2. In essence these are analogous to the time reversal permutations considered by Sussmann [12].

4.2. Permutation-invariant elements. Next we turn to a more detailed investigation of the terms in the power series expansion for the S-trajectories of the preceding section. In particular, we show that such power series expansions possess terms that are independent of the sequence of permutations. In essence, these are terms in the series which cannot be modified by changing the sequence, and so may be thought of as obstructions to local controllability.

The first result exposes the pattern in which invariant terms arise in the series expansion (4.3) under sequences of permutations of a given length. If $\boldsymbol{s} = (s_1, \dots, s_k) \in \mathbb{N}^k$, we set

$$m(\boldsymbol{s}) = \min\{s_i \mid 1 \leq i \leq k\}$$

and define $m_i(\boldsymbol{s})$ inductively by

$$m_0(\boldsymbol{s}) + 2 = \min\{s_i + s_j \mid 1 \leq i, j \leq k,\ i \neq j\}$$

and

$$m_\ell(\boldsymbol{s}) = m_0(\boldsymbol{s}) + \ell m(\boldsymbol{s}).$$

LEMMA 4.1. *Let $\pi \in P_k^\ell$, and let $\mathcal{X}_t^\pi(p)$ be the S-trajectory*

$$\mathcal{X}_t^\pi(p) = (Q_\pi^0 + Q_\pi^1 t + \cdots)_t(p)$$

defined by (4.3). *Then $Q_\pi^0, \dots, Q_\pi^{m_\ell(\boldsymbol{s})}$ are independent of π, and $Q_\pi^0, \dots, Q_\pi^{m(\boldsymbol{s})-2}$ vanish identically.*

Proof. We begin by considering the case $\pi \in P_k^0$. To help with notation we set

$$X_j(t) = \sum_{i=0}^{\infty} A^i(X^{\pi(j)}, Y^{\pi(j)}) t^{(i+1)s_{\pi(j)}},$$

so that the S-trajectory $\mathcal{X}_t^\pi(p)$ defined by (4.2) is the composition of integral curves of the vector fields X_j followed for one unit of time. Thus $\mathcal{X}_t^\pi(p) = (X_1(t))_1 \circ \cdots \circ (X_k(t))_1(p)$, and, using the Campbell–Baker–Hausdorff formula, we have

$$\mathcal{X}_t^\pi(p) = \left(\sum_{j=1}^{k} X_j(t) + \sum_{1 \leq i < j \leq k} \frac{1}{2}[X_i(t), Y_i(t)] + \cdots \right)_1 (p), \tag{4.4}$$

where the additional terms are iterated brackets of the vector fields $X_i(t)$ of degree greater than 2. We note that $\sum_{j=1}^k X_j(t)$ is independent of our choice of $\pi \in P_k^0$. Writing the above vector field explicitly as a power series in t,

$$\mathcal{X}_t^\pi(p) = (Q_\pi^0 t + Q_\pi^1 t^2 + \cdots)_1(p),$$

we see that, from the definition of $X_j(t)$, the lowest power of t with a nonzero coefficient will be $t^{m(\boldsymbol{s})}$, where $m(\boldsymbol{s}) = \min\{s_i \mid 1 \leq i \leq k\}$ as above. In particular, $Q_\pi^0, \dots, Q_\pi^{m(\boldsymbol{s})-2}$ are identically zero. Similarly the lowest power of t with a nonzero coefficient in $\sum_{1 \leq i < j \leq k} \frac{1}{2}[X_i(t), Y_j(t)]$ will be $t^{m_0(\boldsymbol{s})+2}$, so that $Q_\pi^{m_0(\boldsymbol{s})+1}$ is the coefficient of the t which could vary with $\pi \in P_k^0$. From our definition of $m_0(\boldsymbol{s})$ we have

$$\mathcal{X}_t^\pi(p) = (Q_\pi^{m(\boldsymbol{s})-1} t^{m(\boldsymbol{s})-1} + \cdots + Q_\pi^{m_0(\boldsymbol{s})+1} t^{m_0(\boldsymbol{s})+1} + \cdots)_t(p),$$

where $Q_\pi^{m(\boldsymbol{s})-1}, \dots, Q_\pi^{m_0(\boldsymbol{s})}$ are invariant with respect to $\pi \in P_k^0$. This proves the lemma in the case $\ell = 0$. Now suppose that the lemma holds for $\pi \in P_k^\ell$. Let $\pi = (\pi_1, \dots, \pi_{\Gamma(k,\ell)}) \in P_k^{\ell+1}$, where $\pi_i \in P_k^\ell$, and set

$$\mathcal{X}_t^\pi(p) = \mathcal{X}_t^{\pi_1} \circ \cdots \circ \mathcal{X}_t^{\pi_{\Gamma(k,\ell)}}(p).$$

By assumption,

$$\mathcal{X}_t^{\pi_i}(p) = (Q_{\pi_i}^{m(\boldsymbol{s})-1} t^{m(\boldsymbol{s})-1} + Q_{\pi_i}^{m(\boldsymbol{s})} t^{m(\boldsymbol{s})} + \cdots)_t(p),$$

where $Q_{\pi_i}^{m(\boldsymbol{s})-1}, \dots, Q_{\pi_i}^{m_\ell(\boldsymbol{s})}$ are independent of π_i. Setting $Q^j = Q_{\pi_i}^j$ for $j = m(\boldsymbol{s}) - 1, \dots, m_\ell(\boldsymbol{s})$, it follows that $\mathcal{X}_t^{\pi_i}(p) = (\bar{X}_i(t))_t(p)$, where

$$\begin{aligned}\bar{X}_i(t) = {} & Q^{m(\boldsymbol{s})-1} t^{m(\boldsymbol{s})-1} + Q^{m(\boldsymbol{s})} t^{m(\boldsymbol{s})} + \cdots + Q^{m_\ell(\boldsymbol{s})} t^{m_\ell(\boldsymbol{s})} \\ & + Q_{\pi_i}^{m_\ell(\boldsymbol{s})+1} t^{m_\ell(\boldsymbol{s})+1} + \cdots.\end{aligned}$$

As in (4.4), the Campbell–Baker–Hausdorff formula yields an expression for $\mathcal{X}_t^\pi$ with $\bar{X}_i(t)$ replacing $X_i(t)$. Arguing as in the case $\ell = 0$ above, we can conclude that

$$\begin{aligned}\mathcal{X}_t^\pi(p) = \big(& \Gamma(k,\ell) Q^{m(\boldsymbol{s})-1} t^{m(\boldsymbol{s})-1} + \cdots + \Gamma(k,\ell) Q^{m_\ell(\boldsymbol{s})} t^{m_\ell(\boldsymbol{s})} \\ & + (Q_{\pi_1}^{m_\ell(\boldsymbol{s})+1} + \cdots + Q_{\pi_{\Gamma(k,\ell)}}^{m_\ell(\boldsymbol{s})+1}) t^{m_\ell(\boldsymbol{s})+1} + \cdots \\ & + (Q_{\pi_1}^{m_\ell(\boldsymbol{s})+m} + \cdots + Q_{\pi_{\Gamma(k,\ell)}}^{m_\ell(\boldsymbol{s})+m}) t^{m_\ell(\boldsymbol{s})+m} \\ & + Q_\pi^{m_\ell(\boldsymbol{s})+m+1} t^{m_\ell(\boldsymbol{s})+m+1} + \cdots \big)_t(p).\end{aligned}$$

Since $m_\ell(\boldsymbol{s}) + m = m_{\ell+1}(\boldsymbol{s})$ and in the above equation the coefficients of t^i with $i \le m_{\ell+1}(\boldsymbol{s})$ are π-invariant, the induction is complete. □

Let $\pi \in P_k^\ell$. In light of Lemma 4.1 we set

$$Q_{\text{inv}}^i = Q_\pi^i, \quad m_{\ell-1}(\boldsymbol{s}) < i \le m_\ell(\boldsymbol{s}),$$

where $Q_{\text{inv}}^i = Q_{\text{inv}}^i(X^1, Y^1, \dots, X^k, Y^k, \boldsymbol{s})$ depends on X^i, Y^i and $\boldsymbol{s}$ is independent of π. For $\ell = 0$ we set

$$Q_{\text{inv}}^i = Q_\pi^i, \quad i \in \{0, 1, \dots, m_0(\boldsymbol{s})\}.$$

In the case $\boldsymbol{s} = (1, \dots, 1)$ it is straightforward to show that $m_\ell(\boldsymbol{s}) = \ell$ and

$$\begin{aligned}Q_{\text{inv}}^0 &= A^0(X^1, Y^1) + \cdots + A^0(X^k, Y^k), \\ Q_{\text{inv}}^1 &= k!(A^1(X^1, Y^1) + \cdots + A^1(X^k, Y^k)), \\ Q_{\text{inv}}^2 &= (k!)^2(A^2(X^1, Y^1) + \cdots + A^2(X^k, Y^k)) + B,\end{aligned}$$

where B is a linear combination of degree 3 brackets of the vector fields $A^0(X^i, Y^i)$. For our application, the pairs $\{X^i, Y^i\}$ above will be complementary at p so that $A^0(X^i, Y^i)$ and hence B vanish at p.

The following proposition relates the definition of Q_{inv}^i to the S-trajectory corresponding to $\pi \in P_k^\ell$, where $i \le m_\ell(\boldsymbol{s})$.

PROPOSITION 4.2. *For each $\ell \ge 0$ and $\pi \in P_k^\ell$ there corresponds an S-trajectory of the form*

$$\mathcal{X}_t^\pi(p) = \big(\alpha_0 Q_{\text{inv}}^0 + \cdots + \alpha_{m_\ell(\boldsymbol{s})} Q_{\text{inv}}^{m_\ell(\boldsymbol{s})} t^{m_\ell(\boldsymbol{s})} + Q_\pi^{m_\ell(\boldsymbol{s})+1} t^{m_\ell(\boldsymbol{s})+1} + \cdots\big)_t(p),$$

where $\alpha_i > 0$, $i \in \{0, 1, \dots, m_\ell(\boldsymbol{s})\}$.

Proof. The proof of Lemma 4.1 contains this result with a slight change of notation using the subscript "inv" to keep track of the vector fields invariant with respect to the appropriate collection of permutations. □

Remark 4.3. In the case of a single-input affine system $\dot{x} = f(x)+ug(x)$, consider the sets $\{X^1 = f+g, Y^1 = f-g\}$ and $\{X^2 = f-g, Y^2 = f+g\}$, and take $s_1 = s_2 = 1$ to compute

$$
\begin{aligned}
Q^0_{\text{inv}} &= 4f,\\
Q^2_{\text{inv}} &= \frac{8}{3}\text{ad}_g^2 f,\\
Q^4_{\text{inv}} &= \frac{8}{15}\text{ad}_g^4 f + \frac{56}{45}\text{ad}_g\text{ad}_f^3 g - \frac{496}{45}[\text{ad}_f g, \text{ad}_f^2 g],\\
Q^6_{\text{inv}} &= \frac{1136}{315}[[f,g],\text{ad}_f^4 g] - \frac{119912}{945}[\text{ad}_f^2 g, \text{ad}_f^3 g] + \frac{32}{105}\text{ad}_g^3\text{ad}_f^3 g\\
&\quad - \frac{1024}{315}[\text{ad}_f g, \text{ad}_g^3\text{ad}_f^2 g] + \frac{3376}{315}[\text{ad}_f g, [\text{ad}_f g, \text{ad}_g^2 f]] - \frac{144}{35}[\text{ad}_f^2 g, \text{ad}_g^3 f]\\
&\quad + \frac{16}{315}\text{ad}_g^6 f + \frac{176}{945}[\text{ad}_g^2 f, [g, \text{ad}_f^2 g]] - \frac{176}{945}[g, \text{ad}_f^5 g].
\end{aligned}
$$

These are linear combinations of *bad* vector fields as per [12]. We show in Corollary 4.8 that, for two pairs of complementary vector fields, $Q^\ell_{\text{inv}} = 0$ for ℓ odd. We also remark that the eccentric character of the coefficients in the expressions for the permutation-invariant brackets is a consequence of our use of the Campbell–Baker–Hausdorff formula.

Remark 4.4. In a given example, one may have many more permutation-invariant vector fields than the Q^i_{inv}, which are invariant on essentially the free Lie algebra level.

4.3. Applications to local controllability. In this section we summarize the above developments as they apply to conditions for local controllability. The following result relates the permutation-dependent constructions to the more general constructions of section 3.2.

THEOREM 4.5. *Suppose that* $\{X^i, Y^i\} \subset S$ *for* $i = 1, \dots, k$, $\boldsymbol{s} \in \mathbb{N}^k$, *and* $Q^0_{\text{inv}}(p) = Q^1_{\text{inv}}(p) = \cdots = Q^{m_\ell(\boldsymbol{s})}_{\text{inv}}(p) = 0$. *Then*

1. $Q^{m_\ell(\boldsymbol{s})+1}_\pi \in S^{m_\ell(\boldsymbol{s})+1}_p$ *for all* $\pi \in P^\ell_k$ *and*
2. $Q^{m_\ell(\boldsymbol{s})+1}_{\text{inv}} \in S^{m_\ell(\boldsymbol{s})+1}_p$.

The next three corollaries specialize the theorem to interesting cases. The first deals with the case when all time rescalings are equal to 1. In practice, this will often be the case, but in Remark 4.10 we provide a situation where it is beneficial to allow the more general class of rescalings.

COROLLARY 4.6. *Suppose that* $\{X^i, Y^i\} \subset S$ *for* $i = 1, \dots, k$ *and* $\boldsymbol{s} = (1, \dots, 1)$. *Then*

1. *if* $Q^0_{\text{inv}}(p) = Q^1_{\text{inv}}(p) = \cdots = Q^\ell_{\text{inv}}(p) = 0$, *then* $Q^{\ell+1}_\pi \in S^{\ell+1}_p$ *for all* $\pi \in P^\ell_k$ *and*
2. *if* $\{X^i, Y^i\}$ *are complementary at* p *for* $i = 1, \dots, k$, *then*
   (a) $\pm\text{ad}_{X^i} Y^i \in S^1_p$,
   (b) $2\text{ad}^2_{X^i} Y^i - \text{ad}^2_{Y^i} X^i \in S^2_p$ *and* $\text{ad}^2_{X^i} Y^i \in S^2_p$, *and*
   (c) $\pm\text{ad}_{Y^i}\text{ad}^2_{X^i} Y^i \in S^3_p$ *if* $\sum_{i=1}^k (\text{ad}^2_{X^i} Y^i + \text{ad}^2_{Y^i} X^i)(p) = 0$.

Our next result specializes Theorem 4.5 to two pairs of vector fields.

COROLLARY 4.7. *Suppose that* $\{X^1, Y^1\}, \{X^2, Y^2\} \subset S$, $\boldsymbol{s} = (1,1)$, *and* $Q^0_{\text{inv}}(p) = Q^1_{\text{inv}}(p) = \cdots = Q^\ell_{\text{inv}}(p) = 0$. *Then*

1. $Q^{\ell+1}_{\text{inv}} \in S^{\ell+1}_p$ *and*
2. $-Q^{\ell+1}_{\text{inv}} \in S^{\ell+1}_p$ *and* $Q^{\ell+2}_{\text{inv}} \in S^{\ell+2}_p$ *if* ℓ *is even.*

Finally, we consider the case of a single pair of vector fields. In practice, this simple result is often the most useful, as we shall see in section 5.

COROLLARY 4.8. *Suppose* $\{X, Y\} \subset S$. *The following statements hold:*

1. *if* $\boldsymbol{s} = (1,1)$, *then for the pairs* $\{X^1, Y^1\} = \{X, Y\}$ *and* $\{X^2, Y^2\} = \{Y, X\}$ *we have* $Q^\ell_{\text{inv}} = 0$ *for* ℓ *odd;*
2. *if* $\boldsymbol{s} = (1)$, *then* $Q^\ell_{\text{inv}} = A^\ell(X, Y)$. *In particular,* $A^k(X, Y)(p) = 0$, $k \in \{0, 1, \ldots, \ell\}$, *implies* $A^{\ell+1}(X, Y) \in S^{\ell+1}_p$.

Remark 4.9. We can replace one or more of the pairs $\{X^i, Y^i\}$ in Theorem 4.5 with $\{Y^i, X^i\}$ to generate additional vector fields in $S^{\ell+1}_p$.

Remark 4.10. In Theorem 4.5 the vanishing of the vector fields Q^i_{inv} at p can be replaced by conditions for neutralization resembling those in the existing literature (e.g., [12, 2]). That is, we may ask not that $Q^0_{\text{inv}}(p) = \cdots = Q^\ell_{\text{inv}}(p) = 0$, but that $Q^0_{\text{inv}}(p) = \cdots = Q^{\ell-1}_{\text{inv}}(p) = 0$ and $0 \in \text{co}\{Q^0_{\text{inv}}(p), Q^1_{\text{inv}}(p), \ldots, Q^\ell_{\text{inv}}(p)\}$. More generally, suppose that for $i, j \in \mathbb{N}$ we denote

$$Q^i_{\text{inv}} = Q^i_{\text{inv}}(X^1, Y^1, \ldots, X^k, Y^k, \boldsymbol{s}), \quad \tilde{Q}^j_{\text{inv}} = \tilde{Q}^j_{\text{inv}}(\tilde{X}^1, \tilde{Y}^1, \ldots, \tilde{X}^{\tilde{k}}, \tilde{Y}^{\tilde{k}}, \tilde{\boldsymbol{s}}),$$

and that for a specific $\ell, m \in \mathbb{N}$ we have $Q^\ell_{\text{inv}}(p) + \tilde{Q}^m_{\text{inv}}(p) = 0$. Then we consider the augmented collection of pairs of vector fields

$$\{X^1, Y^1\}, \ldots, \{X^k, Y^k\}, \qquad \{\tilde{X}^1, \tilde{Y}^1\}, \ldots, \{\tilde{X}^{\tilde{k}}, \tilde{Y}^{\tilde{k}}\},$$

and choose $\hat{\boldsymbol{s}} = (ms_1, \ldots, ms_k, \ell\tilde{s}_1, \ldots, \ell\tilde{s}_{\tilde{k}})$. The resulting set of invariant vector fields $\hat{Q}^i_{\text{inv}}$ will have $\hat{Q}^{m\ell-1}_{\text{inv}}$, a positive multiple of $Q^\ell_{\text{inv}} + \tilde{Q}^m_{\text{inv}}$, and for $j < ml - 1$ the vector fields $\hat{Q}^j_{\text{inv}}$ will be linear combinations of $Q^0_{\text{inv}}, \ldots, Q^{\ell-1}_{\text{inv}}, \tilde{Q}^0_{\text{inv}}, \ldots, \tilde{Q}^{m-1}_{\text{inv}}$, and their Lie brackets. The notion of rescaling time to generate new higher-order S-trajectories is inherent in the definition of the sets S^k_p. In other words, if one generates "complementary" variations, they can be scaled to the same order, and then our methodology can be applied to the new variations. Examples 5.3 and 5.4 illustrate this point.

Remark 4.11. Stefani's example [9],

$$\begin{aligned} \dot{x} &= u, \\ \dot{y} &= x, \\ \dot{z} &= x^3 y \end{aligned}$$

in $\mathbb{R}^3$, fits the framework of Corollary 4.7. As noted in Sussmann's paper [12], the Lie brackets in $f = (0, x, x^3 y)$ and $g = (1, 0, 0)$ of degree 3, 4, and 5 vanish at $p = (0, 0, 0)$. Consider $\{X^1 = f+g, Y^1 = f-g\}, \{X^2 = f/2+2g, Y^2 = f/2-2g\} \subset S$ and $\boldsymbol{s} = (1,1)$. Corollary 4.6(2a) implies that $\pm[f, g] \in S^1_p$, while $f \pm g \in S = \text{co}\{f + g, f - g\}$. Thus we can find control variations in the directions $(\pm 1, 0, 0), (0, \pm 1, 0)$. To generate the control variations in the directions $(0, 0 \pm 1)$ we use Corollary 4.6(1). Note that $P^0_2 = \{\pi_1, \pi_2\}$, where $\pi_1(1) = 1, \pi_1(2) = 2$ and $\pi_2(1) = 2, \pi_2(2) = 1$, so that

$$\mathcal{X}^{\pi_1}_t(p) = X^1_t \circ Y^1_t \circ X^2_t \circ Y^2_t(p) = (Q^0_{\text{inv}} + Q^1_{\pi_1} t + Q^2_{\pi_1} t^2 + \cdots)_t(p).$$

However, $Q^0_{\rm inv} = 3f$, which vanishes at p, and Corollary 4.6(1) imply that $Q^1_{\pi_1} \in S^1_p$. However, $Q^1_{\pi_1} = 0$; hence $Q^2_{\pi_1} \in S^2_p$ as a consequence of Proposition 3.6(2). Similarly $Q^2_{\pi_1}, Q^3_{\pi_1}, Q^4_{\pi_1}$ vanish at p, as they consist of linear combinations of Lie brackets in f and g of degree 3, 4, and 5; hence $Q^5_{\pi_1} \in S^5_p$. Likewise $Q^5_{\pi_2} \in S^5_p$. Since $Q^5_{\pi_1}(p) = (0, 0, 21/18)$ and $Q^5_{\pi_2}(p) = (0, 0, -21/18)$, Theorem 3.7 implies local controllability.

Proof (proof of Theorem 4.5). Choose $\pi_j \in P^\ell_k$. Then Proposition 4.2 asserts that there exists an S-trajectory

$$\mathcal{X}^\pi_t(p) = \big(\alpha^j_0 Q^0_{\rm inv} + \cdots + \alpha^j_{m_\ell(\boldsymbol{s})} Q^{m_\ell(\boldsymbol{s})}_{\rm inv} t^{m_\ell(\boldsymbol{s})} + Q^{m_\ell(\boldsymbol{s})+1}_{\pi_j} t^{m_\ell(\boldsymbol{s})+1} + \cdots\big)_t(p).$$

Since $Q^i_{\rm inv}(p)$ vanishes for $0 \le i \le m_\ell(\boldsymbol{s})$, it follows that $Q^{m_\ell(\boldsymbol{s})+1}_{\pi_j} \in S^a_p$ for some $a \in \mathbb{N}$. Here $Q^{m_{\ell+1}(\boldsymbol{s})}_{\pi_j}$ is the coefficient Q_{π_j} of the lowest power of t with the property that Q_{π_j} could vary with $\pi_j \in P^\ell_k$. To determine a we note that, in light of Proposition 3.6(3), $X \in S^a_p$ implies X is a linear combination of brackets of vector fields in S of degrees up to and including $a+1$. To determine the bracket of highest degree in $Q^{m_\ell(\boldsymbol{s})+1}_{\pi_j}$ we can, without loss of generality, assume that $\min\{s_1, \ldots, s_k\} = 1$ and take $s_1 = 1$. (If this is not the case, we can replace s_i with $s_i - (\min\{s_1, \ldots, s_k\} - 1)$ without changing the vector fields Q^j_π.) Then

$$\begin{aligned} Y^1_t \circ X^1_t(p) = \big(&A^0(X^1, Y^1) + A^1(X^1, Y^1)t + A^2(X^1, Y^1)t^2 \\ &+ A^3(X^1, Y^1)t^3 + \cdots\big)_t(p), \end{aligned}$$

which has the consequence that $Q^{m_\ell(\boldsymbol{s})+1}_{\pi_j}$ is a linear combination of brackets of vector fields in S of degrees up to and including $m_\ell(\boldsymbol{s}) + 2$. Thus $a = m_\ell(\boldsymbol{s}) + 1$. Finally, if $P^{\ell+1}_k = \{\pi_1, \ldots, \pi_{\Gamma(k,\ell)}\}$, then we form the S-trajectory

$$\mathcal{X}^\pi_t(p) = \left(\alpha_0 Q^0_{\rm inv} + \cdots + \alpha_{m_\ell(\boldsymbol{s})} Q^{m_\ell(\boldsymbol{s})}_{\rm inv} t^{m_\ell(\boldsymbol{s})} + \sum_{j=1}^{\Gamma(k,\ell)} Q^{m_\ell(\boldsymbol{s})+1}_{\pi_j} t^{m_\ell(\boldsymbol{s})+1} + \cdots\right)_t(p),$$

where $\alpha_i > 0$. Since

$$Q^{m_\ell(\boldsymbol{s})+1}_{\rm inv} = \sum_{j=1}^{\Gamma(k,\ell)} Q^{m_\ell(\boldsymbol{s})+1}_{\pi_j},$$

it follows that $Q^{m_\ell(\boldsymbol{s})+1}_{\rm inv} \in S^{\ell+1}_p$. □

Before proving the corollaries to Theorem 4.5, we establish some technical lemmas.

LEMMA 4.12. *Suppose that P, Q are vector fields on $\Omega \subset \mathbb{R}^n$. Then, for t sufficiently small, the integral curve $Q_t \circ P_t(p) = (\sum_{\ell=0}^\infty M^\ell(P, Q))_t(p)$ for vector fields $M^\ell(P, Q)$ with the following properties:*

1. *$M^\ell(P, Q) = (-1)^\ell M^\ell(Q, P)$;*
2. *if $P_t = \sum_{i=0}^\infty A^i_1 t^i$ and $Q_t = \sum_{i=0}^\infty A^i_2 t^i$, then $M^\ell(P, Q)$ has a power series expansion in t whose coefficients are Lie brackets of the vector fields A^i_j of degree $\ell + 1$.*

Proof. The existence of the vector fields $M^\ell(Q, P)$ follows from the Campbell–Baker–Hausdorff formula, and we have $M^0(P, Q) = P + Q = M^0(Q, P)$ and $M^1(P, Q) = \frac{1}{2}\mathrm{ad}_P Q = -\frac{1}{2}\mathrm{ad}_Q P = -M^1(Q, P)$. Thus $P + Q$ is a linear combination of the vector fields A^i_j, which we call Lie brackets of degree 1, while $M^1(P, Q)$ is a linear combination of the vector fields of the form $[A^i_1, A^j_2]$, which are Lie brackets of degree

2. Thus 1 and 2 hold for $\ell = 0, 1$. We now establish 4.12. Set $M^k = M^k(P,Q)$ and $\bar{M}^k = M^k(Q,P)$. Suppose that $M^j = (-1)^j \bar{M}^j$ for $j < \ell$. For $j = \ell$ the Campbell–Baker–Hausdorff formula [13] asserts that

$$(\ell+1)M^\ell = \frac{1}{2}[P-Q, M^{\ell-1}] + \sum_{\substack{p\geq 1\\ 2p\leq \ell}} K_{2p} V_p(P,Q)$$

with

$$V_p(P,Q) = \sum_{\substack{k_1,\dots,k_{2p}>0\\ k_1+\cdots+k_{2p}=\ell}} [M^{k_1-1}, [M^{k_2-1}, \dots, [M^{k_{2p}-1}, P+Q]\dots]].$$

Hence

$$(\ell+1)\bar{M}^\ell = \frac{1}{2}[Q-P, \bar{M}^{\ell-1}] + \sum_{\substack{p\geq 1\\ 2p\leq \ell}} K_{2p} \bar{V}_p(P,Q)$$

with

$$\bar{V}_p(P,Q) = \sum_{\substack{k_1,\dots,k_{2p}>0\\ k_1+\cdots+k_{2p}=\ell}} [\bar{M}^{k_1-1}, [\bar{M}^{k_2-1}, \dots, [\bar{M}^{k_{2p}-1}, P+Q]\cdots]].$$

By our induction hypothesis we know that $M^j = (-1)^j \bar{M}^j$ for $j < \ell$; thus $\frac{1}{2}[Q-P, \bar{M}^{\ell-1}] = (-1)^\ell \frac{1}{2}[P-Q, M^{\ell-1}]$ and

$$\begin{aligned}
\bar{V}_p(P,Q) &= \sum_{\substack{k_1,\dots,k_{2p}>0\\ k_1+\cdots+k_{2p}=\ell}} [(-1)^{k_1-1} M^{k_1-1}, [\cdots, [(-1)^{k_{2p}-1} M^{k_{2p}-1}, P+Q]\dots]] \\
&= \sum_{\substack{k_1,\dots,k_{2p}>0\\ k_1+\cdots+k_{2p}=\ell}} (-1)^{k_1+\cdots+k_{2p}-2p} [M^{k_1-1}, [\cdots, [M^{k_{2p}-1}, P+Q]\dots]] \\
&= (-1)^\ell V_p(P,Q)
\end{aligned}$$

since $(-1)^{k_1+\cdots+k_{2p}-2p} = (-1)^\ell (-1)^{2p} = (-1)^\ell$. This implies that $M^\ell(P,Q) = (-1)^\ell M^\ell(Q,P)$.

To establish point 2 we note that it holds for $\ell = 0, 1$. Suppose that assertion 2 holds for $j < \ell$. Thus $M^{\ell-1}(P,Q)$ has a power series expansion in t whose coefficients are Lie brackets of the vector fields A^i_j of degree ℓ. Now $P-Q$ has a power series expansion in t whose coefficients are Lie brackets of the vector fields A^i_j of degree 1; hence, in the above formula for $M^\ell(P,Q)$, the term $[P-Q, M^{\ell-1}]$ is a combination of Lie brackets of the vector fields A^i_j of degree $\ell+1$. The remaining terms in $M^\ell(P,Q)$ involve the vector fields $V_p(P,Q)$. By our induction hypothesis the vector fields M^{k_i-1} in $V_p(P,Q)$ involve Lie brackets of the vector fields A^i_j of degree k_i. Since $P+Q$ involve Lie brackets of the vector fields A^i_j of degree 1, it follows that $[M^{k_1-1}, [M^{k_2-1}, \dots, [M^{k_{2p}-1}, P+Q]\dots]]$ has a power series expansion in t whose coefficients are Lie brackets of the vector fields A^i_j of degree $k_1+\cdots+k_{2p}+1 = \ell+1$. This completes the induction. □

LEMMA 4.13. *Suppose that* $\{X^1, Y^1\}, \{X^2, Y^2\} \subset S$ *and* $\boldsymbol{s} = (1,1)$. *Then* Q^ℓ_{inv} *is a linear combination of Lie brackets of odd degree of the vector fields* $A^i_j = A^i(X^j, Y^j)$ *for all* $\ell \geq 0$.

In particular, $Q^\ell_{\text{inv}}(X^1,Y^1,X^2,Y^2,\boldsymbol{s}) = (-1)^\ell Q^\ell_{\text{inv}}(Y^1,X^1,Y^2,X^2,\boldsymbol{s})$.

Proof. We begin by examining the S-trajectories which correspond to permutation in P_2^0. By definition, $P_2^0 = \{\pi_1,\pi_2\}$, where $\pi_1(1) = 1, \pi_1(2) = 2$ and $\pi_2(1) = 2, \pi_2(2) = 1$. Then

$$\mathcal{X}_t^{\pi_1}(p) = X_t^1 \circ Y_t^1 \circ X_t^2 \circ Y_t^2(p) = \left(\sum_{i=0}^\infty A_1^i t^i\right)_t \circ \left(\sum_{i=0}^\infty A_2^i t^i\right)_t (p),$$

where $A_1^i = A^i(X^1,Y^1)$ and $A_2^i = A^i(X^2,Y^2)$. Set $P = \sum_{i=0}^\infty A_1^i t^i, Q = \sum_{i=0}^\infty A_2^i t^i$ so P and Q are power series in t whose coefficients are Lie brackets of the vector fields A_j^i of degree 1 (odd). Thus $\mathcal{X}_t^{\pi_1}(p) = P_t \circ Q_t(p)$ and, in light of Lemma 4.12, there exist vector fields $M^i(P,Q)$ such that $\mathcal{X}_t^{\pi_1}(p) = (\sum_{i=0}^\infty M^i(P,Q))_t(p)$. Since $\mathcal{X}_t^{\pi_2}(p) = Q_t \circ P_t(p)$, we have

$$\mathcal{X}_t^{\pi_2}(p) = \left(\sum_{i=0}^\infty M^i(Q,P)\right)_t (p) = \left(\sum_{i=0}^\infty (-1)^i M^i(P,Q)\right)_t (p),$$

where $M^\ell(P,Q)$ is a power series in t whose coefficients are Lie brackets of the vector fields A_j^i of degree $\ell+1$. We let $M^{\text{odd}}(P,Q) = \sum_{i=0}^\infty M^{2i}(P,Q)$ and $M^{\text{even}}(P,Q) = \sum_{i=0}^\infty M^{2i+1}(P,Q)$ so that $M^{\text{odd}}(P,Q)$ ($M^{\text{even}}(P,Q)$) is a power series in t whose coefficients are Lie brackets of the vector fields A_j^i of odd (even) degree. Furthermore $M^{\text{odd}}(P,Q) = M^{\text{odd}}(Q,P)$, while $M^{\text{even}}(P,Q) = -M^{\text{even}}(Q,P)$. We now explore the same issues for $P_2^1 = \{\hat\pi_1,\hat\pi_2\}$, where $\hat\pi_1 = (\pi_1,\pi_2)$ and $\hat\pi_2 = (\pi_2,\pi_1)$ for the permutations $\pi_1,\pi_2 \in P_2^0$ defined above. Setting $\hat P = \sum_{i=0}^\infty M^i(P,Q)$ and $\hat Q = \sum_{i=0}^\infty M^i(Q,P)$, we have, as above,

$$\mathcal{X}_t^{\hat\pi_1}(p) = \mathcal{X}_t^{\pi_1} \circ \mathcal{X}_t^{\pi_2}(p) = \hat P_t \circ \hat Q_t(p).$$

From Lemma 4.12 there now exist vector fields $\hat M^\ell(\hat P,\hat Q)$ such that $\hat M^\ell(\hat Q,\hat P) = (-1)^\ell \hat M^\ell(\hat P,\hat Q)$; hence

$$\mathcal{X}_t^{\hat\pi_1}(p) = \left(\sum_{\ell=0}^\infty \hat M^\ell(\hat P,\hat Q)\right)_t (p), \quad \mathcal{X}_t^{\hat\pi_2}(p) = \left(\sum_{\ell=0}^\infty (-1)^\ell \hat M^\ell(\hat P,\hat Q)\right)_t (p). \tag{4.5}$$

We now establish that the vector fields $\hat M^\ell(\hat P,\hat Q)$ are power series in t whose coefficients are Lie brackets of the vector fields A_j^i of odd degree. We showed above that we have $\hat P = \sum_{i=0}^\infty M^i(P,Q) = M^{\text{odd}}(P,Q) + M^{\text{even}}(P,Q)$, while $\hat Q = \sum_{i=0}^\infty M^i(Q,P) = M^{\text{odd}}(P,Q) - M^{\text{even}}(P,Q)$. From the Campbell–Baker–Hausdorff formula we know that $\hat M^0 = \hat P + \hat Q = 2M^{\text{odd}}(P,Q)$, a power series in t whose coefficients are Lie brackets of the vector fields A_j^i of odd degree. Also

$$\hat M^1 = \frac{1}{2}[\hat P,\hat Q] = [M^{\text{even}}(P,Q), M^{\text{odd}}(P,Q)].$$

Since $M^{\text{even}}(P,Q)$ is composed of Lie brackets of A_j^i of even degree and $M^{\text{odd}}(P,Q)$ is composed of Lie brackets of A_j^i of odd degree, it follows that $\hat M^1$ is composed of Lie brackets of A_j^i of odd degree. Now suppose that this holds for $\hat M^2, \hat M^3, \dots, \hat M^{\ell-1}$. From the Campbell–Baker–Hausdorff formula,

$$(\ell+1)\hat M^\ell = \frac{1}{2}[2M^{\text{even}}(P,Q), M^{\ell-1}] + \sum_{\substack{p\ge 1\\ 2p\le \ell}} K_{2p} V_p(P,Q)$$

with

$$V_p(P,Q) = \sum_{\substack{k_1,\dots,k_{2p}>0 \\ k_1+\cdots+k_{2p}=\ell}} [\hat{M}^{k_1-1},[\hat{M}^{k_2-1},\dots,[\hat{M}^{k_{2p}-1},2M^{\mathrm{odd}}(P,Q)]\cdots]].$$

By our induction hypothesis and the fact that $M^{\mathrm{even}}(P,Q)$ is composed of Lie brackets of A_j^i of even degree, we see that $[2M^{\mathrm{even}}(P,Q),M^{\ell-1}]$ is composed of Lie brackets of A_j^i of odd degree. Looking at the terms in $V_p(P,Q)$, we note that

$$[\hat{M}^{k_1-1},\dots,[\hat{M}^{k_{2p}-1},2M^{\mathrm{odd}}(P,Q)]\cdots]$$

has an even number of terms M^{k_i-1}, and $M^{\mathrm{odd}}(P,Q)$ is composed of Lie brackets of A_j^i of odd degree; it follows that $\hat{M}^\ell$ is composed of Lie brackets of A_j^i of odd degree. We can now repeat the initial argument to show that if $P_2^2=\{\pi_1,\pi_2\}$, then there exist vector fields $M^{\mathrm{even}}(P,Q)$ composed of Lie brackets of A_j^i of even degree and $M^{\mathrm{odd}}(P,Q)$ composed of Lie brackets of A_j^i of odd degree such that

$$\mathcal{X}_t^{\pi_1}(p) = \left(\sum_{i=0}^{\infty} M^i(P,Q)\right)_t(p) = (M^{\mathrm{odd}}(P,Q)+M^{\mathrm{even}}(P,Q))_t(p)$$

and

$$\begin{aligned}\mathcal{X}_t^{\pi_2}(p) &= \left(\sum_{i=0}^{\infty} M^i(Q,P)\right)_t(p) = \left(\sum_{i=0}^{\infty}(-1)^i M^i(P,Q)\right)_t(p)\\ &= (M^{\mathrm{odd}}(P,Q)-M^{\mathrm{even}}(P,Q))_t(p).\end{aligned}$$

We simply repeat the above steps for $P_2^3,P_2^4,\dots$ to conclude that the vector fields $\hat{M}^\ell(\hat{P},\hat{Q})$ are power series in t whose coefficients are Lie brackets of the vector fields A_j^i of odd degree.

We now are in a position to verify that Q_{inv}^ℓ is a linear combination of Lie brackets of the vector fields A_j^i of odd degree. We begin by choosing any $\pi\in P_2^k$. Then we know from above that the corresponding S-trajectory $\mathcal{X}_t^\pi(p) = (M^{\mathrm{odd}}(P,Q)+M^{\mathrm{even}}(P,Q))_t(p)$, where $M^{\mathrm{even}}(P,Q)$ is composed of Lie brackets of A_j^i of even degree and $M^{\mathrm{odd}}(P,Q)$ is composed of Lie brackets of A_j^i of odd degree. In the case where k is odd we showed that $M^{\mathrm{even}}(P,Q)=0$. Suppose $k=1$. Then $\mathcal{X}_t^\pi(p) = (M^{\mathrm{odd}}(P,Q))_t(p)$, where $M^{\mathrm{odd}}(P,Q)$ is a power series in t whose coefficients are Lie brackets of the vector fields A_j^i of odd degree. Thus there exist vector fields $Q_{0,1}^{\mathrm{odd}},Q_{1,1}^{\mathrm{odd}},\dots$, which are linear combinations of Lie brackets of the vector fields A_j^i of odd degree, such that

$$\mathcal{X}_t^\pi(p) = (Q_{0,1}^{\mathrm{odd}}+Q_{1,1}^{\mathrm{odd}}t+Q_{2,1}^{\mathrm{odd}}t^2+\cdots)_t(p).$$

However, from Proposition 4.2 we have

$$\mathcal{X}_t^\pi(p)F = (\alpha_0 Q_{\mathrm{inv}}^0+\alpha_1 Q_{\mathrm{inv}}^1 t+Q_\pi^2 t^2+\cdots)_t(p).$$

This means that $\alpha_0 Q_{\mathrm{inv}}^0 = Q_{0,1}^{\mathrm{odd}},\alpha_1$, $Q_{\mathrm{inv}}^1 = Q_{1,1}^{\mathrm{odd}}$, and Q_π^2 are linear combinations of Lie brackets of the vector fields A_j^i of odd degree. Next we consider the case where $\pi\in P_2^2=\{\pi_1,\pi_2\}$. Then

$$\mathcal{X}_t^{\pi_1}(p) = (M^{\mathrm{odd}}(P,Q)+M^{\mathrm{even}}(P,Q))_t(p),$$

where $M^{\text{odd}}(P,Q)$ (resp., $M^{\text{even}}(P,Q)$) is a power series in t whose coefficients are Lie brackets of the vector fields A^i_j of odd (resp., even) degree. Thus there exist vector fields $Q^{\text{odd}}_{0,1}, Q^{\text{odd}}_{1,1}, \dots$, and $Q^{\text{even}}_{0,1}, Q^{\text{even}}_{1,1}, \dots$, which are linear combinations of Lie brackets of the vector fields A^i_j of odd and even degrees, respectively, such that

$$\mathcal{X}^{\pi_1}_t(p) = ((Q^{\text{odd}}_{0,1} + Q^{\text{even}}_{0,1}) + (Q^{\text{odd}}_{1,1} + Q^{\text{even}}_{1,1})t + (Q^{\text{odd}}_{2,1} + Q^{\text{even}}_{2,1})t^2 + \cdots)_t(p).$$

However, from Proposition 4.2 we have

$$\mathcal{X}^{\pi_1}_t(p) = (\alpha_0 Q^0_{\text{inv}} + \alpha_1 Q^1_{\text{inv}} t + \alpha_2 Q^2_{\text{inv}} t^2 + Q^3_{\pi_1} t^3 + \cdots)_t(p).$$

Similarly, using the expansion for $\mathcal{X}^{\pi_2}_t(p)$ in (4.5),

$$\begin{aligned}\mathcal{X}^{\pi_2}_t(p) &= (\alpha_0 Q^0_{\text{inv}} + \alpha_1 Q^1_{\text{inv}} t + \alpha_2 Q^2_{\text{inv}} t^2 + Q^3_{\pi_2} t^3 + \cdots)_t(p)\\ &= ((Q^{\text{odd}}_{0,1} - Q^{\text{even}}_{0,1}) + (Q^{\text{odd}}_{1,1} - Q^{\text{even}}_{1,1})t + (Q^{\text{odd}}_{2,1} - Q^{\text{even}}_{2,1})t^2 + \cdots)_t(p).\end{aligned}$$

Since Q^ℓ_{inv} is invariant with respect to our choice of permutation in $P^{\ell+1}_2$, we can conclude that $Q^{\text{even}}_{0,1} = Q^{\text{even}}_{1,1} = Q^{\text{even}}_{2,1} = 0$. This in turn implies that Q^0_{inv}, Q^1_{inv}, and Q^2_{inv} are linear combinations of Lie brackets of the vector fields A^i_j of odd degree. It is straightforward to show by induction that this is the case for all Q^ℓ_{inv}.

Finally we show that

$$Q^\ell_{\text{inv}}(X^1, Y^1, X^2, Y^2) = (-1)^\ell Q^\ell_{\text{inv}}(Y^1, X^1, Y^2, X^2).$$

Using Lemma 4.12 with $P = X, Q = Y$, we conclude that $A^i(X,Y) = (-1)^i A^i(Y,X)$. The vector fields A^i_j enter into our S-trajectory in the power series $\sum_{i=0}^\infty A^i(X^j, Y^j)t^i$. Since Q^ℓ_{inv} is the coefficient of t^ℓ in a power series expansion of a similar S-trajectory, we can conclude that Q^ℓ_{inv} is a linear combination of iterated Lie brackets

$$B = [A^{i_1}_{j_1}, [A^{i_2}_{j_2}, \dots, [A^{i_{2k}}_{j_{2k}}, A^{i_{2k+1}}_{j_{2k+1}}] \dots]]$$

of an odd number of A^i_j's, where $j_m \in \{1,2\}$ and $i_1 + \cdots + i_{2k+1} = \ell - 2k$. In light of Lemma 4.12 with $Q = Y, P = X$, we know that if i_m is even, then $A^{i_m}_j(Y^j, X^j) = A^{i_m}_j(X^j, Y^j)$, and if i_m is odd, then $A^{i_m}_j(Y^j, X^j) = -A^{i_m}_j(X^j, Y^j)$ for $j \in \{1,2\}$. If ℓ is even, then there must be an even number of integers in $\{i_1, \dots, i_{2k+1}\}$ which are odd, and hence B does not change sign when X^i and Y^i are interchanged. This completes the proof. □

Proof of Corollary 4.6. Suppose that $\boldsymbol{s} = (1, \dots, 1)$. Then 4.6 follows from Theorem 4.5 and the observation that in the case $\boldsymbol{s} = (1, \dots, 1)$ we have $m_i(\boldsymbol{s}) = i$. Suppose that the subsets $\{X^i, Y^i\} \subset S$ are complementary at p for $i = 1, \dots, k$. From Remark 3.3 (or from the definition of $A^1(X^i, Y^i)$) we know that $\text{ad}_{X^i} Y^i \in S^1_p$. Also $\{X^i, Y^i\}$ complementary at p implies $\{Y^i, X^i\}$ complementary at p; hence $-\text{ad}_{X^i} Y^i \in S^1_p$. This gives part 2(a) of the corollary. To establish 2(b) we can use Lemma 4.13 with the choices $X^1 = X^i, Y^1 = Y^i, X^2 = Y^i, Y^2 = X^i$ to conclude $Q^1_{\text{inv}}(X^1, Y^1, X^2, Y^2, \boldsymbol{s}) = -Q^1_{\text{inv}}(X^2, Y^2, X^1, Y^1, \boldsymbol{s})$. Since Q^1_{inv} is invariant with respect to permutations of $\{1,2\}$, we conclude that $Q^1_{\text{inv}}(X^1, Y^1, X^2, Y^2, \boldsymbol{s}) = 0$. As a result of Theorem 4.5, we have $Q^2_\pi \in S^2_p$ for all $\pi \in P^1_2$. One can easily check from the definition that $Q^2_{\text{inv}} = (\text{ad}^2_{X^i} Y^i + \text{ad}^2_{Y^i} X^i)/6$, while $Q^2_\pi = 2\text{ad}^2_{X^i} Y^i - \text{ad}^2_{Y^i} X^i$ for all $\pi \in P^1_2$. Finally, if we reverse X^i and Y^i, we get $2\text{ad}^2_{Y^i} X^i - \text{ad}^2_{X^i} Y^i$, and $\text{co}\{2\text{ad}^2_{Y^i} X^i - \text{ad}^2_{X^i} Y^i, 2\text{ad}^2_{X^i} Y^i - \text{ad}^2_{Y^i} X^i\}$ contains a positive multiple of $\text{ad}^2_{X^i} Y^i$;

hence $\mathrm{ad}^2_{X^i}Y^i \in S^2_p$. To complete the proof we must show that 2(c) holds. Here we simply augment our set of complementary vector fields by adding in k additional pairs, namely those of the form (Y^i, X^i). Arguing as in the proof of 2(b) above, we find that $Q^1_{\mathrm{inv}} = 0$, $Q^2_{\mathrm{inv}} = \sum_{i=1}^k(\mathrm{ad}^2_{X^i}Y^i + \mathrm{ad}^2_{Y^i}X^i)$; hence $Q^2_{\mathrm{inv}}(p) = 0$ and $Q^3_{\mathrm{inv}} = \sum_{i=1}^k \mathrm{ad}_{Y^i}\mathrm{ad}^2_{X^i}Y^i$. Thus Theorem 4.5 implies $\sum_{i=1}^k \mathrm{ad}_{Y^i}\mathrm{ad}^2_{X^i}Y^i \in S^3_p$. Now we note that reversing X^i and Y^i in $\mathrm{ad}_{Y^i}\mathrm{ad}^2_{X^i}Y^i$ gives the negative of this vector field. In this way we can isolate each term in the above sum and conclude that $\pm\mathrm{ad}_{Y^i}\mathrm{ad}^2_{X^i}Y^i \in S^3_p$. □

Proof of Corollary 4.7. Suppose that $\{X^1, Y^1\}, \{X^2, Y^2\} \subset S$, $\boldsymbol{s} = (1, 1)$, and $Q^0_{\mathrm{inv}}(p) = Q^1_{\mathrm{inv}}(p) = \cdots = Q^\ell_{\mathrm{inv}}(p) = 0$. We begin by establishing assertion 4.7. We have $Q^{\ell+1}_\pi \in S^{\ell+1}_p$ by Corollary 4.6 for any $\pi \in P^\ell_2$. Since $Q^{\ell+1}_{\mathrm{inv}}$ is a linear combination of the vector fields $Q^{\ell+1}_\pi$ using positive coefficients, it follows that $Q^{\ell+1}_{\mathrm{inv}} \in S^{\ell+1}_p$. Alternatively, from Proposition 4.2, there is an S-trajectory of the form

$$\mathcal{X}_t(p) = \big(\alpha_0 Q^0_{\mathrm{inv}} + \cdots + \alpha_\ell Q^\ell_{\mathrm{inv}} t^\ell + \alpha_{\ell+1} Q^{\ell+1}_{\mathrm{inv}} t^{\ell+1} + Q^{\ell+2}_{\mathrm{inv}} t^{\ell+2} + \cdots\big)_t(p),$$

where $\alpha_i > 0$. Since $Q^0_{\mathrm{inv}}(p) = Q^1_{\mathrm{inv}}(p) = \cdots = Q^\ell_{\mathrm{inv}}(p) = 0$, we have $Q^{\ell+1}_{\mathrm{inv}} \in S^{\ell+1}_p$.

To establish 4.7 we note that

$$\begin{aligned} Q^{\ell+1}_{\mathrm{inv}}(X^1, Y^1, X^2, Y^2, \boldsymbol{s}) &= (-1)^{\ell+1} Q^{\ell+1}_{\mathrm{inv}}(Y^1, X^1, Y^2, X^2, \boldsymbol{s}) \\ &= -Q^{\ell+1}_{\mathrm{inv}}(X^1, Y^1, X^2, Y^2, \boldsymbol{s}) \end{aligned}$$

as a consequence of Lemma 4.13 and the assumption that $\ell + 1$ is odd. Similarly

$$Q^{\ell+2}_{\mathrm{inv}}(X^1, Y^1, X^2, Y^2, \boldsymbol{s}) = Q^{\ell+2}_{\mathrm{inv}}(Y^1, X^1, Y^2, X^2, \boldsymbol{s}).$$

Thus we can proceed as above using $\{Y^1, X^1\}, \{Y^2, X^2\} \subset S$, $\boldsymbol{s} = (1, 1)$ instead of $\{X^1, Y^1\}, \{X^2, Y^2\} \subset S$, to form an S-trajectory

$$\hat{\mathcal{X}}_t(p) = \big(\alpha_0 Q^0_{\mathrm{inv}} + \cdots + \alpha_\ell Q^\ell_{\mathrm{inv}} t^\ell - \alpha_{\ell+1} Q^{\ell+1}_{\mathrm{inv}} t^{\ell+1} + Q^{\ell+2}_{\mathrm{inv}} t^{\ell+2} + \cdots\big)_t(p),$$

and conclude that

$$-Q^{\ell+1}_{\mathrm{inv}} = -Q^{\ell+1}_{\mathrm{inv}}(X^1, Y^1, X^2, Y^2, \boldsymbol{s}) = Q^{\ell+1}_{\mathrm{inv}}(Y^1, X^1, Y^2, X^2, \boldsymbol{s}) \in S^{\ell+1}_p.$$

Finally, we form the S-trajectory

$$\hat{\mathcal{X}}_t \circ \mathcal{X}_{2t} \circ \hat{\mathcal{X}}_t(p) = \big(4\alpha_0 Q^0_{\mathrm{inv}} + \cdots + 4\alpha_\ell Q^\ell_{\mathrm{inv}} t^\ell + 4Q^{\ell+2}_{\mathrm{inv}} t^{\ell+2} + \cdots\big)_t(p)$$

and note that $Q^0_{\mathrm{inv}}(p) = Q^1_{\mathrm{inv}}(p) = \cdots = Q^\ell_{\mathrm{inv}}(p) = 0$ implies $Q^{\ell+2}_{\mathrm{inv}} \in S^{\ell+2}_p$. □

Proof of Corollary 4.8. The proof relies on Lemma 4.13 together with the fact that reversing the roles of $\{X^1, Y^1\}$ and $\{X^2, Y^2\}$ is, in this case, the same as interchanging X^i and Y^i. Thus permutation-invariance means that Q^ℓ_{inv} vanishes for ℓ odd, establishing 4.8. If $\boldsymbol{s} = (1)$, then P^ℓ_1 consists of the single permutation $1 \mapsto 1$, so $Q^\ell_{\mathrm{inv}} = A^\ell(X, Y)$. Then 4.8 follows from Corollary 4.6. □

5. Examples. In the examples, the sets S of vector fields are not convex. As noted in section 3.1, we can replace S by its convex hull without affecting local controllability. Certain of these examples may be treated using existing techniques in the literature. Therefore, such examples should be regarded as being illustrative of our theory, rather than as presenting new ideas. However, we might mention that we do not know of a theory that will cover Example 5.5.

Example 5.1. As in [10], we consider the system $S = \{X, Y, Z\}$ in the plane where, in local coordinates (x, y),

$$X = (1, 0), \quad Y = (-1, x^2), \quad Z = (0, -1).$$

Then

$$[X, Y] = (0, 2x), \quad [X, [X, Y]] = (0, 2),$$
$$[Y, [X, Y]] = (0, -2), \quad [Z, Y] = [Z, X] = (0, 0).$$

For $p = (0, 0)$ we have $L^0(S, p) = \mathbb{R} \times \{0\}$, and hence $Z_p^0 = \{X, Y\}$ and $S_p^1 = \{[X, Y], [Y, X]\}$. This implies that $S(p) = \{(1, 0), (-1, 0), (0, -1)\}$, $S_p^1(p) = \{(0, 0)\}$, and it follows that $L^1(S, p) = L^0(S, p)$ and $Z_p^1 = \{X, Y, [X, Y], [Y, X]\}$. Since $X(p) + Y(p) = (0, 0)$ the subset $\{X, Y\} \subset S$ is complementary at p. From Corollary 4.6(2b) we have $\mathrm{ad}_X^2 Y = (0, 2) \in S_p^2$. Thus

$$\mathrm{co}(\{(1, 0), (-1, 0), (0, -1), (0, 2)\}) \subset \mathrm{co}(S(p) \cup S_p^1(p) \cup S_p^2(p))$$

and $0 \in \mathrm{int}(\mathrm{co}(S(p) \cup S_p^1(p) \cup S_p^2(p)))$. Local controllability at p then follows from Theorem 3.7.

Example 5.2. Consider the affine system

$$\begin{aligned} \dot{x} &= u_1, \\ \dot{y} &= u_2, \\ \dot{z} &= x^2 - y^4, \end{aligned} \tag{5.1}$$

with $|u_a| \leq 1$ for $a = 1, 2$ and $p = (x(0), y(0), z(0)) = (0, 0, 0)$. Here the system model is of the form

$$\dot{x}(t) = f_0(x(t)) + u_1(t) f_1(x(t)) + u_2(t) f_2(x(t)),$$

where f_0, f_1, f_2 are smooth vector fields on $\mathbb{R}^3$ which, in local coordinates, are defined by $f_0 = (0, 0, x^2 - y^4)$, $f_1 = (1, 0, 0)$, and $f_2 = (0, 1, 0)$. The nonzero Lie brackets are

$$\begin{aligned} &\mathrm{ad}_{f_1} f_0 = (0, 0, 2x), && \mathrm{ad}_{f_1}^2 f_0 = (0, 0, 2), \\ &\mathrm{ad}_{f_2} f_0 = (0, 0, -4y^3), && \mathrm{ad}_{f_2}^2 f_0 = (0, 0, -12y^2), \\ &\mathrm{ad}_{f_2}^3 f_0 = (0, 0, -24y), && \mathrm{ad}_{f_2}^4 f_0 = (0, 0, -24), \end{aligned}$$

while $\mathrm{ad}_{f_1}^3 f_0 = \mathrm{ad}_{f_0}^k \mathrm{ad}_{f_1}^j f_0 = (0, 0, 0)$ for $j, k \geq 1$ and $\mathrm{ad}_{f_2}^5 f_0 = \mathrm{ad}_{f_0}^k \mathrm{ad}_{f_2}^j f_0 = (0, 0, 0)$ for $j, k \geq 1$. The tangent space to $\mathbb{R}^3$ at p is spanned by $f_1(p)$, $f_2(p)$, and $[f_1, [f_1, f_0]](p)$; hence the first-order sufficient condition Theorem 2.1 cannot be employed. The generalization of Hermes' condition, Theorem 7.3 of [12], does not apply because the "bad" bracket $\mathrm{ad}_{f_1}^2 f_0$ is not expressible in terms of "good" and "bad" brackets of the required orders. On the other hand, the drift vector field f_0 vanishes at p, so that $\{X^1, Y^1\} = \{f_0 + f_1, f_0 - f_1\}$ is complementary at p, as is $\{X^2, Y^2\} = \{f_0 + f_2, f_0 - f_2\}$. In light of (4.1), we have $A^0(X^1, Y^1)(p) = A^1(X^1, Y^1)(p) = (0, 0, 0)$, while $A^2(X^1, Y^1)(p)$ is a positive multiple of $(0, 0, 1)$. Corollary 4.8 lets us conclude that $(0, 0, 1) \in S_p^2(p)$. Similarly $A^i(X^2, Y^2)(p) = (0, 0, 0)$ for $i = 0, 1, 2, 3$, and $A^4(X^2, Y^2)(p)$ is a positive multiple of $(0, 0, -1)$, so that $(0, 0, -1) \in S_p^4(p)$, as a consequence of Corollary 4.8. Finally, we note that $f_0(p) \pm f_1(p) = (\pm 1, 0, 0) \in S(p)$ and

$f_0(p) \pm f_2(p) = (0, \pm 1, 0) \in S(p)$; hence $0 \in \operatorname{int}(\operatorname{co}(S(p) \cup S_p^1(p) \cup \cdots \cup S_p^4(p)))$. Thus system (5.1) is l.c. as a consequence of Theorem 3.7.

The next example illustrates the weakening of the hypotheses of Theorem 4.5 described in Remark 4.10.

Example 5.3. Consider the system $S = \{W, X, Y\}$ in $\mathbb{R}^3$ where, in local coordinates (x, y, z),

$$W = (0, 0, -1), \quad X = (1, z, 0), \quad Y = (-1, 0, x^2).$$

Then

$$[X, Y] = (0, -x^2, 2x), \quad [X, [X, Y]] = (0, -4x, 2),$$
$$[Y, [Y, X]] = (0, -2x, 2), \quad [Y, [X, [X, Y]]] = (0, 4, 4x).$$

We take $p = (0, 0, 0)$. Since $(X + Y)(p) = 0$, we have $\{X, Y\}$ complementary at p. In light of (4.1) and Corollary 4.8, Q^i_{inv} is a positive multiple of $A^i(X, Y)$, and we have the S-trajectory

$$\begin{aligned} X_t \circ Y_t(p) &= \left(A^0(X,Y) + A^1(X,Y)t + A^2(X,Y)t^2 + A^3(X,Y)t^3 + \cdots\right)_t(p) \\ &= \left((X+Y) + \frac{1}{2}\operatorname{ad}_X Y t + \frac{1}{12}(\operatorname{ad}_Y^2 X + \operatorname{ad}_X^2 Y)t^2 \right. \\ &\qquad \left. -\frac{1}{24}\operatorname{ad}_Y \operatorname{ad}_X^2 Y t^3 + \cdots\right)_t (p). \end{aligned}$$

Here $A^0(X, Y)(p) = (X + Y)(p) = 0$ and $A^1(X, Y)(p) = \frac{1}{2}\operatorname{ad}_X Y(p) = 0$. Thus $A^2(X, Y) = \frac{1}{12}(\operatorname{ad}_Y^2 X + \operatorname{ad}_X^2 Y) \in S_p^2$ by Corollary 4.8. We note that

$$W \in S, \quad A^2 = \frac{1}{12}(\operatorname{ad}_Y^2 X + \operatorname{ad}_X^2 Y) \in S_p^2, \quad \left(\frac{1}{3}W + A^2\right)(p) = (0, 0, 0),$$

where $A^i = A^i(X, Y)$. As in Remark 4.10, we consider the pairs $\{\frac{1}{6}W, \frac{1}{6}W\}, \{X, Y\} \subset \operatorname{co}(S)$ and take $\boldsymbol{s} = (3, 1)$. Here $k = 2$, and thus $P_2^0 = \{\pi_1, \pi_2\}$, where $\pi_1(1) = 1, \pi_1(2) = 2, \pi_2(1) = 2, \pi_2(2) = 1$. The S-trajectories (4.2) corresponding to π_1, π_2 are

$$\begin{aligned} \mathcal{X}_t^{\pi_1}(p) &= \left(\frac{1}{6}W\right)_{t^3} \circ \left(\frac{1}{6}W\right)_{t^3} \circ X_t \circ Y_t(p) \\ &= \left(\frac{1}{3}W\right)_{t^3} \circ (A^0 + A^1 t + A^2 t^2 + \cdots)_t(p) \\ &= \left(A^0 + A^1 t + \left(A^2 + \frac{1}{3}W\right)t^2 + \left(A^3 - \frac{1}{6}[A^0, W]\right)t^3 + \cdots\right)_t (p) \\ &= (Q^0_{\pi_1} + Q^1_{\pi_1} t + Q^2_{\pi_1} t^2 + \cdots)_t(p) \end{aligned}$$

and

$$\begin{aligned} \mathcal{X}_t^{\pi_2}(p) &= \left(A^0 + A^1 t + \left(A^2 + \frac{1}{3}W\right)t^2 + \left(A^3 + \frac{1}{6}[A^0, W]\right)t^3 + \cdots\right)_t (p) \\ &= (Q^0_{\pi_2} + Q^1_{\pi_2} t + Q^2_{\pi_2} t^2 + \cdots)_t(p). \end{aligned}$$

Here $m(\boldsymbol{s}) = 1, m_0(\boldsymbol{s}) = 2, m_1(\boldsymbol{s}) = 3, m_2(\boldsymbol{s}) = 4$ and Lemma 4.1 implies that $Q^0_\pi, Q^1_\pi, Q^2_\pi$ are constant functions of $\pi \in P_2^0$, which is shown explicitly above. From

our definition it follows that $Q^0_{\rm inv} = A^0$, $Q^1_{\rm inv} = A^1$, and $Q^2_{\rm inv} = A^2 + \frac{1}{3}W$. Similarly if $\pi = (\pi_1, \pi_2) \in P^1_2$, we have

$$\begin{aligned}\mathcal{X}^\pi_t(p) &= \mathcal{X}^{\pi_1}_t \circ \mathcal{X}^{\pi_2}_t(p)\\ &= \left(2A^0 + 2A^1 t + 2\left(A^2 + \frac{1}{3}W\right)t^2 + 2A^3t^3 + \cdots\right)_t(p);\end{aligned}$$

hence $Q^3_{\rm inv} = 2A^3 = -\frac{1}{12}{\rm ad}_Y{\rm ad}^2_X Y$. Since $Q^0_{\rm inv}$, $Q^1_{\rm inv}$, and $Q^2_{\rm inv}$ each vanish at p, Theorem 4.5 implies that

$$Q^3_{\rm inv} = -\frac{1}{12}{\rm ad}_Y{\rm ad}^2_X Y = \left(0, -\frac{1}{3}, -\frac{1}{3}x\right) \in S^3_p.$$

Interchanging X and Y and repeating the previous steps, we can conclude that $-\frac{1}{12}{\rm ad}_X{\rm ad}^2_Y X = (0, \frac{1}{3}, \frac{1}{3}x) \in S^3_p$. In summary, $S(p)$ contains the vectors

$$W(p) = (0, 0, -1), \quad X(p) = (1, 0, 0), \quad Y(p) = (-1, 0, 0);$$

$S^2_p(p)$ contains the vector $\frac{1}{12}({\rm ad}^2_Y X + {\rm ad}^2_X Y)(p) = (0, 0, \frac{1}{3})$; and $S^3_p(p)$ contains the vectors

$$-\frac{1}{12}{\rm ad}_Y{\rm ad}^2_X Y(p) = \left(0, -\frac{1}{3}, 0\right), \quad -\frac{1}{12}{\rm ad}_X{\rm ad}^2_Y X(p) = \left(0, \frac{1}{3}, 0\right).$$

Thus

$$\begin{aligned}{\rm co}\left(\left\{(\pm 1, 0, 0), \left(0, \pm\frac{1}{3}, 0\right), \left(0, 0, \frac{1}{3}\right), (0, 0, -1)\right\}\right) \subset {\rm co}(S(p) \cup S^1_p(p)\\ \cup S^2_p(p) \cup S^3_p(p))\end{aligned}$$

and $0 \in {\rm int}({\rm co}(S(p) \cup S^1_p(p) \cup S^2_p(p) \cup S^3_p(p)))$. Local controllability at p follows from Theorem 3.7. This example illustrates the use of time rescaling to generate new higher-order S-trajectories (see Remark 4.10).

Example 5.4. Here is a control affine system which has a "bad" bracket that can be neutralized as in [2]:

$$\begin{aligned}\dot{x} &= yz + u_1,\\ \dot{y} &= -xz + u_2,\\ \dot{z} &= -u_2,\end{aligned}$$

with $|u_i| \le 1$, $i = 1, 2$, and $p = (0, 0, 0)$. Here $f = (yz, -xz, 0)$, $g_1 = (1, 0, 0)$, $g_2 = (0, 1, -1)$, and the brackets are

$$\begin{gathered}[f, g_1] = (0, z, 0), \quad [f, g_2] = (y - z, -x, 0),\\ [g_1, [f, g_1]] = (0, 0, 0), \quad [g_2, [f, g_2]] = (2, 0, 0), \quad [g_1, [f, g_2]] = (0, -1, 0).\end{gathered}$$

Motivated by Remark 4.10, we will show that the *bad* bracket $[g_2, [f, g_2]]$ can be neutralized. To this end we set

$$\begin{aligned}S &= \{f + ag_1 + bg_2 \mid -1 \le a, b \le 1\},\\ W &= f + g_1, \quad X = f + g_2, \quad Y = f - g_2,\end{aligned}$$

and consider the pairs $\{\frac{1}{3}W, \frac{1}{3}W\}, \{X, Y\} \subset \mathrm{co}(S)$ and $\boldsymbol{s} = (3, 1)$. With $P_2^0 = \{\pi_1, \pi_2\}$ and $A^i = A^i(X, Y)$ as defined in Example 5.3, the S-trajectory (4.2) corresponding to π_1 is

$$\begin{aligned}
\mathcal{X}_t^{\pi_1}(p) &= \Big(\frac{1}{3}W\Big)_{t^3} \circ \Big(\frac{1}{3}W\Big)_{t^3} \circ X_t \circ Y_t(p) \\
&= \Big(\frac{2}{3}W\Big)_{t^3} \circ (A^0 + A^1 t + A^2 t^2 + \cdots)_t(p) \\
&= \Big(A^0 + A^1 t + \Big(A^2 + \frac{2}{3}W\Big)t^2 + \Big(A^3 - \frac{1}{3}[A^0, W]\Big)t^3 \\
&\quad + \Big(A^4 - \frac{1}{3}[A^1, W] + \frac{1}{18}[A^0, [A^0, W]]\Big)t^4 + \cdots\Big)_t(p) \\
&= (Q_{\pi_1}^0 + Q_{\pi_1}^1 t + Q_{\pi_1}^2 t^2 + \cdots)_t(p).
\end{aligned}$$

Here $A^2 = \frac{1}{3}\mathrm{ad}_{g_2}^2 f$ does not vanish at p but is neutralized in the above S-trajectory, as $A^2 + \frac{2}{3}W$ does vanish at p. It is straightforward to check that $Q_{\pi_1}^0, \ldots, Q_{\pi_1}^3$ vanish at p, while

$$Q_{\pi_1}^4(p) = -\frac{1}{3}[A^1, W](p) = -\frac{1}{3}[g_1, [f, g_2]](p) = \Big(0, \frac{1}{3}, 0\Big).$$

From our definition of S_p^m (or from Proposition 3.6(2) it follows that $(0, \frac{1}{3}, 0) \in S_p^4(p)$. Now we can repeat the above construction with X and Y interchanged to conclude that $(0, -\frac{1}{3}, 0) \in S_p^4(p)$. Since $f \pm g_1, f \pm g_2 \in S$, we have

$$\begin{aligned}
\mathrm{co}\left(\left\{\pm(1,0,0), \pm(0,1,-1), \pm\left(0, \frac{1}{3}, 0\right)\right\}\right) \subset \mathrm{co}(S(p) \cup S_p^1(p) \cup S_p^2(p) \\
\cup S_p^3(p) \cup S_p^4(p))
\end{aligned}$$

and $0 \in \mathrm{int}(\mathrm{co}(S(p) \cup \cdots \cup S_p^4(p)))$. Local controllability at p follows from Theorem 3.7.

Example 5.5. We consider the system on $\mathbb{R}^3$ defined by

$$\begin{aligned}
\dot{x} &= u_1, \\
\dot{y} &= u_2, \\
\dot{z} &= x^2(1 + \tfrac{1}{2}u_2),
\end{aligned}$$

and with $(u_1, u_2) \in U = [-\alpha, \alpha]^2$. We take as our reference point $p = (0, 0, 0)$. For $\alpha < 2$ the system is obviously not STLC (small-time locally controllable) from p ($\dot{z} > 0$ in this case). Let us show that this system *is* controllable if the controls are allowed to be sufficiently large. Some relevant Lie brackets for this system are

$$\begin{gathered}
[f, g_1] = (0, 0, -2x), \quad [f, g_2] = (0, 0, 0), \quad [g_1, g_2] = (0, 0, x), \\
[f, [f, g_1]] = [f, [f, g_2]] = (0, 0, 0), \quad [g_1, [f, g_1]] = (0, 0, -2), \\
[g_2, [f, g_2]] = [g_2, [f, g_1]] = [g_2, [g_1, g_2]] = (0, 0, 0), \quad [g_1, [g_1, g_2]] = (0, 0, 1).
\end{gathered}$$

We define two complementary sets $\{X^1, X^2\} = \{f + \alpha g_1, f - \alpha g_1\}$ and $\{Y^1, Y^2\} = \{f - \alpha g_2, f + \alpha g_2\}$. By Corollary 4.6(2b) we have $\mathrm{ad}_{X^1}^2 X^2(p) = -2\alpha^2[g_1, [f, g_1]](p) \in S_p^2$. Also consider $\pi = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right) \in P_2^0$. By Proposition 4.2 we have

$$\mathcal{X}_t^{\pi}(p) = (Q_{\mathrm{inv}}^0 + tQ_{\pi}^1 + t^2 Q_{\pi}^2 + \cdots)_t(p),$$

where a direct calculation using the Campbell–Baker–Hausdorff formula yields

$$\begin{aligned}
Q^0_{\text{inv}} &= 4f,\\
Q^1_\pi &= 2\alpha[f,g_2] - 2\alpha[f,g_1] - \alpha^2[g_1,g_2],\\
Q^2_\pi &= \alpha[f,[f,g_2]] + \alpha[f,[f,g_1]] + \frac{1}{2}\alpha^2[g_1,[f,g_2]] + \frac{1}{2}\alpha^2[g_2,[f,g_1]]\\
&\quad - \frac{5}{6}\alpha^2[g_2,[f,g_2]] + \frac{1}{2}\alpha^3[g_2,[g_1,g_2]] - \frac{5}{6}\alpha^2[g_1,[f,g_1]] - \frac{1}{2}\alpha^3[g_1,[g_1,g_2]].
\end{aligned}$$

Since $Q^0_{\text{inv}}(p) = 0$, by Corollary 4.6(1), we have $Q^1_\pi \in S^1_p$. Furthermore, since $Q^1_\pi(p) = 0$, by Proposition 3.6(2) we have $Q^2_\pi \in S^2_p$. One can then see that, provided that α is sufficiently large (to be exact, if $\alpha > \frac{10}{3}$), then we have $0 \in \text{int}(\text{co}(S(p) \cup S^2_p))$. Small-time local controllability of this example for the sufficiently large control set now follows from Theorem 3.7. The lower bound of $\frac{10}{3}$ on the size of the control set to ensure small-time local controllability is undoubtedly not sharp.

Acknowledgments. We would like to thank the anonymous reviewers whose comments significantly improved the paper. One reviewer, in particular, read the paper very carefully, and noticed a couple of significant errors which have now been corrected.

REFERENCES

[1] A. A. Agrachev and R. V. Gamkrelidze, *Local controllability and semigroups of diffeomorphisms*, Acta Appl. Math., 32 (1993), pp. 1–57.

[2] R. M. Bianchini and G. Stefani, *Controllability along a trajectory: A variational approach*, SIAM J. Control Optim., 31 (1993), pp. 900–927.

[3] H. Hermes and M. Kawski, *Local controllability of a single input, affine system*, in Nonlinear Analysis and Applications, Lecture Notes in Pure and Appl. Math., 109, Marcel Dekker, New York, 1987, pp. 235–248.

[4] H. Frankowska, *Local controllability of control systems with feedback*, J. Optim. Theory Appl., 60 (1989), pp. 277–296.

[5] M. Kawski, *A necessary condition for local controllability*, in Differential Geometry: The Interface Between Pure and Applied Mathematics, Contemp. Math. 68, AMS, Providence, RI, 1987, pp. 143–155.

[6] M. Kawski, *High-order small-time local controllability*, in Nonlinear Controllability and Optimal Control, Monogr. Textbooks Pure Appl. Math. 133, Marcel Dekker, New York, 1990, pp. 431–467.

[7] M. Kawski, *High-order conditions for local controllability in practice*, in Proceedings of the Symposium on the Mathamatical Theory of Networks and Systems (MTNS 91), Kobe, Japan, 1991.

[8] J.-P. Serre, *Lie Algebras and Lie Groups*, Lecture Notes in Math. 1500, Springer-Verlag, New York, Heidelberg, Berlin, 1992.

[9] G. Stefani, *Local controllability of nonlinear systems: An example*, Systems Control Lett., 6 (1985), pp. 123–125.

[10] H. J. Sussmann, *A sufficient condition for local controllability*, SIAM J. Control Optim., 16 (1978), pp. 790–802.

[11] H. J. Sussmann, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.

[12] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

[13] V. S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representations*, Graduate Texts in Math. 102, Springer-Verlag, New York, Heidelberg, Berlin, 1985 (reprint of 1974 edition by Prentice-Hall).

[14] F. W. Warner, *Foundations of Differentiable Manifolds and Lie Groups*, Graduate Texts in Math. 94, 2nd ed., Springer-Verlag, New York, Heidelberg, Berlin, 1983.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 325–340

ON A REPRESENTATION OF THE LIMIT OCCUPATIONAL MEASURES SET OF A CONTROL SYSTEM WITH APPLICATIONS TO SINGULARLY PERTURBED CONTROL SYSTEMS*

VLADIMIR GAITSGORY†

Abstract. A representation of the limit occupational measures set of a control system in terms of the vector function defining the system's dynamics is established. Applications in averaging of singularly perturbed control systems are demonstrated.

Key words. singularly perturbed control systems, occupational measures, averaging method, limit occupational measures sets, approximation of slow motions

AMS subject classifications. 34E15, 34C29, 34A60, 93C70

DOI. 10.1137/S0363012903424186

1. Introduction and preliminaries.

Under certain conditions, the set of occupational measures generated by admissible controls and corresponding solutions of a control system converges (as the length of the time interval tends to infinity) to a limit set. If this limit set is independent of the initial conditions within some subset of the state space, it is called a limit occupational measures set (LOMS) of the system.

Criteria for the existence of the LOMS were discussed in [24], [25], where it was used as a tool for analysis of singularly perturbed control systems (SPCS).

In this paper, we give a natural representation for the LOMS which enhances its applications in SPCS. Our main results are stated in Theorem 2.1. We establish that, under the assumptions made, the convex hull of a union of occupational measures sets converges to a convex and compact set of probability measures defined in (2.6) (Theorem 2.1(i)) and that the LOMS of the control system is equal to this set if it exists (Theorem 2.1(ii)). We also give necessary and sufficient conditions for the existence of the LOMS (Theorem 2.1(iii)).

The paper consists of six sections. Theorem 2.1 is stated in section 2 and proved in sections 4–6. Applications in SPCS are discussed in section 3.

Singularly perturbed problems of control and optimization have been studied intensively in both deterministic and stochastic settings (see [1], [2], [3], [4], [5], [6], [7], [8], [11], [12], [14], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [36], [37], [38], [39], [40], [41], [42], [45], [46], [48] and the references therein).

Originally, the most common approaches to SPCS, especially in the deterministic case, were related to an approximation of the SPCS by the systems obtained via equating the singular perturbations parameter to zero (with further application of the boundary layer method (see [37], [44]) for an asymptotical description of the fast dynamics). This type of approach was successfully applied to a number of important classes of problems (see [30], [31], [38] and also [17], [36], [39], [45] for some recent results obtained in this direction).

*Received by the editors March 10, 2003; accepted for publication (in revised form) December 1, 2003; published electronically June 25, 2004. This work was supported by Australian Research Council Discovery-Project grant DP0346099 and IREX grant X00106494.

http://www.siam.org/journals/sicon/43-1/42418.html

†School of Mathematics, University of South Australia, The Mawson Lakes Campus, Mawson Lakes SA 5095, Australia (v.gaitsgory@unisa.edu.au).

Various averaging type approaches allowing a consideration of more general classes of SPCS, in which the equating of the small parameter to zero may not lead to a right approximation, were developed in [2], [3], [4], [5], [6], [7], [11], [20], [21], [22], [23], [24], [25], [26], [27], [41], [46].

In [24] and [25], in particular (see also [3], [5], [46] for related results), the slow trajectories were approximated by the solutions of the averaged system in which the controls are measure-valued and take their values in the LOMS of the associated system (that is, the system that would describe the fast dynamics if the slow state variables were "frozen"). The paper continues this line of research by establishing that the LOMS allows a representation in terms of the vector function defining the right-hand side of the associated system.

Let us introduce some notation and definitions which are used throughout the paper. Given a compact metric space W, $\mathcal{B}(W)$ will stand for the σ-algebra of its Borel subsets and $\mathcal{P}(W)$ will denote the set of probability measures defined on $\mathcal{B}(W)$. The set $\mathcal{P}(W)$ will be treated as a compact metric space with a metric ρ, which is consistent with its weak convergence topology (see, e.g., [13]). A sequence $\gamma^k \in \mathcal{P}(W)$ converges to $\gamma \in \mathcal{P}(W)$ in this metric if and only if

$$\lim_{k\to\infty} \int_W q(w)\gamma^k(dw) \;=\; \int_W q(w)\gamma(dw) \tag{1.1}$$

for any continuous $q(w) : W \to \mathbb{R}^1$. There are many ways of defining such a metric ρ. In this paper, we will use the following definition: $\forall \gamma', \gamma'' \in \mathcal{P}(W)$,

$$\rho(\gamma', \gamma'') \stackrel{\text{def}}{=} \sum_{l=1}^{\infty} \frac{1}{2^l} \left| \int_W q_l(w)\gamma'(dw) - \int_W q_l(w)\gamma''(dw) \right|, \tag{1.2}$$

where $q_l(\cdot)$, $l = 1, 2, \ldots$, is a sequence of Lipschitz continuous functions which is dense in the unit ball of $C(W)$ (the space of continuous functions on W). Using the metric ρ, one can define the Hausdorff metric ρ_H on the set of subsets of $\mathcal{P}(W)$ as follows: $\forall\, \Gamma_i \subset \mathcal{P}(W)$, $i = 1, 2$,

$$\rho_H(\Gamma_1, \Gamma_2) \stackrel{\text{def}}{=} \max \left\{ \sup_{\gamma\in\Gamma_1} \rho(\gamma, \Gamma_2), \sup_{\gamma\in\Gamma_2} \rho(\gamma, \Gamma_1) \right\}, \tag{1.3}$$

where $\rho(\gamma, \Gamma_i) \stackrel{\text{def}}{=} \inf_{\gamma'\in\Gamma_i} \rho(\gamma, \gamma')$. It can be verified (see, e.g., Lemma П2.4, p. 205 in [22]) that, with the definition of the metric ρ as in (1.2),

$$\rho_H(co\Gamma_1, co\Gamma_2) \le \rho_H(\Gamma_1, \Gamma_2), \tag{1.4}$$

where *co* stands for the convex hull of the corresponding set.

In what follows, we will deal with the convergence in the Hausdorff metric of sets in $\mathcal{P}(W)$ defined as unions of occupational measures. Given a measurable function $w(t) : [0, S] \to W$, the occupational measure $p^{w(\cdot)} \in \mathcal{P}(W)$ generated by this function is defined by taking

$$p^{w(\cdot)}(Q) \stackrel{\text{def}}{=} \frac{1}{S}\text{meas}\left\{ t \;\middle|\; w(t) \in Q \right\} \qquad \forall Q \in \mathcal{B}(W),$$

where meas $\{\cdot\}$ stands for the Lebesgue measure on $[0, S]$.

2. Main theorem. Consider a control system

$$\dot{y}(\tau) \;=\; f(u(\tau), y(\tau)), \quad \tau \in [0, S], \tag{2.1}$$

where the function $f(u, y) : U \times \mathbb{R}^m \to \mathbb{R}^m$ is continuous in (u, y) and satisfies Lipschitz conditions in y, U is a compact metric space, and the controls are Lebesgue measurable functions $u(\tau) : [0, S] \to U$.

Let Y be a compact subset of $\mathbb{R}^m$ and $Y^\delta \stackrel{\text{def}}{=} Y + \delta B$, where δ is a positive number and B is the closed unit ball in $\mathbb{R}^m$, and let us introduce the following definition and assumptions.

DEFINITION. *A pair $(u(\tau), y(\tau))$ is called admissible (δ-admissible) for system (2.1) on the interval $[0, S]$ if $u(\tau)$ is a control, $y(\tau)$ is the corresponding solution of (2.1), and $y(\tau) \in Y$ ($y(\tau) \in Y^\delta$) $\forall \tau \in [0, S]$.*

Assumption I. For any initial condition $y(0) \in Y$, there exists a control $u(\tau)$ such that the corresponding solution of (2.1) does not leave Y on $[0, S]$ for any $S > 0$.

Assumption II. For any Lipschitz continuous function $g(u, y) : U \times \mathbb{R}^m \to \mathbb{R}^1$,

$$\left| \frac{1}{S} \sup_{(u(\cdot), y(\cdot))} \int_0^S g(u(\tau), y(\tau)) d\tau \;-\; \frac{1}{S} \sup_{(u^\delta(\cdot), y^\delta(\cdot))} \int_0^S g(u^\delta(\tau), y^\delta(\tau)) d\tau \right| \stackrel{\text{def}}{=} \mu_g(\delta, S) \;\to\; 0 \tag{2.2}$$

as $\delta \to 0$ and $S \to \infty$, where the *sup*s in the above expression are, respectively, over all admissible pairs and over all δ-admissible pairs which satisfy the condition $y^\delta(0) \in Y$.

Assumption I is equivalent to the assumption that the viability kernel of Y is equal to Y. It is satisfied, for example, if, for any $y \in Y$, there exists $u \in U$ such that $f(u, y) = 0$. More general sufficient (and necessary) conditions for this assumption to be satisfied can be found in [9], [10].

Note that if Assumption I is replaced by a stronger assumption that Y is a forward invariant set—that is, all solutions of (2.1) obtained with the measurable controls $u(\tau) : [0, S] \to U$ $\forall S > 0$ do not leave Y—then Assumption II is satisfied automatically. In this case, all δ-admissible pairs satisfying the condition $y^\delta(0) \in Y$ are admissible and (2.2) is valid with $\mu_g(\delta, S) \equiv 0$.

Let $(u(\tau), y(\tau)) : [0, S] \to U \times Y$ be an admissible pair and let $p^{(u(\cdot), y(\cdot))} \in \mathcal{P}(U \times Y)$ be the occupational measure generated by this pair. Denote by $\Gamma(S, y)$ and $\Gamma(S, Y)$ the sets of occupational measures defined by the equations

$$\Gamma(S, y) \stackrel{\text{def}}{=} \bigcup_{(u(\tau),\ y(\tau))} \left\{ p^{(u(\cdot), y(\cdot))} \right\}, \qquad \Gamma(S, Y) \stackrel{\text{def}}{=} \bigcup_{y \in Y} \left\{ \Gamma(S, y) \right\}, \tag{2.3}$$

where the first union is over all admissible pairs of (2.1) satisfying the initial conditions $y(0) = y$ and the second is over all initial conditions

$$y(0) \;=\; y \;\in\; Y. \tag{2.4}$$

DEFINITION. *A convex and compact set $\Gamma \subset \mathcal{P}(U \times Y)$ is called the LOMS of system (2.1) on Y if there exists a function $\nu(S)$, $\lim_{S \to \infty} \nu(S) = 0$, such that*

$$\rho_H(\Gamma(S, y), \Gamma) \;\le\; \nu(S) \quad \forall y \;\in\; Y. \tag{2.5}$$

In Theorem 2.1 below, we relate the LOMS of system (2.1) on Y to the set $W \subset \mathcal{P}(U \times Y)$ defined by the equation

$$W \stackrel{\text{def}}{=} \left\{ \gamma \mid \gamma \in \mathcal{P}(U \times Y);\ \int_{U \times Y} (\phi'(y))^T f(u, y) \gamma(du, dy) = 0 \ \forall \phi(\cdot) \in C^1 \right\}, \tag{2.6}$$

where C^1 is the space of continuously differentiable functions $\phi(y) : \mathbb{R}^m \to \mathbb{R}^1$ and $\phi'(y)$ is the vector column of partial derivatives (the gradient) of $\phi(y)$. Note that, as can be easily verified, the set W is convex and compact in the weak convergence topology of $\mathcal{P}(U \times Y)$.

THEOREM 2.1. *Let Assumptions* I *and* II *be satisfied. Then the following hold:*

(i) *The estimate*

$$\rho_H(co\Gamma(S,Y),W) \;\le\; \bar{\nu}(S) \tag{2.7}$$

is valid for some $\bar{\nu}(S)$, $\lim_{S\to\infty}\bar{\nu}(S)=0$.

(ii) *If the LOMS* Γ *of system* (2.1) *on* Y *exists, it is equal to the set* W*:*

$$\Gamma \;=\; W. \tag{2.8}$$

(iii) *The LOMS* Γ *of system* (2.1) *on* Y *exists if and only if*

$$\rho_H(\Gamma(S,y'),\Gamma(S,y'')) \;\le\; \hat{\nu}(S) \quad \forall y', y'' \in Y, \tag{2.9}$$

for some $\hat{\nu}(S)$, $\lim_{S\to\infty}\hat{\nu}(S)=0$.

Proof. Statements (i) and (ii) of the theorem are proved in sections 5–6. Statement (iii) is proved in section 4. □

3. On applications in SPCS. Let us consider an SPCS defined on the interval $[0,T]$ $(T>0)$ by the equations

$$\epsilon\dot{y}(t) = f(u(t),y(t),z(t)), \qquad y(0)=y_0, \tag{3.1}$$

$$\dot{z}(t) = g(u(t),y(t),z(t)), \qquad z(0)=z_0, \tag{3.2}$$

where $\epsilon>0$ is a small parameter; $f : U\times\mathbb{R}^m\times\mathbb{R}^n \to \mathbb{R}^m$, $g : U\times\mathbb{R}^m\times\mathbb{R}^n \to \mathbb{R}^n$ are continuous vector functions satisfying Lipschitz conditions in z and y; U is a compact metric space, and the controls are measurable functions satisfying the inclusion $u(t)\in U$.

Along with (3.1)–(3.2), let us consider the associated system

$$\dot{y}(\tau) = f(u(\tau),y(\tau),z), \quad \tau\in[0,S], \tag{3.3}$$

in which (in contrast to (3.1)) z is a vector of fixed parameters: $z=\text{const}$. The controls in (3.3) are measurable functions satisfying the inclusion $u(\tau)\in U$.

Assume that, $\forall z$ from a sufficiently large area $Z\subset\mathbb{R}^n$, the solutions of the associated system (3.3) with the initial conditions in a compact set $Y\subset\mathbb{R}^m$ do not leave this set $\forall\tau\ge 0$ (that is, Y is forward invariant with respect to the solutions of (3.3)). Assume also that the solutions of the SPCS (3.1)–(3.2) do not leave $Y\times Z'$ for $t\in[0,T]$, where Z' is a compact set belonging to the interior of Z.

Let $u(\tau)\in U$ be a control and $y(\tau)\in Y$ be the solution of the associated system (3.3), obtained with this control and the initial condition $y(0)=y$. Denote by $\gamma^{(u(\cdot),y(\cdot))}\in\mathcal{P}(U\times Y)$ the occupational measure generated by the pair $(u(\cdot),y(\cdot)) : [0,S]\to U\times Y$ and denote by $\Gamma(z,S,y)\subset\mathcal{P}(U\times Y)$ the union of all such occupational measures.

Assume that the LOMS $\Gamma(z)$ of the associated system exists, that is,

$$\lim_{S\to\infty}\rho_H(\Gamma(z,S,y),\Gamma(z))=0,$$

with the convergence being uniform with respect to $(y, z) \in Y \times Z$. Note that, by Theorem 2.1(ii), $\Gamma(z) = W(z)$ (the latter is defined in (2.6), with the dependence on z being due to the fact that the vector function $f(\cdot)$ includes the dependence on z).

Define $\tilde{g}(\gamma, z) : \mathcal{P}(U \times Y) \to \mathbb{R}^n$ by the equation

$$\tilde{g}(\gamma, z) \stackrel{\text{def}}{=} \int_{U \times Y} g(u, y, z)\gamma(du, dy)$$

and consider the *averaged* system

$$\dot{z}(t) = \tilde{g}(\gamma(t), z(t)), \qquad z(0) = z_0, \tag{3.4}$$

in which the controls are Lebesgue measurable functions $\gamma(\cdot) : [0, T] \to \mathcal{P}(U \times Y)$ satisfying the inclusion

$$\gamma(t) \in W(z(t)). \tag{3.5}$$

The following result is a corollary of Theorem 2.1(ii) and Theorem 4.2 in [24].

COROLLARY 3.1. *Let the assumptions made above be satisfied. Also let the multivalued map* $V_g(\cdot) : Z \to 2^{R^n}$,

$$V_g(z) \stackrel{\text{def}}{=} \bigcup_{\gamma \in W(z)} \{\tilde{g}(\gamma, z)\},$$

be Lipschitz continuous. Then the following hold:

(i) *Corresponding to any solution* $(z^{sp}(t), y^{sp}(t))$ *of* (3.1)–(3.2) *there exists a solution* $z^{av}(t)$ *of* (3.4) *such that*

$$\max_{t \in [0,T]} ||z^{sp}(t) - z^{av}(t)|| \le \mu(\epsilon), \qquad \lim_{\epsilon \to 0} \mu(\epsilon) = 0. \tag{3.6}$$

(ii) *Corresponding to any solution* $z^{av}(t)$ *of* (3.4) *there exists a solution* $(z^{sp}(t), y^{sp}(t))$ *of* (3.1)–(3.2) *which satisfies* (3.6).

Proof. The proof follows from Theorem 4.2 in [24] with the replacement of $\Gamma(z)$ by $W(z)$ (see also Theorem 2.6 in [25] and related results in [3]). □

Sufficient conditions for the assumptions used in Corollary 3.1 to be valid have been discussed in [24], [25], where it was noticed, in particular, that these assumptions (including the existence of the LOMS and the Lipschitz continuity of $V_g(z)$) are satisfied if there exist positive definite matrices C and D such that, for any $u \in U$, any $y^1, y^2 \in R^m$, and any $z \in Z$,

$$(f(u, y^1, z) - f(u, y^2, z))^T C(y^1 - y^2) \le -(y^1 - y^2)^T D(y^1 - y^2).$$

The existence of such C and D can be guaranteed, for example, if $f(u, y, z) = A(z)y + B(z)u$, where $A(z)$ and $B(z)$ are matrices functions of the corresponding dimensions and the eigenvalues of $A(z)$ have negative real parts $\forall z \in Z$.

Let $G(\cdot) : \mathbb{R}^n \to \mathbb{R}^1$ be a continuous function. From Corollary 3.1 it follows that the optimal value of the problem

$$\inf_{(z^{sp}(\cdot), y^{sp}(\cdot))} G(z^{sp}(T)) \stackrel{\text{def}}{=} G^\star_\epsilon, \tag{3.7}$$

where inf is over the solutions of (3.1)–(3.2), converges (as ϵ tends to zero) to the optimal value of the problem

$$\inf_{z^{av}(\cdot)} G(z^{av}(T)) \stackrel{\text{def}}{=} G^\star, \tag{3.8}$$

where inf is over the solutions of (3.4). That is,

$$\lim_{\epsilon\to 0} G^\star_\epsilon = G^\star. \tag{3.9}$$

Also, it can be shown that a near optimal solution of (3.7) can be constructed on the basis of the optimal (or near optimal) solution of (3.8) (see details in [23], [24], [25], [27]).

Note that statement (i) of Corollary 3.1 remains valid even in the event the assumption about the existence of the LOMS is not satisfied, with (3.9) being replaced in this case by a weaker statement that

$$\liminf_{\epsilon\to 0} G^\star_\epsilon \geq G^\star.$$

The validity of this can be established via a straightforward extension of the averaging techniques used in [23], [24], [25], [27] in combination with Theorem 2.1(i).

In conclusion, let us observe that a numerical analysis of the solutions of (3.4) satisfying the inclusion (3.5) can be based on the fact that the set $W(z)$ allows the representation in the form of a countable system of equations:

$$W(z) = \left\{ \gamma \mid \gamma \in \mathcal{P}(U\times Y); \int_{U\times Y} (\phi_i'(y))^T f(u,y,z)\gamma(du,dy) = 0\ ,\ i = 1,2,\dots \right\}\ . \tag{3.10}$$

Here, $\{\phi_i(\cdot)\}$ is a sequence of continuously differentiable functions such that any function $\phi(\cdot)\in C^1$ and its gradient $\phi'(\cdot)$ can be simultaneously approximated on Y by linear combinations of functions from $\{\phi_i\}$ and their corresponding gradients. (An example of such a sequence is the sequence of the monomials $y_1^{i_1}\dots y_m^{i_m}$, $i_1,\dots,i_m = 0,1,\dots$, where y_j $(j=1,\dots,m)$ stands for the jth component of y; see, e.g., [33].) To numerically approximate the solutions of the averaged system (3.4)–(3.5) one may need to truncate the system of equations in (3.10) and subsequently approximate the resulting set by the set of measures supported on a grid. The details of such a procedure will be studied in a different paper.

4. Sets of time averages and proof of Theorem 2.1(iii). Let $\delta_0 > 0$ be fixed and $q_l(u,y): U\times Y^{\delta_0}\to\mathbb{R}^1$, $l=1,2,\dots$, be a sequence of Lipschitz continuous functions which is dense in the space of continuous functions on $U\times Y^{\delta_0}$. Let

$$h(u,y) \ = \ (q_1(u,y),\dots,q_j(u,y)),\ j=1,2,\dots, \tag{4.1}$$

and let $V_h(S,y)$ be the set of time averages defined by the equation

$$V_h(S,y) \stackrel{\text{def}}{=} \bigcup_{(u(\cdot),\ y(\cdot))} \left\{ \frac{1}{S}\int_0^S h\big(u(\tau),y(\tau)\big)d\tau \right\}, \tag{4.2}$$

where the union is over all admissible pairs of (2.1) satisfying the initial conditions (2.4).

The proof of Theorem 2.1(iii) is based on the following proposition.

PROPOSITION 4.1. *Let Assumption* I *be satisfied. Then the following hold:*

(i) *The LOMS* Γ *of system* (2.1) *on* Y *exists if and only if, for every vector function* $h(u,y)$ *defined in* (4.1), *there exist a convex and compact set* V_h *and a function* $\nu_h(S)$, $\lim_{S\to\infty}\nu_h(S)=0$, *such that*

$$d_H(V_h(S,y),V_h) \ \leq\ \nu_h(S) \quad \forall y\ \in\ Y, \tag{4.3}$$

where $d_H(\cdot,\cdot)$ is the Hausdorff metric defined on the bounded subsets of the corresponding finite dimensional space by the Euclidean norm.

(ii) *For a given $h(u,y)$ as in* (4.1)*, a convex and compact set V_h satisfying* (4.3) *exists if and only if*

$$d_H(V_h(S,y'),V_h(S,y'')) \;\le\; \hat{\nu}_h(S) \quad \forall y', y'' \;\in\; Y \tag{4.4}$$

for some $\hat{\nu}_h(S)$ tending to zero as S tends to infinity.

Proof. The proof of Proposition 4.1(i) is similar to that of Theorem 3.1 in [24] (see also Corollary 3.7 in [25] and the more general result in [7]). The proof of the "if" statement in Proposition 4.1(ii) follows exactly the same steps as that of Proposition 3.2 in [26] (see also [21], [22], and [7]), where this statement was proved for the case when Y is forward invariant with respect to the system (2.1). The proof of the "only if" statement is obvious. □

Proof of Theorem 2.1(iii). If (2.9) is valid, then the estimate (4.4) is true for every $h(u,y)$ defined in (4.1). Hence, by Proposition 4.1(ii), for every such $h(u,y)$, there exists a convex and compact set V_h satisfying (4.3). This, by Proposition 4.1(i), implies the existence of the LOMS. Thus, the "if" statement in Theorem 2.1(iii) is proved. The proof of the "only if" statement is obvious. □

To conclude this section, let us show that Assumption II can be equivalently reformulated in terms of convergence to zero of the Hausdorff metric between the sets of time averages defined below. For $0 < \delta \le \delta_0$, let

$$V_h^\delta(S,y) \stackrel{\text{def}}{=} \bigcup_{(u^\delta(\cdot),\; y^\delta(\cdot))} \left\{ \frac{1}{S} \int_0^S h\big(u^\delta(\tau), y^\delta(\tau)\big) d\tau \right\}, \tag{4.5}$$

where, in contrast to (4.2), the union is over all δ-admissible pairs of (2.1) satisfying the initial conditions (2.4). Denote

$$V_h(S,Y) \stackrel{\text{def}}{=} \bigcup_{y\in Y} \{V_h(S,y)\}, \qquad V_h^\delta(S,Y) \stackrel{\text{def}}{=} \bigcup_{y\in Y} \left\{V_h^\delta(S,y)\right\}.$$

The following lemma is used in the proof of Theorem 2.1(i) (see section 5 below).

LEMMA 4.2. *Assumption* II *is equivalent to that, for any $h(\cdot)$ as in* (4.1)*,*

$$d_H(coV_h(S,Y), coV_h^\delta(S,Y)) \;\stackrel{\text{def}}{=}\; \bar{\nu}_h(\delta,S) \;\to\; 0 \tag{4.6}$$

as $\delta \to 0$ and $S \to \infty$.

Proof. Let $\Psi_V(\cdot)$ stand for the support function of a set $V \subset R^j$. That is, for any $\eta \in R^j$, $\Psi_V(\eta) \stackrel{\text{def}}{=} \sup_{v\in V} \eta^T v$. If Assumption II is satisfied, then, taking $g(u,y) = \eta^T h(u,y)$ in (2.2), one can obtain that the function $\mu_g(\delta,S)$ defined by the equation

$$|\Psi_{coV_h(S,Y)}(\eta) - \Psi_{coV_h^\delta(S,Y)}(\eta)| = |\Psi_{V_h(S,Y)}(\eta) - \Psi_{V_h^\delta(S,Y)}(\eta)| \;\stackrel{\text{def}}{=}\; \mu_g(\delta,S) \tag{4.7}$$

tends to zero as $\delta \to 0$ and $S \to \infty$. Using a standard argument based on the separability of convex sets, one can verify that (4.7) implies (4.6). Thus, the validity of (4.6) is implied by the validity of Assumption II.

Now let (4.6) be satisfied for any $h(\cdot)$ constructed as in (4.1). Then $\mu_g(\delta,S)$ in (4.7) tends to zero as δ tends to zero and S tends to infinity. By taking all but one component of η to be equal to zero in (4.7), one can verify that (2.2) is valid for any $g(u,y) = q_l(u,y)$, $l = 1,2,\ldots$. Since the sequence $q_l(u,y)$, $l = 1,2,\ldots$, is dense in $C(U \times Y^\delta)$, it implies that (2.2) is valid for any continuous (and, in particular, Lipschitz continuous) $g(\cdot)$. This completes the proof of the lemma. □

5. Proofs of Theorem 2.1(i) and Theorem 2.1(ii).

Proof of Theorem 2.1(i). To prove the required statement, one needs to establish the validity of the following two inequalities:

$$\sup_{\gamma\in W}\rho(\gamma, co\Gamma(S,Y)) \le \bar\nu(S), \tag{5.1}$$

$$\sup_{\gamma\in co\Gamma(S,Y)}\rho(\gamma, W) \le \bar\nu(S). \tag{5.2}$$

Let us first prove the validity of (5.2). It is straightforward to verify that from the convexity of W it follows that

$$\sup_{\gamma\in co\Gamma(S,Y)}\rho(\gamma, W) = \sup_{\gamma\in \Gamma(S,Y)}\rho(\gamma, W).$$

Hence, to prove (5.2) it is enough to show that the function $\bar\nu(S)$ defined by the equation

$$\bar\nu(S) \stackrel{\text{def}}{=} \sup_{\gamma\in\Gamma(S,Y)}\rho(\gamma, W) \tag{5.3}$$

tends to zero as S tends to infinity. Assume this is not the case. Then there exist a positive number δ, a sequence $S^k \to \infty$, and sequences $y^k \in Y$ and $\gamma^k \in \Gamma(S^k, y^k)$ such that $\rho(\gamma^k, W) \ge \delta$, $k = 1, 2, \dots$. Without loss of generality one may assume that there exists $\lim_{k\to\infty}\gamma^k \stackrel{\text{def}}{=} \gamma \in \mathcal{P}(U\times Y)$ (since $\mathcal{P}(U\times Y)$ is compact). From the continuity of the metric it follows that

$$\rho(\gamma, W) \ge \delta. \tag{5.4}$$

By the definition of the convergence in $\mathcal{P}(U\times Y)$ (see (1.1)),

$$\lim_{k\to\infty}\int_{U\times Y}(\phi\prime(y))^T f(u,y)\gamma^k(du,dy) \;=\; \int_{U\times Y}(\phi\prime(y))^T f(u,y)\gamma(du,dy) \tag{5.5}$$

for any $\phi \in C^1$. Also, from the fact that $\gamma^k \in \Gamma(S^k, y^k)$, it follows that there exists an admissible pair $(u^k(\tau), y^k(\tau))$ (for system (2.1) on the interval $[0, S^k]$) such that

$$\int_{U\times Y}(\phi\prime(y))^T f(u,y)\gamma^k(du,dy) \;=\; \frac{1}{S^k}\int_0^{S^k}(\phi\prime(y^k(\tau)))^T f(u^k(\tau), y^k(\tau))d\tau.$$

The second integral is apparently equal to

$$\frac{\phi(y^k(S^k)) - \phi(y^k(0))}{S^k}$$

and tends to zero as S^k tends to infinity (since $y^k(\tau) \in Y\ \forall\ \tau \in [0, S^k]$ and Y is a compact set). This and (5.5) imply that

$$\int_{U\times Y}(\phi\prime(y))^T f(u,y)\gamma(du,dy) \;=\; 0\ \ \forall\phi\in C^1 \quad\Rightarrow\quad \gamma\ \in\ W.$$

The latter contradicts (5.4) and, hence, $\bar\nu(s)$ defined in (5.3) tends to zero as S tends to infinity. This proves (5.2).

Let us now prove the validity of inequality (5.1). Let $\hat{r}$ be a positive number such that Y is contained in the interior of $\hat{r}B \stackrel{\text{def}}{=} \hat{B}$ (as above, B is the closed unit ball in $\mathbb{R}^m$) and let $\psi(y) : \mathbb{R}^m \to [0,1]$ be a continuously differentiable function such that

$$\psi(y) \;=\; 1 \quad \forall y \;\in Y, \qquad \psi(y) \;=\; 0 \quad \forall y \;\in \mathbb{R}^m/\hat{B}. \tag{5.6}$$

Define the function $F(u,y) : U \times \mathbb{R}^m \to \mathbb{R}^m$ by the equation

$$F(u,y) \;\stackrel{\text{def}}{=}\; \psi(y) f(u,y). \tag{5.7}$$

Note that, by (5.6),

$$F(u,y) \;=\; f(u,y) \quad \forall (u,y) \;\in U \times Y, \qquad F(u,y) = 0 \quad \forall (u,y) \;\in U \times \mathbb{R}^m/\hat{B}. \tag{5.8}$$

Let $\bar{C}(U \times \mathbb{R}^m)$ be the space of bounded continuous functions on $U \times \mathbb{R}^m$ taking values in $\mathbb{R}^1$, and let $\hat{C}^1$ be the space of continuously differentiable functions on $\mathbb{R}^m$ taking values in $\mathbb{R}^1$ and vanishing at infinity. Note that, since Y is a compact set, one can replace C^1 by $\hat{C}^1 \subset C^1$ in (2.6) without adding new elements to the set W. Define a linear operator $A : \hat{C}^1 \to \bar{C}(U \times \mathbb{R}^m)$ by the equation

$$(A\phi)(u,y) \;\stackrel{\text{def}}{=}\; (\phi'(y))^T F(u,y) \quad \forall \phi \;\in \hat{C}^1.$$

This operator satisfies the conditions of Theorem 4.1 in [43], namely, the following:

(i) $\hat{C}^1$, the domain of A, is an algebra and is dense in the space $\hat{C}$ of continuous functions on $\mathbb{R}^m$ which vanish at infinity (this is an immediate consequence of the Stone–Weierstrass theorem; see, e.g., Theorem IV.6.16 in [15]).

(ii) For each $\phi \in \hat{C}^1$ and $u \in U$, $(A\phi)(u,\cdot) \stackrel{\text{def}}{=} (\phi'(\cdot))^T F(u,\cdot) \in \hat{C}$.

(iii) For each $\phi \in \hat{C}^1$,

$$\lim_{||y|| \to \infty} \max_{u \in U} (A\phi)(u,y) \;=\; \lim_{||y|| \to \infty} \max_{u \in U} (\phi'(y))^T F(u,y) \;=\; 0.$$

(iv) For each $u \in U$, the operator $A_u \phi \stackrel{\text{def}}{=} (A\phi)(u,\cdot)$ satisfies the positive maximum principle, i.e., if $\phi(y^*) = \sup_y \phi(y) > 0$, then

$$(A_u\phi)(y^*) \;=\; (\phi'(y^*))^T F(u,y^*) \;\leq\; 0.$$

Note that (ii) and (iii) are satisfied because of (5.8) and that (iv) follows from the fact that $\phi'(y^*) = 0$.

Let us consider now an arbitrary $\gamma \in W$ and extend its definition to the Borel subsets of $U \times \mathbb{R}^m$ by taking $\gamma(Q) \stackrel{\text{def}}{=} \gamma(Q \cap (U \times Y))$ $\forall Q \in \mathcal{B}(U \times \mathbb{R}^m)$. By (5.8) and (2.6),

$$\int_{U \times \mathbb{R}^m} (A\phi)(u,y)\gamma(du,dy) \;=\; \int_{U \times \mathbb{R}^m} (\phi'(y))^T F(u,y)\gamma(du,dy)$$

$$= \int_{U \times Y} (\phi'(y))^T F(u,y)\gamma(du,dy) \;=\; \int_{U \times Y} (\phi'(y))^T f(u,y)\gamma(du,dy) \;=\; 0$$

$\forall \phi \in \hat{C}^1$. From Theorem 4.1 in [43] it follows that there exist a probability space $(\Omega, \mathcal{F}, P)$, a filtration $\{\mathcal{F}_\tau\}$ of σ-subalgebras of $\mathcal{F}$, and a $\mathcal{P}(U) \times \mathbb{R}^m$-valued random process $(\lambda(\tau), y(\tau)) \;=\; (\lambda(\tau,\omega), y(\tau,\omega))$ which satisfies the following conditions:

(a) $(\lambda(\cdot), y(\cdot))$ is $\{\mathcal{F}_\tau\}$-progressive and stationary with

$$E[\lambda(\tau)(D_1)\chi_{D_2}(y(\tau))] = \gamma(D_1 \times D_2) \quad \forall \tau \ \geq \ 0, \tag{5.9}$$

where D_1 and D_2 are arbitrary Borel subsets of U and $\mathbb{R}^m$, respectively, and $\chi_{D_2}(\cdot)$ is the indicator function of D_2.

(b) For every $\phi \in \hat{C}^1$, $\phi(y(\tau)) - \int_0^\tau \int_U (\phi'(y(s)))^T F(u, y(s))\lambda(s)(du)ds$ is an $\{\mathcal{F}_\tau\}$-martingale.

A further characterization of the pair $(\lambda(\tau,\omega), y(\tau,\omega))$ is given by the following lemma.

LEMMA 5.1. *There exists a subset Δ of Ω such that $P(\Delta) = 0$ and such that, $\forall \omega \in \Omega/\Delta$, the pair $(\lambda(\tau,\omega), y(\tau,\omega))$ satisfies the equation*

$$\dot{y}(\tau,\omega) \ = \ \bar{F}(\lambda(\tau,\omega), y(\tau,\omega)) \tag{5.10}$$

for almost all $\tau \in [0, S]$ $(\forall S > 0)$, where

$$\bar{F}(\lambda, y) \ \stackrel{\text{def}}{=} \int_U F(u, y)\lambda(du). \tag{5.11}$$

Proof. Proof of the lemma is given in section 6 below. □

Let τ_i, $i = 1, 2, \ldots$, stand for a sequence of all rational numbers belonging to the interval $[0, S]$. By (5.9),

$$P\{\omega \mid y(\tau_i, \omega) \ \in Y\} \ = \ E[\chi_Y(y(\tau_i))] \ = \ E[\lambda(\tau_i)(U)\chi_Y(y(\tau_i))] \ = \ \gamma(U \times Y) \ = \ 1.$$

That is, for every i there exists a subset Δ_i of Ω such that $P(\Delta_i) = 0$ and such that

$$y(\tau_i, \omega) \in Y \ \forall \omega \in \Omega/\Delta_i \quad \Rightarrow \quad y(\tau_i, \omega) \in Y, \ i = 1, 2, \ldots, \ \forall \omega \in \Omega/(\cup_i \Delta_i). \tag{5.12}$$

From the fact that $y(\tau,\omega)$ satisfies (5.10) it follows that $y(\cdot,\omega)$ is continuous (in fact, absolutely continuous) for $\omega \in \Omega/\Delta$. Thus, the inclusions (5.12) and the fact that Y is compact imply that

$$y(\tau,\omega) \ \in Y \ \ \forall \tau \in [0, S], \quad \forall \omega \in \Omega/\bar{\Delta}, \tag{5.13}$$

where $\bar{\Delta} \stackrel{\text{def}}{=} \Delta \cup (\cup_i \Delta_i)$, with

$$P(\bar{\Delta}) \ = \ 0. \tag{5.14}$$

Note that, by (5.8) and (5.11), equation (5.10) is equivalent to

$$\dot{y}(\tau,\omega) \ = \ \bar{f}(\lambda(\tau,\omega), y(\tau,\omega)) \tag{5.15}$$

for $\omega \in \Omega/\bar{\Delta}$, where

$$\bar{f}(\lambda, y) \ \stackrel{\text{def}}{=} \int_U f(u, y)\lambda(du). \tag{5.16}$$

Consider the control system

$$\dot{y}(\tau) \ = \ \bar{f}(\lambda(\tau), y(\tau)), \tag{5.17}$$

where $\lambda(\tau)$ is a relaxed control, that is, a Lebesgue measurable function $\lambda(\tau) : [0, S] \to \mathcal{P}(U)$ (see [47]). A pair $(\lambda(\tau), y(\tau))$ will be called admissible (for system (5.17) on

the interval $[0, S]$) if $\lambda(\tau)$ is a relaxed control, $y(\tau)$ is the corresponding solution of (5.17), and $y(\tau) \in Y \ \forall \tau \in [0, S]$. Let $h(u, y)$ be as in (4.1) and

$$\bar{h}(\lambda, y) \stackrel{\text{def}}{=} \int_U h(u, y)\lambda(du). \tag{5.18}$$

Consider the set of the time averages

$$\bar{V}_h(S, y) \stackrel{\text{def}}{=} \bigcup_{(\lambda(\cdot),\ y(\cdot))} \left\{ \frac{1}{S} \int_0^S \bar{h}\big(\lambda(\tau), y(\tau)\big) d\tau \right\}, \tag{5.19}$$

where the union is over all admissible pairs of (5.17) satisfying the initial conditions (2.4).

By the relaxation theorem (see, e.g., Theorem 10.4.4, p. 402 in [10]),

$$\bar{V}_h(S, y) \ \subset \ \mathrm{cl} V_h^{\delta}(S, y) \ \ \forall \delta \ > \ 0, \tag{5.20}$$

where cl stands for the closure of the set and $V_h^{\delta}(S, y)$ is defined in (4.5). Since $y(\tau, \omega)$ satisfies (5.13) and (5.15), from (5.20) and (4.6) it follows that, $\forall \omega \in \Omega / \bar{\Delta}$,

$$\begin{aligned} &\frac{1}{S} \int_0^S \bar{h}\big(\lambda(\tau, \omega), y(\tau, \omega)\big) d\tau \ \in \ \bar{V}_h(S, y(0, \omega)) \\ &\subset \ \mathrm{cl} V_h^{\delta}(S, Y) \ \subset \bar{co} V_h(S, Y) \ + \ \bar{\nu}_h(\delta, S) B_j, \end{aligned}$$

where $\bar{co}$ is the closed convex hull of the corresponding set and B_j is the closed unit ball in $\mathbb{R}^j$ (j is the dimension of the vector function $h(\cdot)$). Using now (5.14), one obtains from here that

$$\frac{1}{S} \int_0^S E[\bar{h}\big(\lambda(\tau, \omega), y(\tau, \omega)\big)] d\tau \ \in \ \bar{co} V_h(S, Y) \ + \ \bar{\nu}_h(\delta, S) B_j. \tag{5.21}$$

From (5.9), however, it follows that

$$E[\bar{h}\big(\lambda(\tau, \omega), y(\tau, \omega)\big)] \ = \ \int_{U \times Y} h(u, y) \gamma(du, dy). \tag{5.22}$$

Consequently, by (5.21),

$$\begin{aligned} &\int_{U \times Y} h(u, y) \gamma(du, dy) \ \in \ \bar{co} V_h(S, Y) \ + \ \bar{\nu}_h(\delta, S) B_j \\ \Rightarrow \quad &\int_{U \times Y} h(u, y) \gamma(du, dy) \ \in \ \bar{co} V_h(S, Y) \ + \ \bar{\nu}_h(S) B_j, \end{aligned}$$

where

$$\bar{\nu}_h(S) \stackrel{\text{def}}{=} \limsup_{\delta \to 0} \bar{\nu}_h(\delta, S).$$

Note that from the fact that $\bar{\nu}_h(\delta, S) \to 0$ as $\delta \to 0$ and $S \to \infty$ it follows that $\bar{\nu}_h(S) \to 0$ as $S \to \infty$. Since γ is an arbitrary element of W, one can conclude that

$$\bigcup_{\gamma \in W} \left\{ \int_{U \times Y} h(u, y) \gamma(du, dy) \right\} \ \subset \ \bar{co} V_h(S, Y) \ + \ \bar{\nu}_h(S) B_j. \tag{5.23}$$

The set $V_h(S,Y)$ allows the representation

$$V_h(S,Y) \;=\; \bigcup_{\gamma\in\Gamma(S,Y)}\Big\{\int_{U\times Y} h(u,y)\gamma(du,dy)\Big\}.$$

Hence,

$$\bar{co}V_h(S,Y) = \bar{co}\bigcup_{\gamma\in\Gamma(S,Y)}\Big\{\int_{U\times Y} h(u,y)\gamma(du,dy)\Big\}$$

$$= \bigcup_{\gamma\in\bar{co}\Gamma(S,Y)}\Big\{\int_{U\times Y} h(u,y)\gamma(du,dy)\Big\}.$$

That is, (5.23) can be rewritten in the form

$$\bigcup_{\gamma\in W}\Big\{\int_{U\times Y} h(u,y)\gamma(du,dy)\Big\} \;\subset\; \bigcup_{\gamma\in\bar{co}\Gamma(S,Y)}\Big\{\int_{U\times Y} h(u,y)\gamma(du,dy)\Big\} \;+\; \bar{\nu}_h(S)B_j.$$

Since it is true for any $h(u,y)$ as in (4.1), the validity of (5.1), with some $\bar{\nu}(S)$ tending to zero as S tends to infinity, follows from Lemma 3.5 in [25]. This completes the proof of Theorem 2.1(i). □

Proof of Theorem 2.1(ii). If the LOMS Γ exists, then, by (1.4),

$$\rho_H(co\Gamma(S,Y),\Gamma) \;=\; \rho_H(co\Gamma(S,Y),co\Gamma) \;\le\; \rho_H(\Gamma(S,Y),\Gamma) \;\le\; \nu(S),$$

where $\nu(S)$ is from (2.5). This and (2.7) imply equation (2.8) (because both Γ and W are compact). □

6. Proof of Lemma 5.1. The proof is divided into three steps. First, it is established that the random processes

$$J_i(\tau) \stackrel{\text{def}}{=} y_i(\tau) - \int_0^\tau \bar{F}_i(\lambda(s),y(s))ds\,, \qquad K_i(\tau) \stackrel{\text{def}}{=} y_i^2(\tau) - 2\int_0^\tau y_i(s)\bar{F}_i(\lambda(s),y(s))ds, \tag{6.1}$$

$i=1,\dots,m$, are $\{\mathcal{F}_\tau\}$-martingales, where $y_i(\cdot),\bar{F}_i(\cdot)$, are the ith components of $y(\cdot)$ and $\bar{F}(\cdot)$, respectively. That is, for $\tau>\sigma\ge 0$,

$$E[\,J_i(\tau)\mid\mathcal{F}_\sigma\,] = J_i(\sigma), \qquad E[\,K_i(\tau)|\mathcal{F}_\sigma\,] = K_i(\sigma). \tag{6.2}$$

Second, it is shown that the processes $J_i^2(\tau)$, $i=1,\dots,m$, are $\{\mathcal{F}_\tau\}$-martingales. That is, for $\tau>\sigma\ge 0$,

$$E[\,J_i^2(\tau)\mid\mathcal{F}_\sigma\,] = J_i^2(\sigma), \quad i=1,\dots,m. \tag{6.3}$$

Finally, the statement of the lemma is proved on the basis of (6.3).

Let us verify the validity of (6.2). Let $N>0$ and $\psi_N(\theta):[0,\infty)\to[0,\infty)$ be a continuously differentiable function such that

$$\psi_N(\theta)=1 \quad \forall\theta\in[0,N^2]\,, \qquad \psi_N(\theta)=0 \quad \forall\theta\in[N^2+1\,,\infty),$$

and $\psi_N(\theta)\in[0,1]$ $\forall\theta\in(N^2,N^2+1)$. Define $\phi_{i,N}(\cdot)\in\hat{C}^1$ by the equation $\phi_{i,N}(y)\stackrel{\text{def}}{=} y_i\psi_N(||y||^2)$, where $||y||$ is the Euclidean norm of y. Note that $|\phi_{i,N}(y)|\le|y_i|$ and that $\phi_{i,N}(y)=y_i$ for $||y||\le N$.

By condition (b), the process

$$J_{i,N}(\tau) \stackrel{\text{def}}{=} \phi_{i,N}(y(\tau)) - \int_0^\tau \phi'_{i,N}(y(s))^T \bar{F}_i(\lambda(s), y(s))ds$$

is an $\{\mathcal{F}_\tau\}$-martingale. Hence,

$$E[\, J_{i,N}(\tau) \mid \mathcal{F}_\sigma \,] = J_{i,N}(\sigma), \quad \tau > \sigma. \tag{6.4}$$

It can be seen, however, that for N large enough,

$$E[\,|J_{i,N}(\tau) - J_i(\tau)|\,] = E[\,|\phi_{i,N}(y(\tau)) - y_i(\tau)|\,] \le 2E[\,|y_i(\tau)|\chi_{Q_N}(y(\tau))\,]$$

$$\le 2\sqrt{E[\,|y_i(\tau)|^2]}\sqrt{E[\,\chi_{Q_N}(y(\tau))\,]} \le 2\sqrt{E[\,|y_i(\tau)|^2]}\sqrt{\gamma(U \times Q_N)} = 0,$$

where $\chi_{Q_N}(\cdot)$ is the indicator function of the set $Q_N \stackrel{\text{def}}{=} \{y \mid ||y|| > N\}$ and (5.8), (5.11) as well as (5.9) and the fact that $\gamma(U \times Y) = 1$ have been used. Thus, for sufficiently large N, $J_{i,N}(\tau) = J_i(\tau)$ a.s. and, hence, (6.4) implies the validity of the first equation in (6.2). The validity of the second equation in (6.2) is verified in a similar way (by using the test functions $\phi_{i,N}(y) \stackrel{\text{def}}{=} y_i^2\psi_N(||y||^2)$).

Let us now prove (6.3). Using the second equation in (6.2), one can write

$$E[J_i^2(\tau) \mid \mathcal{F}_\sigma] - J_i^2(\sigma) = E[J_i^2(\tau) - K_i(\tau) \mid \mathcal{F}_\sigma] - (J_i^2(\sigma) - K_i(\sigma))$$

$$= E\left[-2\int_0^\tau (y_i(\tau) - y_i(s))\bar{F}_i(\lambda(s), y(s))ds + \left(\int_0^\tau \bar{F}_i(\lambda(s), y(s))ds\right)^2 \mid \mathcal{F}_\sigma\right]$$

$$- \left(-2\int_0^\sigma (y_i(\sigma) - y_i(s))\bar{F}_i(\lambda(s), y(s))ds + \left(\int_0^\sigma \bar{F}_i(\lambda(s), y(s))ds\right)^2\right)$$

$$= E\left[-2(y_i(\tau) - y_i(\sigma))\int_0^\sigma \bar{F}_i(\lambda(s), y(s))ds - 2\int_\sigma^\tau (y_i(\tau) - y_i(s))\bar{F}_i(\lambda(s), y(s))ds \mid \mathcal{F}_\sigma\right]$$

$$+ E\left[2\int_\sigma^\tau \bar{F}_i(\lambda(s), y(s))ds \int_0^\sigma \bar{F}_i(\lambda(s), y(s))ds + \left(\int_\sigma^\tau \bar{F}_i(\lambda(s), y(s))ds\right)^2 \mid \mathcal{F}_\sigma\right].$$

Note that, by the first equation in (6.2),

$$E\left[(y_i(\tau) - y_i(\sigma))\int_0^\sigma \bar{F}_i(\lambda(s), y(s))ds \mid \mathcal{F}_\sigma\right]$$

$$- E\left[\int_\sigma^\tau \bar{F}_i(\lambda(s), y(s))ds \int_0^\sigma \bar{F}_i(\lambda(s), y(s))ds \mid \mathcal{F}_\sigma\right] = 0.$$

Hence, to complete the proof of (6.3), it is now sufficient to show that

$$E\left[\int_\sigma^\tau (y_i(\tau) - y_i(s))\bar{F}_i(\lambda(s), y(s))ds|\mathcal{F}_\sigma\right] = \frac{1}{2}E\left[\left(\int_\sigma^\tau \bar{F}_i(\lambda(s), y(s))ds\right)^2 |\mathcal{F}_\sigma\right]. \tag{6.5}$$

Using again the fact that $J_i(\cdot)$ is a martingale, one can obtain that

$$E\left[\int_\sigma^\tau (y_i(\tau)-y_i(s))\bar{F}_i(\lambda(s),y(s))ds \mid \mathcal{F}_\sigma\right]$$

$$=\int_\sigma^\tau E[\,(y_i(\tau)-y_i(s))\bar{F}_i(\lambda(s),y(s)) \mid \mathcal{F}_\sigma]ds$$

$$=\int_\sigma^\tau E[\,E[\,(y_i(\tau)-y_i(s)) \mid \mathcal{F}_s]\bar{F}_i(\lambda(s),y(s)) \mid \mathcal{F}_\sigma]ds$$

$$=\int_\sigma^\tau E\left[\left(\int_s^\tau \bar{F}_i(\lambda(s'),y(s'))ds'\right)\bar{F}_i(\lambda(s),y(s)) \mid \mathcal{F}_\sigma\right]ds$$

$$=E\left[\int_\sigma^\tau\left(\int_s^\tau \bar{F}_i(\lambda(s'),y(s'))ds'\right)\bar{F}_i(\lambda(s),y(s))ds \mid \mathcal{F}_\sigma\right].$$

Since

$$E\left[\int_\sigma^\tau\left(\int_s^\tau \bar{F}_i(\lambda(s'),y(s'))ds'\right)\bar{F}_i(\lambda(s),y(s))ds \mid \mathcal{F}_\sigma\right]$$

$$=E\left[\int_\sigma^\tau\left(\int_\sigma^s \bar{F}_i(\lambda(s'),y(s'))ds'\right)\bar{F}_i(\lambda(s),y(s))ds \mid \mathcal{F}_\sigma\right]$$

and the sum of the left- and the right-hand sides in the above equation is equal to $E[\,(\int_\sigma^\tau \bar{F}_i(\lambda(s),y(s))ds)^2 \mid \mathcal{F}_\sigma]$, it follows that (6.5) is valid and, thus, (6.3) is established.

From (6.3) and the first equation in (6.2), it follows that, for $\tau>\sigma\ge 0$,

$$E[\,(J_i(\tau)-J_i(\sigma))^2 \mid \mathcal{F}_\sigma]=0 \;\Rightarrow\; E[\,(J_i(\tau)-J_i(\sigma))^2]=0, \tag{6.6}$$

$i=1,\dots,m$. By Kolmogorov's continuity theorem (see, e.g., Theorem 1.10, p. 23 in [35]), there exists a continuous version of the $\mathbb{R}^m$-valued process $J(\cdot)\stackrel{\text{def}}{=}\{J_i(\cdot)\}$, $i=1,\dots,m$. For this version, (6.6) implies that

$$J(\tau)=J(0)\quad\forall\tau\in[0,S],\quad\forall S>0.$$

In accordance with our notation (see (6.1)), the latter is equivalent to

$$y(\tau)=y(0)+\int_0^\tau \bar{F}(\lambda(s),y(s))ds,$$

which, in turn, is equivalent to (5.10). This completes the proof of the lemma. □

Acknowledgment. The author wants to express his gratitude to J.-B. Lasserre for useful discussions enhancing the progress in writing this paper and also to V. Borkar for referring the author to the result of R. H. Stockbridge (see [43]) as well as to some other results in the theory of control martingale problems, which were used in the proof of the main theorem.

REFERENCES

[1] E. Altman and V. Gaitsgory, *Asymptotic optimization of a nonlinear hybrid system governed by a Markov decision process*, SIAM J. Control Optim., 35 (1997), pp. 2070–2085.

[2] O. Alvarez and M. Bardi, *Viscosity solutions methods for singular perturbations in deterministic and stochastic control*, SIAM J. Control Optim., 40 (2001), pp. 1159–1188.

[3] Z. Artstein, *The chattering limit of singularly perturbed optimal control problems*, in Proceedings of CDC-2000, Control and Decision Conference, Sydney, 2000, pp. 564–569.
[4] Z. Artstein, *An occupational measure solution to a singularly perturbed optimal control problem*, Control Cybernet., 31 (2002), pp. 623–642.
[5] Z. Artstein and V. Gaitsgory, *Tracking fast trajectories along a slow dynamics: A singular perturbations approach*, SIAM J. Control Optim., 35 (1997), pp. 1487–1507.
[6] Z. Artstein and V. Gaitsgory, *The value function of singularly perturbed control systems*, Appl. Math. Optim., 41 (2000), pp. 425–445.
[7] Z. Artstein and V. Gaitsgory, *Convergence to convex compact sets in infinite dimensions*, J. Math. Anal. Appl., 284 (2003), pp. 471–480.
[8] Z. Artstein and A. Leizarowitz, *Singularly perturbed control systems with one-dimensional fast dynamics*, SIAM J. Control Optim., 41 (2002), pp. 641–658.
[9] J.-P. Aubin, *Viability Theory*, Birkhäuser Boston, Boston, 1991.
[10] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.
[11] F. Bagagiolo and M. Bardi, *Singular perturbation of a finite horizon problem with state-space constraints*, SIAM J. Control Optim., 36 (1998), pp. 2040–2060.
[12] A. Bensoussan, *Perturbation Methods in Optimal Control*, John Wiley, New York, 1989.
[13] D.P. Bertsekas and S.E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
[14] F. Colonius and R. Fabri, *Controllability for systems with slowly varying parameters*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 207–216.
[15] N. Danford and J.T. Schwartz, *Linear Operators*, Interscience, New York, 1958.
[16] A. Dontchev, *Time-scale decomposition of the reachable set of constrained linear systems*, Math. Control Signals Systems, 5 (1992), pp. 327–340.
[17] A. Dontchev, T. Donchev, and I. Slavov, *A Tichonov-type theorem for singularly perturbed differential inclusion*, Nonlinear Anal., 26 (1996), pp. 1547–1554.
[18] T. Donchev and I. Slavov, *Averaging method for one-sided Lipschitz differential inclusions with generalized solutions*, SIAM J. Control Optim., 37 (1999), pp. 1600–1613.
[19] J.A. Filar, V. Gaitsgory, and A.B. Haurie, *Control of singularly perturbed hybrid stochastic systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 179–190.
[20] O.P. Filatov and M.M. Hapaev, *Averaging of Systems of Differential Inclusions*, Moscow University Publishing House, Moscow, 1998 (in Russian).
[21] V. Gaitsgory, *Use of the averaging method in control problems*, Differential Equations, 22 (1986), pp. 1290–1299 (translated from Russian).
[22] V. Gaitsgory, *Control of Systems with Fast and Slow Motions*, Nauka, Moscow, 1991 (in Russian).
[23] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.
[24] V. Gaitsgory and A. Leizarowitz, *Limit occupational measures set for a control system and averaging of singularly perturbed control systems*, J. Math. Anal. Appl., 233 (1999), pp. 461–475.
[25] V. Gaitsgory and M.-T. Nguyen, *Multiscale singularly perturbed control systems: Limit occupational measures sets and averaging*, SIAM J. Control Optim., 41 (2002), pp. 954–974.
[26] G. Grammel, *Averaging of singularly perturbed systems*, Nonlinear Anal., 28 (1997), pp. 1851–1865.
[27] G. Grammel, *Periodic near optimal control*, J. Math. Anal. Appl., 248 (2000), pp. 124–144.
[28] Y. Kabanov and S. Pergamenshikov, *On convergence of attainability sets for controlled two-scale stochastic linear systems*, SIAM J. Control Optim., 35 (1997), pp. 134–159.
[29] Y. Kabanov and S. Pergamenshchikov, *Two-Scale Stochastic Systems*, Springer-Verlag, Berlin, 2003.
[30] P.V. Kokotovíc, *Applications of singular perturbation techniques to control problems*, SIAM Rev., 26 (1984), pp. 501–550.
[31] P.V. Kokotovíc, H.K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, New York, 1986.
[32] H.J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser Boston, Boston, 1990.
[33] J.G. Llavona, *Approximation of Continuously Differentiable Functions*, North–Holland Math. Stud. 130, North–Holland, Amsterdam, 1986.
[34] A. Leizarowitz, *Order reduction is invalid for singularly perturbed control problems with a vector fast variable*, Math. Control Signals Systems, 15 (2002), pp. 101–119.
[35] R.S. Liptser and A.N. Shiryev, *Statistic of Random Processes* I. *General Theory*, Springer-Verlag, New York, 1977.

[36] S.D. Naidu, *Singular perturbations and time scales in control theory and applications: An overview*, Dynam. Contin. Discrete Impuls. Systems, Series B: Applications and Algorithms, 9 (2002), pp. 233–278.
[37] R.E. O'Malley, Jr., *Introduction to Singular Perturbations*, Academic Press, New York, 1974.
[38] R.E. O'Malley, Jr., *Singular perturbations and optimal control*, in Mathematical Control Theory, W.A. Copel, ed., Lecture Notes in Math. 680, Springer-Verlag, Berlin, 1978, pp. 170–218.
[39] M. Quincampoix and H. Zhang, *Singular perturbations in non-linear optimal control systems*, Differential Integral Equations, 8 (1995), pp. 931–944.
[40] A.A. Pervozvansky and V. Gaitsgory, *Theory of Suboptimal Decisions*, Kluwer Academic, Dordrecht, The Netherlands, 1988.
[41] V.A. Plotnikov, A.V. Plotnikov, and A.N. Vituk, *Differential Equations with Multivalued Right-Hand Sides: Asymptotic Methods*, AstroPrint, Odessa, Russia, 1999 (in Russian).
[42] S.P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhäuser Boston, Boston, 1994.
[43] R.H. Stockbridge, *Time-average control of a martingale problem. Existence of a stationary solution*, Ann. Probab., 18 (1990), pp. 190–205.
[44] A.B. Vasil'eva and V.F. Butuzov, *Asymptotic Expansions of Solutions of Singularly Perturbed Equations*, Nauka, Moscow, 1973 (in Russian).
[45] V. Veliov, *A generalization of Tichonov theorem for singularly perturbed differential inclusions*, J. Dynam. Control Systems, 3 (1997), pp. 1–28.
[46] A. Vigodner, *Limits of singularly perturbed control problems with statistical limits of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.
[47] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[48] G.G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications. A Singular Perturbation Approach*, Springer-Verlag, New York, 1997.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 341–356

BOUNDARY FEEDBACK STABILIZATION OF THE UNDAMPED EULER–BERNOULLI BEAM WITH BOTH ENDS FREE*

FAMING GUO† AND FALUN HUANG‡

Abstract. In this paper, we are concerned with a boundary feedback system of a class of nonuniform undamped Euler–Bernoulli beam with both ends free. We give some sufficient conditions and some necessary conditions for the system to have exponential stability. Our method is based on the operator semigroup technique, the multiplier technique, and the contradiction argument of the frequency domain method.

Key words. C_0-semigroup, Euler–Bernoulli beam, exponential stability, multiplier technique, frequency domain method

AMS subject classifications. 93C20, 35B37, 35B40

DOI. 10.1137/S0363012901380961

1. Introduction. The exponential stability of the boundary feedback system of an Euler–Bernoulli beam with one or two ends fixed has been studied extensively during the past two decades, but little attention has been paid to the case of the beam with both ends free. However, a long and thin object flying in the sky can be considered as a elastic beam with both ends free, and the boundary feedback stabilization of this beam is of great interest, both in control theory and in engineering practice. In this paper we shall consider this kind of problem. More precisely, we consider a undamped nonuniform Euler–Bernoulli beam of length L with both ends free, and its transverse vibration can be described by the following boundary feedback system:

(1.1)
$$\begin{cases} \rho(x)\dfrac{\partial^2}{\partial t^2}\omega(x,t) + \dfrac{\partial^2}{\partial x^2}\left(EI(x)\dfrac{\partial^2}{\partial x^2}\omega(x,t)\right) = 0, & (x,t)\in(0,L)\times R^+, \\ EI(x)\dfrac{\partial^2}{\partial x^2}\omega(x,t)\Big|_{x=L} = -\dfrac{\partial}{\partial x}\left(EI(x)\dfrac{\partial^2}{\partial x^2}\omega(x,t)\right)\Big|_{x=L} = 0, & t\in R^+, \\ EI(x)\dfrac{\partial^2}{\partial x^2}\omega(x,t)\Big|_{x=0} = k_1\dfrac{\partial}{\partial x}\omega(0,t) + k_2\dfrac{\partial^2}{\partial t\partial x}\omega(0,t), & t\in R^+, \\ -\dfrac{\partial}{\partial x}\left(EI(x)\dfrac{\partial^2}{\partial x^2}\omega\right)\Big|_{x=0} = k_3\omega(0,t) + k_4\dfrac{\partial}{\partial t}\omega(0,t), & t\in R^+, \\ \omega(x,0) = \omega_0(x), \quad \dfrac{\partial}{\partial t}\omega(x,0) = \omega_1(x), & x\in(0,L), \end{cases}$$

*Received by the editors April 4, 2001; accepted for publication (in revised form) May 22, 2003; published electronically June 25, 2004. This research was supported by the National Natural Science Foundation of China.

http://www.siam.org/journals/sicon/43-1/38096.html

†School of Applied Mathematics, University of Electronic Science and Technology of China, Chengdu, 610054, People's Republic of China (famingguo@163.com).

‡Mathematical College, Sichuan University, Chengdu, 610064, People's Republic of China.

where $\omega(x,t)$ denotes the transverse displacement of the beam at position x and time t, $\rho(x)$ its mass density at position x, $EI(x)$ its flexural rigidity, and $k_j \geq 0$ $(j = 1,2,3,4)$ the feedback coefficients. The bending moment at $x = 0$, $EI\frac{\partial^2}{\partial x^2}\omega|_{x=0}$, is controlled by the linear feedback of rotation angle and angular velocity, and the shear force at $x = 0$, $-\frac{\partial}{\partial x}\big(EI\frac{\partial^2}{\partial x^2}\omega\big)|_{x=0}$, is controlled by the linear feedback of displacement and velocity. We refer to [3], [5], [6], and [30] for the precise description of the problem and for more technical details.

In this paper we are interested in the following feedback stabilization problem: *Under what conditions on k_j $(j = 1,2,3,4)$, does the energy $E(t)$ (see* (2.2) *for its definition) of the system* (1.1) *exponentially decay?*

In 1980 a class of Euler–Bernoulli beams with structural damping was proposed for large flexible structures by Chen and Russell [6] and Russell [14], and the exponential stability of an Euler–Bernoulli beam with structural damping has been discussed at length, beginning with [6] by Chen and Russell and continuing with the more general results [13], [14] by Huang. We refer to Chen et al. [4] and Kim [15] for the beam with viscous damping, and to Chen, Liu, and Liu [7] and Liu and Liu [23], [24] for the beam with Kelvin–Voigt damping. For the boundary feedback stabilization of an undamped Euler–Bernoulli beam that is clamped at one end and free at the other end, Littman and Markus [22] proved the strong stabilization together with the lack of exponential stabilization under velocity feedback (also see Rao [27]); Conrad and Morgül [9] used the energy multiplier method to show the exponential stabilization under linear boundary feedback control $(-\alpha\omega_t(1,t)+\omega_{xxxt}(1,t))$. More recently, Guo [11] obtained the exponential stabilization of the nonuniform Euler–Bernoulli beam under linear boundary feedback control by the Riesz basis approach. See references therein for further articles.

In the literature, various techniques have been developed to address the stabilization/controllability problem for distributed parameter systems. The spectral method is a useful method for the one space dimensional problem (see Guo [11]). For the high dimensional system with constant coefficients, the Hilbert uniqueness method (HUM), introduced by Lions in [21], is a powerful tool. For the counterpart with variable coefficients, Yao [31] has introduced the Riemann geometry method. We refer to Bardos, Lebeau, and Rauch [2], Fursikov and Imanuvilov [10], and Zhang [32] for other related methods.

However, the known methods are not easily adapted to our stabilization problem. Indeed, in system (1.1), the beam equation uses variable coefficients, both ends of the beam are free, and the boundary conditions are more complex than those in the literature. These restrictions certainly introduce some technical difficulties. Therefore, we need to develop a new method to solve our problem. Our approach is based on the operator semigroup technique, the multiplier technique with the contradiction argument of a frequency domain method. Recall that multiplier techniques were developed in the work of Lions [21], Lagnese [18], Komornik [16], [17], Lasieka [19], Lasieka and Triggiani [20], and Zuazua [33], [34] for various PDEs and control problems. On the other hand, the frequency domain method is based on the boundedness on the imaginary axis of the resolvent of a C_0-semigroup generator to establish the exponential stability of the C_0-semigroup on a Hilbert space (see Huang [12] and Prüss [26]).

This paper is organized as follows. In section 2, we will state our main results. In section 3, we show the well-posedness of the system and derive some spectral properties of the underlying semigroup. Finally, in sections 4 and 5, we will prove our main results.

2. Statement of the main results. Throughout this paper, we need the following natural hypothesis:

$$\rho(\cdot) \in C^1[0,L], \quad EI(\cdot) \in C^2[0,L]; \quad \rho(x) > 0, \quad EI(x) > 0, \quad x \in [0,L]. \tag{2.1}$$

Denote the energy of system (1.1) by

$$E(t) = \frac{1}{2}\left[\int_0^L \left(EI\left|\frac{\partial^2}{\partial x^2}\omega\right|^2 + \rho\left|\frac{\partial}{\partial t}\omega\right|^2\right)dx + k_1\left|\frac{\partial}{\partial x}\omega(0,t)\right|^2 + k_3|\omega(0,t)|^2\right], \tag{2.2}$$

where $k_1|\frac{\partial}{\partial x}\omega(0,t)|^2 + k_3|\omega(0,t)|^2$ represents the energy of the rigid motion of the elastic system. Simple calculations yield that

$$\begin{aligned}\frac{d}{dt}E(t) = &\int_0^L \left[EI\frac{\partial^2}{\partial x^2}\omega\frac{\partial^3}{\partial t\partial x^2}\omega + \rho\frac{\partial}{\partial t}\omega\frac{\partial^2}{\partial t^2}\omega\right]dx \\ &+ k_1\frac{\partial}{\partial x}\omega(0,t)\frac{\partial^2}{\partial t\partial x}\omega(0,t) + k_3\omega(0,t)\frac{\partial}{\partial t}\omega(0,t)\end{aligned} \tag{2.3}$$

and

$$\begin{aligned}\int_0^L \rho\frac{\partial}{\partial t}\omega\frac{\partial^2}{\partial t^2}\omega dx &= -\int_0^L \frac{\partial}{\partial t}\omega\frac{\partial^2}{\partial x^2}\left(EI\frac{\partial^2}{\partial x^2}\omega\right)dx \\ &= -\frac{\partial}{\partial t}\omega(0,t)\left(k_3\omega(0,t) + k_4\frac{\partial}{\partial t}\omega(0,t)\right) \\ &\quad + \int_0^L \frac{\partial^2}{\partial t\partial x}\omega\frac{\partial}{\partial x}\left(EI\frac{\partial^2}{\partial x^2}\omega\right)dx \\ &= -\frac{\partial}{\partial t}\omega(0,t)\left(k_3\omega(0,t) + k_4\frac{\partial}{\partial t}\omega(0,t)\right) \\ &\quad - \frac{\partial^2}{\partial t\partial x}\omega(0,t)\left(k_1\frac{\partial}{\partial x}\omega(0,t) + k_2\frac{\partial^2}{\partial t\partial x}\omega(0,t)\right) \\ &\quad - \int_0^L \frac{\partial^3}{\partial t\partial x^2}\omega\left(EI\frac{\partial^2}{\partial x^2}\omega\right)dx.\end{aligned} \tag{2.4}$$

From (2.3) and (2.4) we obtain

$$\frac{d}{dt}E(t) = -k_4\left|\frac{\partial}{\partial t}\omega(0.t)\right|^2 - k_2\left|\frac{\partial^2}{\partial t\partial x}\omega(0,t)\right|^2, \tag{2.5}$$

which means that $k_2 \geq 0$ and $k_4 \geq 0$ are necessary for the energy $E(t)$ to be not increasing.

It is easy to see that $k_1 \neq 0$ and $k_3 \neq 0$ are necessary for the energy $E(t)$ to uniformly exponentially decay. In fact, if $k_1 = 0$, then the beam can be rotated rigidly, and $\omega(x,t) = x, x \in [0,L]$ is an eigenvector belonging to eigenvector $\lambda = 0$. Likewise, if $k_3 = 0$, then the beam can be translated rigidly, and $\omega(x,t) \equiv 1, x \in [0,L]$ is an eigenvector belonging to eigenvalue $\lambda = 0$.

For simplicity, we will denote $\frac{\partial}{\partial x}u$ and $\frac{\partial}{\partial t}u$ by u' and $\dot{u}$, respectively. Let $H = L^2_\rho(0,L)$ with the norm $\|u\| = (\int_0^L \rho(x)|u(x)|^2dx)^{\frac{1}{2}}$ and $H^k = H^k(0,L), k \geq 1$, where $H^k(0,L)$ is the Sobolev space of order k, and let $H_E = H^2(0,L)$ with the equivalent norm

$$\|u\|_{H_E} = \left(k_1|u'(0)|^2 + k_3|u(0)|^2 + \int_0^L EI|u''|^2dx\right)^{\frac{1}{2}} \quad (k_1, k_3 > 0).$$

Define $\mathcal{H} = H_E \oplus H$ with the norm

$$\|(u,v)\|_{\mathcal{H}} = (\|u\|^2_{H_E} + \|v\|^2_H)^{\frac{1}{2}}.$$

To formulate (1.1) as an abstract Cauchy problem on $\mathcal{H}$, we define a linear operator $\mathcal{A}$ as follows:

$$D(\mathcal{A}) = \Bigg\{(u,v) \in H^4 \oplus H^2| \quad EI\frac{d^2}{dx^2}u \in H^2,\, EI\frac{d^2}{dx^2}u\bigg|_{x=L} = -\frac{d}{dx}\left(EI\frac{d^2}{dx^2}u\right)\bigg|_{x=L} = 0,$$

$$EI\frac{d^2}{dx^2}u\bigg|_{x=0} = k_1u'(0) + k_2v'(0), -\frac{d}{dx}\left(EI\frac{d^2}{dx^2}u\right)\bigg|_{x=0} = k_3u(0) + k_4v(0)\Bigg\} \tag{2.6}$$

and

$$\mathcal{A}\begin{pmatrix}u\\v\end{pmatrix} = \begin{pmatrix}0 & I\\ -A & 0\end{pmatrix}\begin{pmatrix}u\\v\end{pmatrix}, \quad (u,v) \in D(\mathcal{A}), \tag{2.7}$$

where

$$A = \rho^{-1}\frac{d^2}{dx^2}\left(EI\frac{d^2}{dx^2}\cdot\right)$$

and $D(A) = \{u \in H^2|$ there exists $v \in H^2$ such that $(u,v) \in D(\mathcal{A})\}$. Then the system (1.1) can be formulated as the following Cauchy problem on $\mathcal{H}$:

$$\begin{pmatrix}\dot{u}\\ \dot{v}\end{pmatrix} = \mathcal{A}\begin{pmatrix}u\\v\end{pmatrix}, \quad \begin{pmatrix}u(0)\\v(0)\end{pmatrix} = \begin{pmatrix}w_0\\w_1\end{pmatrix}.$$

Now we can state our main results as follows.

THEOREM 2.1. *Let* (2.1) *hold and* $k_j > 0, j = 1,2,3,4$. *Then the* C_0*-semigroup* $e^{t\mathcal{A}}$ *is uniformly exponentially stable; i.e., there exist constants* $M, \omega > 0$ *such that*

$$\|e^{t\mathcal{A}}\| \leq Me^{-\omega t}, \quad t \geq 0.$$

THEOREM 2.2. *Assume that* $k_1, k_3 > 0$ *and that the* C_0*-semigroup* $e^{t\mathcal{A}}$ *decays uniformly exponentially. Then* $k_2 > 0$ *and* $k_4 > 0$.

3. Preliminaries. In this section, we will prove that $(\mathcal{A}, D(\mathcal{A}))$ generates a C_0-contraction semigroup on $\mathcal{H}$, which shows the well-posedness of system (1.1), and give some spectral properties of the generator $\mathcal{A}$.

THEOREM 3.1. *Let $k_1, k_2, k_3, k_4 > 0$. Then $\mathcal{A}$ is the infinitesimal generator of a C_0-contraction semigroup $e^{t\mathcal{A}}$ on $\mathcal{H}$.*

Proof. It is easy to see that $\mathcal{A}$ is densely defined in $\mathcal{H}$. Furthermore, for any $(u, v) \in D(\mathcal{A})$, we have

$$\begin{aligned} \operatorname{Re}\left\langle \mathcal{A}\begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \right\rangle_{\mathcal{H}} &= \operatorname{Re}\left\langle \begin{pmatrix} v \\ -Au \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \right\rangle_{\mathcal{H}} \\ &= \operatorname{Re}[\langle v, u\rangle_{H_E} - \langle Au, v\rangle_H], \end{aligned} \tag{3.1}$$

where

$$(v, u)_{H_E} = k_1 v'(0)\overline{u'(0)} + k_3 v(0)\overline{u(0)} + \int_0^L EIv''(x)\overline{u''(x)}dx \tag{3.2}$$

and

$$\begin{aligned} (Au, v)_H &= \int_0^L \frac{d^2}{dx^2}(EIu'')\overline{v}dx \\ &= k_3 u(0)\overline{v(0)} + k_4|v(0)|^2 + k_1 u'(0)\overline{v'(0)} + k_2|v'(0)|^2 + \int_0^L EIu''\overline{v''}dx. \end{aligned} \tag{3.3}$$

Substituting (3.2) and (3.3) into (3.1), we obtain

$$\operatorname{Re}\left\langle \mathcal{A}\begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \right\rangle_{\mathcal{H}} = -k_4|v(0)|^2 - k_2|v'(0)|^2,$$

which implies that $\mathcal{A}$ is dissipative in $\mathcal{H}$.

Finally, we show that $\lambda = 0 \in \rho(\mathcal{A})$ (the resolvent set of $\mathcal{A}$). For any $(f, g) \in \mathcal{H}$, we are going to solve the following equation:

$$\mathcal{A}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \quad (u, v) \in D(\mathcal{A}). \tag{3.4}$$

This implies

$$\begin{cases} v &= f, \\ Au &= -g, \end{cases} \tag{3.5}$$

and hence

$$\frac{d^2}{dx^2}(EIu'') = -\rho g. \tag{3.6}$$

Integrating from x to L and using the boundary conditions at $x = L$, we have

$$\frac{d}{dx}(EIu'') = \int_x^L \rho(z)g(z)dz, \tag{3.7}$$

and consequently,

$$EIu'' = -\int_x^L \int_y^L \rho(z)g(z)dzdy, \tag{3.8}$$

or equivalently,

$$u''(x) = -\frac{1}{EI(x)} \int_x^L \int_y^L \rho(z)g(z)dzdy, \quad x \in [0, L]. \tag{3.9}$$

Therefore, we have

$$u(x) = u(0) + xu'(0) - \int_0^x \int_0^y \frac{1}{EI(y_1)} \int_{y_1}^L \int_{y_2}^L \rho(z)g(z)dzdy_2dy_1dy, \quad x \in [0, L]. \tag{3.10}$$

Let $x = 0$ in (3.7); we can obtain

$$-(k_3u(0) + k_4v(0)) = \int_0^L \rho(z)g(z)dz,$$

which, using (3.5), yields

$$u(0) = -\frac{1}{k_3} \left(k_4 f(0) + \int_0^L \rho(z)g(z)dz \right). \tag{3.11}$$

Similarly, let $x = 0$ in (3.8); we have

$$u'(0) = -\frac{1}{k_1} \left(k_2 f'(0) + \int_0^L \int_y^L \rho(z)g(z)dzdy \right). \tag{3.12}$$

From (3.10), (3.11), and (3.12) we can assert that

$$u(x) = -\frac{1}{k_3} \left[k_4 f(0) + \int_0^L \rho(z)g(z)dz - \frac{x}{k_1} \left(k_2 f'(0) + \int_0^L \int_y^L \rho(z)g(z)dzdy \right) \right]$$

$$- \int_0^x \int_0^y \frac{1}{EI(y_2)} \int_{y_2}^L \int_{y_3}^L \rho(z)g(z)dzdy_3dy_2dy_1, \qquad x \in [0, L]. \tag{3.13}$$

Since hypothesis (2.1) is satisfied, we can easily deduce that $u \in H^2$ and that $(u, v) \in D(\mathcal{A})$ is the unique solution of (3.4) ($v = f$). Hence, $\lambda = 0 \in \rho(\mathcal{A})$. Finally, from the above discussion and the Lumer–Phillips theorem [25, Theorem 1.4.3], it follows that $\mathcal{A}$ generates a C_0-contraction semigroup. The proof has been completed. □

PROPOSITION 3.2. *Assume that $k_j > 0$, $j = 1, 2, 3, 4$. Then $iR \subset \rho(\mathcal{A})$.*

Proof. From the proof of Theorem 3.1 we have $\lambda = 0 \in \rho(\mathcal{A})$ and we can prove

$$\{\lambda \in \sigma(\mathcal{A}) : \mathrm{Im}\lambda \neq 0\} \subset \sigma_p(A)$$

in a similar way as in [7, Lemma 4.1]. Therefore it suffices to show $i\omega \notin \sigma_p(\mathcal{A})$. Indeed, if it is not true, then there exists $\omega \in R$, $\omega \neq 0$, such that $i\omega \in \sigma_p(\mathcal{A})$. Hence, there exists $(u, v) \in D(\mathcal{A})$, $(u, v) \neq 0$, such that

$$(i\omega - \mathcal{A}) \begin{pmatrix} u \\ v \end{pmatrix} = 0,$$

which implies

$$i\omega u = v \tag{3.14}$$

and

$$Au + i\omega v = 0. \tag{3.15}$$

From (3.14) and (3.15) it follows that

$$\frac{d^2}{dx^2}\left(EI\frac{d^2}{dx^2}u\right) - \omega^2\rho u = 0, \tag{3.16}$$

and consequently,

$$\begin{aligned}
0 &= \int_0^L \frac{d^2}{dx^2}\left(EI\frac{d^2}{dx^2}u\right)\overline{u}dx - \omega^2\int_0^L \rho|u|^2dx \\
&= k_3|u(0)|^2 + i\omega k_4|u(0)|^2 - \int_0^L \frac{d}{dx}\left(EI\frac{d^2}{dx^2}u\right)\frac{d}{dx}\overline{u}dx - \omega^2\int_0^L \rho|u|^2dx \\
&= k_3|u(0)|^2 + i\omega k_4|u(0)|^2 + k_1\left|\frac{d}{dx}u(0)\right|^2 + i\omega k_2\left|\frac{d}{dx}u(0)\right|^2 \\
&\quad + \int_0^L EI\left|\frac{d^2}{dx^2}u\right|^2 dx - \omega^2\int_0^L \rho|u|^2dx.
\end{aligned}$$

Using $\omega \neq 0$, we conclude that

$$\begin{cases}
k_3|u(0)|^2 + k_1\left|\dfrac{d}{dx}u(0)\right|^2 + \displaystyle\int_0^L EI\left|\frac{d^2}{dx^2}u\right|^2 dx - \omega^2\int_0^L \rho|u|^2dx = 0, \\
k_4|u(0)|^2 + k_2\left|\dfrac{d}{dx}u(0)\right|^2 = 0.
\end{cases} \tag{3.17}$$

Since hypothesis (2.1) is satisfied, we have $u \in H^4[0,L]$ and hence $u \in C^3[0,L]$. Moreover, it follows from (3.17) that

$$\begin{cases}
u(0) = u'(0) = 0, \\
\displaystyle\int_0^L EI|u''|^2dx = \omega^2\int_0^L \rho|u|^2dx,
\end{cases} \tag{3.18}$$

which implies that

$$v(0) = v'(0) = 0.$$

Therefore, using the boundary conditions at $x = 0$, we get

$$EI(0)u''(0) = k_1u(0) + k_2v(0) = 0$$

and

$$(EIu'')'|_{x=0} = EI'(0)u''(0) + EI(0)u'''(0) = k_3u'(0) + k_4v'(0) = 0,$$

which, together with $EI(0) > 0$, implies that

$$u''(0) = u'''(0) = 0. \tag{3.19}$$

From (3.18) and (3.19), it follows that u is a solution of the following equation:

$$\begin{cases} \dfrac{d^2}{dx^2}\left(EI\dfrac{d^2}{dx^2}u\right) - \omega^2\rho u = 0, \\ \\ u(0) = u'(0) = u''(0) = u'''(0) = 0. \end{cases} \tag{3.20}$$

The uniqueness theorem of ODEs shows that $u \equiv 0$ and $v = i\omega\rho u \equiv 0$. This is in contradiction with $(u, v) \neq 0$, and the proof is completed. □

4. Proof of Theorem 2.1. Clearly, if $w = w(x, t)$ is the solution of the system (1.1), then

$$\left(w, \frac{\partial}{\partial t}w\right) = e^{t\mathcal{A}}(w_0, w_1) \quad \text{and} \quad \left\|\left(w, \frac{\partial}{\partial t}w\right)\right\|_{\mathcal{H}}^2 = 2E(t), \qquad t > 0.$$

Hence, the uniformly exponential decay of the energy $E(t)$ is equivalent to the uniform exponential stability of C_0-semigroup $e^{t\mathcal{A}}$. It follows from Proposition 3.2 and the frequency domain results (see [12], [26]) that we need only to prove

$$\sup_{\lambda \in iR} \|(\lambda - \mathcal{A})^{-1}\| < +\infty. \tag{4.1}$$

If (4.1) is not true, using $iR \subset \rho(\mathcal{A})$, then we can deduce that there exist $\lambda_n = i\omega_n \in iR, (u_n, v_n) \in D(\mathcal{A})$ such that

$$\|(u_n, v_n)\|_{\mathcal{H}} = 1, \quad |\omega_n| \to \infty (n \to \infty) \tag{4.2}$$

and

$$(i\omega_n - \mathcal{A})(u_n, v_n) = (f_n, g_n) \to 0 \quad \text{in} \quad \mathcal{H}. \tag{4.3}$$

From (4.3), it follows that

$$i\omega_n u_n - v_n = f_n \to 0 \quad \text{in} \quad H_E, \tag{4.4}$$

$$\frac{d^2}{dx^2}\left(EI\frac{d^2}{dx^2}u_n\right) + i\omega_n\rho v_n = \rho g_n \to 0 \quad \text{in} \quad H. \tag{4.5}$$

By (4.4), we have

$$\|f_n\|_{H_E}^2 = k_1|f_n'(0)|^2 + k_3|f_n(0)|^2 + \int_0^L EI|f_n''|^2 dx \to 0,$$

which yields that

$$f_n'(0) \to 0, \quad f_n(0) \to 0. \tag{4.6}$$

Again by (4.4), we have that $(f_n, u_n)_H = i\omega_n\|u_n\|_H^2 - (v_n, u_n)_H \to 0$, and consequently,

$$\|u_n\|_H^2 = \frac{1}{i\omega_n}[(f_n, u_n)_H + (v_n, u_n)_H] \to 0. \tag{4.7}$$

Moreover, we have

$$\begin{aligned}
(g_n, u_n)_H &= \int_0^L \frac{d^2}{dx^2}\left(EI\frac{d^2}{dx^2}u_n\right)\overline{u_n}dx + i\omega_n\int_0^L \rho v_n\overline{u_n}dx \\
&= k_3|u_n(0)|^2 + k_4 v_n(0)\overline{u_n(0)} - \int_0^L \frac{d}{dx}\left(EI\frac{d^2}{dx^2}u_n\right)\overline{u_n'}dx + i\omega_n(v_n,u_n)_H \\
&= k_3|u_n(0)|^2 + k_4 v_n(0)\overline{u_n(0)} + k_1|u_n'(0)|^2 \\
&\qquad + k_2 v_n'(0)\overline{u_n'(0)} + \int_0^L EI|u_n''|^2dx + i\omega_n(v_n,u_n)_H \\
&= \|u_n\|_{H_E}^2 + k_4(i\omega_n u_n(0) - f_n(0))\overline{u_n(0)} \\
&\qquad + k_2(i\omega_n u_n'(0) - f_n'(0))\overline{u_n'(0)} + i\omega_n(v_n,u_n)_H \\
&= \|u_n\|_{H_E}^2 + k_4 i\omega_n|u_n(0)|^2 + k_2 i\omega_n|u_n'(0)|^2 \\
&\qquad - k_4 f_n(0)\overline{u_n(0)} - k_2 f_n'(0)\overline{u_n'(0)} + i\omega_n(v_n,u_n)_H \\
&\to 0. \qquad (4.8)
\end{aligned}$$

Therefore, from (4.6), (4.8), $\mathrm{Re}(v_n,u_n)_H \to 0$, and $|\omega_n| \to 0$, we can deduce that

$$u_n(0) \to 0, \quad u_n'(0) \to 0. \qquad (4.9)$$

From (4.4), it follows that

$$(f_n,u_n)_{H_E} = (i\omega_n u_n - v_n, u_n)_{H_E} = i\omega_n\|u_n\|_{H_E}^2 - \overline{(u_n,v_n)}_{H_E} \to 0,$$

and consequently,

$$\mathrm{Re}(u_n,v_n)_{H_E} \to 0, \quad \|u_n\|_{H_E}^2 + \mathrm{Im}\left(u_n, \frac{v_n}{\omega_n}\right)_{H_E} \to 0. \qquad (4.10)$$

Since

$$\begin{aligned}
&\left(\rho^{-1}\left[\frac{d^2}{dx^2}\left(EI\frac{d^2}{dx^2}u_n\right) + i\omega_n v_n\right], v_n\right)_H \qquad (4.11)\\
&\quad = (u_n,v_n)_{H_E} + k_4|v_n(0)|^2 + k_2|v_n'(0)|^2 + i\omega_n\|v_n\|_H^2 \to 0,
\end{aligned}$$

we can easily deduce that

$$\mathrm{Re}(u_n,v_n)_{H_E} + k_4|v_n(0)|^2 + k_2|v_n'(0)|^2 \to 0.$$

Combining this together with (4.10), we get

$$v_n(0) \to 0, \quad v_n'(0) \to 0. \qquad (4.12)$$

Without loss of generality, let $\|v_n\|^2 \to \gamma \in [0,1]$; we obtain that $\|u_n\|_{H_E}^2 \to 1-\gamma$. From (4.11) it follows that

$$\mathrm{Im}\left(u_n, \frac{v_n}{\omega_n}\right)_{H_E} + \|v_n\|_H^2 \to 0; \qquad (4.13)$$

combining this with (4.10) implies $\gamma = \frac{1}{2}$.

Furthermore, if

$$\begin{pmatrix}\xi_n(x,t)\\ \eta_n(x,t)\end{pmatrix} = e^{t\mathcal{A}}\begin{pmatrix}u_n(x)\\ v_n(x)\end{pmatrix},$$

then

$$\begin{aligned}\left\|\begin{pmatrix}\xi_n(.,t)\\ \eta_n(.,t)\end{pmatrix} - e^{t\omega_n i}\begin{pmatrix}u_n\\ v_n\end{pmatrix}\right\|_{\mathcal{H}} &= \left\|e^{t(\mathcal{A}-i\omega_n)}\begin{pmatrix}u_n\\ v_n\end{pmatrix} - \begin{pmatrix}u_n\\ v_n\end{pmatrix}\right\|_{\mathcal{H}}\\ &= \left\|\int_0^t e^{\tau(\mathcal{A}-i\omega_n)}(\mathcal{A}-i\omega_n I)\begin{pmatrix}u_n\\ v_n\end{pmatrix} d\tau\right\|_{\mathcal{H}}\\ &\le \int_0^t \|e^{\tau(\mathcal{A}-i\omega_n)}\| \left\|(\mathcal{A}-i\omega_n I)\begin{pmatrix}u_n\\ v_n\end{pmatrix}\right\|_{\mathcal{H}} d\tau\\ &\le t\left\|(\mathcal{A}-i\omega_n)\begin{pmatrix}u_n\\ v_n\end{pmatrix}\right\|_{\mathcal{H}}.\end{aligned}$$

Therefore, we deduce that

$$\sup_{t\in[0,T]}\left\|\begin{pmatrix}\xi_n(.,t)\\ \eta_n(.,t)\end{pmatrix} - e^{i\omega_n t}\begin{pmatrix}u_n\\ v_n\end{pmatrix}\right\|_{\mathcal{H}} \to 0 \tag{4.14}$$

for any $T>0$. Note that $|e^{i\omega_n t}u_n(0)| = |u_n(0)| \to 0$ and $|e^{i\omega_n t}u_n'(0)| = |u_n'(0)| \to 0$; together with the definition of $\|(.,.)\|_{\mathcal{H}}$ and (4.9), we can assert that

$$\xi_n(0,t)\to 0, \quad \frac{\partial}{\partial x}\xi_n(0,t)\to 0 \tag{4.15}$$

uniformly for t in compact intervals of R^+.

On the other hand, we have

$$\begin{aligned}\left\|\begin{pmatrix} i\omega_n\xi_n-\eta_n\\ \rho^{-1}\frac{\partial^2}{\partial x^2}(EI\frac{\partial^2}{\partial x^2}\xi_n)+i\omega_n\eta_n\end{pmatrix}\right\|_{\mathcal{H}} &= \left\|(i\omega_n-\mathcal{A})\begin{pmatrix}\xi_n\\ \eta_n\end{pmatrix}\right\|_{\mathcal{H}}\\ &= \left\|(i\omega_n-\mathcal{A})e^{t\mathcal{A}}\begin{pmatrix}u_n\\ v_n\end{pmatrix}\right\|_{\mathcal{H}}\\ &= \left\|e^{t\mathcal{A}}(i\omega_n-\mathcal{A})\begin{pmatrix}u_n\\ v_n\end{pmatrix}\right\|_{\mathcal{H}}\\ &\le \left\|(i\omega_n-\mathcal{A})\begin{pmatrix}u_n\\ v_n\end{pmatrix}\right\|_{\mathcal{H}} \to 0\end{aligned} \tag{4.16}$$

uniformly for $t\in[0,\infty)$. Therefore,

$$(i\omega_n\xi_n-\eta_n,\xi_n)_{H_E} = i\omega_n\|\xi_n\|_{H_E} - \overline{(\xi_n,\eta_n)} \to 0$$

uniformly for $t\in[0,\infty)$. This yields

$$\mathrm{Re}(\xi_n,\eta_n)_{H_E}\to 0 \quad \text{uniformly for} \quad t\in[0,\infty).$$

Similar to (4.11), we can get

$$\begin{aligned}&\left(\rho^{-1}\frac{\partial^2}{\partial x^2}\left(EI\frac{\partial^2}{\partial x^2}\xi_n\right)+i\omega_n\eta_n,\eta_n\right)_H\\ &\quad= (\xi_n,\eta_n)_{H_E} + k_4|\eta_n(0,t)|^2 + k_2|\eta_n'(0,t)|^2 + i\omega_n\|\eta_n\|_H^2 \to 0\end{aligned}$$

uniformly for t in compact intervals of R^+, and consequently,

$$\eta_n(0,t) \to 0, \qquad \eta_n'(0,t) \to 0 \tag{4.17}$$

uniformly for t in compact intervals of R^+. Since $(\xi_n, \eta_n) \in D(\mathcal{A})$, using (4.15) and (4.16), we obtain

$$EI(0)\frac{\partial^2}{\partial x^2}\xi_n(0,t) = k_1\xi_n'(0,t) + k_2\eta_n'(0,t) \to 0$$

and

$$\begin{aligned} -\frac{\partial}{\partial x}\left(EI\frac{\partial^2}{\partial x^2}\xi_n\right)\Big|_{x=0} &= \frac{d}{dx}EI(0)\frac{\partial^2}{\partial x^2}\xi_n(0,t) + EI(0)\frac{\partial^3}{\partial x^3}\xi_n(0,t) \\ &= k_3\xi_n(0,t) + k_4\eta_n(0,t) \to 0 \end{aligned}$$

uniformly for t in compact intervals of R^+. Hence

$$\frac{\partial^2}{\partial x^2}\xi_n(0,t) \to 0, \qquad \frac{\partial^3}{\partial x^3}\xi_n(0,t) \to 0 \tag{4.18}$$

uniformly for t in compact intervals of R^+. Moreover,

$$\begin{aligned} &EI(L)\frac{\partial^2}{\partial x^2}\xi_n(L,t) = 0, \\ &-\frac{\partial}{\partial x}\left(EI\frac{\partial^2}{\partial x^2}\xi_n(x,t)\right)\Big|_{x=L} = \left(\frac{\partial}{\partial x}EI\frac{\partial^2}{\partial x^2}\xi_n(x,t) + EI\frac{\partial^3}{\partial x^3}\xi_n(x,t)\right)\Big|_{x=L} = 0, \end{aligned}$$

which implies that

$$\frac{\partial^2}{\partial x^2}\xi_n(L,t) = 0, \qquad \frac{\partial^3}{\partial x^3}\xi_n(L,t) = 0, \quad t > 0. \tag{4.19}$$

Furthermore, it follows from (4.16) that

$$i\omega_n\xi_n - \eta_n = h_n \to 0$$

in H_E uniformly for t, and hence, it holds in H. Therefore,

$$\xi_n(.,t) = \frac{1}{i\omega_n}(h_n + \eta_n) \to 0 \tag{4.20}$$

in H uniformly for t.

Now let $M\xi = 2(x-L)e^{-\theta x}\frac{\partial}{\partial x}\xi$, where θ is such a positive constant that $\theta\rho - \rho' \geq 0$ and $3\theta EI + (EI)' \geq 0$. It is easy to see that hypothesis (2.1) guarantees the existence of θ. For $T > 0$, we have

$$\begin{aligned} 0 &= \operatorname{Re}\int_0^T\int_0^L \left[\rho\frac{\partial^2}{\partial t^2}\xi_n + \frac{\partial^2}{\partial x^2}\left(EI\frac{\partial^2}{\partial x^2}\xi_n\right)\right] M\overline{\xi_n}dxdt \\ &= \operatorname{Re}\left\{\left[\int_0^L \rho\frac{\partial}{\partial t}\xi_n M\overline{\xi_n}dx\right]_0^L - \int_0^T\int_0^L \rho\frac{\partial}{\partial t}\xi_n M\frac{\partial}{\partial t}\overline{\xi_n}dxdt\right. \\ &\quad + \int_0^T\left[\frac{\partial}{\partial x}\left(EI\frac{\partial^2}{\partial x^2}\xi_n\right)M\overline{\xi_n}\right]_0^L dt - \int_0^T\left[\left(EI\frac{\partial^2}{\partial x^2}\xi_n\right)\frac{\partial}{\partial x}M\overline{\xi_n}\right]_0^L dt \\ &\quad \left. + \int_0^T\int_0^L EI\frac{\partial^2}{\partial x^2}\xi_n\frac{\partial^2}{\partial x^2}M\overline{\xi_n}dxdt\right\}. \end{aligned} \tag{4.21}$$

Since

$$\frac{\partial}{\partial t}\xi_n = \eta_n$$

and

$$\mathrm{Re}\left(\rho\frac{\partial}{\partial t}\xi_n M\frac{\partial}{\partial t}\overline{\xi_n}\right) = \mathrm{Re}\left(\rho\frac{\partial}{\partial t}\xi_n 2(x-L)e^{-\theta x}\frac{\partial^2}{\partial x\partial t}\overline{\xi_n}\right) = e^{-\theta x}(x-L)\rho\frac{\partial}{\partial x}|\eta_n|^2,$$

we can deduce

$$\begin{aligned}
&\int_0^T\int_0^L \rho\frac{\partial}{\partial t}\xi_n M\frac{\partial}{\partial t}\overline{\xi_n}dxdt\\
&\quad= \int_0^T\int_0^L (x-L)\rho e^{\theta x}\frac{\partial}{\partial x}|\eta_n|^2 dxdt\\
&\quad= \int_0^T [(x-L)\rho e^{-\theta x}|\eta_n|^2]_0^L dt - \int_0^T\int_0^L e^{-\theta x}[\rho+(L-x)(\theta\rho-\rho')]|\eta_n|^2 dxdt\\
&\quad= \int_0^T L\rho(0)|\eta_n(0,t)|^2 dt - \int_0^T\int_0^L e^{-\theta x}[\rho+(L-x)(\theta\rho-\rho')]|\eta_n|^2 dxdt. \qquad (4.22)
\end{aligned}$$

From the definition of M, we have

$$\begin{aligned}
\frac{\partial^2}{\partial x^2}M\overline{\xi_n} &= \frac{\partial^2}{\partial x^2}\left[2(x-L)e^{-\theta x}\frac{\partial}{\partial x}\overline{\xi_n}\right]\\
&= \frac{\partial}{\partial x}\left[2e^{-\theta x}\frac{\partial}{\partial x}\overline{\xi_n} - 2\theta(x-L)e^{-\theta x}\frac{\partial}{\partial x}\overline{\xi_n} + 2(x-L)e^{-\theta x}\frac{\partial^2}{\partial x^2}\overline{\xi_n}\right]\\
&= -2\theta e^{-\theta x}\frac{\partial}{\partial x}\overline{\xi_n} + 2e^{-\theta x}\frac{\partial^2}{\partial x^2}\overline{\xi_n} - 2\theta e^{-\theta x}\frac{\partial}{\partial x}\overline{\xi_n} + 2\theta^2(x-L)e^{-\theta x}\frac{\partial}{\partial x}\overline{\xi_n}\\
&\quad - 2\theta(x-L)e^{-\theta x}\frac{\partial^2}{\partial x^2}\overline{\xi_n} + 2e^{-\theta x}\frac{\partial^2}{\partial x^2}\overline{\xi_n}\\
&\quad - 2\theta(x-L)e^{-\theta x}\frac{\partial^2}{\partial x^2}\overline{\xi_n} + 2(x-L)e^{-\theta x}\frac{\partial^3}{\partial x^3}\overline{\xi_n}\\
&= -[4\theta+2\theta^2(L-x)]e^{-\theta x}\frac{\partial}{\partial x}\overline{\xi_n} + 4e^{-\theta x}[1+\theta(L-x)]\frac{\partial^2}{\partial x^2}\overline{\xi_n}\\
&\quad + 2(x-L)e^{-\theta x}\frac{\partial^3}{\partial x^3}\overline{\xi_n}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\int_0^T\int_0^L EI\frac{\partial^2}{\partial x^2}\xi_n\frac{\partial^2}{\partial x^2}M\overline{\xi_n}dxdt &= -\int_0^T\int_0^L EI\frac{\partial^2}{\partial x^2}\xi_n\ 2\theta[2+\theta(L-x)]e^{-\theta x}\frac{\partial}{\partial x}\overline{\xi_n}dxdt\\
&\quad + \int_0^T\int_0^L 4e^{-\theta x}[1+\theta(L-x)]EI\left|\frac{\partial^2}{\partial x^2}\xi_n\right|^2 dxdt\\
&\quad + \int_0^T\int_0^L (x-L)e^{-\theta x}EI\frac{\partial}{\partial x}\left|\frac{\partial^2}{\partial x^2}\xi_n\right|^2 dxdt,
\end{aligned}$$

or equivalently,

(4.23)

$$\begin{aligned}
&\int_0^T \int_0^L (x-L)e^{-\theta x} EI \frac{\partial}{\partial x} \left| \frac{\partial^2}{\partial x^2}\xi_n \right|^2 dxdt \\
&= \int_0^T \left[(x-L)e^{-\theta x} EI \left| \frac{\partial^2}{\partial x^2}\xi_n \right|^2 \right]_0^L dt - \int_0^T \int_0^L \frac{d}{dx}[(x-L)e^{-\theta x} EI] \left| \frac{\partial^2}{\partial x^2}\xi_n \right|^2 dxdt \\
&= EI(0)L \int_0^T \left| \frac{\partial^2}{\partial x^2}\xi_n(0,t) \right|^2 dt \\
&\quad - \int_0^T \int_0^L e^{-\theta x}[(1+\theta(L-x))EI - (L-x)(EI)'] \left| \frac{\partial^2}{\partial x^2}\xi_n \right|^2 dxdt.
\end{aligned}$$

We also have

$$\begin{aligned}
&\int_0^T \int_0^L EI \frac{\partial^2}{\partial x^2}\xi_n \frac{\partial^2}{\partial x^2} M\overline{\xi_n} dxdt \\
&= EI(0)L \int_0^T \left| \frac{\partial^2}{\partial x^2}\xi_n(0,t) \right|^2 dt - \int_0^T \int_0^L \left(EI \frac{\partial^2}{\partial x^2}\xi_n \right) 2\theta(2+\theta(L-x))e^{-\theta x} \frac{\partial}{\partial x}\overline{\xi_n} dxdt \\
&\quad + \int_0^T \int_0^L e^{-\theta x}[3EI + 3\theta(L-x)EI + (L-x)(EI)'] \left| \frac{\partial^2}{\partial x^2}\xi_n \right|^2 dxdt.
\end{aligned}$$

From (4.21), (4.22), and (4.23), it follows that

$$\begin{aligned}
&e^{-\theta L} \int_0^T \int_0^L \left(EI \left| \frac{\partial^2}{\partial x^2}\xi_n \right|^2 + \rho|\eta_n|^2 \right) dxdt \\
&\quad \le \int_0^T \int_0^L e^{-\theta x}[3EI + 3\theta(L-x)EI + (L-x)(EI)'] \left| \frac{\partial^2}{\partial x^2}\xi_n \right|^2 dxdt \\
&\qquad + \int_0^T \int_0^L e^{-\theta x}[\rho + (L-x)(\theta\rho - \rho')]|\eta_n|^2 dxdt \\
&\quad = \mathrm{Re} \left\{ - \left[\int_0^L \rho \frac{\partial}{\partial t}\xi_n M\overline{\xi_n} dx \right]_0^T - \int_0^T \left[\frac{\partial}{\partial x} \left(EI \frac{\partial^2}{\partial x^2}\xi_n \right) M\overline{\xi_n} \right]_0^L dt \right. \\
&\qquad\quad + \int_0^T \left[EI \frac{\partial^2}{\partial x^2}\xi_n \frac{\partial}{\partial x} M\overline{\xi_n} \right]_0^L dt \\
&\qquad\quad + \int_0^T L\rho(0)|\eta_n(0,t)|^2 dt - EI(0)L \int_0^T \left| \frac{\partial^2}{\partial x^2}\xi_n(0,t) \right|^2 dt \\
&\qquad\quad \left. + \int_0^T \int_0^L EI \frac{\partial^2}{\partial x^2}\xi_n 2\theta(2+\theta(L-x))e^{-\theta x} \frac{\partial}{\partial x}\overline{\xi_n} dxdt \right\}. \qquad (4.24)
\end{aligned}$$

Finally, we are going to show that the right-hand side of (4.22) converges to zero as $n \to \infty$. Indeed, from (4.15), (4.17), and (4.18), we obtain

$$\int_0^T L\rho(0)|\eta_n(0,t)|^2 dt \to 0, \quad EI(0)L\int_0^T \left|\frac{\partial^2}{\partial x^2}\xi_n(0,t)\right|^2 dt \to 0 \tag{4.25}$$

and

$$\begin{aligned} &-\int_0^T \left[\frac{\partial}{\partial x}\left(EI\frac{\partial^2}{\partial x^2}\xi_n\right)M\overline{\xi_n}\right]_0^L dt + \int_0^T \left[EI\frac{\partial^2}{\partial x^2}\xi_n\frac{\partial}{\partial x}M\overline{\xi_n}\right]_0^L dt \\ &\quad = \int_0^T \frac{\partial}{\partial x}\left(EI\frac{\partial^2}{\partial x^2}\xi_n\right)\Big|_{x=0}\left(-2L\frac{\partial}{\partial x}\overline{\xi_n(0,t)}\right)dt \\ &\qquad - \int_0^T EI(0)\frac{\partial^2}{\partial x^2}\xi_n(0,t)\left[2(1+L\theta)\frac{\partial}{\partial x}\overline{\xi_n(0,t)} - 2L\frac{\partial^2}{\partial x^2}\overline{\xi_n(0,t)}\right]dt \\ &\qquad \to 0. \end{aligned} \tag{4.26}$$

Since $\|\xi_n(.,t)\|_{H_E} < 1$ and $\|\xi_n(.,t)\|_H \to 0$ uniformly for $t \in [0,T]$, we can easily deduce that

$$\left\|\frac{\partial}{\partial x}\xi_n(0,t)\right\|_H \le C\left\|\frac{\partial^2}{\partial x^2}\xi_n(.,t)\right\|_H^{\frac{1}{2}}\|\xi_n\|_H^{\frac{1}{2}} \to 0 \ \text{ uniformly for } \ t \in [0,T]$$

by the Hardy–Littlewood inequality. Consequently,

$$2\theta\int_0^T\int_0^L EI\frac{\partial^2}{\partial x^2}\xi_n[2+\theta(L-x)]e^{-\theta x}\frac{\partial}{\partial x}\overline{\xi_n}dxdt \to 0 \tag{4.27}$$

and

$$\int_0^L \rho\frac{\partial}{\partial t}\xi_n M\overline{\xi_n}dx|_0^T = \int_0^L \rho\eta_n 2(x-L)e^{-\theta x}\frac{\partial}{\partial x}\overline{\xi_n}dx|_0^T \to 0. \tag{4.28}$$

Using (4.25)–(4.28), we can assert that the right-hand side of (4.24) converges to zero as $n \to \infty$. Hence

$$\int_0^T\int_0^L \left(EI\left|\frac{\partial^2}{\partial x^2}\xi_n\right|^2 + \rho|\eta_n|^2\right)dxdt \to 0. \tag{4.29}$$

However, using $\|u_n\|_{H_E}^2 + \|v_n\|_H^2 = 1$, it follows from (4.14), (4.15), and (4.17) that

$$\begin{aligned} &\lim_{n\to\infty}\int_0^T\int_0^L \left(EI\left|\frac{\partial^2}{\partial x^2}\xi_n\right|^2 + \rho|\eta_n|^2\right)dxdt \\ &= \lim_{n\to\infty}\int_0^T\int_0^L \left(EI\left|\frac{\partial^2}{\partial x^2}e^{i\omega_n t}u_n\right|^2 + \rho|e^{i\omega_n t}v_n|^2\right)dxdt \\ &= \lim_{n\to\infty}\int_0^T\int_0^L \left(EI\left|\frac{\partial^2}{\partial x^2}u_n\right|^2 + \rho|v_n|^2\right)dxdt = T, \end{aligned}$$

which is in contradiction with (4.29), and the proof is completed. □

5. Proof of Theorem 2.2. First it is impossible that $k_2 = k_4 = 0$. Indeed, it follows from $k_2 = k_4 = 0$ and (2.5) that

$$\begin{aligned} E(t) &= \frac{1}{2}\left\| e^{t\mathcal{A}} \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix} \right\|_{\mathcal{H}}^2 \\ &\equiv E(0) \\ &= \frac{1}{2}\left[\int_0^L \left(EI \left| \frac{d^2}{dx^2}\omega_0 \right|^2 + \rho|\omega_1|^2 \right) dx + k_1|\omega_0'(0)|^2 + k_3|\omega_0(0)|^2 \right], \quad t \geq 0, \end{aligned}$$

for any $(\omega_0, \omega_1) \in D(\mathcal{A})$. Thus the energy $E(t)$ of the system (1.1) does not uniformly exponentially decay.

Next, let $k_2 = 0$ and $k_4 > 0$. We can define

$$H_E^0 = \{u \in H_E |\ u(0) = 0\}, \quad \mathcal{H}^0 = H_E^0 \oplus H, \quad \mathcal{A}^0 = \mathcal{A}|_{\mathcal{H}^0}.$$

That is,

$$D(\mathcal{A}^0) = \{(u, v) \in D(\mathcal{A}) \cap \mathcal{H}^0 |\ \mathcal{A}(u, v) \in \mathcal{H}^0\}$$

and

$$\mathcal{A}^0(u, v) = \mathcal{A}(u, v), \quad (u, v) \in D(\mathcal{A}^0).$$

By the same argument as in the proof of Theorem 3.1, we can conclude that $\mathcal{A}^0$ is dissipative and $0 \in \rho(\mathcal{A}^0)$. Hence, $\mathcal{A}^0$ is m-dissipative and generates a C_0-contraction semigroup $e^{t\mathcal{A}^0}$. Since $e^{t\mathcal{A}^0}(\omega_0, \omega_1) \in D(\mathcal{A}^0)$, the same proof of (2.5) yields

$$\frac{d}{dt}\left(\frac{1}{2}\left\| e^{t\mathcal{A}^0} \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix} \right\|_{\mathcal{H}^0}^2 \right) = \frac{d}{dt} E(t) = 0,$$

which implies that C_0-semigroup $e^{t\mathcal{A}^0}$ is an isometric semigroup. Therefore the energy of the system (1.1) does not exponentially decay. Finally, let $k_2 > 0$ and $k_4 = 0$; we only need to define $H_E^0 = \{u \in H_E |\ u'(0) = 0\}$ and the similar proof follows. This concludes the proof of the theorem. □

Acknowledgment. The authors would like to thank the anonymous referees for their valuable comments and suggestions.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.
[3] G. Chen, M. C. Defour, A. M. Krall, and G. Payre, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.
[4] G. Chen, S. A. Fulling, F. J. Narcowich, and S. Sun, *Exponential decay of energy of evolution equations with locally distributed dampings*, SIAM J. Appl. Math., 51 (1991), pp. 266–301.
[5] G. Chen, S. G. Krantz, D. W. Ma, C. E. Wayne, and H. H. West, *The Euler-Bernoulli beam equation with boundary energy dissipation*, in Operator Methods for Optimal Control Problems, Sung J. Lee, ed., Marcel Dekker, New York, 1988, pp. 67–96.
[6] G. Chen and D. L. Russell, *A mathematical model for linear elastic systems with structure damping*, Quart. Appl. Math., 39 (1982), pp. 433–454.

[7] S. Chen, K. Liu, and Z. Liu, *Spectrum and stability for elastic systems with global or local Kelvin–Voigt damping*, SIAM J. Appl. Math., 59 (1998), pp. 651–668.
[8] F. Conrad, *Stabilization of beams by pointwise feedback control*, SIAM J. Control Optim., 28 (1990), pp. 423–437.
[9] F. Conrad and Ö. Morgül, *On the stabilization of a flexible beam with a tip mass*, SIAM J. Control Optim., 36 (1998), pp. 1962–1986.
[10] A. V. Fursikov and O. Y. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes 34, Research Institute of Mathematics, Seoul National University, Seoul, Korea, 1996.
[11] B.-Z. Guo, *Riesz basis property and exponential stability of controlled Euler–Bernoulli beam equations with variable coefficients*, SIAM J. Control Optim., 40 (2002), pp. 1905–1923.
[12] F. L. Huang, *Characteristic conditions for exponential stability of linear dynamical systems in Hilbert space*, Ann. Differential Equations, 1 (1985), pp. 1348–1356.
[13] F. L. Huang, *Some problems for linear elastic systems with damping*, Acta Math. Sci., 6 (1986), pp. 101–107.
[14] F. Huang, *On the mathematical model for linear elastic systems with analytic damping*, SIAM J. Control Optim., 26 (1988), pp. 714–724.
[15] J. U. Kim, *Exponential decay of the energy of a one-dimensional nonhomogeneous medium*, SIAM J. Control Optim., 29 (1991), pp. 368–380.
[16] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, Res. Appl. Math., Masson, Paris, 1994.
[17] V. Komornik, *Rapid boundary stabilization of linear distributed systems*, SIAM J. Control Optim., 35 (1997), pp. 1591–1613.
[18] J. Lagnese, *Uniform stabilization of a nonlinear beam by nonlinear boundary feedback*, J. Differential Equations, 50 (1991), pp. 355–388.
[19] I. Lasiecka, *Exponential decay rates for the solutions of Euler-Bernoulli equations with boundary dissipation occurring in the moment only*, J. Differential Equations, 95 (1992), pp. 169–182.
[20] I. Lasiecka and R. Triggiani, *Sharp trace estimates of solutions to Kirchhoff and Euler-Bernoulli equations*, Appl. Math. Optim., 28 (1993), pp. 277–306.
[21] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.
[22] W. Littman and L. Markus, *Stabilization of a hybrid system of elasticity by feedback boundary damping*, Ann. Mat. Pura Appl., 152 (1988), pp. 281–330.
[23] K. Liu and Z. Liu, *Exponential decay of energy of the Euler–Bernoulli beam with locally distributed Kelvin–Voigt damping*, SIAM J. Control Optim., 36 (1998), pp. 1086–1098.
[24] K. S. Liu and Z. Y. Liu, *Boundary stabilization of a nonhomogenous beam with rotary inertia at the tip*, J. Comput. Appl. Math., 114 (2000), pp. 1–10.
[25] A. Pazy, *Semigroup of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[26] J. Prüss, *On the spectrum of C_0-semigroups*, Trans. Amer. Math. Soc., 283 (1984), pp. 847–857.
[27] B. Rao, *Uniform stabilization of a hybrid system of elasticity*, SIAM J. Control Optim., 33 (1995), pp. 440–454.
[28] R. Rebarber, *Exponential stability of coupled beams with dissipative joints: A frequency domain approach*, SIAM J. Control Optim., 33 (1995), pp. 1–28.
[29] D. L. Russell, *Controllability and stabilizatiblity theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
[30] D. L. Russell, *Mathematical models for the elastic beam and their control-theoretic properties*, in Semigroups Theory and Applications, Pitman Research Notes 152, Pitman, Boston, 1986, pp. 177–217.
[31] P.-F. Yao, *On the observability inequalities for exact controllability of wave equations with variable coefficients*, SIAM J. Control Optim., 37 (1999), pp. 1568–1599.
[32] X. Zhang, *Explicit observability inequalities for the wave equation with lower order terms by means of Carleman inequalities*, SIAM J. Control Optim., 39 (2000), pp. 812–834.
[33] E. Zuazua, *Exponential decay for the semilinear wave equation with locally distributed damping*, Comm. Partial Differential Equations, 15 (1990), pp. 205–235.
[34] E. Zuazua, *Exponential decay for the semilinear wave equation with localized damping in unbounded domains*, J. Math. Pures Appl., 70 (1991), pp. 513–529.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 357–376

THE PRIMAL-DUAL ACTIVE SET METHOD FOR NONLINEAR OPTIMAL CONTROL PROBLEMS WITH BILATERAL CONSTRAINTS*

KAZUFUMI ITO† AND KARL KUNISCH‡

Abstract. The primal-dual active set method has proved to be an efficient numerical tool in the context of diverse applications. So far it has been investigated mainly for linear problems. This paper is devoted to the study of global convergence of the primal-dual active set method for nonlinear problems with bilateral constraints. Utilizing the close relationship between the primal-dual active set method and semismooth Newton methods, local superlinear convergence of the method is investigated as well.

Key words. primal-dual active set method, semismooth Newton method, optimal control, bilateral constraints

AMS subject classifications. 49M15, 49M29, 90C26, 90C46

DOI. 10.1137/S0363012902411015

1. Introduction.

This paper is devoted to the study of algorithms for solving problems of the type

$$\begin{aligned} &\min \tfrac{1}{2}\,|T(u)-z|^2 \;+\; \tfrac{\alpha}{2}\,|u|^2 \\ &\text{over } u \text{ subject to } \;\varphi \le u \le \psi, \end{aligned} \tag{1.1}$$

where T is a nonlinear operator between Hilbert spaces and α, z, φ, and ψ are given. A typical motivation for considering (1.1) is given by optimal control problems, where T is the control-to-output mapping. Depending on the type of dynamics and the way in which the control enters into the equation, T may be affine or nonlinear. The inequalities in (1.1) then describe bilateral constraints on the class of admissible controls. Without the presence of the constraints, (1.1) has received a considerable amount of attention over the last 20 years, and numerical methods for solving (1.1) have achieved a high level of sophistication. The presence of the infinite dimensional constraints, however, complicates the problem significantly, and analytical as well as numerical issues are not yet completely resolved.

It should be mentioned that discretizing (1.1) results in a finite dimensional problem, for which numerical algorithms are readily available. However, this approach cannot avoid the infinite dimensional problem, since grid refinement must be addressed. Moreover, in the process of discretization, important properties such as smoothing and compactness of T and existence of integrable Lagrange multiplier associated with inequality constraints are hidden. Such properties, however, have a significant effect on the behavior of any fine-tuned algorithm. We refer, e.g., to [IK2] for a discussion on the regularizing properties of T depending on whether control or state constraints or problems of obstacle or control-of-obstacle type are described by T. Smoothing

*Received by the editors July 17, 2002; accepted for publication (in revised form) December 1, 2003; published electronically June 25, 2004.
http://www.siam.org/journals/sicon/43-1/41101.html
†Department of Mathematics, North Carolina State University, Raleigh, NC (kit@math.ncsu.edu).
‡Institut für Mathematik, Karl-Franzens-Universität Graz, A-8010 Graz, Austria (karl.kunisch@kfunigraz.ac.at). The work of this author was supported in part by the Fonds zur Förderung der wissenschaftlichen Forschung under SFB 03, Optimierung und Kontrolle.

properties of T which hold for many optimal control problems will play an essential role in our analysis of (1.1).

We turn to a brief description of the primal-dual active set strategy and the contributions of this paper. The primal-dual active set strategy is an iterative algorithm that on the basis of the current primal variable u^k and the current Lagrange multiplier λ^k associated to the constraints in (1.1) predicts new active sets $A^+_{k+1} = A^+_{k+1}(u^k, \lambda^k)$ and $A^-_{k+1} = A^-_{k+1}(u^k, \lambda^k)$ and requires solving the equality constrained problems

$$\left\{ \begin{array}{l} \min \frac{1}{2} \, |T(u) - z|^2 \; + \; \frac{\alpha}{2} \, |u|^2 \\ \text{over } u \text{ with } \; u = \psi \;\; \text{on} \;\; A^+_{k+1}, \quad u = \varphi \;\; \text{on} \;\; A^-_{k+1}. \end{array} \right. \tag{1.2}$$

Here A^+ and A^- refer to the sets that are active at the upper, respectively, lower, bound. In case T is linear, the solution to (1.2) is characterized by a linear equation on the complement of $A^+_{k+1} \cup A^-_{k+1}$. In our work the strategy for choosing A^+_{k+1} and A^-_{k+1} is based on convex analysis techniques [IK1, BIK]. Due to the simple nature of the box constraints in (1.1), this strategy is related to strategies already used in [B, HI]. We shall further comment on this below.

To put the algorithm described above into a more general context, it will be convenient to recall that the first order optimality condition for (1.1) is given by

$$\left\{ \begin{array}{l} \alpha \, u + T'(u)^*(T(u) - z) + \lambda = 0, \\ \lambda = \max(0, \lambda + \alpha \, (u - \psi)) + \min(0, \lambda + \alpha \, (u - \varphi)). \end{array} \right. \tag{1.3}$$

We shall argue in section 6 that this is equivalent to

$$\begin{aligned} \alpha(u - \psi) + \max(0, \alpha \, \psi + T'(u)^*(T(u) - z)) \\ + \min(0, -T'(u)^*(T(u) - z) - \alpha \, \varphi) = 0. \end{aligned} \tag{1.4}$$

Note that (1.4) involves the term $p(u) := -T'(u)^*(T(u) - z)$, which is commonly referred to as the adjoint state in optimal control. Solving (1.1) by means of optimality conditions amounts to solving (1.4), which involves two types of nonlinearities of very different nature: the simple but nondifferentiable max and min operations and the nonlinear mapping T, which is typically smooth (the precise conditions will be formulated below) and possibly highly nonlinear. This difference between the nonlinearities max/min and $u \mapsto p(u)$ motivates us to separately consider their linearization in Newton-type iterative approaches. Linearizing $p(u)$ at the current iterate u^k and leaving the nondifferentiable functions unchanged results in

$$\begin{aligned} \alpha(u - \psi) + \max(0, \alpha \, \psi - p(u^k) - p'(u^k)(u - u^k)) \\ + \min(0, p(u^k) + p'(u^k)(u - u^k) - \alpha \, \varphi) = 0, \end{aligned} \tag{1.5}$$

where

$$p'(u^k)\delta \, u = -T'(u^k)^* T'(u^k)\delta \, u - (T(u^k) - z, T''(u^k)(\delta \, u, \cdot)). \tag{1.6}$$

This can be considered as a projected Newton iteration, as we shall show at the end of this section, and it is also closely related to the sequential quadratic programming (SQP) approach to (1.1), which has frequently been considered in the literature; see, e.g., [TV] and the references given there. In the SQP approach to (1.1) both u and $y = T(u)$ are considered as independent variables, and at each iteration level a quadratic approximation to the cost in (1.1) is minimized subject to the constraints $\varphi \le u \le \psi$.

Alternatively, one can take generalized derivatives of the nondifferentiable functions in (1.4), leaving the nonlinearity $p(u)$ unchanged. The resulting auxiliary problems are smooth and can be solved with an appropriate standard method. This partial (semismooth) Newton method is the focus of this paper. In fact, we shall start our analysis with the primal-dual active set strategy and only later argue its equivalence to the partial (semismooth) Newton method. Finally, of course, both nonlinearities max/min and $p(u)$ can be "linearized" simultaneously, resulting in a semismooth Newton method. The semismooth Newton method for general-purpose nonlinear finite dimensional optimization problems has been well studied; see, for instance, [LPR] and the references given there. Much less is known about such methods in infinite dimensions, specifically in the context of optimal control; see, however, [HIK, U]. We shall address the semismooth Newton approach to (1.4) in section 7.

We have studied the primal-dual active set strategy in several previous papers [IK1, BIK, HIK]. The significant difference in the present contribution compared with earlier work is the treatment of bilateral constraints and of nonlinear rather than only linear operators. As in earlier work the global convergence proofs are based on a properly chosen merit functional. The merit functional utilized in [BIK] is not appropriate for analyzing bilateral constraints. Therefore we were forced to find a new merit-functional that has the additional advantage over the one in [BIK] that the details of the convergence proof are more transparent.

Let us briefly describe the contents of this paper. In section 2 we give a precise problem statement and description of the primal-dual active set algorithm. Section 3 is devoted to a general framework for establishing convergence of the algorithm from arbitrary initial data. The applicability of the framework for linear and nonlinear problems is described in sections 4 and 5, respectively. In section 6 we give sufficient conditions for local superlinear convergence of the algorithm. In the context of local convergence we address in section 7 superlinear convergence of the semismooth Newton method, when both types of nonlinearities max/min and $p(u)$ in (1.4) are linearized. While the focus of this paper is an analytical one, we nevertheless present in section 8 a numerical example illustrating some effects of treating the nonlinearities in (1.4) separately rather than directly applying a generalized chain rule.

Let us comment on related contributions. In [H] the primal-dual active set algorithm is extended to bilateral constraints with convergence proofs based on the merit functional from [BIK]. This necessitates further assumptions and modifications of the algorithm, which can be avoided with the new merit functional. The primal-dual active set algorithm is closely related to the Bertsekas projected Newton method, which was developed for finite dimensional problems in [B] and extended to optimal control problems in [KS]. The main difference between the two algorithms is that Bertsekas' algorithm is a feasible algorithm, while ours is not. We comment further on this in Remark 7.4, where we also indicate how the local convergence proofs of this paper can be utilized for the Bertsekas method. The dual algorithm developed in [HI] for optimal control of ordinary differential equations is closely related to the primal-dual active set method. The concepts for the local convergence proofs based on generalized equations in [HI] and semismooth Newton methods in our work are significantly different. There are no global convergence results in [HI]. In [HH] the dual algorithm was modified by the introduction of backtracking steps, resulting in guaranteed convergence for finite dimensional problems. For a detailed comparison of the two methods, see [BK]. As mentioned, semismooth Newton methods were recently extended to nonlinear problems. One of the peculiarities in the application of such

methods to (1.1) is the lack of generalized differentiability (in a sense to be specified in section 7) of the max/min operations between L^p spaces. In [U] this difficulty is compensated by an additional smoothing step introduced to the semismooth Newton method. In our work the lack of generalized differentiability is incorporated in the form in which the optimality condition is expressed (see (1.3), (1.4)) as well as a regularity condition for the mapping $u \to T(u)$.

We close this introduction by establishing the announced relationship between the SQP method applied to (1.1) and (1.5). Let y represent an independent variable, denote by μ the Lagrange multiplier associated to the constraint $y - T(u) = 0$, and introduce the Lagrangian

$$L(y, u, \mu) = \tfrac{1}{2}\,|y - z|^2 + \tfrac{\alpha}{2}\,|u|^2 + (\mu, T(u) - y).$$

Given (y^k, u^k, μ^k) the new correction is obtained as the solution to

$$\text{(1.7)} \qquad \begin{cases} \min\ L'(y^k, u^k, \mu^k)(\delta y, \delta u) + \frac{1}{2}L''(y^k, u^k, \mu^k)((\delta y, \delta u), (\delta y, \delta u)) \\ \text{subject to} \\ T(u^k) + T'(u^k)\delta u\ =\ y^k + \delta y, \quad \varphi \le u_k + \delta u \le \psi, \end{cases}$$

over $(\delta y, \delta u)$, where primes denote differentiation with respect to (y, u). Denoting by $\delta\mu$ and $\bar\lambda$ the Lagrange multipliers to the two constraints in (1.7), and setting

$$(y^{k+1}, u^{k+1}, \mu^{k+1}) = (y^k, u^k, \mu^k) + (\delta y, \delta u, \delta \mu)$$

the optimality condition for (1.7) is given by

$$y^{k+1} = T(u^k) + T'(u^k)\delta u,$$

$$\mu^{k+1} = y^{k+1} - z,$$

$$\alpha\, u^{k+1} + T'(u^k)^*\mu^{k+1} + (\mu^k, T''(u^k)(\delta u, \cdot)) + \bar\lambda = 0,$$

$$\bar\lambda = \max(0, \bar\lambda + (u^{k+1} - \psi)) + \min(0, \bar\lambda + (u^{k+1} - \varphi)).$$

From the first three equations we get

$$\alpha\, u^{k+1} + T'(u^k)^*(T(u^k) + T'(u^k)\delta u - z) + (y^k - z, T''(u^k)(\delta u, \cdot)) + \bar\lambda = 0.$$

This, together with the last equation in the optimality condition, implies that

$$\text{(1.8)} \qquad \alpha(u^{k+1} - \psi) + \max(0, \alpha\,\psi - \omega^k) + \min(0, \omega^k - \alpha\,\varphi),$$

where

$$\text{(1.9)} \qquad \omega^k = -T'(u^k)^*(T(u^k) + T'(u^k)\delta u - z) - (y^k - z, T''(u^k)(\delta u, \cdot)).$$

Comparing to (1.5)–(1.6), we observe that the only difference between the SQP method and the Newton method applied to the smooth operator T in (1.4) occurs in the second summand of the second derivative. In the SQP method y^k is computed from the linearized equation $y^k = T(u^{k-1}) + T'(u^{k-1})\delta u)$, whereas $T(u^k)$ in (1.6) requires the computation of the nonlinear operator T at u^k. Concerning the term $(T(u^k) - z, T''(u^k)(\delta u, \cdot))$ in (1.6) it is worthwhile to observe that

$$\begin{aligned} &(T(u^k) - z, T''(u^k)(\delta u, v)) \\ &\quad = -(T'(u^k)^*(T(u^k) - z), T^{-1}(T(u^k))''(T'(u^k)\delta u, T'(u^k)v)) \\ &\quad = (p(u^k), T^{-1}(T(u^k))''(T'(u^k)\delta u, T'(u^k)v)). \end{aligned}$$

In applications to optimal control of differential equations $T(u)$, respectively, $T'(u)v$, require the solution of a nonlinear, respectively linearized, differential equation.

2. Problem statement and algorithm. Let Y and U be Hilbert spaces with $U = L^2(\Sigma)$, where Σ a bounded measurable set in $\mathbb{R}^n$, and let $T\colon U \to Y$ be a, possibly nonlinear, continuously differentiable, injective, mapping with Fréchet derivative denoted by T'. Further let $\varphi, \psi \in U$ with $\varphi < \psi$ a.e. in Σ. For $\alpha > 0$ and $z \in Y$ consider

$$\min_{\varphi \le u \le \psi} J(u) = \tfrac{1}{2}\,|T(u) - z|_Y^2 + \tfrac{\alpha}{2}\,|u|_U^2. \tag{2.1}$$

We refer to (2.1) as the bilaterally constrained problem. When the constraint $\varphi \le u$ is not present, formally obtained by setting $\varphi = -\infty$, we have the unilaterally constrained problem. The necessary optimality condition for (2.1) is given by

$$(\alpha u + T'(u)^*(T(u) - z), \tilde{u} - u)_U \ge 0 \quad \text{for all } \varphi \le \tilde{u} \le \psi.$$

A direct computation shows that this is equivalent to the existence of $\lambda \in U$ such that

$$\begin{cases} \alpha u + T'(u)^*(T(u) - z) + \lambda = 0, \\ \lambda = 0 \text{ on } I, \quad \lambda \ge 0 \text{ on } A^+, \quad \lambda \le 0 \text{ on } A^-, \end{cases}$$

where $I = \{x : \varphi(x) < u(x) < \psi(x)\}$, $A^+ = \{x : u(x) = \psi(x)\}$, and $A^- = \{x : u(x) = \varphi(x)\}$, in the a.e. sense. The second condition can be equivalently expressed as

$$\lambda = \max(0, \lambda + \alpha(u - \psi)) + \min(0, \lambda + \alpha(u - \varphi)),$$

where $c > 0$ is arbitrarily fixed. For our work the choice $c = \alpha$ is convenient and results in the first order optimality system,

$$\begin{cases} \alpha u + T'(u)^*(T(u) - z) + \lambda = 0, \\ \lambda = \max(0, \lambda + \alpha(u - \psi)) + \min(0, \lambda + \alpha(u - \varphi)), \end{cases} \tag{2.2}$$

where $(u, \lambda) \in U \times U$, and max as well as min are interpreted as pointwise a.e. operations.

We next specify the primal-dual active set algorithm. The iteration index is denoted by k and an initial choice (u^0, λ^0) is assumed to be available.

Primal-dual active set algorithm.

(i) Given (u^k, λ^k), determine

$$A_{k+1}^+ = \{x\colon (\lambda^k + \alpha(u^k - \psi))(x) > 0\},$$

$$\begin{aligned} I_{k+1} &= \{x\colon (\lambda^k + \alpha(u^k - \psi))(x) \le 0 \le \lambda^k + \alpha(u^k - \varphi)(x)\}, \\ A_{k+1}^- &= \{x\colon (\lambda^k + \alpha(u^k - \varphi))(x) < 0\}. \end{aligned}$$

(ii) Determine (u^{k+1}, λ^{k+1}) from

$$u^{k+1} = \psi \ \text{ on } \ A_{k+1}^+, \quad u^{k+1} = \varphi \ \text{ on } \ A_{k+1}^-, \quad \lambda^{k+1} = 0 \ \text{ on } \ I_{k+1},$$

and

$$\alpha\, u^{k+1} + T'(u^{k+1})^*(T(u^{k+1}) - z) + \lambda^{k+1} = 0. \tag{2.3}$$

Note that the equations for (u^{k+1}, λ^{k+1}) in step (ii) of the algorithm constitute the necessary optimality condition for the auxiliary problem

$$\left\{ \begin{array}{l} \min \frac{1}{2} \, |T(u) - z|_Y^2 + \frac{\alpha}{2} \, |u|_U^2 \ \text{ over } \ u \in U \\ \text{subject to } \quad u = \psi \ \text{ on } \ A_{k+1}^+, \quad u = \varphi \ \text{ on } \ A_{k+1}^-. \end{array} \right. \tag{2.4}$$

For the global convergence analysis which will be given in the following sections we require the primal-adjoint structure of the optimality system that arises if the action of T is given as the solution of an equation. This situation typically arises in optimal control and parameter estimation problems. Thus we consider the case where $y = T(u)$ is given as the solution of an equation of the form

$$H(y) = u.$$

Assume that for every $u \in U$ this equation has a unique solution $y = T(u) \in X$, where X is dense subset of Y, and that $u \to T(u)$ is continuous from u to X. Assume further that $H : X \to U$ is C^1 and that $H'(y)$ has a bounded inverse $H'(y)^{-1} \in \mathcal{L}(U, Y)$. Then,

$$T'(u) = (H'(y))^{-1} \text{ and } T'(u)^* = (H'(y)^*)^{-1} \in \mathcal{L}(Y, U) \qquad \text{for } u \in U,$$

where $y = T(u)$ and the adjoint of $H'(y)$ is taken as an operator with domain in Y to U. Consequently (2.2) can equivalently be expressed as

$$\left\{ \begin{array}{l} H(y) = u, \\ \\ H'(y)^* p = -(y - z), \\ \\ \alpha \, u - p + \lambda = 0, \\ \\ \lambda = \max(0, \lambda + (u - \psi)) + \min(0, \lambda + (u - \varphi)), \end{array} \right. \tag{2.5}$$

where $p = -T'(u)^*(T(u) - z)$ is the adjoint state. Analogously, for $u^{k+1} \in U$, setting $y^{k+1} = T(u^{k+1})$, $p^{k+1} = -T'(u^{k+1})^*(T(u^{k+1}) - z)$, (2.3) can equivalently be expressed as

$$\left\{ \begin{array}{l} H(y^{k+1}) = \left\{ \begin{array}{lll} \psi & \text{on} & A_{k+1}^+, \\ \frac{1}{\alpha} \, p^{k+1} & \text{on} & I_{k+1}, \\ \varphi & \text{on} & A_{k+1}^-, \end{array} \right. \\ \\ H'(y^{k+1})^* p^{k+1} = -(y^{k+1} - z), \\ \\ \alpha \, u^{k+1} - p^{k+1} + \lambda^{k+1} = 0. \end{array} \right. \tag{2.6}$$

Example 2.1. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain in $\mathbb{R}^n$ with Lipschitz continuous boundary Γ and $\widetilde{\Omega} \subset \Omega$, $\widetilde{\Gamma} \subset \Gamma$ measurable subsets. Let $E_{\widetilde{\Gamma}} \colon L^2(\widetilde{\Gamma}) \to L^2(\Gamma)$ be the extension-by-zero operator and $R_{\widetilde{\Omega}} \colon L^2(\Omega) \to L^2(\widetilde{\Omega})$ the canonical restriction operator. Define $L \colon L^2(\Gamma) \to L^2(\Omega)$ as the solution operator to the inhomogeneous Neumann boundary value $L \, u = y$, i.e., y is the solution to

$$(\nabla y, \nabla v)_\Omega + (y, v)_\Omega = (u, v)_\Gamma \quad \text{for all } \ v \in H^1(\Omega), \tag{2.7}$$

and then $T = R_{\widetilde{\Omega}} L\, E_{\widetilde{\Gamma}} \colon L^2(\widetilde{\Gamma}) \to L^2(\widetilde{\Omega})$. Then $T^* \colon L^2(\widetilde{\Omega}) \to L^2(\widetilde{\Gamma})$ is given by

$$T^* = R_{\widetilde{\Gamma}} L^*\, E_{\widetilde{\Omega}}$$

with $R_{\widetilde{\Gamma}}$ the restriction operator to $\widetilde{\Gamma}$ and $E_{\widetilde{\Omega}}$ the extension operator from $\widetilde{\Omega}$ to Ω by zero. Further, the adjoint $L^* \colon L^2(\Omega) \to L^2(\Gamma)$ of L is given by $L^* z = \tau_\Gamma p$, where τ_Γ is the Dirichlet trace operator from $H^1(\Omega)$ to $L^2(\Gamma)$ and p is the solution to

$$(\nabla p, \nabla w)_\Omega + (p, w)_\Omega = (z, w)_\Omega \quad \text{for all} \quad w \in H^1(\Omega). \tag{2.8}$$

3. Global convergence. In this section we give conditions that guarantee convergence of the primal-dual active set strategy for linear and certain nonlinear operators T from arbitrary initial data. The convergence proof is based on an appropriately defined functional that decays when evaluated along the iterates of the algorithm. An a priori estimate for the adjoint variable p in (2.6) will play an essential role. In this section we shall assume this estimate to hold. In the following sections the estimate will be investigated separately for the linear and nonlinear cases.

To specify the condition alluded to above let us consider two consecutive iterates of the algorithm. For every $k = 1, 2, \ldots$, the sets A_k^+, A_k^-, and I_k give a mutually disjoint decomposition of Σ. According to (i) and (ii) in the form (2.6) we find

$$H(y^{k+1}) - H(y^k) = u^{k+1} - u^k = \begin{cases} R^k_{A^+} & \text{on} \quad A^+_{k+1}, \\ \frac{1}{\alpha}(p^{k+1} - p^k) + R^k_I & \text{on} \quad I_{k+1}, \\ R^k_{A^-} & \text{on} \quad A^-_{k+1} \end{cases} \tag{3.1}$$

and

$$H'(y^{k+1})^* p^{k+1} - H'(y^k)^* p^k + y^{k+1} - y^k = 0, \tag{3.2}$$

where the residual R^k is given by

$$R^k_{A^+} = \begin{cases} 0 & \text{on} \quad A^+_k \cap A^+_{k+1}, \\ \psi - \frac{1}{\alpha} p^k = \psi - u^k < 0 & \text{on} \quad I_k \cap A^+_{k+1}, \\ \psi - \varphi < \frac{1}{\alpha} \lambda_k & \text{on} \quad A^-_k \cap A^+_{k+1}, \end{cases} \tag{3.3}$$

$$R^k_I = \begin{cases} \frac{1}{\alpha} \lambda^k = \frac{1}{\alpha} p^k - \psi \le 0 & \text{on} \quad A^+_k \cap I_{k+1}, \\ 0 & \text{on} \quad I_k \cap I_{k+1}, \\ \frac{1}{\alpha} \lambda^k = \frac{1}{\alpha} p^k - \varphi \ge 0 & \text{on} \quad A^-_k \cap I_{k+1}, \end{cases} \tag{3.4}$$

$$R^k_{A^-} = \begin{cases} \varphi - \psi > \frac{1}{\alpha} \lambda_k & \text{on} \quad A^+_k \cap A^-_{k+1}, \\ \varphi - \frac{1}{\alpha} p^k = \varphi - u^k > 0 & \text{on} \quad I_k \cap A^-_{k+1}, \\ 0 & \text{on} \quad A^-_k \cap A^-_{k+1}. \end{cases} \tag{3.5}$$

Let R^k denote the function defined on Ω whose restrictions to A^+_{k+1}, I_{k+1}, A^-_{k+1} coincide with $R^k_{A^+}$, R^k_I, and $R^k_{A^-}$.

We shall utilize the following a priori estimate:

$$\begin{cases} \text{There exists} \ \ \rho < \alpha \ \ \text{such that} \\ \\ |p^{k+1} - p^k|_U < \rho\, |R^k|_U \ \ \text{for every} \ \ k = 1, 2, \ldots. \end{cases} \tag{3.6}$$

The convergence proof will be based on the following merit functional $M: U \times U \to R$ given by

$$M(u,\lambda) = \alpha^2 \int_\Sigma (|(u-\psi)^+|_U^2 + |(\varphi - u)^+|_U^2)\, dx + \int_{\mathcal{A}^+(u)} |\lambda^-|_U^2\, dx + \int_{\mathcal{A}^-(u)} |\lambda^+|_U^2\, dx,$$

where $\mathcal{A}^+(u) = \{x\colon u \geq \psi\}$ and $\mathcal{A}^-(u) = \{x\colon u \leq \varphi\}$. For pairs $(u,\lambda) \in U \times U$ satisfying

$$\lambda(u-\psi)(\varphi - u)(x) = 0 \quad \text{for a.a. } x \in \Sigma, \tag{3.7}$$

at most one of the integrands of M can be strictly positive at $x \in \Sigma$.

THEOREM 3.1. *Assume that* (3.6) *holds for the iterates of the primal-dual active set strategy. Then* $M(u^{k+1},\lambda^{k+1}) \leq \alpha^{-2}\rho^2 M(u^k,\lambda^k)$ *for every* $k = 1,\dots$. *Moreover there exist* $(u^*,y^*,p^*,\lambda^*) \in U \times Y \times U \times U$ *with* $(y^*,p^*) \in$ *range* $(T(u^*))\times$ *range* $(T'(u^*)^*)$ *such that* $\lim_{k\to\infty}(u^k,p^k,\lambda^k) = (u^*,p^*,\lambda^*)$ *and* (u^*,y^*,p^*,λ^*) *satisfies* (2.5).

Proof. From (2.6) we have

$$\begin{aligned} \lambda^{k+1} &= p^{k+1} - \alpha\psi \quad \text{on} \quad A_{k+1}^+, \\ u^{k+1} &= \tfrac{1}{\alpha}\, p^{k+1} \quad \text{on} \quad I_{k+1}, \\ \lambda^{k+1} &= p^{k+1} - \alpha\varphi \quad \text{on} \quad A_{k+1}^-. \end{aligned}$$

Using step (ii) of the algorithm in the form of (2.6) implies that

$$\lambda^{k+1} = p^{k+1} - p^k + p^k - \alpha\psi = p^{k+1} - p^k + \begin{cases} \lambda^k > 0 & \text{on} \quad A_k^+ \cap A_{k+1}^+, \\ \alpha(u^k - \psi) > 0 & \text{on} \quad I_k \cap A_{k+1}^+, \\ \alpha u^k + \lambda^k - \alpha\psi \geq 0 & \text{on} \quad A_k^- \cap A_{k+1}^+, \end{cases}$$

and therefore

$$|\lambda^{k+1,-}(x)| \leq |p^{k+1}(x) - p^k(x)| \quad \text{for } x \in A_{k+1}^+. \tag{3.8}$$

Analogously one derives

$$|\lambda^{k+1,+}(x)| \leq |p^{k+1}(x) - p^k(x)| \quad \text{for } x \in A_{k+1}^-. \tag{3.9}$$

Moreover

$$\begin{aligned} u^{k+1} - \psi &= \frac{1}{\alpha}(p^{k+1} - p^k + p^k) - \psi \\ &= \frac{1}{\alpha}(p^{k+1} - p^k) + \begin{cases} \frac{1}{\alpha}\lambda^k \leq 0 & \text{on} \quad A_k^+ \cap I_{k+1}, \\ u^k - \psi \leq 0 & \text{on} \quad I_k \cap I_{k+1}, \\ \frac{1}{\alpha}\lambda^k + u - \psi \leq 0 & \text{on} \quad A_k^- \cap I_{k+1}, \end{cases} \end{aligned}$$

which implies that

$$|(u^{k+1} - \psi)^+(x)| \;\leq\; \tfrac{1}{\alpha}\, |p^{k+1}(x) - p^k(x)| \quad \text{for } x \in I_{k+1}. \tag{3.10}$$

Analogously one derives that

$$|(\varphi - u^{k+1})^+(x)| \;\leq\; \tfrac{1}{\alpha}\, |p^{k+1}(x) - p^k(x)| \quad \text{for } x \in I_{k+1}. \tag{3.11}$$

Due to (ii) of the algorithm we have that

$$(u^{k+1}-\psi)^+ = (\varphi - u^{k+1})^+ = 0 \quad \text{on} \quad A_{k+1}^+ \cup A_{k+1}^-,$$

which together with (3.10)–(3.11) implies that

$$|(u^{k+1}-\psi)^+(x)| + |(\varphi - u^{k+1})^+(x)| \le \tfrac{1}{\alpha}|p^{k+1}(x) - p^k(x)| \quad \text{for } x \in \Sigma. \tag{3.12}$$

From (3.8), (3.9) and since $\varphi < \psi$ a.e. on Σ we find

$$|\lambda^{k+1,-}(x)| \;\le\; |p^{k+1}(x) - p^k(x)| \quad \text{for} \quad x \in \mathcal{A}^+(u^{k+1}) \tag{3.13}$$

and

$$|\lambda^{k+1,+}(x)| \;\le\; |p^{k+1}(x) - p^k(x)| \quad \text{for} \quad x \in \mathcal{A}^-(u^{k+1}). \tag{3.14}$$

Combining (3.11)–(3.14) implies that

$$M(u^{k+1}, \lambda^{k+1}) \;\le\; \int_\Sigma |p^{k+1}(x) - p^k(x)|^2 dx. \tag{3.15}$$

Since (3.6) is supposed to hold we have

$$M(u^{k+1}, \lambda^{k+1}) \;\le\; \rho^2 |R^k|_U^2.$$

Moreover, from (3.3)–(3.5) we deduce that

$$|R^k|_U^2 \;\le\; \alpha^{-2} M(u^k, \lambda^k), \tag{3.16}$$

and consequently

$$M(u^{k+1}, \lambda^{k+1}) \;\le\; \alpha^{-2}\rho^2 M(u^k, \lambda^k) \quad \text{for} \quad k = 1, 2, \ldots. \tag{3.17}$$

From (3.6), (3.16), and (3.17) it follows that $|p^{k+1} - p^k|_U \le (\frac{\rho}{\alpha})^k \rho |R^0|_U$. Thus there exists $p^* \in U$ such that $\lim_{k\to\infty} p^k = p^*$.

Note that for $k \ge 1$

$$A_{k+1}^+ = \{x\colon p^k(x) > \alpha\,\psi(x)\}, \quad I_{k+1} = \{x\colon \alpha\,\varphi(x) \le p^k(x) \le \alpha\,\psi(x)\},$$

$$A_{k+1}^- = \{x\colon p^k(x) < \alpha\,\varphi(x)\},$$

and hence

$$\lambda^{k+1} = \max(0, p^k - \alpha\,\psi) + \min(0, p^k - \alpha\,\varphi) + (p^{k+1} - p^k)_{\chi_{A_{k+1}^+ \cup A_{k+1}^-}}.$$

Since $\lim_{k\to\infty}(p^{k+1} - p^k) = 0$ and $\lim_{k\to\infty} p^k$ exists, it follows that there exists $\lambda^* \in U$ such that $\lim_{k\to\infty} \lambda^k = \lambda^*$, and

$$\lambda^* = \max(0, p^* - \alpha\,\psi) + \min(0, p^* - \alpha\,\varphi). \tag{3.18}$$

From the last equation in (2.6) it follows that there exists u^* such that $\lim_{k\to\infty} u^k = u^*$ and $\alpha\,u^* - p^* + \lambda^* = 0$. Combined with (3.18) the triple (u^*, p^*, λ^*) satisfies the complementarity condition given by the second equation in (2.2). Passing to the limit with respect to k in (2.3) we obtain that the first equation in (2.1) is satisfied by (u^*, λ^*). Setting $y^* = T(u^*)$ we find that $(u^*, y^*, p^*, \lambda^*)$ satisfies (2.5). □

4. The linear case. In this section we give sufficient conditions for (3.6) to hold if $T \in \mathcal{L}(U, Y)$ is an injective operator with dense range in Y. If we define $H = T^{-1}$, then H is closed and (3.1)–(3.2) can be expressed as

$$\left\{ \begin{array}{l} H(y^{k+1} - y^k) = R^k + \frac{1}{\alpha}(p^{k+1} - p^k)\chi_{I_{k+1}}, \\ H^*(p^{k+1} - p^k) + y^{k+1} - y^k = 0. \end{array} \right. \tag{4.1}$$

THEOREM 4.1. *If T is linear and $\|T\|^2_{\mathcal{L}(U,Y)} < \alpha$, then* (3.6) *holds.*

Proof. Since $y^{k+1} - y^k \in \text{range } (H)$ we can apply H to the second equation in (4.1) and obtain

$$H\,H^*(p^{k+1} - p^k) + \frac{1}{\alpha}(p^{k+1} - p^k)\chi_{I_{k+1}} = -R^k. \tag{4.2}$$

Taking the inner product with respect to $p^{k+1} - p^k$ implies

$$|H^*(p^{k+1} - p^k)|^2_Y \;\le\; |R^k|\,|p^{k+1} - p^k|_U.$$

Since $|p^{k+1} - p^k|_U \le |T^*H^*(p^{k+1} - p^k)|_U \le \|T\|_{\mathcal{L}(U,Y)}|H^*(p^{k+1} - p^k)|_Y$, the claim follows. □

Remark 4.1. In [BIK] the condition $\|T\|^2_{\mathcal{L}(U,Y)} \le \frac{\alpha}{2}$ was obtained for global convergence by a different technique for the unilateral case. Note that in applications to optimal control problems, T represents the solution operator to a differential equation. For elliptic or parabolic equations T is smoothing and it is reasonable to assume that $\|T\|$ is relatively small. (For $T = (-\Delta)^{-1}$ with Dirichlet boundary conditions on the unit square in R^2, we have $\|T\|^2 = \frac{1}{4\pi^4}$.)

The smallness condition of Theorem 4.1 is not required if appropriate structural properties can be utilized. For example, in finite dimensions, if T is an M matrix, then global convergence was obtained in [HIK]. Sufficient conditions for global convergence if T is a P-matrix are also given in [HIK]. Global convergence of infinite dimensional obstacle problems was analyzed in [IK2] exploiting the maximum principle.

5. A class of nonlinear problems. This section is devoted to an analysis of specific nonlinear problems for which (3.6) can be satisfied. Let $\Sigma = \Omega \subset \mathbb{R}^n$, $n = 2$ or 3, with smooth boundary $\partial\,\Omega$. Further let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a monotone mapping with locally Lipschitzian derivative, satisfying $\phi(0) = 0$, and such that the substitution operator determined by ϕ maps $H^1(\Omega)$ into $L^2(\Omega)$. In the notation of section 2 we choose $U = Y = L^2(\Omega)$ and define T as the solution operator to

$$\left\{ \begin{array}{lll} H(y) = -\Delta y + \phi(y) = u & \text{in} & \Omega, \\ y = 0 & \text{on} & \partial\Omega, \end{array} \right. \tag{5.1}$$

where Δ denotes the Laplacian. The adjoint equation is given by

$$\left\{ \begin{array}{lll} H'(y)^*p = -\Delta p + \phi'(y)p = -(y - z) & \text{in} & \Omega, \\ p = 0 & \text{on} & \partial\Omega. \end{array} \right. \tag{5.2}$$

Let (u^0, λ^0) be an arbitrary initialization and let $\widetilde{U} = \{u^k\colon k = 1, \dots\}$ denote the set of iterates generated by the algorithm. Since these iterates are solutions to the auxiliary problem (2.4) it follows that for every $\bar{\alpha} > 0$ the set $\widetilde{U}$ is bounded in $L^2(\Omega)$ uniformly with respect to $\alpha \ge \bar{\alpha}$.

By monotone operator theory and regularity theory of elliptic partial differential equations it follows that the set of primal states $\{y^k = y(u^k)\colon k = 1, \dots\}$ and adjoint

states $\{p^k = p(y(u^k))\colon k = 1, \dots\}$ are bounded subsets of $L^\infty(\Omega)$. Let C denote this bound and let L_C denote the Lipschitz constant of ϕ on the ball $B_C(0)$ with center 0 and radius C in $\mathbb{R}$. Let $H_0^1(\Omega) = \{u \in H^1(\Omega) : u = 0 \text{ on } \partial\Omega\}$ be the Hilbert space endowed with norm $|\nabla u|_{L^2}$ and let κ stand for the embedding constant from $H_0^1(\Omega)$ into $L^2(\Omega)$.

PROPOSITION 5.1. *Assume that* $0 < \frac{(1+C\,L_C)\,\kappa^4}{\alpha-(1+C\,L_C)\,\kappa^4} < 1$. *Then* (3.6) *holds for the mapping* T *determined by the solution operator to* (5.1).

Proof. For the case under consideration, (3.1) and (3.2) can be expressed as

$$-\Delta(y^{k+1} - y^k) + \phi(y^{k+1}) - \phi(y^k) = \frac{1}{\alpha}\,(p^{k+1} - p^k)\chi_{I^{k+1}} + R^k, \tag{5.3}$$

$$-\Delta(p^{k+1} - p^k) + \phi'(y^{k+1})(p^{k+1} - p^k) + (\phi'(y^{k+1}) - \phi'(y^k))p^k + y^{k+1} - y^k = 0. \tag{5.4}$$

Taking the inner product of (5.3) with $y^{k+1} - y^k$ we have, using monotonicity of ϕ,

$$|y^{k+1} - y^k|_1 \le \frac{\kappa^2}{\alpha}\,|(p^{k+1} - p^k)|_1 + |R^k|_{-1}, \tag{5.5}$$

where $|\cdot|_1$ and $|\cdot|_{-1}$ denote the norms in $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, respectively. Note that $\phi'(y^{k+1}) \ge 0$. Hence from (5.4) we find

$$\begin{aligned} |p^{k+1} - p^k|_1^2 &\le C\,L_C|y^{k+1} - y^k|_{L^2}|p^{k+1} - p^k|_{L^2} + |(y^{k+1} - y^k, p^{k+1} - p^k)| \\ &\le (1 + C\,L_C)\,\kappa^2|y^{k+1} - y^k|_1|p^{k+1} - p^k|_1. \end{aligned}$$

Thus,

$$|p^{k+1} - p^k|_1 \;\le\; (1 + C\,L_C)\,\kappa^2\,|y^{k+1} - y^k|_1$$

and hence from (5.5)

$$|y^{k+1} - y^k|_1 \;\le\; \frac{\alpha}{\alpha - (1 + C\,L_C)\,\kappa^4}|R^k|_{-1}.$$

It thus follows that

$$|p^{k+1} - p^k|_{L^2} \;\le\; \frac{\alpha\,(1 + C\,L_C)\,\kappa^4}{\alpha - (1 + C\,L_C)\,\kappa^4}\;|R^k|_{L^2}.$$

This implies (3.6) with

$$\rho \;= \frac{(1 + C\,L_C)\,\kappa^4}{\alpha - (1 + C\,L_C)\,\kappa^4}. \qquad \square$$

6. Local superlinear convergence. In this section we derive sufficient conditions for superlinear convergence. Our analysis is based on expressing the primal-dual algorithm as a partial semismooth Newton algorithm for solving the optimality system (2.2). We refer to the procedure as partial semismooth Newton approach, since only the nonlinearity due to the max operation is linearized whereas the mapping $u \to T(u)$ is not. Observe that the second equation in (2.2) can equivalently be expressed as

$$\lambda = \max(0, \lambda + c(u - \psi)) + \min(0, \lambda + c(u - \varphi)) \quad \text{for every } c > 0. \tag{6.1}$$

Choosing $\alpha = c$, the necessary optimality condition can equivalently be formulated as

$$\mathcal{F}(u) = \alpha(u-\psi) + \max(0, \alpha\,\psi - p(u)) + \min(0, p(u) - \alpha\,\varphi) = 0, \tag{6.2}$$

where

$$p(u) = -T'(u)^*(T(u) - z).$$

Aside from the initialization, the iterative step of the algorithm can be equivalently expressed as

$$\begin{aligned}\alpha(u^{k+1}-\psi) - G^k_{\max}(p(u^{k+1}) - p(u^k)) + G^k_{\min}(p(u^{k+1}) - p(u^k))\\ + \max(0, \alpha\,\psi - p(u^k)) + \min(0, p(u^k) - \alpha\,\varphi) = 0,\end{aligned} \tag{6.3}$$

where

$$G^k_{\max}\phi = \begin{cases} 0 & \text{on} \quad A^+_{k+1} = \{x\colon p^k - \alpha\,\psi > 0\},\\ \phi & \text{on} \quad I^\psi_{k+1} = \{x\colon p^k - \alpha\,\psi \le 0\},\end{cases}$$

$$G^k_{\min}\phi = \begin{cases} \phi & \text{on} \quad A^-_{k+1} = \{x\colon p^k - \alpha\,\varphi < 0\},\\ 0 & \text{on} \quad I^\varphi_{k+1} = \{x\colon p^k - \alpha\,\varphi \ge 0\},\end{cases}$$

and $p^k = p(u^k)$.

Henceforth let u^* denote a solution to (2.1), set $y^* = T(u^*)$, and let λ^* be such that (u^*, λ^*) satisfy (2.1). Further, let $\{u^k\}_{k=1}^\infty$ denote the sequence of iterates which, together with u^* are assumed to be contained in $\mathcal{N}$, introduced in section 2. Note that (u^*, λ^*) satisfy

$$\begin{aligned}&\alpha\,u^* - p(u^*) + \lambda^* = 0,\\ &\alpha(u^* - \psi) + \max(0, \alpha\,\psi - p(u^*)) + \min(0, p(u^*) - \alpha\,\varphi) = 0.\end{aligned} \tag{6.4}$$

Combined with (6.3) we obtain

$$\begin{aligned}&\alpha(u^{k+1} - u^*) - G^k_{\max}(p(u^{k+1}) - p(u^*)) + G^k_{\min}(p(u^{k+1}) - p(u^*))\\ &= \max(0, \alpha\,\psi - p(u^*)) - \max(0, \alpha\,\psi - p(u^k)) + G^k_{\max}(p(u^*) - p(u^k))\\ &+ \min(0, p(u^*) - \alpha\,\varphi) - \min(0, p(u^k) - \alpha\,\varphi) - G^k_{\min}(p(u^*) - p(u^k)) =: R(u^k).\end{aligned} \tag{6.5}$$

Thus, given u^k, the new iterate u^{k+1} satisfies

$$\alpha(u - u^*) - G^I(p(u) - p(u^*)) = R, \tag{6.6}$$

where $R = R(u^k)$, and G^I is the characteristic function of the set $\{x\colon \alpha\,\varphi(x) \le p(u^k)(x) \le \alpha\,\psi(x)\}$. Equivalently, (6.6) can be expressed as

$$\begin{cases} H(y) - H(y^*) = \frac{1}{\alpha}\,G^I(p - p(u^*)) + \frac{1}{\alpha}\,R,\\ H'(y)^*p - H'(y^*)^*p(u^*) + y - y^* = 0,\end{cases} \tag{6.7}$$

with $R = R(u^k)$, $I = I(p(u^k))$, where y and u are related by $y = T(u)$.

In the statement of the following conditions, which will be used to establish local superlinear convergence, $N(u^*)$ denotes a neighborhood of u^* in $L^2(\Sigma)$.

$$\begin{cases} \text{There exists } q > 2, \text{ a neighborhood } N(u^*) \text{ and } L > 0 \text{ such that}\\ |p(u) - p(u^*)|_{L^q(\Sigma)} \le L|u - u^*|_{L^2(\Sigma)} \quad \text{for all } u \in N(u^*),\end{cases} \tag{6.8}$$

where $p(u) = -T'(u)^*(T(u) - z)$.

$$\text{(6.9)} \qquad \begin{cases} \text{There exists a constant } C \text{ independent of } I \text{ and } R \text{ such that} \\ \text{for every solution } u \text{ to (6.6) with } u \in N(u^*) \\ |u - u^*|_{L^2(\Sigma)} \le C|R|_{L^2(\Sigma)}. \end{cases}$$

Condition (6.9) will be applied with $u = u^k$ and $R = R(u^k)$, so that existence of u satisfying (6.6), respectively (y, p) satisfying (6.7), is a priori established. Throughout the remainder of the paper we require that $\varphi, \psi \in L^q(\Sigma)$.

LEMMA 6.1. *Assume that* (6.8) *is satisfied. Then*

$$|\max(0, \alpha\,\psi - p(u^* + h)) - \max(0, \alpha\,\psi - p(u^*)) + G_{\max}^{\psi(p(u^*+h))}(p(u^* + h) - p(u^*))|_{L^2(\Sigma)} = o(|h|_{L^2(\Sigma)}).$$

Proof. The mapping $u \to \max(0, u)$ is Newton differentiable for every $q > 2$, i.e.,

$$\text{(6.10)} \qquad |\max(0, u + v) - \max(0, u) - G(u + v)v|_{L^2(\Sigma)} = o(|v|_{L^q(\Sigma)}),$$

where

$$G(u + v)(x) = \begin{cases} 1 & \text{if } (u + v)(x) \ge 0, \\ 0 & \text{if } (u + v)(x) < 0; \end{cases}$$

see [HIK, U]. Applying (6.10) with $u = \alpha\,\psi - p(u^*)$ and $u + v = \alpha\,\psi - p(u^* + h)$ implies

$$|\max(0, \alpha\,\psi - p(u^* + h)) - \max(0, \alpha\,\psi - p(u^*)) + G_{\max}^{\psi(p(u^*+h))}(p(u^* + h) - p(u^*))|_{L^2(\Sigma)} = o(|p(u^* + h) - p(u^*)|_{L^q(\Sigma)}).$$

The claim follows with (6.8). □

THEOREM 6.2. *Assume that the iterates u^k converge to a solution u^* of* (2.1) *and that* (6.8) *is satisfied. If in addition* (6.9) *holds, then u^k converges superlinearly.*

Proof. In case (6.9) is satisfied, we have for all sufficiently large k

$$|u^{k+1} - u^*|_U \;\le\; C|R(u^k)|_U.$$

Using Lemma 6.1 with $h = u^k - u^*$, and an analogous estimate for the min operation, it follows from 6.5 that

$$|u^{k+1} - u^*|_U \;\le\; o(|u^k - u^*|_U). \qquad \square$$

PROPOSITION 6.3. *If T is linear, then* (6.9) *is satisfied.*

Proof. Let I be an arbitrary subset of Σ and let $R \in L^2(\Sigma)$. Denote by A the complement of I in Σ. For $\phi \in L^2(\Sigma)$ let $\phi_I = \phi\,\chi_I$ and $\phi_A = \phi\,\chi_A$. From (6.6) we have

$$\text{(6.11)} \qquad |(u - u^*)_A|_U \;\le\; \tfrac{1}{\alpha}|R_A|_U,$$

where u is any solution to (6.6). Moreover,

$$\alpha(u - u^*)_I - (p(u) - p(u^*))_I \;=\; R_I$$

and hence

$$\alpha(u - u^*)_I + (T^*T(u - u^*))\chi_I \;=\; R_I.$$

Taking the inner product with $(u-u^*)_I$ we find

$$\alpha|(u-u^*)_I|_U^2 + |T(u-u^*)_I|_Y^2 \;\le\; \tfrac{1}{2}|T(u-u^*)_I|_Y^2 \;+\; \tfrac{1}{2}|T(u-u^*)_A|_Y + \tfrac{\alpha}{2}|(u-u^*)_I|_U^2 \;+\; \tfrac{1}{\alpha}|R_I|_U^2,$$

and therefore

$$|(u-u^*)_I|_U^2 + |T(u-u^*)_I|_Y^2 \;\le\; \tfrac{1}{\alpha^2}\|T\|^2|R_A|_U^2 + \tfrac{1}{\alpha}|R_I|_U^2.$$

Combined with (6.11) this gives the desired result. □

In the following propositions we turn to providing conditions that imply (6.9) in case T is nonlinear. The abbreviations $y^* = y(u^*)$ and $p^* = p(u^*)$ will be used.

PROPOSITION 6.4. *Assume that $p : U \to U$ is Lipschitz continuous with Lipschitz constant γ. If $\alpha > \gamma$, then* (6.9) *is satisfied.*

Proof. From (6.6),

$$(\alpha - \gamma)\,|u-u^*|_U \le |R|_U,$$

and (6.9) follows. □

PROPOSITION 6.5. *Assume that there exists a neighborhood $\mathcal{N}(u^*)$ and constants $c > 0$ and $\delta > 0$ such that*

$$((H'(y) - H'(y^*))^* p^*, y - y^*)_U + |y-y^*|_Y^2 \ge \delta\,|y-y^*|_Y^2 \tag{6.12}$$

and

$$|p-p^*|_U \le c\,|y-y^*|_Y \tag{6.13}$$

for $u \in \mathcal{N}(u^*)$, where $y = T(u)$ and $p = -T'(u)^*(T(u)-z)$. If further

$$(H(y) - H(y^*) - H'(y)(y-y^*), p-p^*)_U \le \epsilon\,|y-y^*|_Y^2, \tag{6.14}$$

where $\epsilon \to 0$ as $|u-u^*| \to 0$, then (6.9) *is satisfied.*

Proof. From (6.7),

$$H'(y)(y-y^*) + H(y) - H(y^*) - H'(y)(y-y^*) = \tfrac{1}{\alpha}G^I(p-p^*) + \tfrac{1}{\alpha}R,$$

$$H'(y)^*(p-p^*) + (H'(y) - H'(y^*))^* p^* + y - y^* = 0.$$

Multiplying the first equation by $-(p-p^*)$ and the second equation by $y-y^*$ and then summing them, we have

$$((H'(y) - H'(y^*))^* p^*, y-y^*)_U + |y-y^*|_Y^2 + \tfrac{1}{\alpha}|(p-p^*)_I|_U^2$$

$$\le (H(y) - H(y^*) - H'(y)(y-y^*), p-p^*)_U - \tfrac{1}{\alpha}(R, p-p^*).$$

The assumptions of the proposition imply the existence of a constant C and a neighborhood $N(u^*)$ such that

$$|p-p^*|_U \le C\,|R|_U$$

for all $u \in N(u^*)$. The conclusion now follows from (6.2) and (6.4). □

If H is twice continuously Fréchet differentiable in a neighborhood of y^* and $H''(y^*)p^*$ is small, then (6.12) in Proposition 6.4 is satisfied. This smallness condition

can be implied by smallness of $y^* - z$, for example. The solution u^* itself and hence y^* depend on α, and the question arises whether by appropriate choice of α it can be guaranteed that $y^* - z$ is sufficiently small. This problem is addressed by assuming that there exists a feasible control $\bar{u}$ such that $T(\bar{u}) - z$ is sufficiently small. Then, it can be shown that under appropriate conditions, the smallness condition for $y^* - z$ holds for all sufficiently small α.

Remark 6.1. We demonstrate the applicability of Proposition 6.5. For this purpose we return to the class of problems described at the beginning of section 5 and assume in addition that $\phi \in C^2$. Let B denote a bounded subset of $L^2(\Omega)$. Then $y = y(u)$ and $p = p(u)$ with $u \in B$ are bounded sets in $H^2(\Omega) \cap H_0^1(\Omega) \cap L^\infty(\Omega)$. We have

$$-\Delta(y - y^*) + \phi(y) - \phi(y^*) = u - u^*,$$

$$-\Delta(p - p^*) + \phi'(y)(p - p^*) + (\phi'(y) - \phi'(y^*))p^* + y - y^* = 0.$$

Since ϕ is monotone, it follows from the first equation with the arguments of Proposition 5.1 that

$$|y - y^*|_1 \leq \kappa^2 \, |u - u^*|_{L^2},$$

and from the second equation we deduce that

$$|p - p^*|_1 \leq (\kappa + M|p^*|_{L^\infty}) \, |y - y^*|_{L^2}$$

for a constant M independent of $u \in B$. Thus, (6.13) holds. Moreover, since $|(\phi'(y) - \phi'(y^*))p^* + y - y^*|_{L^2} \to 0$ as $|u - u^*|_{L^2} \to 0$, it follows that $|p - p^*|_{H^2} \to 0$ for $|u - u^*| \to 0$ as well. To argue (6.14) observe that

$$(H(y) - H(y^*) - H'(y)(y - y^*), p - p^*)$$

$$= \int_0^1 (\phi''(y + t(y - y^*))(y - y^*)^2, p - p^*) \, dt \leq \bar{M} \, |p - p^*|_{L^\infty} ||y - y^*|_{L^2}^2,$$

for a constant $\bar{M}$ independent of $u \in B$. Moreover $|p - p^*|_{L^\infty} \to 0$ as $|u - u^*|_{L^2} \to 0$ and hence (6.14) holds. Finally note that

$$|(\phi'(y) - \phi'(y^*))p^*|_{L^2} \leq \hat{M} \, |y^* - z|_{L^2} |y - y^*|_{L^2},$$

where $\hat{M}$ is independent of $\alpha > 0$ and thus (6.12) follows.

7. Semismooth Newton method. So far we have analyzed global convergence as well as local superlinear convergence of the partial semismooth Newton method for solving (2.1). In this section we consider the semismooth Newton method for (2.1). The starting point is the optimality condition in the form (6.2). Given u^0, the semismooth Newton iteration is determined by

$$\begin{aligned} &\alpha(u^{k+1} - \psi) - G_{\max}^k \, p'(u^k)(u^{k+1} - u^k) + G_{\min}^k \, p'(u^k)(u^{k+1} - u^k) \\ &\qquad + \max(0, \alpha\,\psi - p(u^k)) + \min(0, p(u^k) - \alpha\,\varphi) = 0, \end{aligned} \tag{7.1}$$

where $G_{\max}^k$ and $G_{\min}^k$ are as given for (6.3). Combined with (6.4) the error equation is found to be

$$\begin{cases} \alpha(u^{k+1} - u^*) - G_{\max}^k p'(u_k)(u^{k+1} - u^*) + G_{\min}^k p'(u_k)(u^{k+1} - u^*) \\ = -\max(0, \alpha\,\psi - p(u^k)) + \max(0, \alpha\,\psi - p(u^*)) - G'_{\max} p'(u_k)(u^k - u^*) \\ -\min(0, p(u^k) - \alpha\,\varphi) + \min(0, p(u^*) - \alpha\,\varphi) + G'_{\min} p'(u^k - u^*) = \widetilde{R}. \end{cases} \tag{7.2}$$

Thus, given u^k the next iterate u^{k+1} is the solution to

$$\alpha(u-u^*) - G^{I_k} p'(u_k)(u-u^*) = \widetilde{R}, \tag{7.3}$$

where $\widetilde{R} = \widetilde{R}(u^k)$ and G^{I_k} is the characteristic function of the set $I_k = \{x\colon \alpha\,\varphi(x) \le p(u^k)(x) \le \alpha\,\psi(x)\}$. Equation (7.3) is the analogue to (6.6), which characterized the error equation for the partial semismooth Newton method. From (7.3) it follows that the convergence rate of u^k to u^* is determined by $\widetilde{R}$ and invertibility of the family of operators

$$M(u, I) = \alpha\, I - G^I p'(u)$$

as elements of $\mathcal{L}(L^2(\Sigma))$, where $u \in L^2(\Sigma)$ and I is a measurable subset of Ω. With respect to the latter we shall utilize the following assumption:

$$\begin{cases} \text{There exists a neighborhood } \bar{N}(u^*) \text{ of } u^* \text{ and a constant } C > 0 \\ \text{such that} \\ \|M^{-1}(u, I)\|_{\mathcal{L}(L^2(\Sigma))} \le C \quad \text{for all } u \in U(u^*) \ \text{ and } \ I \subset \Omega. \end{cases} \tag{7.4}$$

If we assume p to be C^1 and the conditions of either Proposition 6.4 or 6.5 are satisfied, then (7.4) holds. This follows from minor modifications of the proofs of these results. Turning to the estimate of $\widetilde{R}$ we observe that it involves the generalized derivatives of the composite mappings $u \to \max(0, \alpha\,\psi - p(u))$ and $u \to \min(0, p(u) - \alpha\,\varphi)$. We recall the notion of Newton differentiability from [HIK] and prove a chain rule.

Let X, Z be Banach spaces, $D \subset X$ an open subset, and $\omega \subset \mathbb{R}^n$ bounded.

Definition 7.1. *$F\colon D \subset X \to Z$ is called Newton differentiable at x^* if there exists a neighborhood $N(x^*) \subset D$ and a family of mappings $G\colon N(x^*) \to \mathcal{L}(X, Z)$ such that $\{\|G(x)\|\colon x \in N(x^*)\}$ is bounded and*

$$\lim_{|h|_X \to 0} \frac{1}{|h|_X} |F(x^* + h) - F(x^*) - G(x^* + h)h|_Z = 0. \tag{A}$$

Lemma 7.2 (chain rule). *Let $g\colon D \subset L^p(\omega) \to L^q(\omega)$, $1 \le p < q < \infty$, be continuously Fréchet differentiable at $y^* \in D$ and let $\phi\colon L^q(\omega) \to L^p(\omega)$ be Newton differentiable at $g(y^*)$ with a generalized gradient G. Then $F = \phi(g)$ is Newton differentiable at y^* with a generalized gradient given by $G(g)g' \in \mathcal{L}(L^p(\omega), L^p(\omega))$.*

Proof. Let U be a convex neighborhood of y^* in $L^p(\Omega)$ such that g' is continuous in U and $g(U)$ is contained in the neighborhood $N(g(y^*))$ according to the definition of Newton differentiability of ϕ at $g(y^*)$. Further let U and $C > 0$ be such that

$$\|G(g(y))\|_{\mathcal{L}(L^q, L^p)} \le C \quad \text{for all } y \in U$$

and

$$\|g'(y)\|_{\mathcal{L}(L^p, L^q)} \le C \quad \text{for all } y \in U.$$

Let $h \in L^p(\Omega)$ be such that $y^* + h \in U$. We shall utilize that

$$g(y^* + h) = g(y^*) + \int_0^1 g'(y^* + sh)h\, ds \tag{7.5}$$

and $g(y^* + h) \in N(g(y^*))$. Newton differentiability of ϕ at $g(y^*)$ implies that

(7.6)
$$\lim_{|h|_{L^p}\to 0} \frac{1}{|h|_{L^p}} \left|\phi(g(y^* + h)) - \phi(g(y^*)) - G(g(y^* + h)) \int_0^1 g'(y + sh)h \, ds\right|_{L^p} = 0.$$

Here we use (7.5) and the fact that

$$\left|\int_0^1 g'(y + sh)h \, ds\right|_{L^q} \le \int_0^1 \|g'(y + sh)\|_{\mathcal{L}(L^p,L^q)} ds |h|_{L^p} \le C|h|_{L^p}.$$

Let $\epsilon > 0$ be arbitrary. Due to continuity of g' at y^*, there exists a convex neighborhood $U_\epsilon \subset U$ of y^* such that $\|g'(y) - g'(y^*)\|_{\mathcal{L}(L^p,L^q)} < \epsilon$ for all $y \in U_\epsilon$. Consequently

$$\begin{aligned}
&\tfrac{1}{|h|_{L^p}} \left|G(g(y^* + h))\Big(\int_0^1 g'(y^* + sh)h \, ds - g'(y^* + h)h\Big)\right|_{L^p} \\
&\quad \le \tfrac{C}{|h|_{L^p}} \int_0^1 \|g'(y^* + sh) - g'(y^* + h)\|_{\mathcal{L}(L^p,L^q)} ds |h|_{L^p} \\
&\quad \le C\,\epsilon \quad \text{for all } y \in U_\epsilon.
\end{aligned}$$

Combining this estimate with (7.6) we find

$$\lim_{|h|_{L^p}\to 0} \frac{1}{|h|_{L^p}} |\phi(g(y^* + h)) - \phi(g(y^*)) - G(g(y^* + h))g'(y^* + h)|_{L^p} \le C\,\epsilon. \qquad \square$$

THEOREM 7.3. *If* (7.4) *holds and* $u \to p(u) = -(T(u)')^*(T(u) - z)$ *is continuously Fréchet differentiable from neighborhood* $N(u^*) \subset L^2(\Sigma)$ *of* u^* *to* $L^q(\Sigma)$ *for some* $q > 2$, *then the semismooth Newton iteration* (7.1) *converges superlinearly provided* $|u^0 - u^*|_U$ *is sufficiently small.*

Proof. The mappings $u \to \max(0, u)$ and $u \to \min(0, u)$ are Newton differentiable at every u from $L^q(\Sigma)$ to $L^2(\Sigma)$ for each $q > 2$. The claim follows from (7.3), (7.4), and Lemma 7.2 with $p = 2$ applied to $\widetilde{R}$. $\square$

Remark 7.1. For the results of sections 6 and 7, the requirement $\varphi < \psi$ can be relaxed to $\varphi \le \psi$ a.e. in Σ. However, we require $\varphi, \psi \in L^q(\Sigma)$.

Remark 7.2. While the analysis of semismoothness in function space is rather recent, it has a long tradition in finite dimensional spaces. We refer to the pioneering work [M] for semismooth functionals and its extensions, e.g., in [QS, SQ] to R^n-valued functions and nonsmooth versions of Newton's method.

Remark 7.3. The results of sections 6 and 7 depend in an essential manner on the regularity properties of the composite functions $u \to \max(0, \alpha\,\psi - p(u))$ and $u \to \min(0, p(u) - \alpha\,\varphi)$. Recall that $u \to \max(0, u)$ is not Newton differentiable with the indicator function of $\{x\colon u(x) > 0\}$ as generalized derivative from $L^2(\Sigma)$ to itself [HIK], and therefore the smoothing properties that were assumed to hold for $u \to p(u)$ as mapping from $L^2(\Sigma)$ to $L^q(\Sigma)$, for some $q > 2$, are essential. They are satisfied by many optimal control problems related to elliptic and parabolic partial differential equations. The results of section 7 are closely related to the general theory on semismooth Newton methods developed in [U]. In [U] composite functions $\Psi = \psi(F)$ are considered, where F is a differentiable mapping between vector-valued functions

spaces and ψ is a semismooth mapping from R^n to R^k. The generalized derivative of Ψ is given as the composition of the generalized derivative of ψ and F'. The mapping $u \to \max(0, u)$ is not generalized differentiable in the sense of [U] from $L^2(\Sigma)$ to itself. Expressed in the present context, the results in [U] are based on generalized differentiability of $u \to \max(0, \alpha\,\psi - p(u)) + \min(0, p(u) - \alpha\varphi)$ from $L^q(\Sigma)$ to $L^2(\Sigma)$, with $2 < q$, solvability (under appropriate conditions) of the Newton equation in $L^2(\Sigma)$, and an additional smoothing step in the Newton iteration resulting in an update that is again in $L^q(\Sigma)$. Convergence rates are subsequently obtained in $L^q(\Sigma)$. In our work, on the contrary, we utilize smoothing properties of p, which guarantee that the composite mapping $u \to \max(0, \alpha\,\psi - p(u)) + \min(0, p(u) - \alpha\varphi)$ is Newton differentiable.

Remark 7.4. There is a close relationship between the algorithms analyzed in this paper and Bertsekas' projected Newton method for simple constraints; see [B, p. 76 ff]. In the Bertsekas algorithm the new iterate is projected onto the admissible set $U_{\text{ad}} = \{u\colon \varphi \le u \le \psi\}$, before the new system matrix and right-hand side are computed. A more complete comparison is given in [BK]. Since to our knowledge the Bertsekas algorithm was not analyzed in a general context in the infinite dimensional setting, the following observation is of interest. Let P denote the projection in $L^2(\Sigma)$ onto U_{ad}, and assume that for the iterates of the partial (or full) semismooth Newton method we have $|u^{k+1} - u^*| \le \delta_k |u^k - u^*|$ with $\lim_{k\to\infty} \delta_k = 0$. Then the feasible iterates $\tilde{u}^k = P(u^*)$ which are obtained from modifying the algorithm by including a projection step after the Newton update converge superlinearly as well. In fact we have

$$|\tilde{u}^{k+1} - u^*|_{L^2(\Sigma)} = |P(u^{k+1}) - P(u^*)|_{L^2(\Sigma)} \le |u^{k+1} - u^*|_{L^2(\Sigma)} \le \delta_k |\tilde{u}^k - u^*|_{L^2(\Sigma)},$$

as desired.

8. Numerical tests. We give a brief account of some numerical tests that we carried out to solve optimal control problems with unilateral constraints for nonlinear problems of the form

$$\text{(8.1)} \qquad \begin{cases} \min \frac{1}{2}\,|y(u) - z|^2_{L^2(\Omega)} + \frac{\alpha}{2}\,|u|^2_{L^2(\Omega)} \\ -\mu\,\Delta\, y + g(y) = u \ \text{ in } \Omega, \\ y = 0 \ \text{ on } \partial\Omega, \\ u \le \psi \ \text{ in } \Omega. \end{cases}$$

The domain Ω was chosen as the unit square, and the discretization of (8.1) was obtained on the basis of finite differences with a five-point star approximation of the Laplacian and a uniform grid with mesh size $h = \frac{1}{n}$. When applying the primal-dual active set strategy specified at the beginning of section 2, then (2.3) can be expressed as in (2.6) with $H(y) = -\mu\,\Delta\, y + g(y)$. The auxiliary equality constrained problems (2.4) were solved by an SQP method with stopping criterion 10^{-12} for the discrete L^2-norm of the increments. The primal-dual algorithm was initialized by solving (8.1) without the constrained $u \le \psi$ by the SQP method with stopping criterion 10^{-3}.

In the numerical examples that we tested we observed that the primal-dual active set algorithm terminated after finitely many iterations by producing identical active sets in two consecutive iterations. We refer to the corresponding iterate as the solution u_h^* of the discretized problem. This is justified by the fact that such a solution satisfies the discretized form of the optimality system (2.6).

TABLE 1

k	1	2	3	4
q^k	0.3776	0.4082	0.1587	0.0334
$q\text{lam}^{\,k}$	0.1656	0.2462	0.0954	0.0207

The first question that we addressed is whether superlinear convergence can be observed in numerical practice. The problem data are specified by

$$z = e^{x_1} \sin x_2 - 1, \quad \psi = x_1 x_2 - 1, \quad \mu = 0.1, \quad \alpha = 10^{-3},$$
$$g(y) = \sin 3y, \quad \text{and} \quad n = 40.$$

In Table 1 we give the results for

$$q^k = \frac{|u^{k+1} - u_h^*|_{L^2}}{|u^k - u_h^*|_{L^2}}, \quad q\text{lam}^k = \frac{|\lambda^{k+1} - \lambda_h^*|_{L^2}}{|\lambda^k - \lambda_h^*|_{L^2}},$$

where u^0 is the solution from the initialization phase, $u^k, k \geq 1$, the solution of the kth iteration of the primal-dual active set strategy, and u_h^* the solution of the discretized problem as introduced above. Further, λ^k and λ_h^* denote the corresponding Lagrange multipliers and $|\cdot|$ stands for the discrete L^2-norm.

We tested with several other nonlinearities, including $g(y) = y^3$, $g(y) = y^5$, $g(y) = e^y$ and different mesh sizes, and again observed that the algorithm self-terminates due to the coincidence of two consecutive active sets.

In the second class of tests we compared three algorithms. For the sake of the following discussion let us refer to the algorithm that we described at the beginning of this section as the PD-SQP algorithm. This refers to the fact that in an outer loop we utilize the primal-dual active set strategy, and the resulting nonlinear equality constrained problem, see (2.4), is solved by an SQP method. For the second algorithm, referred to as SQP-PD, the order of the two loops is reversed. This is to say, an SQP approach is applied to (8.1) and the resulting inequality constrained linear-quadratic subproblems are solved by the primal-dual active set strategy. This strategy is described in [TV], for example. Here the primal-dual active set strategy allows us to solve the subproblems exactly since again for sufficiently large values of μ and α the algorithm terminates when the active sets coincide in two consecutive iterations. The outer loop is terminated with a stopping criterion measuring the increment in the SQP iteration and is set to 10^{-12} in the examples. The third algorithm arises by applying a semismooth Newton algorithm to the optimality system (2.5) with $H(y) = -\mu\,\Delta\, y + g(y)$. This algorithm, the SSN algorithm, coincides with SQP-PD as well as PD-SQP if in the latter two only one iteration in the inner loops is carried out. The SSN algorithm treats both nonlinearities $g(y)$ as well as $\lambda = \max(0, \lambda + (u - \psi))$ simultaneously. Differently from SQP-PD and PD-SQP, it is not inherent to this algorithm that the nonlinearity $\lambda = \max(0, \lambda + (u - \psi))$ is realized exactly, although depending on the stopping criterion in numerical practice this can occur.

For our comparison the stopping criterion was set to 10^{-12} as for the other algorithms. The comparison itself was made on the basis of the overall number of required solves of linear optimality systems. For test examples with moderate values of μ and α, PD-SQP requires more solves than SQP-PD or SSN. This situation changes for problems with smaller values of α and μ. In view of Theorem 4.1 and Proposition 5.1 it is then less likely that sufficient conditions for global convergence of the primal-dual active set strategy (in SQP-PD as well as in PD-SQP) are satisfied. For such problems

PD-SQP proved to be superior to SQP-PD, with the latter possibly nonterminating in the inner (PD) loop by the criterion of equality of two consecutive active sets. We therefore added the criterion of a maximum of 10 inner iterations to the SQP-PD algorithm. For the problem data given by

$$z = e^{x_1} \sin x_2 - 1,\ \psi = x_1 \cdot x_2 - 1,\ \mu = 10^{-4},\ \alpha = 10^{-4},\ g(y) = y^5,\ n = 40, \tag{8.2}$$

the number of problem solves required by the PD-SQP/SSN/SQP-PD methods is 34/36/66, and with μ, α changed to $\mu = 10^{-3}$, $\alpha = 10^{-5}$ the number of solves is given by 39/51/99. A similar comparison with z, ψ as in (8.1) and $\mu = 10^{-3}$, $\alpha = 10^{-4}$, $g(y) = y^3$ resulted in 27/18/81 required problem solves.

REFERENCES

[B] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[BK] M. Bergounioux and K. Kunisch, *Primal-dual active set strategy for state constrained optimal control problems*, Comput. Optim. Appl., 22 (2002), pp. 193–224.

[BIK] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[H] M. Hintermüller, *A primal-dual active set strategy for bilaterally control constrained optimal control problems*, Quart. Appl. Math., 61 (2003), pp. 131–160.

[HH] W. W. Hager and D. W. Hearn, *Application of the dual active set algorithm to quadratic network problems*, Comp. Optim. Appl., 1 (1993), pp. 349–373.

[HI] W. W. Hager and G. D. Ianculescu, *Dual approximations in optimal control*, SIAM J. Control Optim., 22 (1984), pp. 423–465.

[HIK] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.

[IK1] K. Ito and K. Kunisch, *Augmented Lagrangian methods for nonsmooth convex optimization in Hilbert spaces*, Nonlinear Anal. Theory Meth. Appl., 41 (2000), pp. 573–589.

[IK2] K. Ito and K. Kunisch, *Semismooth Newton methods for variational inequalities of the first kind*, Math. Model. Numer. Anal., 37 (2002), pp. 41–62.

[KS] C. T. Kelley and E. W. Sachs, *Solution of optimal control problems by a pointwise projected Newton method*, SIAM J. Control Optim., 33 (1995), pp. 1731–1757.

[LPR] Z. Q. Luo, J. S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.

[M] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.

[QS] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.

[SQ] J. Sun and L. Qi, *On NCP-functions*, Comput. Optim. Appl., 13 (1999), pp. 201–220.

[TV] F. Tröltzsch and S. Volkwein, *The SQP-method for Control Constrained Optimal Control of the Burgers Equation*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 649–674.

[U] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–841.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 1, pp. 377–378

ERRATUM: NONLINEAR OBSERVER DESIGN IN THE SIEGEL DOMAIN*

ARTHUR J. KRENER† AND MINGQING XIAO‡

Abstract. There is an error in the proof of the main result of our paper [*SIAM J. Control Optim.*, 41 (2002), pp. 932–953]. An additional assumption is needed for the main result to hold. In this erratum, we supply a corrected version of the main result.

DOI. 10.1137/S0363012903435114

In our paper [1], we considered the problem of transforming the real analytic system

$$\dot{x} = f(x) = Fx + \cdots, \tag{0.1}$$

$$y = h(x) = Hx + \cdots \tag{0.2}$$

by a local, analytic change of coordinates

$$z = \theta(x)$$

and an analytic output injection

$$\beta(y)$$

into the system

$$\dot{z} = Az + \beta(y).$$

As shown by Kazantzis and Kravaris [2], this question is of interest because the latter system admits an observer

$$\dot{\hat{z}} = A\hat{z} + \beta(y)$$

with linear error dynamics

$$\dot{\tilde{z}} = A\tilde{z},$$

where $\tilde{z} = z - \hat{z}$. Such a $\theta(x)$ and $\beta(y)$ must satisfy the PDE

$$\frac{\partial \theta}{\partial x}(x) f(x) = A\theta(x) + \beta(h(x)). \tag{0.3}$$

Using the Lyapunov auxiliary theorem, Kazantzis and Kravaris showed, given an analytic β, that this PDE admits a unique solution if all the eigenvalues of F lie in the same half plane, either the left or the right, and the eigenvalues of A are not resonant with those of F.

*Received by the editors September 22, 2003; accepted for publication October 29, 2003; published electronically June 25, 2004.
http://www.siam.org/journals/sicon/43-1/43511.html
†Department of Mathematics, University of California, Davis, CA 95616-8633 (ajkrener@ucdavis.edu). The research of this author was supported in part by NSF 9970998.
‡Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408 (mxiao@math.siu.edu).

In our paper, we claimed (Theorem 2) the existence of solutions to this PDE under considerably weaker hypothesis on the spectrum of F, but there is an error in our proof. An additional assumption is required: that the eigenvalues of the linear part of the system are of type (C, ν) with respect to (w.r.t.) themselves. The correct statement of the main result is as follows.

MAIN THEOREM. *Assume that* $f : \mathbf{R}^n \to \mathbf{R}^n, h : \mathbf{R}^n \to \mathbf{R}^p$, *and* $\beta : \mathbf{R}^p \to \mathbf{R}^n$ *are analytic vector fields with* $f(0) = 0,\ h(0) = 0,\ \beta(0) = 0$, *and* $F = \frac{\partial f}{\partial x}(0),\ H = \frac{\partial h}{\partial x}(0),\ B = \frac{\partial \beta}{\partial y}(0)$. *Suppose*

1. *there exists an invertible* $n \times n$ *matrix* T *so that* $TF = AT - BH$,
2. *there exists a* $C > 0, \nu > 0$ *such that all the eigenvalues of* A *are of type* (C, ν) *w.r.t.* $\sigma(F)$,
3. *there exists a* $C > 0, \nu > 0$ *such that all the eigenvalues of* F *are of type* (C, ν) *w.r.t.* $\sigma(F)$.

Then there exists a unique analytic solution $z = \theta(x)$ *to the PDE* (0.3) *locally around* $x = 0$. *Moreover* $\frac{\partial \theta}{\partial x}(0) = T$, *so* θ *is a local diffeomorphism.*

Proof. Because the eigenvalues of F are of type (C, ν) w.r.t. themselves, it follows from Siegel's theorem [3] that there exists an analytic local change of coordinates which linearizes the dynamics of the system (0.1). Therefore without loss of generality we may assume that the system is of the form

$$\dot{x} = Fx, \tag{0.4}$$

$$y = h(x). \tag{0.5}$$

Then the PDE (0.3) becomes

$$\frac{\partial \theta}{\partial x}(x)Fx = A\theta(x) + \beta(h(x)), \tag{0.6}$$

which has a unique local solution by Theorem 1 of [1]. □

Remarks. With the additional assumption, our result is no longer a generalization of that of Kazantzis and Kravaris [2]. The set of complex vectors $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{C}^n$ which are not of type (C, ν) w.r.t. themselves for any C is a set of measure zero if ν is large enough [3]. Therefore the additional assumption is satisfied by almost all systems.

REFERENCES

[1] A. J. KRENER AND M. XIAO, *Nonlinear observer design in the Siegel domain*, SIAM J. Control Optim., 41 (2002), pp. 932–953.

[2] N. KAZANTZIS AND C. KRAVARIS, *Nonlinear observer design using Lyapunov's auxiliary theorem*, Systems Control Lett., 34 (1998), pp. 241–247.

[3] V. I. ARNOL'D, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, Berlin, 1988.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 379–412

MAX-PLUS EIGENVECTOR METHODS FOR NONLINEAR H_∞ PROBLEMS: ERROR ANALYSIS*

WILLIAM M. MCENEANEY†

Abstract. The H_∞ problem for a nonlinear system is considered. The corresponding dynamic programming equation is a fully nonlinear, first-order, steady-state partial differential equation (PDE), possessing a term which is quadratic in the gradient. The solutions are typically nonsmooth, and further, there is nonuniqueness among the class of viscosity solutions. In the case where one tests a feedback control to see if it yields an H_∞ controller, or where either the controller or disturbance sufficiently dominates, the PDE is a Hamilton–Jacobi–Bellman equation. The computation of the solution of a nonlinear, steady-state, first-order PDE is typically quite difficult. In a companion paper, we developed an entirely new class of methods for obtaining the "correct" solution of such PDEs. These methods are based on the linearity of the associated semigroup over the max-plus (or, in some cases, min-plus) algebra. In particular, solution of the PDE is reduced to solution of a max-plus (or min-plus) eigenvector problem for known unique eigenvalue 0 (the max-plus multiplicative identity). It was demonstrated that the eigenvector is unique and that the power method converges to it. In the companion paper, the basic method was laid out without discussion of errors and convergence. In this paper, we both approach the error analysis for such an algorithm, and demonstrate convergence. The errors are due to both the truncation of the basis expansion and computation of the matrix whose eigenvector one computes.

Key words. nonlinear H_∞ control, dynamic programming, numerical methods, partial differential equations, max-plus algebra

AMS subject classifications. 49L, 93C10, 35B37, 35F20, 65N99, 47D99

DOI. 10.1137/S0363012902414688

1. Introduction.

We consider the H_∞ problem for a nonlinear system. The corresponding dynamic programming equation (DPE) is a fully nonlinear, first-order, steady-state partial differential equation (PDE), possessing a term which is quadratic in the gradient (for background, see [1], [2], [18], [19], [37] among many notable others). The solutions are typically nonsmooth, and further, there are multiple viscosity solutions—that is, one does not even have uniqueness among the class of viscosity solutions (cf. [32], [33]). The computation of the solution of a nonlinear, steady-state, first-order PDE is typically quite difficult, and possibly even more so in the presence of the nonuniqueness mentioned above. Some previous works in the general area of numerical methods for these problems are [3], [7], [8], [15], [17], [22], and the references therein. In the companion paper [25] to this article, the mathematical background and basic algorithm for a class of numerical methods for such PDEs was discussed. This class of methods employs the max-plus linearity of the associated semigroup. It is a completely new class of methods. The approach is appropriate for two classes of PDEs associated with nonlinear (infinite time-horizon) H_∞ problems. The first is the (design) case where one tests a feedback control to see if it yields an H_∞ controller; the corresponding PDE is a Hamilton–Jacobi–Bellman (HJB) equation. In the case where the "optimal" feedback control is being determined as well, the problem takes

*Received by the editors September 17, 2002; accepted for publication (in revised form) October 12, 2003; published electronically July 2, 2004. Research for this work was partially supported by NSF grants DMS-9971546 and DMS-0307229.
http://www.siam.org/journals/sicon/43-2/41468.html
†Department of Mathematics and Department of Mechanical and Aerospace Engineering, University of California, San Diego, La Jolla, CA 92093-0112 (wmceneaney@ucsd.edu, http://math.ucsd.edu/~wmcenean/).

the form of a differential game, and the PDE is, in general, an Isaacs equation. However, if the controller sufficiently dominates the disturbance, the PDE is still an HJB equation, and this is the second case. There is a slight difference in that the second case makes use of the min-plus algebra rather than the max-plus algebra. In this paper, we will consider only the first case, so as to simplify the discussion. However, one can certainly generalize the discussion to the second case.

The recent history of this new class of methods stems from a study of the robust/H_∞ filter ([31], [13], [11], [14], [28]; see also [5], [21], [6]), which has an associated time-dependent, fully nonlinear, first-order PDE. In [12], the linearity of the associated semigroup *over the max-plus algebra* was noted and provided a key ingredient in the development of a numerical algorithm for this filter. This linearity had previously been noted in [24]. A second key ingredient (first noted to our knowledge in [12]) was the development of an appropriate basis for the solution space over the max-plus algebra, i.e., with the max-plus algebra replacing the standard underlying field. (See also [20], [23] for related work.) This reduced the problem of propagation of the solution of the PDE forward in time to max-plus matrix-vector multiplication—with dimension being that of the number of basis functions being used. A key point here is that only a finite number of basis functions are used, and so one needs to determine a bound on the induced errors.

Returning to the (design case) H_∞ problem, the associated steady-state PDE is solved to determine whether this is indeed an H_∞ controller with that disturbance attenuation level. (If there is a nonnegative, locally bounded solution which is zero at the origin, then it is such an H_∞ controller; see, for instance, [1], [36].) The Hamiltonian is concave in the gradient variable. An example of such a PDE is

$$0 = H(x, \nabla W) = -\left[\frac{1}{2\gamma^2}(\nabla W)^T a(x) \nabla W + (f(x))^T \nabla W + l(x)\right],$$

where the notation will be described further below. There are typically multiple solutions of such PDEs—even when one normalizes by requiring $W(0) = 0$. In the linear-quadratic case, two of these solutions correspond to the stable and antistable manifolds associated with the Hamiltonian. The "correct" solution (i.e., the one corresponding to the available storage or value) was characterized in [36] as the smallest, nonnegative viscosity solution which is zero at the origin. In [33], [32], for the class of problems considered here, a specific quadratic growth bound was given which isolated this correct solution as the unique, nonnegative solution satisfying $0 \le W(x) \le C|x|^2$ for a specific C depending on the problem data.

The max-plus–based methods make use of the fact that the solutions are actually fixed points of the associated semigroup, that is,

$$W = S_\tau[W], \tag{1.1}$$

where S_τ is the semigroup with time-step $\tau > 0$. (See (2.7) for a definition of the semigroup.) In this case, one does not actually use the infinitesimal version of the semigroup (the PDE).

The max-plus algebra is a commutative semifield over $\mathbf{R} \cup \{-\infty\}$ with the addition and multiplication operations given by

$$\begin{aligned} a \oplus b &= \max\{a, b\}, \\ a \otimes b &= a + b, \end{aligned} \tag{1.2}$$

where the operations are defined for $-\infty$ in the obvious way. Note that $-\infty$ is the additive identity, and 0 is the multiplicative identity. Note that it is not a field since the additive inverses are missing. Roughly speaking, it can be extended to mimic the inclusion of additive inverses [4], but we do not need that here. Note that since 0 is the multiplicative identity, we can rewrite (1.1) as

$$0 \otimes W = S_\tau[W]. \tag{1.3}$$

In the companion paper [25] (see also the references therein), we showed that S_τ is linear over the max-plus algebra. With this in mind, one then thinks of W as an infinite-dimensional eigenvector (or eigenfunction) for S_τ corresponding to eigenvalue 0. If one approximates W by some finite-dimensional vector of coefficients in a max-plus basis expansion, then (1.3) can be recast as a finite-dimensional max-plus eigenvector equation (approximating the true solution). Thus, the nonlinear PDE problem is reduced to the solution of a (max-plus) linear eigenvector problem. In [25], an algorithm was generated under the assumption that the actual solution was spanned by a finite number of the basis functions. It also assumed exact computation of the finite-dimensional matrix which had the solution as the eigenvector. (Uniqueness of the eigenvalue and eigenvector were proven there.) In order to keep the paper at a reasonable size, further results regarding details of the numerical methods, convergence proofs, and error bounds were delayed to the current paper (although one may note that [26], [27] contain some of the main points).

Since, in reality, the value function would not have a *finite* max-plus expansion in any but the most unusual cases, we must consider the errors introduced by truncation of the expansion. In [16], the question was addressed in a broad sense. In [27], it was shown that as the number of basis functions increased, the approximation obtained by the algorithm converged to the true value function (assuming perfect computation of the matrix whose eigenvector one wants). We will now obtain some error estimates for the size of the errors introduced by this basis truncation. We also consider errors introduced by the approximation of the elements of the matrix corresponding to the H_∞ problem. Finally, these lead us to consider the relative rates at which the spacing between the basis functions and the improvement in the time-propagation errors in the matrix element computations must converge.

First, we need to review some results from [25] and other earlier papers which will be needed here. This is done in section 2. In section 3, we obtain a bound on the size of the errors in the computation of the finite-dimensional matrix, beyond which one cannot guarantee that the method will produce an approximate solution. Then, in section 4, we consider the errors in the solution introduced by truncation of the basis functions. In section 5, we consider the errors in the solution introduced by approximation of the elements of the finite-dimensional matrix. In section 6, we combine these to determine the relative rates at which the spacing between the basis functions and the matrix element errors must go to zero together.

Portions of this paper have appeared previously in [29], [30], [27], and [26], and the last two specifically discuss aspects of the convergence and error analysis.

2. Review of the max-plus based algorithm. In this section, the H_∞ problem class under consideration and accompanying assumptions are given. This is followed by a review of previous results regarding the max-plus–based algorithm which are necessary for the error analysis to follow.

We will consider the infinite time-horizon H_∞ problem in the fixed-feedback case where the control is built into the choice of dynamics. Recall that the case of active

control computation (i.e., the game case) is discussed briefly in [25], [29], and [30]. Consider the system

$$\dot{X} = f(X) + \sigma(X)w, \qquad X(0) = x, \tag{2.1}$$

where X is the state taking values in $\mathbf{R}^m$, f represents the nominal dynamics, the disturbance w lies in $\mathcal{W} \doteq \{w : [0,\infty) \to \mathbf{R}^\kappa : w \in L_2[0,T] \text{ for all } T < \infty\}$, and σ is an $m \times \kappa$ matrix-valued multiplier on the disturbance.

We will make the following assumptions. These assumptions are not necessary but are sufficient for the results to follow. No attempt has been made at this point to formulate tight assumptions. In particular, in order to provide some clear sketches of proofs, we will assume that all the functions f, σ, and l (given below) are smooth, although that is not required for the results. We will assume that there exist $K_f, c \in (0,\infty)$ such that

$$\begin{aligned} &(x-y)^T(f(x)-f(y)) \le -c|x-y|^2 \quad \forall\, x,y \in \mathbf{R}^m, \\ &f(0) = 0, \\ &|f_x(x)| \le K_f \quad \forall\, x \in \mathbf{R}^m. \end{aligned} \tag{A1}$$

Note that this automatically implies the closed-loop stability criterion of H_∞ control. We assume that there exist $M, K_\sigma < \infty$ such that

$$\begin{aligned} &|\sigma(x)| \le M \quad \forall\, x \in \mathbf{R}^m, \\ &|\sigma^{-1}(x)| \le M \quad \forall\, x \in \mathbf{R}^m, \\ &|\sigma_x(x)| \le K_\sigma \quad \forall\, x \in \mathbf{R}^m. \end{aligned} \tag{A2}$$

Here, we, of course, use σ^{-1} to indicate the Moore–Penrose pseudoinverse, and it is implicit in the bound on $\sigma^{-1}(x)$ that σ is uniformly nondegenerate. Let $l(x)$ be the running cost (to which the L_2-norm disturbance penalty will be added). We assume that there exist $\beta, \alpha < \infty$ such that

$$\begin{aligned} &l_{xx}(x) \le \beta \quad \forall\, x \in \mathbf{R}^m, \\ &0 \le l(x) \le \alpha|x|^2 \quad \forall\, x \in \mathbf{R}^m, \end{aligned} \tag{A3}$$

where notation such as $l_{xx} \le \beta$ will be used as a shorthand to indicate that the matrix $l_{xx} - \beta I$ is negative semidefinite. (There is a reason for allowing β to be greater than 2α, which one might notice; see [32].)

The system is said to satisfy an H_∞ attenuation bound (of γ) if there exists $\gamma \in (0,\infty)$ and a locally bounded available storage function (again, also referred to as the value function in what follows), $W(x)$, such that

$$W(x) = \sup_{w \in \mathcal{W}} \sup_{T < \infty} \int_0^T l(X(t)) - \frac{\gamma^2}{2}|w(t)|^2\, dt. \tag{2.2}$$

The corresponding DPE is

$$\begin{aligned} 0 &= -\sup\nolimits_{w \in \mathbf{R}^\kappa} \left\{ [f(x) + \sigma(x)w]^T \nabla W + l(x) - \tfrac{\gamma^2}{2}|w|^2 \right\} \\ &= -\left[\tfrac{1}{2\gamma^2} (\nabla W)^T \sigma(x) \sigma^T(x) \nabla W + f^T(x) \nabla W + l(x) \right]. \end{aligned} \tag{2.3}$$

Since W itself does not appear in (2.3), one can always scale by an additive constant. It will be assumed throughout that we are looking for a solution satisfying $W(0) = 0$.

We will also suppose that the above constants satisfy

$$\frac{\gamma^2}{2M^2} > \frac{\alpha}{c^2}. \tag{A4}$$

We note that there are examples where (A4) fails and the available storage is ∞. Then one has the following result. (See [32, Thms. 2.5 and 2.6] and [33, Thm. 2.5], where the proofs also appear.)

THEOREM 2.1. *There exists a unique continuous viscosity solution of* (2.3) *in the class*

$$0 \le W(x) \le c\frac{(\gamma-\delta)^2}{2M^2}|x|^2 \tag{2.4}$$

for sufficiently small $\delta > 0$. *Further, this unique continuous viscosity solution is given by*

$$W(x) = \lim_{T\to\infty} V(T,x) = \sup_{T<\infty} V(T,x), \tag{2.5}$$

where V *is the value of the finite time horizon problem with dynamics* (2.1) *and payoff and value*

$$\begin{aligned} J(T,x,w) &= \int_0^T l(X(t)) - \frac{\gamma^2}{2}|w(t)|^2\,dt, \\ V(T,x) &= \sup_{w\in\mathcal{W}} J(T,x,w). \end{aligned} \tag{2.6}$$

Define the semigroup

$$S_\tau[W(\cdot)](x) = \sup_{w\in\mathcal{W}}\Big\{\int_0^\tau l(X(t)) - \frac{\gamma^2}{2}|w(t)|^2\,dt + W(X(\tau))\Big\}, \tag{2.7}$$

where X satisfies (2.1). The next result demonstrates that we may solve the problem by obtaining the fixed point of the semigroup. See [25, Thm. 3.2] and the accompanying proof (or alternatively [30, Thm. 3.2 and proof]).

THEOREM 2.2. *For any* $\tau \in [0,\infty)$, W *given by* (2.5) *satisfies* $S_\tau[W] = W$, *and further, it is the unique solution in the class* (2.4).

The following key result is proved in [25, p. 1153] as well as in earlier references such as [30, Thm. 3.3] and [12, pp. 689–690]. However, to the best of the author's knowledge, the first statement of the result is due to Maslov [24].

THEOREM 2.3. *The solution operator,* S_τ, *is linear in the max-plus algebra.*

As noted in the introduction, the above linearity is a key to the development of the algorithms. A second key is the use of the space of semiconvex functions and a max-plus basis for the space. A function ϕ is *semiconvex* if for every $R < \infty$, there exists C_R such that $\widehat{\phi}(x) \doteq \phi(x) + (C_R/2)|x|^2$ is convex on the ball $B_R \doteq \{x \in \mathbf{R}^m : |x| < R\}$. The infimum over such C_R will be known as the semiconvexity constant for ϕ over B_R. We denote the space of semiconvex functions by $\mathcal{S}$. (The scalar C_R may sometimes be replaced by a symmetric, positive definite matrix where the condition becomes $\phi(x) + (1/2)x^T C_R x$ being convex; the case will be clear from the context.) Let $0 < R < \widehat{R}$, and suppose that ϕ is semiconvex over $B_{\widehat{R}}(0)$ with constant $C_{\widehat{R}}$. Then ϕ is Lipschitz over $B_R(0)$, with some constant L_R. See, for instance, [10] for a proof. Therefore any $\phi \in \mathcal{S}$ must be semiconvex and Lipschitz with some constants

C_R and L_R over any ball $B_R(0)$. Consequently, we define the notation $\mathcal{S}^R_{C,L}$ to be the set of $\phi : \overline{B_R}(0) \to \mathbf{R}$ such that ϕ is semiconvex and Lipschitz over $\overline{B_R}(0)$ with constants C and L, respectively. For simplicity of notation, we will henceforth use the notation $\overline{B}_\rho$ for the closed ball $\overline{B_\rho}(0)$ for any $\rho \in (0,\infty)$. It is essential that the value, W, be semiconvex, and that is given by the next result. The proof appears on pages 1154–1155 in [25].

THEOREM 2.4. *W lies in $\mathcal{S}$; for any $R < \infty$, there exist $C_R, L_R < \infty$ such that $W \in \mathcal{S}^R_{C_R,L_R}$.*

We now turn to the max-plus basis over $\mathcal{S}^R_{C_R,L_R}$. The following theorem is a minor variant of the semiconvex duality result given in [12]. It is derived from convex duality [34], [35] in a straightforward manner. There is a change from [12] in that a scalar constant there is replaced by a symmetric matrix C such that $C - C_R I > 0$, where I is the (usual algebra) identity matrix. This replacement allows more freedom in the actual numerical implementation.

THEOREM 2.5. *Let $\phi \in \mathcal{S}$. In particular, let $C_R, L_R \in (0,\infty)$ be the semiconvexity and Lipschitz constants, respectively, for ϕ over $\overline{B}_R$. Let C be a symmetric, positive definite matrix such that $C - C_R I > 0$. Let $D_R \geq R + |C^{-1}|L_R$, where $|C^{-1}|$ indicates the matrix norm of C^{-1}. (In particular, one may take $D_R = R + L_R/C_R$.) Then for all $x \in \overline{B}_R$,*

(2.8)
$$\phi(x) = \max_{\widetilde{x} \in \overline{B}_{D_R}} \Big[-\frac{1}{2}(x - \widetilde{x})^T C (x - \widetilde{x}) + a_{\widetilde{x}} \Big] = \max_{\widetilde{x} \in \mathbf{R}^m} \Big[-\frac{1}{2}(x - \widetilde{x})^T C (x - \widetilde{x}) + a_{\widetilde{x}} \Big],$$

where

$$a_{\widetilde{x}} = - \max_{x \in \overline{B}_R} \Big[-\frac{1}{2}(x - \widetilde{x})^T C (x - \widetilde{x}) - \phi(x) \Big]. \tag{2.9}$$

COROLLARY 2.6. *Let C_R, L_R, D_R be as in Theorem 2.5. Let $\phi \in \mathcal{S}^R_{C',L'}$, where in this case, C' may be a symmetric, positive definite matrix such that $C - C' > 0$, and $R + |C^{-1}|L' \leq D_R$. Then, (2.8), (2.9) hold.*

Remark 2.7. $R + |C^{-1}|L_R$ may be replaced by $|C^{-1}|\overline{L}_R$, where $\overline{L}_R$ is a Lipschitz constant for $\tilde{\phi}(x) \doteq \phi(x) + \frac{1}{2}x^T C x$ over $\overline{B}_R$. Note that $\overline{L}_R \leq L_R + |C|R$.

Let $\phi \in \mathcal{S}^R_{C_R,L_R}$. Let $\{x_i\}$ be a countable, dense set over $\overline{B}_{D_R}$, and let symmetric $C - C_R I > 0$, where (again) $C_R > 0$ is a semiconvexity constant for ϕ over $\overline{B}_R$. Define

$$\psi_i(x) \doteq -\frac{1}{2}(x - x_i)^T C (x - x_i) \quad \forall\, x \in \overline{B}_R$$

for each i. It may occasionally be handy to extend the domain beyond $\overline{B}_R$ by letting $\psi_i(x) = -\infty$ for $x \notin \overline{B}_R$. Then, using Theorem 2.5, one finds (see [12, pp. 695–698])

$$\phi(x) = \bigoplus_{i=1}^{\infty} [a_i \otimes \psi_i(x)] \quad \forall x \in \overline{B}_R, \qquad \text{where} \qquad a_i \doteq - \max_{x \in \mathbf{R}^m} [\psi_i(x) - \phi(x)]. \tag{2.10}$$

This is a countable max-plus basis expansion for ϕ. More generally, the set $\{\psi_i\}$ forms a max-plus basis for the space $\mathcal{S}^R_{C_R,L_R}$. We now have the following.

THEOREM 2.8. *Given $R < \infty$, there exist semiconvexity and Lipschitz constants constant $C_R, L_R < \infty$ such that $W \in \mathcal{S}^R_{C_R,L_R}$. Let $C - C_R I > 0$ and $\{x_i\}$ be dense*

over $\overline{B}_{D_R}$*, and define the basis* $\{\psi_i\}$ *as above. Then*

$$W(x) = \bigoplus_{i=1}^{\infty} [a_i \otimes \psi_i(x)] \quad \forall\, x \in \overline{B}_R, \tag{2.11}$$

where

$$a_i \doteq -\max_{x \in \overline{B}_R} [\psi_i(x) - W(x)]. \tag{2.12}$$

For the remainder of the section, fix any $\tau \in (0, \infty)$. We also assume throughout this section that one may choose C such that $C - C_R I > 0$ and such that

$$S_\tau[\psi_i] \in \mathcal{S}^R_{C',L'} \ \forall i, \tag{A5}$$

where $C - C' > 0$ and $R + |C^{-1}|L' \leq D_R$. This assures that each $S_\tau[\psi_i]$ has a max-plus basis expansion in terms of the basis $\{\psi_j\}$. We will not discuss this assumption in detail here but simply note that we have verified that this assumption holds for the problems where we have used this max-plus method. We also note that this assumption will need to be replaced by a slightly stricter assumption (A5′) in section 4 for the results there and beyond.

We now proceed to review the basics of the algorithm. In [25], the above theory was developed, and then, rather than proving convergence results for the algorithms, drastic assumptions were made so that the basic concept could be presented, while still keeping the paper to a reasonable length. In [25], [29], [30], it was simply assumed that there was a finite set of basis functions, $\{\psi_i\}_{i=1}^n$, such that W had a finite max-plus basis expansion over $\overline{B}_R$ in those functions, that is,

$$W(x) = \bigoplus_{i=1}^{n} a_i \otimes \psi_i(x), \tag{2.13}$$

and we let $a^T \doteq (a_1\ a_2 \cdots a_n)$, and $B_{j,i} = -\max_{x \in B_R}[\psi_j(x) - S_\tau(\psi_i(x))]$. Let B be the $n \times n$ matrix of elements $B_{j,i}$. Note that B actually depends on τ, but we suppress the dependence in the notation. We made the further drastic assumption that for each $j \in \{1, 2, \ldots, n\}$, $S_\tau[\psi_j]$ also had a finite basis expansion in the same set of basis functions, $\{\psi_i\}_{i=1}^n$, so that

$$S_\tau[\psi_j(x)] = \bigoplus_{i=1}^{n} B_{j,i} \otimes \psi_i(x) \tag{2.14}$$

for all $x \in \overline{B}_R$. Specifically, under (2.13), (2.14) one has the following theorem (see, for instance, [25, Thm. 5.1] or [30, Thm. 5.1]).

THEOREM 2.9. *Suppose expansion* (2.13) *requires* $a_i > -\infty$ *for all* i. $S_\tau[W] = W$ *if and only if* $a = B \otimes a$*, where* $B \otimes a$ *represents max-plus matrix multiplication.*

Continuing with the review, suppose that one has computed B exactly. One must then compute the max-plus eigenvector. We should note that B has a unique max-plus eigenvalue, although possibly many eigenvectors corresponding to that eigenvalue [4]. By the above results, this eigenvalue must be zero. As discussed in [25], [29], [27], one can compute the max-plus eigenvector via the power method; this has the added benefit that the convergence analysis to follow is performed in an analogous way. In the power method, one computes an eigenvector a by

$$a = \lim_{N \to \infty} B^N \otimes \vec{0},$$

where the power is to be understood in the max-plus sense and $\vec{0}$ is the zero vector. Throughout the paper, we let the $\{x_j\}$ be such that $x_1 = 0$, that is, $\psi_1(x) = -\frac{1}{2}x^T Cx$. Since this is simply an approach to arrangement of the basis functions, we do not annotate it as an assumption. The fact that the power method works is encapsulated in the following series of three results which hold under (2.13), (2.14) and are proved in [25, pp. 1158–1160].

LEMMA 2.10. *$B_{1,1} = 0$. Also, there exists $\delta > 0$ such that for all $j \neq 1$, $B_{j,j} \leq -\delta$.*

THEOREM 2.11. *Let $N \in \{1, 2, \ldots, n\}$, $\{k_i\}_{i=1}^{i=N+1}$ such that $1 \leq k_i \leq n$ for all i and $k_{N+1} = k_1$. Suppose we are not in the case $k_i = 1$ for all i. Then*

$$\sum_{i=1}^{N} B_{k_i,k_{i+1}} \leq -\delta.$$

Recall that B has a unique max-plus eigenvalue, although possibly many max-plus eigenvectors corresponding to that eigenvalue [4], and that by the above results, this eigenvalue must be zero (ignoring errors due to approximation).

THEOREM 2.12. *$\lim_{N\to\infty} B^N \otimes 0$ exists, converges in a finite number of steps and satisfies $e = B \otimes e$. Further, this is the unique max-plus eigenvector up to a max-plus multiplicative constant.*

Thus, under the drastic assumptions above, one finds that the power method converges to the unique solution of the eigenvector problem (in a finite number of steps), and that this eigenvector is the finite set of coefficients in the max-plus basis expansion of the value function, W. The next sections will deal with the facts that we actually need to truncate infinite basis expansions, and that the computations of the elements of B are only approximate. Convergence results and error analysis will be performed. This not only will indicate that one can achieve arbitrarily good approximations to W via the above max-plus approach, but also will indicate the rate at which the distance between basis function centers should drop as the time-propagation errors in B drop so as to guarantee convergence. This is somewhat analogous to results for finite difference schemes which indicate the required relative rates at which the time and space step must go to zero for such problems.

For purposes of readability, we briefly *outline* the steps in the max-plus algorithm for (approximate) computation of W over a ball, $\overline{B}_R$.

1. Choose a set of max-plus basis functions of the form $\psi_i(x) = -\frac{1}{2}(x-x_i)^T C(x-x_i)$ where the x_i lie in $\overline{B}_{D_R}$. (In practice however, a rectangular grid has been used.) Choose a "time-step," τ.
2. Compute (approximately) elements of the matrix B given by

$$B_{j,i} = -\max_{x\in\overline{B}_R}[\psi_j(x) - S_\tau(\psi_i(x))].$$

   A reasonably efficient means of computing B is important, and a Runge–Kutta–based approach is indicated in section 5.1.
3. Compute the max-plus eigenvector of B corresponding to max-plus eigenvalue $\lambda = 0$ (i.e., the solution of $e = B \otimes e$). This is obtained from the max-plus power method $a_{k+1} = B \otimes a_k$. This converges exactly in a finite number of steps.
4. Construct the solution approximation from $\widehat{W}(x) \doteq \bigoplus_{i=1}^{n} a_i \otimes \psi_i(x)$ on $B_R(0)$.

3. Allowable errors in computation of B. In this section, we obtain a bound on the maximum allowable errors in the computation of B. If the errors are below this bound, then we can guarantee convergence of the power method to the unique eigenvector. In particular, the guaranteed convergence of the power method relies on Lemma 2.10 and Theorem 2.11 since these imply a certain structure to a directed graph associated with B (see [25], [29]). If there was a sequence $\{k_i\}_{i=1}^{i=N+1}$ such that $1 \le k_i \le n$ for all i and $k_{N+1} = k_1$ such that one does not have $k_i = 1$ for all i and such that

$$\sum_{i=1}^{N} B_{k_i,k_{i+1}} \ge 0$$

then there would be no guarantee of convergence of the power method (nor the ensuing uniqueness result for that matter). In order to determine more exactly, the allowable errors in the computation of the elements of B, we first need to obtain a more exact expression for the δ that appears in Lemma 2.10 and Theorem 2.11, and this will appear in Theorems 3.4 and 3.6. That will be followed by results indicating the allowable error bounds. To begin, one needs the following lemma.

LEMMA 3.1. *Let X satisfy* (2.1) *with initial state $X(0) = x \in \mathbf{R}^m$. Let $K, \tau \in (0, \infty)$, and let $w \in L_2[0, \tau]$. Suppose $\delta > 0$ sufficiently small so that*

$$\delta \le KM^2/[c(1 - e^{-c\tau})], \tag{3.1}$$

where c, M are given in assumptions (A1), (A2). *Then*

$$K|X(\tau) - x|^2 + \delta\|w\|^2_{L_2[0,\tau]} \ge \frac{\delta c}{8M^2}|x|^2\big(1 - e^{-c\tau}\big)^4.$$

Remark 3.2. It may be of interest to note that the assumption on the size of δ does not seem necessary. At one point in the proof to follow, this assumption is used in order to eliminate a case which would lead to a more complex expression on the right-hand side in the result in the lemma statement. If some later technique benefited from not having such an assumption, the lemma proof could be revisited in order to eliminate it. However, at this point, that would seem to be a needless technicality.

Remark 3.3. It is perhaps also worth indicating the intuition behind the inequality obtained in Lemma 3.1. Essentially, it states that, due to the nature of the dynamics of the system, the only way that $|X(\tau) - x|^2$ can be kept small is through input disturbance energy $\|w\|^2$, and so their weighted sum is bounded from below. The dependence on $|x|$ on the right-hand side is indicative of the fact that $|f(x)|$ goes to zero at the origin.

Proof. Note that by (2.1) and assumptions (A1) and (A2),

$$\frac{d}{dt}|X|^2 \le -2c|X|^2 + 2M|X||w| \tag{3.2}$$

$$\le -c|X|^2 + \tfrac{M^2}{c}|w|^2. \tag{3.3}$$

Consequently, for any $t \in [0, \tau]$,

$$|X(t)|^2 \le e^{-ct}|x|^2 + \tfrac{M^2}{c}\int_0^t |w(r)|^2\, dr$$

and so

$$\|w\|^2_{L_2(0,t)} \geq \tfrac{c}{M^2}\Big[|X(t)|^2 - |x|^2\Big] \quad \forall\, t \in [0,\tau]. \tag{3.4}$$

We may suppose

$$|X(t)| \leq \sqrt{1 + (1-e^{-c\tau})^4/2}\,|x| \quad \forall\, t \in [0,\tau]. \tag{3.5}$$

Otherwise by (3.4) and the reverse of (3.5), there exists $t \in [0,\tau]$ such that

$$K|X(\tau)-x|^2 + \delta\|w\|^2_{L_2[0,\tau]} \geq \delta\|w\|^2_{L_2[0,t]} \geq \tfrac{\delta c}{2M^2}(1-e^{-c\tau})^4|x|^2, \tag{3.6}$$

in which case one already has the desired result. Define $\overline{K} \doteq \sqrt{1 + (1-e^{-c\tau})^4/2}$.

Recalling (3.2), and applying (3.5), one has

$$\frac{d}{dt}|X(t)|^2 \leq -2c|X(t)|^2 + 2M\overline{K}|x||w(t)|.$$

Solving this ODI for $|X(t)|^2$, and using the Hölder inequality, yields the bound

$$|X(\tau)|^2 \leq |x|^2 e^{-2c\tau} + \frac{M\overline{K}|x|\|w\|}{\sqrt{c}}\big(1-e^{-4c\tau}\big)^{1/2}. \tag{3.7}$$

This implies

$$|X(\tau)| \leq |x|e^{-c\tau} + \frac{1}{c^{1/4}}\sqrt{M\overline{K}|x|\|w\|}\big(1-e^{-4c\tau}\big)^{1/4}. \tag{3.8}$$

We consider two cases separately. First, we consider the case where $|X(\tau)| \leq |x|$. Then, by (3.8)

$$|X(\tau)-x| \geq |x| - |X(\tau)| \geq |x|(1-e^{-c\tau}) - \frac{1}{c^{1/4}}\sqrt{M\overline{K}|x|\|w\|}\big(1-e^{-4c\tau}\big)^{1/4}. \tag{3.9}$$

Now note that for general $a,b,c \in [0,\infty)$, $a+c \geq b$ implies

$$a^2 \geq \frac{b^2}{2} - c^2. \tag{3.10}$$

By (3.9) and (3.10) (and noting the nonnegativity of the norm),

$$|X(\tau)-x|^2 \geq \max\left\{\frac{1}{2}|x|^2(1-e^{-c\tau})^2 - \frac{M\overline{K}}{\sqrt{c}}|x|\|w\|\big(1-e^{-4c\tau}\big)^{1/2},\, 0\right\},$$

which implies

$$\begin{aligned} K|X(\tau)-x|^2 + \delta\|w\|^2 \geq \max\Big\{ &\frac{K}{2}|x|^2(1-e^{-c\tau})^2 - \frac{KM\overline{K}}{\sqrt{c}}|x|\|w\|\big(1-e^{-4c\tau}\big)^{1/2} \\ &+ \delta\|w\|^2, \delta\|w\|^2\Big\}. \end{aligned} \tag{3.11}$$

The right-hand side of (3.11) is a maximum of two convex quadratic functions of $\|w\|$. The second is monotonically increasing, while the first is positive at $\|w\| = 0$ and initially decreasing. This implies that there are two possibilities for the location of the

minimum of the maximum of the two functions. If the minimum of the first function is to the left of the point where the two functions intersect, then the minimum occurs at the minimum of the first function; alternatively it occurs where the two functions intersect. The minimum of the first function occurs at $\|w\|_{min}$ (where we are abusing notation here, using the *min* subscript on the norm to indicate the value of $\|w\|$ at which the minimum occurs), and this is given by

$$\|w\|_{min} = \frac{KM\overline{K}|x|(1-e^{-4c\tau})^{1/2}}{2\sqrt{c}\delta}. \tag{3.12}$$

The point of intersection of the two functions occurs at

$$\|w\|_{int} = \frac{\sqrt{c}|x|(1-e^{-c\tau})^2}{2M\overline{K}(1-e^{-4c\tau})^{1/2}}. \tag{3.13}$$

The two points coincide when

$$\delta = \frac{KM^2\overline{K}^2(1-e^{-4c\tau})}{c(1-e^{-c\tau})^2} = \frac{KM^2[1+(1-e^{-c\tau})^4/2](1-e^{-4c\tau})}{c(1-e^{-c\tau})^2},$$

and $\|w\|_{int}$ occurs to the left of $\|w\|_{min}$ for δ less than this. It is easy to see that assumption (3.1) implies that δ is less than the value at which the points coincide, and consequently, the minimum of the right-hand side of (3.11) occurs at $\|w\|_{int}$.

Using the value of the right-hand side of (3.11) corresponding to $\|w\|_{int}$, we find that for any disturbance, w,

$$K|X(\tau)-x|^2 + \delta\|w\|^2 \geq \frac{\delta c|x|^2}{4M^2\overline{K}^2}\frac{(1-e^{-c\tau})^4}{(1-e^{-4c\tau})}$$

which, using definition of $\overline{K}$, equals

$$\frac{\delta c|x|^2}{4M^2}\frac{(1-e^{-c\tau})^4}{(1-e^{-4c\tau})[1+(1-e^{-c\tau})^4/2]} \geq \frac{\delta c|x|^2}{8M^2}(1-e^{-c\tau})^4. \tag{3.14}$$

Now we turn to the second case,

$$|X(\tau)| > |x|. \tag{3.15}$$

In this case, (3.15) and (3.8) yield

$$|x|e^{-c\tau} + \frac{1}{c^{1/4}}\sqrt{M\overline{K}|x|\|w\|}\left(1-e^{-4c\tau}\right)^{1/4} > |x|. \tag{3.16}$$

Upon rearrangement, (3.16) yields

$$\|w\| > \frac{\sqrt{c}|x|}{M\overline{K}}\frac{(1-e^{-c\tau})^2}{(1-e^{-4c\tau})^{1/2}}.$$

Consequently, using the definition of $\overline{K}$ and some simple manipulations,

$$\begin{aligned} K|X(\tau)-x|^2 + \delta\|w\|^2 &\geq \frac{\delta c|x|^2(1-e^{-c\tau})^4}{M^2(1-e^{-4c\tau})[1+(1-e^{-c\tau})^4/2]} \\ &\geq \frac{\delta c|x|^2}{2M^2}(1-e^{-c\tau})^4. \end{aligned} \tag{3.17}$$

Combining (3.14) and (3.17) completes the proof. □

Now we turn to how Lemma 3.1 can be used to obtain a more detailed replacement for the δ that appears in Theorems 2.10 and 2.11. Fix $\tau > 0$. Let

$$\hat{\gamma}_0^2 \in \left(\frac{2M^2\alpha}{c^2}, \gamma^2\right), \tag{3.18}$$

and in particular, let $\hat{\gamma}_0^2/2 = \gamma^2/2 - \delta$, where δ is sufficiently small so that

$$2\delta < \gamma^2 - \frac{2M^2\alpha}{c^2}. \tag{3.19}$$

Then all results of section 2 for W hold with γ^2 replaced by $\hat{\gamma}_0^2$, and we denote the corresponding value by $W^{\hat{\gamma}_0}$. In particular, by Theorem 2.8, for any $R < \infty$ there exists semiconvexity constant $C_R^0 < \infty$ for $W^{\hat{\gamma}_0}$ over $\overline{B}_R$, and a Lipschitz constant, L_R^0. Note that the required constants satisfy $C_R^0 < C_R$ (see the proof of Theorem 2.8 as given in [25]). If $L_R^0 > L_R$ sufficiently so that $R + |C^{-1}|L_R^0 > D_R$, we modify our basis to be dense over $\overline{B}_{D_R^0}$, where $D_R^0 \geq R + |C^{-1}|L_R^0$ (and redefine $D_R \doteq D_R^0$ in that case). Then, as before, the set $\{\psi_i\}$ forms a max-plus basis for the space of semiconvex functions over $\overline{B}_R$ with semiconvexity constant, C_R^0, i.e., $\mathcal{S}^R_{C_R^0, L_R^0}$.

For any j, let

$$\overline{x}_j \in \operatorname*{argmax}_{|x| \leq R} \{\psi_j(x) - W^{\hat{\gamma}_0}(x)\}. \tag{3.20}$$

Then for any $x \in \overline{B}_R$,

$$\psi_j(x) - \psi_j(\overline{x}_j) \leq W^{\hat{\gamma}_0}(x) - W^{\hat{\gamma}_0}(\overline{x}_j) - K_0|x - \overline{x}_j|^2, \tag{3.21}$$

where $K_0 > 0$ is the minimum eigenvalue of $C - C_R^0 I > 0$. Note that K_0 depends on $\hat{\gamma}_0$.

THEOREM 3.4. *Let $\hat{\gamma}_0$ satisfy* (3.18). *Let $K = K_0$ satisfy* (3.21) *(where we may take $K_0 > 0$ to be the minimum eigenvalue of $C - C_R^0 I > 0$ if desired). Let $\delta > 0$ satisfy $\delta \leq \frac{\gamma^2}{2} - \frac{\hat{\gamma}_0^2}{2}$ and* (3.1). *Then, for any $j \neq 1$,*

$$B_{j,j} \leq \frac{-\delta c|\overline{x}_j|^2}{8M^2}\left(1 - e^{-c\tau}\right)^4.$$

(Recall that by the choice of ψ_1 as the basis function centered at the origin, $B_{1,1} = 0$; see Lemma 2.10.)

Proof. Let K_0, τ, δ satisfy the assumptions. Then

$$S_\tau[\psi_j](\overline{x}_j) - \psi_j(\overline{x}_j) = \sup_{w \in L_2}\left\{\int_0^\tau l(X(t)) - \frac{\gamma^2}{2}|w(t)|^2\,dt + \psi_j(X(\tau)) - \psi_j(\overline{x}_j)\right\}, \tag{3.22}$$

where X satisfies (2.1) with $X(0) = \overline{x}_j$. Let $\varepsilon > 0$, and w^ε be ε-optimal. Then this implies

$$S_\tau[\psi_j](\overline{x}_j) - \psi_j(\overline{x}_j) \leq \int_0^\tau l(X^\varepsilon(t)) - \frac{\gamma^2}{2}|w^\varepsilon(t)|^2\,dt + \psi_j(X^\varepsilon(\tau)) - \psi_j(\overline{x}_j) + \varepsilon,$$

and by (3.21) and the definition of $\hat{\gamma}_0$

$$\leq \int_0^\tau l(X^\varepsilon(t)) - \frac{\hat{\gamma}_0^2}{2}|w^\varepsilon(t)|^2 - \delta|w^\varepsilon(t)|^2\,dt + W^{\hat{\gamma}_0}(X^\varepsilon(\tau)) - W^{\hat{\gamma}_0}(\overline{x}_j) - K_0|X^\varepsilon(\tau) - \overline{x}_j|^2 + \varepsilon$$

and by Theorem 2.2 (for $W^{\hat{\gamma}_0}$),

$$\leq -\delta\|w^\varepsilon\|^2 - K_0|X^\varepsilon(\tau) - \overline{x}_j|^2 + \varepsilon.$$

Combining this with Lemma 3.1 yields

$$S_\tau[\psi_j](\overline{x}_j) - \psi_j(\overline{x}_j) \leq \frac{-\delta c|\overline{x}_j|^2}{8M^2}\left(1 - e^{-c\tau}\right)^4 + \varepsilon.$$

Since this is true for all $\varepsilon > 0$, one has

$$S_\tau[\psi_j](\overline{x}_j) - \psi_j(\overline{x}_j) \leq \frac{-\delta c|\overline{x}_j|^2}{8M^2}\left(1 - e^{-c\tau}\right)^4. \tag{3.23}$$

But

$$B_{j,j} = \min_{|x|\leq R}\left\{S_\tau[\psi_j](x) - \psi_j(x)\right\}, \tag{3.24}$$

which by (3.23)

$$\leq \frac{-\delta c|\overline{x}_j|^2}{8M^2}\left(1 - e^{-c\tau}\right)^4. \qquad \square$$

Remark 3.5. It is interesting to note that one may modify (3.24) as $B_{j,j} = \min_{x\in\mathbf{R}^m}\{S_\tau[\psi_j](x) - \psi_j(x)\}$ since one has $\psi_j(x) = -\infty$ for $x \notin \overline{B}_R$. One might also note that by the nondegeneracy of σ (assumption (A2)), if any function $\phi > -\infty$ on $\overline{B}_R$, then $S_\tau[\phi] > -\infty$ on $\overline{B}_R$.

THEOREM 3.6. *Let $\hat{\gamma}_0$ satisfy* (3.18). *Let K_0 be as in* (3.21), *and let $\delta > 0$ be given by*

$$\delta = \min\left\{\frac{K_0M^2}{c},\ \frac{\gamma^2}{2} - \frac{\hat{\gamma}_0^2}{2}\right\} \tag{3.25}$$

(which is somewhat tighter than the requirement in the previous theorem). Let $N \in \mathcal{N}$, $\{k_i\}_{i=1}^{i=N+1}$ such that $1 \leq k_i \leq n$ for all i and $k_{N+1} = k_1$. Suppose we are not in the case $k_i = 1$ for all i. Then

$$\sum_{i=1}^N B_{k_i,k_{i+1}} \leq -\max_{k_i}|\overline{x}_{k_i}|^2\frac{\delta c}{8M^2}\left(1 - e^{-cN\tau}\right)^4.$$

Proof. By Theorem 3.4, this is true for $N = 1$. We prove the case $N = 2$. The proof of the general case will then be obvious. First, note the monotonicity of the semigroup in the sense that if $g_1(x) \leq g_2(x)$ for all x, then

$$S_\tau[g_1](x) \leq S_\tau[g_2](x) \quad \forall\, x \in \mathbf{R}^m. \tag{3.26}$$

Suppose either $i \neq 1$ or $j \neq 1$. By definition, $\psi_j(x) + B_{j,i} \leq S_\tau[\psi_i](x)$ for all $x \in \mathbf{R}^m$. Using (3.26) and the max-plus linearity of the semigroup yields

$$S_\tau[\psi_j](x) + B_{j,i} \leq S_{2\tau}[\psi_i](x) \quad \forall\, x,$$

which implies in particular that

$$S_\tau[\psi_j](\overline{x}_i) + B_{j,i} \leq S_{2\tau}[\psi_i](\overline{x}_i). \tag{3.27}$$

Now, employing the same proof as that of Theorem 3.4, but with τ replaced by 2τ (noting that condition (3.1) is satisfied with 2τ replacing τ by our assumption (3.25)), one has as in (3.23)

$$S_{2\tau}[\psi_i](\overline{x}_i) - \psi_i(\overline{x}_i) \leq \frac{-\delta c|\overline{x}_i|^2}{8M^2}\big(1 - e^{-2c\tau}\big)^4. \tag{3.28}$$

Combining (3.27) and (3.28) yields

$$\Big[S_\tau[\psi_j](\overline{x}_i) - \psi_i(\overline{x}_i)\Big] + B_{j,i} \leq \frac{-\delta c|\overline{x}_i|^2}{8M^2}\big(1 - e^{-2c\tau}\big)^4.$$

Using the definition of $B_{i,j}$, this implies

$$B_{i,j} + B_{j,i} \leq \frac{-\delta c|\overline{x}_i|^2}{8M^2}\big(1 - e^{-2c\tau}\big)^4. \tag{3.29}$$

By symmetry, one also has

$$B_{i,j} + B_{j,i} \leq \frac{-\delta c|\overline{x}_j|^2}{8M^2}\big(1 - e^{-2c\tau}\big)^4. \tag{3.30}$$

Combining (3.29) and (3.30) yields

$$B_{i,j} + B_{j,i} \leq -\max\Big\{|\overline{x}_i|^2, |\overline{x}_j|^2\Big\}\frac{\delta c}{8M^2}\big(1 - e^{-2c\tau}\big)^4. \qquad \square$$

The convergence of the power method (described in the previous section) relied on a certain structure of B ($B_{1,1} = 0$ and strictly negative loop sums as described in the assumptions of Theorem 2.11). Combining this with the above result on the size of loop sums, one can obtain a condition which guarantees convergence of the power method to a unique eigenvector corresponding to eigenvalue zero. This is given in the next theorem.

THEOREM 3.7. *Let B be given by $B_{j,i} = -\max_{x \in \overline{B}_R}\big(\psi_j(x) - S_\tau[\psi_i](x)\big)$ for all $i, j \leq n$, and let $\widetilde{B}$ be an approximation of B with $\widetilde{B}_{1,1} = 0$ and such that there exists $\varepsilon > 0$ such that*

$$|\widetilde{B}_{i,j} - B_{i,j}| \leq \max\big\{|\overline{x}_i|^2, |\overline{x}_j|^2\big\}\left(\frac{\delta c}{8M^2}\right)\frac{\big(1 - e^{-c\tau}\big)^4}{n^2} - \varepsilon \quad \forall\, i, j \text{ such that } (i,j) \neq (1,1), \tag{3.31}$$

where

$$\delta = \min\left\{\frac{K_0 M^2}{c},\ \frac{\gamma^2}{2} - \frac{\hat{\gamma}_0^2}{2}\right\}. \tag{3.32}$$

Then the power method applied to $\widetilde{B}$ converges in a finite number of steps to the unique eigenvector $\widetilde{e}$ corresponding to eigenvalue zero, that is,

$$\widetilde{e} = \widetilde{B} \otimes \widetilde{e}.$$

Proof. Let $N \in \mathcal{N}$, and consider a sequence of nodes $\{k_i\}_{i=1}^{N+1}$ with $k_1 = k_{N+1}$. We must show that if we are not in the case $k_i = 1$ for all i, then

$$\sum_{i=1}^{N} \widetilde{B}_{k_i,k_{i+1}} < 0.$$

Suppose $N > n^2$. Then any sequence $\{k_i\}_{i=1}^{N+1}$ with $k_1 = k_{N+1}$ must be composed of subloops of length no greater than n^2. Therefore, it is sufficient to prove the result for $N \le n^2$. Note that by the assumptions and Theorem 3.6,

$$\begin{aligned}\sum_{i=1}^{N} \widetilde{B}_{k_i,k_{i+1}} &\le \sum_{i=1}^{N} B_{k_i,k_{i+1}} + \sum_{i=1}^{N} |\widetilde{B}_{k_i,k_{i+1}} - B_{k_i,k_{i+1}}| \\ &\le -\max_{k_i} |\overline{x}_{k_i}|^2 \frac{\delta c}{8M^2} \left(1 - e^{-cN\tau}\right)^4 + \max_{k_i} |\overline{x}_{k_i}|^2 \frac{\delta c}{8M^2} \left(1 - e^{-cN\tau}\right)^4 (N/n^2) - \varepsilon \\ &\le -\varepsilon.\end{aligned}$$

Then by the same proofs as for Theorem 2.12, the result follows. ∎

Theorem 3.7 will be useful later when we analyze the size of errors introduced by our computational approximation to the elements of B.

If the conditions of Theorem 3.7 are met, then one can ask what the size of the errors in the corresponding eigenvector are. Specifically, if eigenvector $\tilde{e}$ is computed using approximation $\widetilde{B}$, what is a bound on the size of the difference between e (the eigenvector of B) and $\tilde{e}$? The following theorem gives a rough, but easily obtained, bound.

THEOREM 3.8. *Let B be given by $B_{i,j} = -\max_{x \in \overline{B}_R} \big(\psi_j(x) - S_\tau[\psi_i](x)\big)$ for all $i, j \le n$, and let $\widetilde{B}$ be an approximation of B with $\widetilde{B}_{1,1} = 0$ and such that there exists $\varepsilon > 0$ such that*

$$|\widetilde{B}_{i,j} - B_{i,j}| \le \max\big\{|\overline{x}_i|^2, |\overline{x}_j|^2\big\} \left(\frac{\delta c}{8M^2}\right) \frac{\left(1 - e^{-c\tau}\right)^4}{n^\mu} - \varepsilon \quad \forall\, i, j, \tag{3.33}$$

where $\mu \in \{2, 3, 4, \dots\}$ and δ is given by (3.32)*. Then the power method will yield the unique eigenvectors e and $\tilde{e}$ of B and $\widetilde{B}$, respectively, in finite numbers of steps, and*

$$\|e - \tilde{e}\| \doteq \max_i |e_i - \tilde{e}_i| \le \big(D_R\big)^2 \left(\frac{\delta c}{8M^2}\right) \frac{\left(1 - e^{-c\tau}\right)^4}{n^{\mu-2}} - \varepsilon.$$

Proof. By Theorem 3.7, one may use the power method to compute $\tilde{e}$, and so one has that for any $j \le n^2$,

$$\tilde{e}_j = \big[\widetilde{B}^{n^2} \otimes 0\big]_j = \max_{m \le n^2} \big[\widetilde{B}^m \otimes 0\big]_j = \max_{m \le n^2} \max_{\{k_l\}_{l=1}^m,\, k_1 = j} \sum_{l=1}^{m} \widetilde{B}_{k_l,k_{l+1}},$$

where the exponents on $\widetilde{B}$ represent max-plus exponentiation and the bound $m \le n^2$ follows from the fact that under the assumption, the sums around any loop other than

that of the trivial loop, $\widetilde{B}_{1,1} = 0$, are strictly negative. Therefore,

$$\tilde{e}_j \leq \max_{m \leq n^2} \max_{\{k_l\}_{l=1}^m,\, k_1 = j} \left[\sum_{l=1}^{m} \left| \widetilde{B}_{k_l,k_{l+1}} - B_{k_l,k_{l+1}} \right| + \sum_{l=1}^{m} B_{k_l,k_{l+1}} \right],$$

which by the assumption (3.33) and the fact that e is the eigenvector of B,

$$\leq \left(D_R\right)^2 \left(\frac{\delta c}{8M^2} \right) \frac{\left(1 - e^{-c\tau}\right)^4}{n^{\mu-2}} - \varepsilon + e_j.$$

By a symmetrical argument, one obtains

$$|\tilde{e}_j - e_j| \leq \left(D_R\right)^2 \left(\frac{\delta c}{8M^2} \right) \frac{\left(1 - e^{-c\tau}\right)^4}{n^{\mu-2}} - \varepsilon. \qquad \square$$

We remark that by taking ε sufficiently small, and noting that $1 - e^{-c\tau} \leq c\tau$ for nonnegative τ, Theorem 3.8 implies (under its assumptions)

$$\|e - \tilde{e}\| = \max_i |e_i - \tilde{e}_i| \leq \left(D_R\right)^2 \left(\frac{\delta c^5}{8M^2} \right) \frac{\tau^4}{n^{\mu-2}}. \tag{3.34}$$

Also note that aside from the case $i = j = 1$ (recall $B_{1,1} = 0$), one has

$$\min_{i \neq 1} \left\{ |\overline{x}_i|^2 \right\} \leq \max \left\{ |\overline{x}_i|^2, |\overline{x}_j|^2 \right\} \quad \forall\, i, j.$$

Using this, and choosing $\varepsilon > 0$ appropriately, one has the following theorem (where we note the condition on the errors in B is uniform but potentially significantly stricter). The proof is nearly identical to that for Theorem 3.8

THEOREM 3.9. *Let B be as in Theorem* 3.8*, and let $\widetilde{B}$ be an approximation of B with $\widetilde{B}_{1,1} = 0$ and such that*

$$|\widetilde{B}_{i,j} - B_{i,j}| \leq \min_{i \neq 1} \left\{ |\overline{x}_i|^2 \right\} \left(\frac{\delta c}{9M^2} \right) \frac{\left(1 - e^{-c\tau}\right)^4}{n^{\mu}} \quad \forall\, i, j, \tag{3.35}$$

where $\mu \in \{2, 3, 4, \dots\}$ and δ is given by (3.32)*. Then the power method will yield the unique eigenvectors e and $\tilde{e}$ of B and $\widetilde{B}$, respectively, in finite numbers of steps, and*

$$\|e - \tilde{e}\| \leq \min_{i \neq 1} \left\{ |\overline{x}_i|^2 \right\} \left(\frac{\delta c}{9M^2} \right) \frac{\left(1 - e^{-c\tau}\right)^4}{n^{\mu-2}}.$$

A simpler variant on this result may be worth using. Note that for $\tau \in [0, 1/c]$, one has $1 - e^{-c\tau} \geq (c/2)\tau$. Then by a proof again nearly identical to that of Theorem 3.8, one has the following.

THEOREM 3.10. *Suppose $\tau \leq 1/c$. Let B be as in Theorem* 3.8*, and let $\widetilde{B}$ be an approximation of B with $\widetilde{B}_{1,1} = 0$ and such that*

$$|\widetilde{B}_{i,j} - B_{i,j}| \leq \min_{i \neq 1} \left\{ |\overline{x}_i|^2 \right\} \left(\frac{\delta c^5}{9(16)M^2} \right) \frac{\tau^4}{n^{\mu}} \quad \forall\, i, j, \tag{3.36}$$

where $\mu \in \{2, 3, 4, \dots\}$ and δ is given by (3.32)*. Then the power method will yield the unique eigenvectors e and $\tilde{e}$ of B and $\widetilde{B}$, respectively, in finite numbers of steps, and*

$$\|e - \tilde{e}\| \leq \min_{i \neq 1} \left\{ |\overline{x}_i|^2 \right\} \left(\frac{\delta c^5}{9(16)M^2} \right) \frac{\tau^4}{n^{\mu-2}}.$$

This variant is included since the simpler right-hand sides might simplify analysis.

4. Convergence and truncation errors. In this section we consider the approximation due to using only a finite number of functions in the max-plus basis expansion. It will be shown that as the number of functions increases (in a reasonable way), the approximate solution obtained by the eigenvector computation of section 2 converges from below to the value function, W. Error bounds will also be obtained.

4.1. Convergence. This subsection contains a quick proof that the errors due to truncation of the basis go to zero as the number of basis functions increases (in a reasonable way). No specific error bounds are obtained; those require the more tedious analysis of the next subsection.

Note that in this subsection a slightly different notation for the indexing and numbers of basis functions in the sets of basis functions is used. This will make the proof simpler. *This alternate notation appears only in this subsection.* Specifically, let us have the sets of basis functions indexed by n; that is, the *sets* are indexed by n. Let the cardinality of the nth set be $\mathcal{I}^{(n)}$. For each n, let $\mathcal{X}^{(n)} \doteq \{x_i^{(n)}\}_{i=1}^{\mathcal{I}^{(n)}}$ and $\mathcal{X}^{(n)} \subset \mathcal{X}^{(n+1)}$. For instance, in the one-dimensional case, one might have $\mathcal{X}^{(1)} = \{0\}$, $\mathcal{X}^{(2)} = \{-1/2, 0, 1/2\}$, $\mathcal{X}^{(3)} = \{-3/4, -1/2, -1/4, 0, 1/4, 1/2, 3/4\}$, and so on. Further, we will let the basis functions be given by $\psi_i^{(n)} \doteq -\frac{1}{2}(x - x_i^{(n)})^T C(x - x_i^{(n)})$, and consider the sets of basis functions $\Psi^{(n)} \doteq \{\psi_i^{(n)} : i \in \mathcal{I}^{(n)}\}$. Then define the approximations to the semigroup operator S_τ by

$$S_\tau^{(n)}[\phi](x) \doteq \bigoplus_{i=1}^{\mathcal{I}^n} a_i^{(n)} \otimes \psi_i^{(n)}(x), \tag{4.1}$$

where

$$a_i^{(n)} \doteq -\max_x \left[\psi_i^{(n)}(x) - S_\tau[\phi](x)\right]. \tag{4.2}$$

In other words, $S_\tau^{(n)}$ is the result of the application of the S_τ followed by the truncation due to a finite number of basis functions. More specifically, if one defines $\mathcal{T}^{(n)}[\phi](x) = \bigoplus_{i=1}^{\mathcal{I}^n} a_i^{(n)} \otimes \psi_i^{(n)}(x)$ with the $a_i^{(n)}$ given by (4.2), then $S_\tau^{(n)}[\phi] = \mathcal{T}^{(n)} \circ S_\tau[\phi]$. Also, let $\mathcal{Y}^{(n)} = \{\phi : \overline{B}_R(0) \to \mathbf{R} \mid \exists\{a_i^{(n)}\}$ such that $\phi(x) = \bigoplus_{i=1}^{\mathcal{I}^n} a_i^{(n)} \otimes \psi_i^{(n)}(x)$ for all $x \in \overline{B}_R(0)\}$. Then note that for $\phi \in \mathcal{Y}^{(n)}$, one has

$$S_\tau^{(n)}[\phi](x) = \bigoplus_{i=1}^{\mathcal{I}^n} \left[\bigoplus_{j=1}^{\mathcal{I}^n} B_{i,j}^{(n)} \otimes a_j^{(n)}\right] \otimes \psi_i^{(n)}(x), \tag{4.3}$$

where $B_{i,j}^{(n)}$ corresponds to $S_\tau^{(n)}[\phi]$.

Lastly, we use the notation ${S_\tau}^N$ to indicate repeated application of S_τ N times. (Of course, by the semigroup property, ${S_\tau}^N = S_{N\tau}$.) Correspondingly, we use the notation ${S_\tau^{(n)}}^N$ to indicate the application of $S_\tau^{(n)}$ N times.

Define $\phi_0(x) \equiv 0$ and

$$\phi_0^{(n)}(x) \doteq \bigoplus_{i=1}^{\mathcal{I}^n} {a^0}_i^{(n)} \otimes \psi_i^{(n)}(x), \qquad {a^0}_i^{(n)} \doteq -\max_x \left[\psi_i^{(n)}(x) - \phi_0(x)\right]. \tag{4.4}$$

It is well known (see [29], [32] among many others) that

$$\lim_{N\to\infty} {S_\tau}^N[\phi_0] = W. \tag{4.5}$$

Also, note that since $\mathcal{X}^{(n)} \subset \mathcal{X}^{(n+1)}$, one has

$$S_\tau^{(n)^N}[\phi_0^{(n)}](x) \leq S_\tau{}^{(n+1)^N}[\phi_0^{(n+1)}](x) \leq S_\tau{}^N[\phi_0](x) \tag{4.6}$$

for all $x \in B_R$.

Note that by (4.4), and the definition of ϕ_0, the corresponding coefficients, $a_i^{0^{(n)}}$, satisfy $a_i^{0^{(n)}} = 0$ for all i. Combining this with Theorem 2.12 and (4.3), one finds that for each n, there exists $\overline{N}(n)$ such that

$$S_\tau^{(n)^N}[\phi_0^{(n)}] = S_\tau^{(n)^{\overline{N}(n)}}[\phi_0^{(n)}] \quad \forall\, N \geq \overline{N}(n). \tag{4.7}$$

Defining

$$W^{(n)^\infty} \doteq S_\tau^{(n)^{\overline{N}(n)}}[\phi_0^{(n)}], \tag{4.8}$$

we further find that the limit is the fixed point. That is,

$$S_\tau^{(n)}[W^{(n)^\infty}] = W^{(n)^\infty}. \tag{4.9}$$

Then, by (4.5), (4.6), and (4.8), we find that

$$W^{(n)^\infty} \text{ is monotonically increasing in } n \tag{4.10}$$

and

$$W^{(n)^\infty} \leq W. \tag{4.11}$$

Therefore, there exists $W^{\infty^\infty} \leq W$ such that

$$W^{(n)^\infty} \uparrow W^{\infty^\infty}, \tag{4.12}$$

and in fact, one can demonstrate equicontinuity of the $W^{(n)^\infty}$ on $\overline{B}_R$ given the assumptions (and consequently uniform convergence).

Under assumption (A5), one can show (see, for instance, Lemma 4.3, although this is more specific than what is needed, or Theorem 3.3 in [27]) that given $\varepsilon > 0$, there exists $n_\varepsilon < \infty$ such that

$$W^{(n)^\infty}(x) = S_\tau^{(n)}[W^{(n)^\infty}](x) \geq S_\tau[W^{(n)^\infty}](x) - \varepsilon$$

for all $x \in B_R$ for any $n \geq n_\varepsilon$. On the other hand, one always has

$$S_\tau^{(n)}[\phi] \leq S_\tau[\phi].$$

Combining these last two inequalities, one obtains

$$W^{(n)^\infty} = S_\tau^{(n)}[W^{(n)^\infty}] \leq S_\tau[W^{(n)^\infty}] \leq S_\tau^{(n)}[W^{(n)^\infty}] + \varepsilon = W^{(n)^\infty} + \varepsilon. \tag{4.13}$$

Combining this with (4.12), one finds the following theorem.

THEOREM 4.1.

$$W^{\infty^\infty} = S_\tau[W^{\infty^\infty}], \tag{4.14}$$

or in other words, W^{∞^∞} is a fixed point of S_τ.

Then, with some more work (see [27], Theorem 3.2), one obtains a convergence theorem.

THEOREM 4.2.

$$W^{\infty^\infty}(x) = W(x) \quad \forall\, x \in \overline{B}_R.$$

4.2. Truncation error estimate. Theorem 4.2 demonstrates convergence of the algorithm to the value function as the basis function density increases. Here we outline one approach to obtaining specific error estimates. The estimates may be rather conservative due to the form of the truncation error bound used; this issue will become more clear below. The main results are in Theorem 4.5 and Remark 4.6. Note that these are only the errors due to truncation to a finite number of basis functions; as noted above, analysis of the errors due to approximation of the entries in the B matrix is discussed further below.

Recall that we choose the basis functions throughout such that $x_1^{(n)} = 0$, or in other words, $\psi_1^{(n)}(x) = \frac{-1}{2}x^T Cx$ for all n. (Note that we return here to the notation where the (n) superscript corresponds to the number of basis functions—as opposed to the more complex notation with cardinality $\mathcal{I}^{(n)}$ which was used in the previous subsection only.) Also, we will use the notation

$$W_{N,\tau}^{(n)}(x) \doteq {S_\tau}^{(n)^N}[\phi_0^{(n)}](x)$$

and we reiterate that the N superscript indicates repeated application of the operator N times. Also, $\phi_0^{(n)}$ is the finite basis expansion of ϕ_0 (with n basis functions).

To specifically set C, we will replace assumption (A5) of section 2 with the following. We assume throughout the remainder of the paper that one may choose matrix $C > 0$ and $\delta' \in (0,1)$ such that with $C' \doteq (1-\delta')C$

$$S_\tau[\psi_i] \in \mathcal{S}^R_{C',L'} \ \forall i, \tag{A5$'$}$$

where $R + |C^{-1}|L' \le D_R$. Again, we do not discuss this assumption in detail, but simply note that we have verified that this assumption holds for the problems we have run. Also note that one could be more general, allowing C' to be a more general positive definite symmetric matrix such that $C - C' > 0$, but we will not include that here. Finally, it should be noted that δ' would depend on τ; as $\tau \downarrow 0$, one would need to take $\delta' \downarrow 0$. Since δ' will appear in the denominator of the error bound of the next lemma (as well as implicitly in the denominator of the fraction on the right-hand side of the error bound in Theorem 4.5), this implies that one does not want to take $\tau \downarrow 0$ as the means for reducing the errors. This will be discussed further in the next section.

The following lemma is a general result about the errors due to truncation when using the above max-plus basis expansion.

LEMMA 4.3. *Let δ', C', L' be as in assumption* (A5$'$)*, and let $\phi \in \mathcal{S}_{R,C'}$ with $\phi(0) = 0$, ϕ differentiable at zero with $\nabla_x \phi(0) = 0$, and $-\frac{1}{2}x^T C'x \le \phi(x) \le \frac{1}{2}\widehat{\mathcal{M}}|x|^2$ for all x for some $\widehat{\mathcal{M}} < \infty$. Let $\{\psi_i\}_{i=1}^n$ consist of basis functions with matrix C, centers $\{x_i\} \subseteq \overline{B}_{D_R}$ such that $C - C'I > 0$, and let $\Delta \doteq \max_{x \in \overline{B}_{D_R}(0)} \min_i |x - x_i|$. Let*

$$\phi^\Delta(x) = \max_i [a_i + \psi_i(x)] \quad \forall\, x \in \overline{B}_R,$$

where

$$a_i = -\max_{x \in \overline{B}_R} [\psi_i(x) - \phi(x)] \quad \forall\, i.$$

Then

$$0 \le \phi(x) - \phi^\Delta(x) \le \begin{cases} |C|\Big[2\widehat{\beta} + 1 + |C|/(\delta' C_R)\Big]|x|\Delta & \text{if } |x| \ge \Delta, \\ \frac{1}{2}\big[\widehat{\mathcal{M}} + |C|\big]|x|\Delta & \text{otherwise,} \end{cases}$$

where $\widehat{\beta}$ is specified in the proof.

Proof. Note that (see [12])

$$\phi(x) = \max_{\widetilde{x}\in\overline{B}_{D_R}} [a(\widetilde{x}) + \psi_{\widetilde{x}}(x)] \quad \forall\, x \in \overline{B}_R(0),$$

where

$$a(\widetilde{x}) = -\max_{x\in\overline{B}_R} [\psi_{\widetilde{x}}(x) - \phi(x)] \quad \forall\, \widetilde{x} \in \overline{B}_{D_R}$$

and

$$\psi_{\widetilde{x}}(x) \doteq -\tfrac{1}{2}(x-\widetilde{x})^T C(x-\widetilde{x}) \quad \forall\, x \in \overline{B}_R(0),\ \widetilde{x} \in \overline{B}_{D_R}.$$

It is obvious that $0 \le \phi(x) - \phi^{\Delta}(x)$, and so we prove the other bound.

Consider any $\overline{x} \in \overline{B}_R$. Then

$$\phi(\overline{x}) = a(\widetilde{x}) + \psi_{\widetilde{x}}(\overline{x}) \tag{4.15}$$

if and only if

$$C(\overline{x} - \widetilde{x}) \in -D_x^- \phi(\overline{x}),$$

where

$$D_x^- \phi(x) = \left\{ p \in \mathbf{R}^m : \ \liminf_{|y-x|\to 0} \frac{\phi(y) - \phi(x) - (y-x)\cdot p}{|y-x|} \ge 0 \right\}.$$

We denote such an $\widetilde{x}$ corresponding to $\overline{x}$ (in (4.15)) as $\overline{\widetilde{x}}$. By the Lipschitz nature of ϕ, one can easily establish that

$$|\overline{\widetilde{x}} - \overline{x}| \le |C^{-1}| L'. \tag{4.16}$$

However, it will be desirable to have a bound where the right-hand side depends linearly on $|\overline{x}|$. (Actually, this may only be necessary for small $\overline{x}$, while (4.16) may be a smaller bound for large $\overline{x}$, but we will obtain it for general $\overline{x}$.) Noting that $\phi \ge -\frac{1}{2}x^T C' x \ge -\frac{1}{2}x^T C x$, one has

$$\tfrac{1}{2}(\overline{x} - \overline{\widetilde{x}})^T C(\overline{x} - \overline{\widetilde{x}}) \le a(\overline{\widetilde{x}}) + \tfrac{1}{2}\overline{x}^T C \overline{x}.$$

Also, since $a(\overline{\widetilde{x}}) + \psi_{\overline{\widetilde{x}}}(\cdot)$ touches ϕ from below at $\overline{x}$, one must have

$$\begin{aligned} \tfrac{1}{2}(\overline{x} - \overline{\widetilde{x}})^T C(\overline{x} - \overline{\widetilde{x}}) - \tfrac{1}{2}(x - \overline{\widetilde{x}})^T C(x - \overline{\widetilde{x}}) &\le a(\overline{\widetilde{x}}) + \tfrac{1}{2}\overline{x}^T C \overline{x} - \tfrac{1}{2}(x - \overline{\widetilde{x}})^T C(x - \overline{\widetilde{x}}) \\ &\le \phi(x) + \tfrac{1}{2}\overline{x}^T C \overline{x} \le \tfrac{1}{2}\widehat{\mathcal{M}}|x|^2 + \tfrac{1}{2}\overline{x}^T C \overline{x} \end{aligned}$$

for all $x \in \overline{B}_R$, where the last inequality is by assumption. Define

$$F(x) \doteq \tfrac{1}{2}(\overline{x} - \overline{\widetilde{x}})^T C(\overline{x} - \overline{\widetilde{x}}) - \tfrac{1}{2}(x - \overline{\widetilde{x}})^T C(x - \overline{\widetilde{x}}) - \tfrac{1}{2}\widehat{\mathcal{M}}|x|^2,$$

and we see that we require $F(x) \le \frac{1}{2}\overline{x}^T C \overline{x}$ for all $x \in \overline{B}_R$. Taking the derivative, we find the maximum of F at $\widehat{x}$ given by

$$\widehat{x} = (C + \widehat{\mathcal{M}} I)^{-1} C \overline{\widetilde{x}} \tag{4.17}$$

and so

$$\widehat{x} - \overline{\widehat{x}} = -\widehat{\mathcal{M}}(C + \widehat{\mathcal{M}}I)^{-1}\overline{\widehat{x}}. \tag{4.18}$$

(In the interests of readability, we ignore the detail of the case where $\widehat{x} \notin B_R(0)$ here.) Therefore, $F(\widehat{x}) \le \frac{1}{2}\overline{x}^T C\overline{x}$ implies

$$\begin{aligned}
(\overline{x} - \overline{\widehat{x}})^T C(\overline{x} - \overline{\widehat{x}}) &\le \overline{\widehat{x}}^T \widehat{\mathcal{M}}(C + \widehat{\mathcal{M}}I)^{-1}C(C + \widehat{\mathcal{M}}I)^{-1}\widehat{\mathcal{M}}\overline{\widehat{x}} \\
&\quad + \overline{\widehat{x}}^T C(C + \widehat{\mathcal{M}}I)^{-1}\widehat{\mathcal{M}}(C + \widehat{\mathcal{M}}I)^{-1}C\overline{\widehat{x}} + \overline{x}^T C\overline{x} \\
&= \widehat{\mathcal{M}}\overline{\widehat{x}}^T \Big[\widehat{\mathcal{M}}(C + \widehat{\mathcal{M}}I)^{-1}C(C + \widehat{\mathcal{M}}I)^{-1} \\
&\qquad + \widehat{\mathcal{M}}(C + \widehat{\mathcal{M}}I)^{-1}\widehat{\mathcal{M}}I(C + \widehat{\mathcal{M}}I)^{-1} \\
&\qquad - \widehat{\mathcal{M}}^2(C + \widehat{\mathcal{M}}I)^{-2} + C(C + \widehat{\mathcal{M}}I)^{-2}C\Big]\overline{\widehat{x}} + \overline{x}^T C\overline{x} \\
&= \widehat{\mathcal{M}}\overline{\widehat{x}}^T \Big[\widehat{\mathcal{M}}C(C + \widehat{\mathcal{M}}I)^{-2} + C(C + \widehat{\mathcal{M}}I)^{-2}C\Big]\overline{\widehat{x}} + \overline{x}^T C\overline{x} \\
&= \overline{\widehat{x}}^T \widehat{\mathcal{M}}C(C + \widehat{\mathcal{M}}I)^{-1}\overline{\widehat{x}} + \overline{x}^T C\overline{x}. \qquad (4.19)
\end{aligned}$$

Noting that C is positive definite symmetric, and writing it as $C = \sqrt{C}\sqrt{C}^T$ where $\sqrt{C} = S\sqrt{\Lambda}$ with S unitary and Λ the matrix of eigenvalues, one may rewrite the first term in the right-hand side of (4.19) as

$$\overline{\widehat{x}}^T \widehat{\mathcal{M}}C(C + \widehat{\mathcal{M}}I)^{-1}\overline{\widehat{x}} = \overline{\widehat{x}}^T \widehat{\mathcal{M}}\tfrac{1}{2}\Big[C(C + \widehat{\mathcal{M}}I)^{-1} + (C + \widehat{\mathcal{M}}I)^{-1}C\Big]\overline{\widehat{x}} = \overline{\widehat{x}}^T \sqrt{C}Q\sqrt{C}^T\overline{\widehat{x}},$$

where

$$Q \doteq \tfrac{1}{2}\widehat{\mathcal{M}}\Big[\sqrt{C}^T(C + \widehat{\mathcal{M}}I)^{-1}\sqrt{C}^{-T} + \sqrt{C}^{-1}(C + \widehat{\mathcal{M}}I)^{-1}\sqrt{C}\Big].$$

Making the change of variables $y = \sqrt{C}^T x$, (4.19) becomes

$$|\overline{y} - \overline{\widehat{y}}|^2 \le \overline{\widehat{y}}^T Q\overline{\widehat{y}} + |\overline{y}|^2.$$

Noting that $\sqrt{C}^T(C + \widehat{\mathcal{M}}I)^{-1}\sqrt{C}^{-T}$ is a similarity transform of $(C + \widehat{\mathcal{M}}I)^{-1}$, one sees that the eigenvalues of Q are the eigenvalues of $\widehat{\mathcal{M}}(C + \widehat{\mathcal{M}}I)^{-1}$. Now, since $(C + \widehat{\mathcal{M}}I)$ is positive definite,

$$(C + \widehat{\mathcal{M}}I) = \overline{S}\,\overline{\Lambda}\,\overline{S}^{-1}$$

with $\overline{\Lambda}$ the diagonal matrix of eigenvalues and $\overline{S}$ the unitary matrix of eigenvectors. Therefore, $\widehat{\mathcal{M}}(C + \widehat{\mathcal{M}}I)^{-1} = \overline{S}(\widehat{\mathcal{M}}\overline{\Lambda}^{-1})\overline{S}^{-1}$, and note that $\beta \doteq \max_i\{\widehat{\mathcal{M}}\overline{\lambda}_i^{-1}\} < 1$ where the $\overline{\lambda}_i$ are the diagonal elements of $\overline{\Lambda}$. Consequently,

$$|\overline{y} - \overline{\widehat{y}}|^2 \le \beta|\overline{\widehat{y}}|^2 + |\overline{y}|^2, \tag{4.20}$$

where $\beta \in (0, 1)$. This implies

$$\begin{aligned}
|\overline{\widehat{y}} - \overline{y}|^2 &\le \beta|\overline{\widehat{y}} - \overline{y} + \overline{y}|^2 + |\overline{y}|^2 \\
&= \beta\Big[|\overline{\widehat{y}} - \overline{y}|^2 + |\overline{y}|^2 + 2(\overline{\widehat{y}} - \overline{y}) \cdot \overline{y}\Big] + |\overline{y}|^2 \\
&\le \beta|\overline{\widehat{y}} - \overline{y}|^2 + (\beta + 1)|\overline{y}|^2 + \beta\Big[\frac{(1 - \beta)/2}{\beta}|\overline{\widehat{y}} - \overline{y}|^2 + \frac{\beta}{(1 - \beta)/2}|\overline{y}|^2\Big],
\end{aligned}$$

which after some rearrangement, yields

$$|\overline{\overline{y}} - \overline{y}|^2 \leq \frac{2(1+\beta^2)}{(1-\beta)^2}|\overline{y}|^2 \tag{4.21}$$

which implies

$$(\overline{x} - \overline{\overline{x}})^T C(\overline{x} - \overline{\overline{x}}) \leq \left[\frac{2(1+\beta^2)}{(1-\beta)^2}\right]\overline{x}^T C\overline{x}.$$

Consequently, there exists $\widehat{\beta} < \infty$ (i.e., $\widehat{\beta} = [|\sqrt{C}|/C_R^0][\sqrt{2(1+\beta^2)}/(1-\beta)]$) such that

$$|\overline{x} - \overline{\overline{x}}| \leq \widehat{\beta}|\overline{x}|. \tag{4.22}$$

Given $\overline{\overline{x}}$, let $\bar{i} \in \operatorname{argmin}_i |x_i - \overline{\overline{x}}|$, and note that

$$|x_{\bar{i}} - \overline{\overline{x}}| \leq \Delta. \tag{4.23}$$

It is easy to see that

$$\begin{aligned}
|\psi_{\overline{\overline{x}}}(x) - \psi_{\bar{i}}(x)| &\leq \tfrac{1}{2}|(x-\overline{\overline{x}})^T C(x-\overline{\overline{x}}) - (x-\overline{\overline{x}})^T C(x-x_{\bar{i}})| \\
&\quad + \tfrac{1}{2}|(x-\overline{\overline{x}})^T C(x-x_{\bar{i}}) - (x-x_{\bar{i}})^T C(x-x_{\bar{i}})| \\
&\leq \tfrac{1}{2}|C|\Big[|\overline{\overline{x}} - x_{\bar{i}}||x-\overline{\overline{x}}| + |\overline{\overline{x}} - x_{\bar{i}}||x-x_{\bar{i}}|\Big] \\
&\leq \tfrac{1}{2}|C|\Big[|\overline{\overline{x}} - x_{\bar{i}}||x-\overline{\overline{x}}| + |\overline{\overline{x}} - x_{\bar{i}}|(|x-\overline{\overline{x}}| + |\overline{\overline{x}} - x_{\bar{i}}|)\Big],
\end{aligned}$$

which by (4.23)

$$\leq |C|\Big[|x-\overline{\overline{x}}|\Delta + \tfrac{1}{2}\Delta^2\Big]. \tag{4.24}$$

Combining (4.22) and (4.24), one finds

$$|\psi_{\overline{\overline{x}}}(\overline{x}) - \psi_{\bar{i}}(\overline{x})| \leq |C|\Big[\widehat{\beta}|\overline{x}|\Delta + \tfrac{1}{2}\Delta^2\Big]. \tag{4.25}$$

Now note that

$$\phi(\overline{x}) - \phi^{\Delta}(\overline{x}) \leq a(\overline{\overline{x}}) + \psi_{\overline{\overline{x}}}(\overline{x}) - [a_{\bar{i}} + \psi_{\bar{i}}(\overline{x})],$$

which by (4.25)

$$\leq |C|\Big[\widehat{\beta}|\overline{x}|\Delta + \tfrac{1}{2}\Delta^2\Big] + a(\overline{\overline{x}}) - a_{\bar{i}}. \tag{4.26}$$

We now deal with the last two terms in this bound. Let

$$\overline{x}_{\bar{i}} \doteq \operatorname*{argmax}_{x \in \overline{B}_R}[\psi_{\bar{i}}(x) - \phi(x)].$$

(Note that we will also skip the technical details of the additional case where $\overline{x}_{\bar{i}}$ lies on the boundary of $\overline{B}_R$.) Then,

$$-C(\overline{x}_{\bar{i}} - x_{\bar{i}}) \in D^- \phi(\overline{x}_{\bar{i}})$$

and

$$-C(\overline{x} - \overline{\overline{x}}) \in D^- \phi(\overline{x}).$$

By the semiconvexity, one has the general result that $p \in D^- \phi(x)$, $q \in D^- \phi(y)$ implies

$$(p - q) \cdot (x - y) \geq -(x - y)^T C'(x - y).$$

Consequently,

$$-(\overline{x}_{\overline{i}} - x_{\overline{i}} + \overline{\overline{x}} - \overline{x})^T C(\overline{x}_{\overline{i}} - \overline{x}) \geq -(\overline{x}_{\overline{i}} - \overline{x})^T C'(\overline{x}_{\overline{i}} - \overline{x}).$$

Recalling that $C' = (1 - \delta')C$, we see that this implies

$$\begin{aligned} -(\overline{x}_{\overline{i}} - \overline{x})^T C(\overline{x}_{\overline{i}} - \overline{x}) + (1 - \delta')(\overline{x}_{\overline{i}} - \overline{x})^T C(\overline{x}_{\overline{i}} - \overline{x}) &\geq -|C|\,|\overline{x}_{\overline{i}} - \overline{x}|\,|x_{\overline{i}} - \overline{\overline{x}}| \\ &\geq -|C|\,|\overline{x}_{\overline{i}} - \overline{x}|\Delta, \end{aligned}$$

or

$$\delta'(\overline{x}_{\overline{i}} - \overline{x})^T C(\overline{x}_{\overline{i}} - \overline{x}) \leq |C|\,|\overline{x}_{\overline{i}} - \overline{x}|\Delta.$$

Noting that $C - C_R I > 0$, this implies

$$|\overline{x}_{\overline{i}} - \overline{x}| \leq \frac{|C|}{\delta' C_R}\Delta. \tag{4.27}$$

Now,

$$\begin{aligned} \overline{\overline{a}} - a_{\overline{i}} &\leq \psi_{\overline{i}}(\overline{x}_{\overline{i}}) - \psi_{\overline{\overline{x}}}(\overline{x}_{\overline{i}}) \\ &= \psi_{\overline{i}}(\overline{x}) - \psi_{\overline{\overline{x}}}(\overline{x}) + [\psi_{\overline{i}}(\overline{x}_{\overline{i}}) - \psi_{\overline{\overline{x}}}(\overline{x}_{\overline{i}})] - [\psi_{\overline{i}}(\overline{x}) - \psi_{\overline{\overline{x}}}(\overline{x})], \end{aligned}$$

which, after cancellation,

$$\begin{aligned} &= \psi_{\overline{i}}(\overline{x}) - \psi_{\overline{\overline{x}}}(\overline{x}) - (\overline{x} - \overline{x}_{\overline{i}})C(x_{\overline{i}} - \overline{\overline{x}}) \\ &\leq |\psi_{\overline{i}}(\overline{x}) - \psi_{\overline{\overline{x}}}(\overline{x})| + |C|\Delta|\overline{x} - \overline{x}_{\overline{i}}|, \end{aligned}$$

which by (4.25) and (4.27)

$$\leq |C|\Big[\widehat{\beta}|\overline{x}| + \big(\tfrac{1}{2} + |C|/(\delta' C_R)\big)\Delta\Big]\Delta. \tag{4.28}$$

Combining (4.26) and (4.28) yields

$$\phi(\overline{x}) - \phi^{\Delta}(\overline{x}) \leq |C|\Big[2\widehat{\beta}|\overline{x}| + \big(1 + |C|/(\delta' C_R)\big)\Delta\Big]\Delta. \tag{4.29}$$

Suppose $|\overline{x}| \geq \Delta$. Then, (4.29) implies

$$\phi(\overline{x}) - \phi^{\Delta}(\overline{x}) \leq |C|\Big[2\widehat{\beta} + 1 + |C|/(\delta' C_R)\Big]|\overline{x}|\Delta, \tag{4.30}$$

which is the first case in right-hand side of the assertion.

Last, suppose $|\overline{x}| < \Delta$. By assumption, there exists $\widehat{\mathcal{M}} < \infty$ such that $\phi(x) \leq \frac{1}{2}\widehat{\mathcal{M}}|x|^2$. Therefore,

$$\phi(\overline{x}) - \phi^{\Delta}(\overline{x}) \leq \tfrac{1}{2}(\widehat{\mathcal{M}} + |C|)|\overline{x}|^2 \leq \tfrac{1}{2}(\widehat{\mathcal{M}} + |C|)|\overline{x}|\Delta,$$

which completes the proof. ∎

The above lemma is a general result about the errors due to truncation with the above max-plus basis expansion. In order to apply this to the problem at hand, one must consider the effect of repeated application of the truncated operator $S_\tau^{(n)}$. Note that $S_\tau^{(n)}$ may be written as the composition of S_τ and a truncation operator, $\mathcal{T}^{(n)}$, where we have

$$\mathcal{T}^{(n)}[\phi] = \phi^\Delta$$

in the notation of the previous lemma, where, in particular, ϕ^Δ was given by

$$\phi^\Delta(x) = \max_i [a_i + \psi_i(x)] \quad \forall\, x \in \overline{B}_R,$$

where

$$a_i = - \max_{x \in \overline{B}_R(0)} [\psi_i(x) - \phi(x)] \quad \forall\, i.$$

In other words, one has the equivalence of notation

$$S_\tau^{(n)}[\phi] = \{\mathcal{T}^{(n)} \circ S_\tau\}[\phi] = \{S_\tau[\phi]\}^\Delta, \tag{4.31}$$

which we shall use freely throughout.

We now proceed to consider how truncation errors accumulate. In order to simplify the analysis, we simply let

$$\mathcal{M}_{C'} \doteq \max\Big\{ |C| \big[2\widehat{\beta} + 1 + |C|/(\delta' C_R)\big],\ \tfrac{1}{2}\big[\widehat{\mathcal{M}} + |C|\big] \Big\}.$$

Fix Δ. We suppose that we have n sufficiently large (with properly distributed basis function centers) so that

$$\max_{x \in \overline{B}_{D_R}} \min_i |x - x_i| \le \Delta.$$

Let ϕ_0 satisfy the conditions on ϕ in Lemma 4.3. (One can simply take $\phi_0 \equiv 0$.) Then, by Lemma 4.3,

$$\phi_0(x) - \mathcal{M}_{C'} |x| \Delta \le \phi_0^{(n)}(x) \le \phi_0(x) \quad \forall\, x \in \overline{B}_R(0). \tag{4.32}$$

Now, for any $x \in \overline{B}_R(0)$, let $w_x^{1,\overline{\varepsilon}}$ be $\overline{\varepsilon}/2$-optimal for $S_\tau[\phi_0](x)$, and let $X_x^{1,\overline{\varepsilon}}$ be the corresponding trajectory. Then,

$$\begin{aligned} 0 &\le S_\tau[\phi_0](x) - S_\tau[\phi_0^{(n)}](x) \\ &\le \phi_0(X_x^{1,\overline{\varepsilon}}(\tau)) - \phi_0^{(n)}(X_x^{1,\overline{\varepsilon}}(\tau)) + \frac{\overline{\varepsilon}}{2}, \end{aligned}$$

which by (4.32)

$$\le \mathcal{M}_{C'} |X_x^{1,\overline{\varepsilon}}(\tau)| \Delta + \frac{\overline{\varepsilon}}{2}. \tag{4.33}$$

Proceeding along, one then finds

$$\begin{aligned} 0 &\le S_\tau[\phi_0](x) - S_\tau^{(n)}[\phi_0^{(n)}](x) \\ &= S_\tau[\phi_0](x) - S_\tau[\phi_0^{(n)}](x) + S_\tau[\phi_0^{(n)}](x) - S_\tau^{(n)}[\phi_0^{(n)}](x), \end{aligned}$$

which by Lemma 4.3, the fact that $S_\tau[\phi_0^{(n)}] \in \mathcal{S}^R_{C',L}$ (by assumption (A5′)), and (4.33)

$$\leq \mathcal{M}_{C'}|X_x^{1,\overline{\varepsilon}}(\tau)|\Delta + \mathcal{M}_{C'}|x|\Delta + \frac{\overline{\varepsilon}}{2}. \tag{4.34}$$

Let us proceed one more step with this approach. For any $x \in \overline{B}_R(0)$, let $w_x^{2,\overline{\varepsilon}}$ be $\overline{\varepsilon}/4$-optimal for $S_\tau[S_\tau[\phi_0]](x)$ (that is $\overline{\varepsilon}/4$-optimal for problem S_τ with terminal cost $S_\tau[\phi_0]$), and let $X_x^{2,\overline{\varepsilon}}$ be the corresponding trajectory. Then, as before,

$$\begin{aligned} 0 &\leq S_{2\tau}[\phi_0](x) - S_\tau[S_\tau^{(n)}\phi_0^{(n)}](x) \\ &= S_\tau[S_\tau[\phi_0]](x) - S_\tau[S_\tau^{(n)}[\phi_0^{(n)}]](x) \\ &\leq S_\tau[\phi_0](X_x^{2,\overline{\varepsilon}}(\tau)) - S_\tau^{(n)}[\phi_0^{(n)}](X_x^{2,\overline{\varepsilon}}(\tau)) + \frac{\overline{\varepsilon}}{4}. \end{aligned} \tag{4.35}$$

Now let

$$w_2^{\overline{\varepsilon}}(t) \doteq \begin{cases} w_x^{2,\overline{\varepsilon}}(t) & \text{if } t \in [0,\tau], \\ w^{1,\overline{\varepsilon}}_{X_x^{2,\overline{\varepsilon}}(\tau)}(t-\tau) & \text{if } t \in (\tau, 2\tau], \end{cases}$$

and let $\overline{X}_x^{2,\overline{\varepsilon}}$ be the corresponding trajectory. Then combining (4.34) and (4.35), one has

$$\begin{aligned} 0 &\leq S_{2\tau}[\phi_0](x) - S_\tau[S_\tau^{(n)}\phi_0^{(n)}](x) \\ &\leq \mathcal{M}_{C'}|\overline{X}_x^{2,\overline{\varepsilon}}(2\tau)|\Delta + \mathcal{M}_{C'}|\overline{X}_x^{2,\overline{\varepsilon}}(\tau)|\Delta + \frac{\overline{\varepsilon}}{2} + \frac{\overline{\varepsilon}}{4}. \end{aligned} \tag{4.36}$$

Applying Lemma 4.3 again, but now using (4.36), one has

$$\begin{aligned} 0 &\leq S_{2\tau}[\phi_0](x) - S_\tau^{(n)}[S_\tau^{(n)}[\phi_0^{(n)}]](x) \\ &= S_\tau[S_\tau[\phi_0]](x) - S_\tau[S_\tau^{(n)}[\phi_0^{(n)}]](x) + S_\tau[S_\tau^{(n)}[\phi_0^{(n)}]](x) - S_\tau^{(n)}[S_\tau^{(n)}[\phi_0^{(n)}]](x) \\ &\leq \mathcal{M}_{C'}|\overline{X}_x^{2,\overline{\varepsilon}}(2\tau)|\Delta + \mathcal{M}_{C'}|\overline{X}_x^{2,\overline{\varepsilon}}(\tau)|\Delta + \mathcal{M}_{C'}|x|\Delta + \frac{\overline{\varepsilon}}{2} + \frac{\overline{\varepsilon}}{4} \\ &= \mathcal{M}_{C'}\Delta \sum_{i=0}^{2} |\overline{X}_x^{2,\overline{\varepsilon}}(i\tau)| + \sum_{i=1}^{2} \frac{\overline{\varepsilon}}{2^i}. \end{aligned} \tag{4.37}$$

It is then clear that, by induction, one obtains the following lemma.

LEMMA 4.4.

$$0 \leq S_{N\tau}[\phi_0](x) - S_\tau^{(n)^N}[\phi_0](x) \leq \mathcal{M}_{C'}\Delta \sum_{i=0}^{N} |\overline{X}_x^{N,\overline{\varepsilon}}(i\tau)| + \sum_{i=1}^{N} \frac{\overline{\varepsilon}}{2^i}, \tag{4.38}$$

where the construction of $\overline{\varepsilon}$-optimal $\overline{X}_x^{N,\overline{\varepsilon}}(\cdot)$ by induction follows in the obvious way as above.

THEOREM 4.5. *Let $\{\psi_i\}_{i=1}^n$, C', and Δ be as in Lemma 4.3. Then, there exists $\overline{m}, \overline{\lambda} \in (0,\infty)$ such that*

$$0 \leq W(x) - W^{(n)^\infty}(x) \leq \mathcal{M}_{C'}\left(\frac{e^{\overline{m}}}{1 - e^{-\overline{\lambda}\tau}}\right)|x|\Delta \quad \forall\, x \in \overline{B}_R(0).$$

Remark 4.6. By Theorem 2.12, there exists $N = N(n) < \infty$ such that

$$W^{(n)^\infty}(x) = W^{(n)^N}(x) \quad \forall\, x \in \overline{B}_R(0),$$

and so Theorem 4.5 also implies

$$0 \le W(x) - W^{(n)^N}(x) \le \mathcal{M}_{C'}\left(\frac{e^{\overline{m}}}{1 - e^{-\overline{\lambda}\tau}}\right)|x|\Delta \quad \forall\, x \in \overline{B}_R(0)$$

for $N \ge N(n)$.

Proof. Let $\overline{\varepsilon} \in (0,1)$. Fix ϕ_0 and x. For each $N < \infty$, construct $w_N^{\overline{\varepsilon}}(\cdot)$ as above along with the corresponding $\overline{X}_x^{N,\overline{\varepsilon}}$. Let $w_\infty^{\overline{\varepsilon}}(t) = w_N^{\overline{\varepsilon}}(t)$ if $t \in [0, N\tau]$, and similarly, $\overline{X}_x^{\infty,\overline{\varepsilon}}(t) = \overline{X}_x^{N,\overline{\varepsilon}}(t)$ if $t \in [0, N\tau]$. Then, by [32] (see also [33]), there exists $\widetilde{K} < \infty$ (independent of $\overline{\varepsilon} \in (0,1)$) such that

$$\|w_\infty^{\overline{\varepsilon}}\|_{L_2(0,N\tau)} \le \widetilde{K}(1 + |x|^2)$$

for all $N < \infty$. Consequently, using assumptions (A1) and (A2), there exist $\overline{m}, \overline{\lambda} \in (0, \infty)$ such that

$$|\overline{X}_x^{\infty,\overline{\varepsilon}}(t)| \le |x| e^{\overline{m} - \overline{\lambda} t} \quad \forall\, t \in [0, \infty). \tag{4.39}$$

Then, by Lemma 4.4 and (4.39),

$$\begin{aligned} 0 \le S_{N\tau}[\phi_0](x) - S_\tau^{(n)^N}[\phi_0](x) &\le \mathcal{M}_{C'}\Delta|x| e^{\overline{m}} \sum_{i=0}^{N} e^{-\overline{\lambda} i\tau} + \sum_{i=1}^{N} \frac{\overline{\varepsilon}}{2^i} \\ &\le \mathcal{M}_{C'}\Delta|x| \frac{e^{\overline{m}}}{1 - e^{-\overline{\lambda}\tau}} + \overline{\varepsilon}. \end{aligned}$$

Since this is true for all $N \in \mathcal{N}$, and $S_\tau^{(n)^N}[\phi_0](x) = S_\tau^{(n)^{N+1}}[\phi_0](x) = W^{(n)^\infty}(x)$ for all $N \ge N(n)$, one obtains the result by taking the limit as $N \to \infty$ and then as $\overline{\varepsilon} \downarrow 0$. □

Last, we note that for τ sufficiently small, where

$$\tau \le 1/\overline{\lambda} \tag{4.40}$$

is sufficient (so that $\overline{\lambda}\tau/2 \le (1 - e^{-\overline{\lambda}\tau})$), one has

$$0 \le W(x) - W^{(n)^\infty}(x) \le \mathcal{M}_{C'}\Delta\Big(\frac{e^{\overline{m}}}{1 - e^{-\overline{\lambda}\tau}}\Big)|x| \le K_1|x|(\Delta/\tau) \tag{4.41}$$

with

$$K_1 \doteq 2\mathcal{M}_{C'} e^{\overline{m}}/\overline{\lambda}. \tag{4.42}$$

5. Errors in the approximation of B. In the previous section, we considered the errors due to truncation while assuming that B and, consequently, the eigenvector e were computed exactly. Of course, as discussed in section 3, there is an allowable upper limit for errors in the elements of B, below which one can guarantee the convergence of the power method. The errors in B also translate into errors in the eigenvector and consequently the approximate solution as discussed in sections 3

and 6. In this section, we consider a power series (in t) for $V(t,x) \doteq S_t[\psi_i](x)$, where we recall $B_{j,i} = -\max_{x\in\overline{B}_R(0)}[-\psi_j(x) - S_\tau[\psi_i](x)]$. With the power series for $V(t,x) = S_t[\psi_i](x)$ truncated at some level, $t^{n'-1}$ (for each i), we obtain a relationship between n', τ, and basis function density which guarantees that the errors in B do not exceed the allowable bounds obtained in section 3. In addition to the errors incurred by truncation of the power series, there may be errors in the computation of the terms in the series themselves. In subsection 5.1, one particular method for computing the power series terms to sufficient accuracy is given.

As noted above, one approach to the computation of B is a Taylor series (in t) approximation to $S_t[\psi_i](x)$. More specifically, letting $V(t,x) = S_t[\psi_i](x)$, so that V satisfies

$$\begin{aligned} V_t &= f\cdot\nabla V + l + \frac{1}{2\gamma^2}\nabla V^T\sigma\sigma^T\nabla V, \\ V(0,x) &= \psi_i(x), \end{aligned} \tag{5.1}$$

one may approximate V as

$$V(t,x) = V_0(x) + V_1(x)t + \frac{1}{2}V_2(x)t^2 + \cdots. \tag{5.2}$$

Here $V_0(x) = \psi_i(x)$ and V_1 is the right-hand side of (5.1) with ψ_i replacing V. Specifically,

$$V_1(x) = f\psi_{ix} + l + \frac{1}{2\gamma^2}\psi_{ix}a\psi_{ix},$$

where $a = \sigma\sigma^T$ and we drop the gradient/vector notation for simplification here and below. The higher order terms are computed by differentiating (5.1) at $t = 0$. Of course this process requires some smoothness for V. The following is well known, and so we only sketch a proof.

THEOREM 5.1. *Given $R' < \infty$ and $n' \in \mathcal{N}$, there exists $\tau' > 0$ such that $V \in C^{n'}((0,\tau')\times B_{R'}(0))$.*

Proof. The result for C^2 can be found, for instance, in [9] as well as many earlier works (see the references in [9] as well as [15]). In order to obtain continuity of higher derivatives, one simply differentiates (5.2) and applies the same technique. For example, the partial $V_{x_l}(t,x)$ satisfies

$$\begin{aligned} U_t &= \Big[f_{x_l}V_x + l_{x_l} + V_x a_{x_l}V_x\Big] + \Big[f + 2V_x a\Big]U_x, \\ U(0,x) &= \psi_{ix_l}(x). \end{aligned}$$

Note that τ' may depend on n'. ∎

Fix some $R', n' < \infty$. Let τ' be given by Theorem 5.1. We assume $\tau < \min\{\tau', 1, 1/c\}$ (where the motivation for the bounds of 1 and $1/c$ appear in (5.8) and (5.11) below) and $\widetilde{R} < R'$. Then we may approximate V over $(0,\tau)\times\overline{B}_{\widetilde{R}}(0)$ by

$$\widetilde{V}(t,x) = V_0(x) + V_1(x)t + V_2(x)\frac{t^2}{2} + \cdots + V_{n'-1}(x)\frac{t^{n'-1}}{(n'-1)!}. \tag{5.3}$$

Letting

$$M_{R',n'} \doteq \max_{(t,x)\in[0,\tau]\times\overline{B}_{\widetilde{R}}(0)} |V_{t^{(n')}}(t,x)|,$$

one has

$$|V(t,x) - \widetilde{V}(t,x)| \leq M_{R',n'} \frac{\tau^{n'}}{(n')!} \quad \forall\, (t,x) \in [0,\tau] \times \overline{B}_{\widetilde{R}}(0). \tag{5.4}$$

Now define the corresponding approximation to B by

$$\widetilde{B}_{j,i} = - \max_{x \in \overline{B}_{\widetilde{R}}(0)} \left\{ \psi_j(x) - \widetilde{V}(\tau, x) \right\}. \tag{5.5}$$

By (5.4) and (5.5), one has

$$|B_{j,i} - \widetilde{B}_{j,i}| \leq M_{R',n'} \frac{\tau^{n'}}{(n')!}. \tag{5.6}$$

Comparing (5.6) with Theorem 3.10, one finds that a sufficient condition for the convergence of the power method (using $\widetilde{B}$ computed from approximation $\widetilde{V}$) is that $\tau \leq 1/c$ and that for some $\mu \in \{2, 3, 4, \dots\}$ ($\mu = 2$ is the weakest condition)

$$M_{R',n'} \frac{\tau^{n'}}{(n')!} \leq \left[\min_{i \neq 1} |\overline{x}_i|^2 \right] \left(\frac{\delta c^5}{9(16) M^2} \right) \frac{\tau^4}{n^\mu}.$$

Note that the $\tau \leq 1/c$ condition can be removed by using Theorems 3.8 and 3.9 instead of Theorem 3.10.

Since computation of $\widetilde{B}_{j,i}$ requires the maximization operation, below we will introduce an approximation for $\widetilde{B}_{j,i}$, to be denoted by $\widehat{B}_{j,i}$ (where the maximum may only be computed approximately rather than exactly). Suppose further that

$$|\widetilde{B}_{j,i} - \widehat{B}_{j,i}| \leq M_{R',n'} \frac{\tau^{n'}}{(n')!}. \tag{5.7}$$

Then, by (5.7), (5.6) with Theorem 3.10, one finds that a sufficient condition for the convergence of the power method (using $\widehat{B}$) is that $\tau \leq 1/c$ and that for some $\mu \in \{2, 3, 4, \dots\}$ ($\mu = 2$ is the weakest condition)

$$2 M_{R',n'} \frac{\tau^{n'}}{(n')!} \leq \left[\min_{i \neq 1} |\overline{x}_i|^2 \right] \left(\frac{\delta c^5}{9(16) M^2} \right) \frac{\tau^4}{n^\mu}, \tag{5.8}$$

and so a sufficient condition is

$$\tau^{n'-4} \leq \left[\min_{i \neq 1} |\overline{x}_i|^2 \right] \left(\frac{\delta c^5 (n')!}{9(32) M^2 M_{R',n'}} \right) \frac{1}{n^\mu}. \tag{5.9}$$

Suppose a rectangular grid of evenly spaced basis function centers with N_D center-points per dimension, and recall that ψ_1 is centered at the origin which implies N_D is odd. (Perhaps it should be noted that this is conservative in that we are considering a rectangular grid encompassing $\overline{B}_{D_R}$ rather than just those basis functions centered in the sphere itself.) This implies $\min_{i \neq 1} |\overline{x}_i|^2 = 4 D_R^2 / (N_D - 1)^2$, and (5.9) becomes

$$\tau^{n'-4} \leq \left(\frac{D_R^2 \delta c^5 (n')!}{9(8) M^2 M_{R',n'}} \right) \left(\frac{1}{N_D} \right)^{m\mu} \left(\frac{1}{N_D - 1} \right)^2,$$

which implies a sufficient condition is

$$\tau^{n'-4} \le \left(\frac{D_R^2 \delta c^5 (n')!}{9(8) M^2 M_{R',n'}}\right)\left(\frac{1}{N_D}\right)^{m\mu+2} \doteq \widetilde{\mathcal{M}}_{R',n'}\left(\frac{1}{N_D}\right)^{m\mu+2}, \tag{5.10}$$

where we recall that m is the dimension of the state space.

Therefore, if one fixes $\tau < \min\{1, 1/c\}$, then it is sufficient that

$$n' \ge 4 + \frac{\log \widetilde{\mathcal{M}}_{R',n'} + (m\mu+2)\log(1/N_D)}{\log \tau}. \tag{5.11}$$

Alternatively, one may, without loss of generality, require $\widetilde{\mathcal{M}}_{R',n'} \ge 1$ in which case (noting that $\log \tau < 0$ since $\tau < 1$) (5.11) yields the sufficient condition

$$n' \ge 4 + \frac{(m\mu+2)\log(1/N_D)}{\log \tau}, \tag{5.12}$$

in which case the lower bound on n' scales like $\log(1/N_D)$. We remark that this sufficient condition may be quite conservative.

5.1. A method for computing B. As noted above, one would not typically have a closed-form expression for the $B_{j,i}$ or even the $\widetilde{B}_{j,i}$ terms, and we denote the approximation of $\widetilde{B}$ by $\widehat{B}$. In this subsection, we indicate some specifics of a numerical method for the approximation. This is not essential to the paper, but we felt that it was useful to sketch an approximation technique so as to concretely indicate one approach to this subproblem.

The approach taken was to define

$$\widetilde{X}_{j,i}(t) \doteq \operatorname{argmax}\{\psi_j(x) - \widetilde{V}(t,x)\},$$

where $\widetilde{V}$ is given by (5.3) (i.e., the truncated power series expansion of $S_t[\psi_i](x)$), and then to propagate $\widetilde{X}_{j,i}$ as the solution of an ODE forward from $t = 0$ to τ via a Runge–Kutta method. One difficulty is that $\widetilde{X}_{j,i}(t)$ diverges as $t \downarrow 0$. In order to remedy this, and also remedy unbounded derivatives as $t \downarrow 0$, we replace $\psi_j(x)$ by $\psi^\tau_{j,i}(t,x)$, where

$$\psi^\tau_{j,i}(t,x) \doteq -\frac{1}{2}(x - \xi(t))^T[(C + \overline{\delta}(1 - t/\tau))I](x - \xi(t)), \tag{5.13}$$

where

$$\xi(t) \doteq x_i + (t/\tau)(x_j - x_i), \tag{5.14}$$

and $\overline{\delta} > 0$. Then one may define

$$\widetilde{X}^\tau_{j,i}(t) \doteq \operatorname*{argmax}_x\{\psi^\tau_{j,i}(t,x) - \widetilde{V}(t,x)\}, \tag{5.15}$$

and note that

$$\widetilde{X}^\tau_{j,i}(\tau) = \widetilde{X}_{j,i}(\tau) = \operatorname{argmax}\{\psi_j(x) - \widetilde{V}(t,x)\}.$$

Since $\widetilde{X}^\tau_{j,i}(t)$ is the argmax at each time $t \in [0,\tau]$, this implies

$$[\psi^\tau_{j,i}]_x(t, \widetilde{X}^\tau_{j,i}(t)) - \widetilde{V}_x(t, \widetilde{X}^\tau_{j,i}(t)) = 0$$

for all $t \in [0,\tau]$. Differentiating with respect to time, implies

$$\Big[[\psi^\tau_{j,i}]_{xx}(t,\widetilde{X}^\tau_{j,i}(t)) - \widetilde{V}_{xx}(t,\widetilde{X}^\tau_{j,i}(t))\Big]\dot{\widetilde{X}}^\tau_{j,i}(t) + \Big[[\psi^\tau_{j,i}]_{tx}(t,\widetilde{X}^\tau_{j,i}(t)) - \widetilde{V}_{tx}(t,\widetilde{X}^\tau_{j,i}(t))\Big] = 0,$$

or,

$$\dot{\widetilde{X}}^\tau_{j,i}(t) = \Big[[\psi^\tau_{j,i}]_{xx}(t,\widetilde{X}^\tau_{j,i}(t)) - \widetilde{V}_{xx}(t,\widetilde{X}^\tau_{j,i}(t))\Big]^{-1}\Big[[\psi^\tau_{j,i}]_{tx}(t,\widetilde{X}^\tau_{j,i}(t)) - \widetilde{V}_{tx}(t,\widetilde{X}^\tau_{j,i}(t))\Big]. \tag{5.16}$$

The initial state for (5.16) is

$$\begin{aligned}\widetilde{X}^\tau_{j,i}(0) &= \operatorname*{argmax}_x\{\psi^\tau_{j,i}(0,x) - \widetilde{V}(0,x)\}\\ &= \operatorname*{argmax}_x\{-\tfrac{1}{2}(x-x_i)^T(C+\overline{\delta}I)(x-x_i) - \psi_i(x)\} = x_i.\end{aligned}$$

Note that

$$\Big[[\psi^\tau_{j,i}]_{xx}(0,x) - \widetilde{V}_{xx}(0,x)\Big] = -[C+\overline{\delta}I] + C = -\overline{\delta}I, \tag{5.17}$$

which is negative definite, and

$$\Big[[\psi^\tau_{j,i}]_{xx}(\tau,x) - \widetilde{V}_{xx}(\tau,x)\Big] = -C - \widetilde{V}_{xx}(\tau,x) \tag{5.18}$$

would be negative definite on $\overline{B}_R$ by assumption (A5′) if approximation $\widetilde{V}(\tau,\cdot)$ were replaced by $S_\tau[\psi_i]$. Also,

$$\widetilde{X}^\tau_{j,i}(0) = x_i \in \overline{B}_{D_R}, \quad \text{and} \quad \widetilde{X}^\tau_{j,i}(\tau) \in \overline{B}_R \tag{5.19}$$

if approximation $\widetilde{V}(\tau,\cdot)$ is replaced by $S_\tau[\psi_i]$. This suggests the following assumption (which is only used for this approach to computing B). Suppose there exists $\hat{\delta} > 0$ such that

$$\begin{aligned}&\Big[[\psi^\tau_{j,i}]_{xx}(t,x) - \widetilde{V}_{xx}(t,x)\Big] + \hat{\delta}I < 0 \quad \forall\, |x| \le \hat{g}(t),\, \forall t \in [0,\tau],\\ &|\widetilde{X}^\tau_{j,i}(t)| \le \hat{g}(t) \quad \forall\, t \in [0,\tau],\end{aligned} \tag{5.20}$$

where $g : [0,\tau] \to \mathbf{R}$ is any function such that $\hat{g}(0) = D_R$, $\hat{g}(\tau) = R$ and $\hat{g}$ is monotonically decreasing. Note that, by (5.17)–(5.19), the conditions are satisfied at both endpoints ($t = 0$ and $t = \tau$) when $\widetilde{V}(\tau,\cdot)$ is replaced by $S_\tau[\psi_i]$. Consequently, this may not be significantly more restrictive than the general assumptions, and for the purposes of sketching this particular approach to computing B, let us assume (5.20). Note that this guarantees the existence of the inverse in (5.16) and further that $\widetilde{X}^\tau_{j,i}(\tau)$ is the unique maximizer in $\overline{B}_R$.

Analytical expressions for the right-hand side of (5.16) can be obtained from (5.3) and (5.13). (These can be used to generate sufficient conditions that guarantee (5.20), but these are likely much too conservative.) Thus, one merely needs to propagate the n-dimensional ODE (5.16) forward to time τ. A Runge–Kutta method may be used for this and the resulting approximate solution is denoted by $\widehat{X}^\tau_{j,i}$. The approximation of the elements of $\widehat{B}$ are then given by

$$\begin{aligned}\widehat{B}_{j,i} &= -\Big\{\psi^\tau_{j,i}(\widehat{X}^\tau(\tau)) - \widetilde{V}(\tau,\widehat{X}^\tau(\tau))\Big\}\\ &= -\Big\{\psi_j(\widehat{X}^\tau(\tau)) - \widetilde{V}(\tau,\widehat{X}^\tau(\tau))\Big\}.\end{aligned} \tag{5.21}$$

Note that

(5.22) the number of steps in the Runge–Kutta algorithm must be controlled so that (5.7) is satisfied.

6. Error summary. The error analyses of the previous three sections will now be combined. In particular, the errors due to truncation and the errors in computation of B will be combined to produce overall error bounds (6.8), (6.9). A condition required for the algorithm to work (assuming one uses the power series of section 5 for computation of B) is also obtained.

Theorems 3.7 to 3.10 provided sufficient conditions for the power method step to converge to the max-plus eigenvector. Employing the simplest condition (but also the strictest), that of Theorem 3.10, convergence of the power method with approximation $\widehat{B}$ to B is guaranteed if

$$|\widehat{B}_{i,j} - B_{i,j}| \le \min_{i\neq 1}\{|\overline{x}_i|^2\}\left(\frac{\delta c^5}{9(16)M^2}\right)\frac{\tau^4}{n^\mu} \quad \forall\, i,j, \tag{6.1}$$

where $\mu \in \{2,3,4,\dots\}$ and δ is given by (3.32). Note that we are assuming $\tau \le \min\{1, 1/c, \tau'\}$ as in section 5 (as well as all assumptions including (A5′) and technical conditions (3.1), (3.19) which appear in section 3). Then, Theorem 3.10 implies a resulting error bound for the max-plus eigenvector given by

$$\|e - \hat{e}\| \doteq \max_i |e_i - \hat{e}_i| \le \min_{i\neq 1}\{|\overline{x}_i|^2\}\left(\frac{\delta c^5}{9(16)M^2}\right)\frac{\tau^4}{n^{\mu-2}}, \tag{6.2}$$

where $\hat{e}$ corresponds to $\widehat{B}$. We remark that slightly different error estimates (under slightly different conditions) are also given in Theorems 3.8 and 3.9.

Suppose we adopt the notation $\widehat{W}(x) \doteq \bigoplus_{i=1}^n \hat{e}_i \otimes \psi_i(x)$ and $W^f(x) \doteq \bigoplus_{i=1}^n e_i \otimes \psi_i(x)$ so that W^f corresponds to the finite expansion with zero error in the computation/approximation of B. Then, by (6.2),

$$\begin{aligned}\|\widehat{W} - W^f\| \doteq \max_{|x|\le R} |\widehat{W}(x) - W^f(x)| &\le \min_{i\neq 1}\{|\overline{x}_i|^2\}\left(\frac{\delta c^5}{9(16)M^2}\right)\frac{\tau^4}{n^{\mu-2}} \\ &= \min_{i\neq 1}\{|\overline{x}_i|^2\}\left(\frac{\delta c^5\tau^4}{9(16)M^2}\right)\left(\frac{1}{N_D}\right)^{m(\mu-2)},\end{aligned} \tag{6.3}$$

where again, N_D is the number of centers of basis functions per dimension of the state space with a rectangular, evenly spaced grid of centers. It should be recalled that the basis functions are such that ψ_1 is centered at the origin ($\overline{x}_1 = 0$), and so N_D is odd. (Perhaps one should note that we are being sloppy here by using the number of basis functions corresponding to covering the entire rectangle which encloses the sphere $\overline{B}_{D_R}$, although only those with centers covering the sphere itself are required for the bound. Consequently, the above bound is conservative.) Also, with the evenly spaced basis function centers, (6.3) can be written as

$$\|\widehat{W} - W^f\| \le \left(\frac{D_R^2\delta c^5\tau^4}{9(4)M^2}\right)\left(\frac{1}{N_D}\right)^{m(\mu-2)}\left(\frac{1}{N_D - 1}\right)^2. \tag{6.4}$$

Using the approach of section 5, (6.1) is satisfied if

$$\tau^{n'-4} \le \widetilde{\mathcal{M}}_{R',n'}\left(\frac{1}{N_D}\right)^{m\mu+2}, \tag{6.5}$$

where $\widetilde{\mathcal{M}}_{R',n'}$ is given by (5.10) and n' is the number of terms (including zeroth order) in the Taylor series, *and* if (5.22) is satisfied.

This does not account for the truncation errors induced by using only a finite number of basis functions. Let W be the true value function (see section 2). Then, by (4.41),

$$|W(x) - W^f(x)| \leq K_1 \frac{|x|}{\tau}\left(\frac{2D_R}{N_D - 1}\right) \quad \forall\, x \in \overline{B}_R, \tag{6.6}$$

where K_1 is given by (4.42), $2D_R/(N_D - 1) = \Delta$, and τ satisfies (4.40); τ now must satisfy

$$\tau \leq \min\{1, 1/c, 1/\overline{\lambda}, \tau'\}. \tag{6.7}$$

The error bound (6.6) is not without drawbacks. In particular, τ appears in the denominator. However, it does not seem possible with the techniques of this paper to remove that term. This is the reason for concentrating in section 5 on fixed τ with increasing n' as the means for reducing errors.

Combining (6.4) and (6.6), the total error bound (assuming convergence of the power method—for which (6.5) and (5.22) form a sufficient condition—and $\tau \leq \min\{1, 1/c, 1/\overline{\lambda}, \tau'\}$) is given by

$$|W(x) - \widehat{W}(x)| \leq \left(\frac{D_R^2 \delta c^5 \tau^4}{9(4) M^2}\right)\left(\frac{1}{N_D}\right)^{m(\mu-2)}\left(\frac{1}{N_D - 1}\right)^2 + K_1 \frac{|x|}{\tau}\left(\frac{2D_R}{N_D - 1}\right),$$

which for $N_D \geq 3$

$$\leq \left(\frac{D_R^2 \delta c^5 \tau^4}{18 M^2}\right)\left(\frac{1}{N_D}\right)^{m(\mu-2)+2} + K_1 \frac{|x|}{\tau}\left(\frac{2D_R}{N_D}\right). \tag{6.8}$$

Since the best error rate in the last term in $1/N_D$, we take $\mu = 2$, and find in that case

$$|W(x) - \widehat{W}(x)| \leq \left[\frac{D_R^2 \delta c^5 \tau^4}{18 M^2} + 2K_1 D_R \frac{|x|}{\tau}\right]\left(\frac{1}{N_D}\right). \tag{6.9}$$

That is, the total error goes down linearly in $(1/N_D)$. Note that this rate is constrained by the fact that the solutions are only viscosity solutions—which may have discontinuous first derivatives. It is conjectured that with smooth solutions, the rate would instead be $(1/N_D)^2$.

This assumes that conditions (6.5) and (5.22) are met as well as (6.7). Also, as in section 5, one may prefer to write (6.5) as

$$n' \geq 4 + \frac{\log \widetilde{\mathcal{M}}_{R',n'} + (m\mu + 2)\log(1/N_D)}{\log \tau},$$

or assuming without loss of generality that $\widetilde{\mathcal{M}}_{R',n'} \geq 1$, one has the less tight but clearer bound of

$$n' \geq 4 + \frac{(m\mu + 2)\log(1/N_D)}{\log \tau},$$

in which case the lower bound on n' scales like $\log(1/N_D)$. From this, one sees for instance that doubling N_D would typically imply the addition of

$$\lceil\Big(\frac{(2m+2)\log(1/2)}{\log\tau}\Big) = \lceil\Big(\frac{(2m+2)\log 2}{\log(1/\tau)}\Big)$$

to n', where $\lceil(z)$ indicates the smallest integer greater than or equal to z. Again, this assumes the use of the Taylor series/Runge–Kutta approach of section 5 toward the approximation of B. Alternate approaches may yield different conditions.

Remark 6.1. All error bounds are actually conceived as the errors which may be achieved with given computer effort. A key underlying assumption of this paper is that all the elements of B are computed. This requires substantial effort since the number of terms in B is the square of the number of basis functions. In practice, it has been observed that elements of B, $B_{i,j}$, corresponding to basis function pairs where $|x_i - x_j|$ is large generally, do not contribute at all to the resulting eigenvector (recall that this is the max-plus algebra). By not computing these terms, one could greatly reduce the computations. This is a question for further study which lies beyond the bounds of the current paper.

Acknowledgments. The author would like to thank Professors Wendell H. Fleming and Matthew R. James for helpful discussions, both during the author's visit to Australian National University and subsequently. The author also thanks the referees for helpful comments.

REFERENCES

[1] T. Basar and P. Bernhard, H_∞*–Optimal Control and Related Minimax Design Problems*, Birkhäuser Boston, Boston, 1991.

[2] J. A. Ball and J. W. Helton, H_∞ *control for nonlinear plants: Connections with differential games*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 956–962.

[3] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, 1997.

[4] F. L. Baccelli, G. Cohen, G. J. Olsder, and J.-P. Quadrat, *Synchronization and Linearity*, John Wiley, Chichester, UK, 1992.

[5] R. K. Boel, M. R. James, and I. R. Petersen, *Robustness and risk sensitive filtering*, in Proceedings of the 36th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1997, pp. 2273–2278.

[6] G. Didinsky, T. Basar, and P. Bernhard, *Structural properties of minimax policies of a class of differential games arising in nonlinear* H_∞*–control and filtering*, in Proceedings of the 32nd IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1993, pp. 184–189.

[7] P. Dupuis and M. Boué, *Markov chain approximations for deterministic control problems with affine dynamics and quadratic cost in the control*, SIAM J. Numer. Anal., 36 (1999), pp. 667–695.

[8] P. Dupuis and A. Szpiro, *Convergence of the optimal feedback policies in a numerical method for a class of deterministic optimal control problems*, SIAM J. Control Optim., 40 (2001), pp. 393–420.

[9] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[10] W. H. Fleming, *Functions of Several Variables*, Springer-Verlag, New York, 1977.

[11] W. H. Fleming and W. M. McEneaney, *Robust limits of risk sensitive nonlinear filters*, Math. Control Signals Systems, 14 (2001), pp. 109–142.

[12] W. H. Fleming and W. M. McEneaney, *A max-plus-based algorithm for a Hamilton–Jacobi–Bellman equation of nonlinear filtering*, SIAM J. Control Optim., 38 (2000), pp. 683–710.

[13] W. H. Fleming, *Deterministic nonlinear filtering*, Ann. Scuola Norm. Sup. Pisa, Cl. Sci. 4, 25 (1997), pp. 435–454.

[14] W. H. Fleming and W. M. McEneaney, *Risk sensitive and robust nonlinear filtering*, in Proceedings of the 36th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1997, pp. 1088–1093.
[15] W. H. Fleming and H. M. Soner, *Controlled Markov Process and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[16] E. Gallestey, M. R. James, and W. M. McEneaney, *Max–plus approximation methods in partially observed* H_∞ *control*, in Proceedings of the 38th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, pp. 3011–3016.
[17] M. Hardt, J. W. Helton and K. Kreutz-Delgado, *Numerical solution of nonlinear* H_2 *and* H_∞ *control problems with application to jet engine compressors*, Control Systems Technology, (2000), pp. 98–111.
[18] J. W. Helton and M. R. James, *Extending* H_∞ *Control to Nonlinear Systems*, SIAM, Philadelphia, 1999.
[19] M. R. James, *A partial differential inequality for dissipative nonlinear systems*, Systems Control Lett., 21 (1993) pp. 315–320.
[20] V. N. Kolokoltsov and V. P. Maslov, *Idempotent Analysis and Its Application*, Math. Appl. 401, Kluwer, Dordrecht, The Netherlands, 1997.
[21] A. J. Krener, *Necessary and sufficient conditions for worst case (*H_∞*) control and estimation*, J. Math. Systems Estim. Control, 4 (1994), pp. 485–488.
[22] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York, 1992.
[23] G. L. Litvinov, V. P. Maslov, and G. B. Shpiz, *Idempotent functional analysis: An algebraic approach*, Math. Notes, 69 (2001) pp. 696–729.
[24] V. P. Maslov, *On a new principle of superposition for optimization problems*, Russian Math. Surveys, 42 (1987), pp. 43–54.
[25] W. M. McEneaney, *Max-plus eigenvector representations for solution of nonlinear* H_∞ *problems: Basic concepts*, IEEE Trans. Automat. Control, 48 (2003), pp. 1150–1163.
[26] W. M. McEneaney, *Error analysis of a max-plus algorithm for a first-order HJB equation*, in Stochastic Theory and Control, Lecture Notes in Control and Inform. Sci. 280, B. Pasik-Duncan, ed., Springer-Verlag, Berlin, 2002, pp. 335–351.
[27] W. M. McEneaney, *Convergence and error analysis for a max–plus algorithm*, in Proceedings of the 39th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 2000, pp. 1194–1199.
[28] W. M. McEneaney, *Robust/game-theoretic methods in filtering and estimation*, in Proceedings of the Symposium on Advances in Enterprise Control, 1999, pp. 3–10.
[29] M. Horton and W. M. McEneaney, *Computation of max–plus eigenvector representations for nonlinear* H_∞ *value functions*, in 1999 American Control Conference, pp. 1400–1404.
[30] W. M. McEneaney and M. Horton, *Max–plus eigenvector representations for nonlinear* H_∞ *value functions*, in Proceedings of the 37th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1998, pp. 3506–3511.
[31] W. M. McEneaney, *Robust/*H_∞ *filtering for nonlinear systems*, Systems Control Lett., 33 (1998), pp. 315–325.
[32] W. M. McEneaney, *A uniqueness result for the Isaacs equation corresponding to nonlinear* H_∞ *control*, Math. Control Signals Systems, 11 (1998), pp. 303–334.
[33] W. M. McEneaney, *Elimination of troublesome disturbances with application to representation results for* H_∞ *control DPEs*, in Proceedings of the Seventh International Symposium on Dynamic Games and Applications, Shonan Village, Japan, 1996, pp. 662–671.
[34] R. T. Rockafellar, *Conjugate Duality and Optimization*, Regional Conference Series in Applied Mathematics 16, SIAM, Philadelphia, 1974.
[35] R. T. Rockafellar and R. J. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[36] P. Soravia, H_∞ *control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.
[37] A. J. van der Schaft, L_2*–gain analysis of nonlinear systems and nonlinear state feedback* H_∞ *control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 413–436

STABILIZABILITY OF STOCHASTIC LINEAR SYSTEMS WITH FINITE FEEDBACK DATA RATES*

GIRISH N. NAIR† AND ROBIN J. EVANS†

Abstract. Feedback control with limited data rates is an emerging area which incorporates ideas from both control and information theory. A fundamental question it poses is how low the closed-loop data rate can be made before a given dynamical system is impossible to stabilize by *any* coding and control law. Analogously to source coding, this defines the smallest error-free data rate sufficient to achieve "reliable" control, and explicit expressions for it have been derived for linear time-invariant systems without disturbances. In this paper, the more general case of finite-dimensional linear systems with process and observation noise is considered, the object being mean square state stability. By inductive arguments employing the *entropy power inequality* of information theory, and a new quantizer error bound, an explicit expression for the infimum stabilizing data rate is derived, under very mild conditions on the initial state and noise probability distributions.

Key words. stochastic control, communication theory, source coding, entropy

AMS subject classifications. 93E15, 94A05, 94A17, 94A29

DOI. 10.1137/S0363012902402116

1. Introduction.

Communications and control have traditionally been areas with little common ground. For the most part communications theory is concerned with the reliable transmission of information from one point to another, and is relatively indifferent to its specific purpose or whether it is eventually fed back to the source. Control theory, in contrast, is concerned mainly with using information in a feedback loop to achieve some performance objective, and usually assumes that limitations in the communications links are not significant enough to affect performance drastically.

The reasons usually given for this mutual indifference are, first, that a communications system is generally used for a broad range of purposes and can rarely be designed to match a particular objective and, secondly, that to explicitly model communication limitations would complicate controller synthesis. However, in recent years emerging applications such as microelectromechanical systems, mobile telephone power control, and networked industrial control systems have begun to cross the boundary between these disciplines. In these applications, the aim is to control a dynamical system consisting of many separate components connected by a digital communication network. Although the total available capacity in bits per second may be large, each component is effectively allocated only a small portion. This can introduce significant quantization errors and delays, due to the low resolution and finite transmission time of each discrete-valued, digital *symbol*. Quantization resolution can be improved at the expense of delay and vice-versa, but nonetheless there remains an upper bound on the amount of *information*, in some sense, that may be exchanged per unit time. Clearly, by designing coders and decoders that are matched to the dynamical system

*Received by the editors February 5, 2002; accepted for publication (in revised form) January 24, 2004; published electronically July 2, 2004. Supported by Australian Research Council grants DP0345044 and DP0210197.

http://www.siam.org/journals/sicon/43-2/40211.html

†Department of Electrical and Electronic Engineering, University of Melbourne, VIC 3010, Australia (g.nair@ee.mu.oz.au, r.evans@ee.mu.oz.au). The first author was also supported by Melbourne Early Career Researcher and Melbourne Research grants.

and controllers, a more economical use of communication resources ought to be possible. Conversely, closed-loop performance should improve by matching the feedback laws to the specific coding and decoding schemes used.

The first step towards gaining a comprehensive understanding of these issues is to analyze the simplest possible network topology, consisting of one controller and dynamical system connected by a feedback loop with a given data rate in bits per unit time. In view of the limited communication resource, a natural question is: what is the smallest data rate above which there exists a coding and control law that stabilizes the system? This is analogous to Shannon's *source coding* theory, which seeks to determine the smallest data rate above which a given random process can be reliably communicated, i.e., with arbitrarily small error, by some coder and decoder [25, 7]. However, despite this analogy, Shannon's theory has generally not been fruitful in real-time control systems since its reliance on arbitrarily long *block* coders entails arbitrarily long delays. While this can be overcome by recursive coders such as *delta* and *differential pulse code* modulators, stationarity or ergodicity are still assumed [17]. Although this may be justified in communications, it does not always suit the unstable dynamics encountered in control.

In recent years, somewhat more progress on this topic has been made in the control literature. Beginning with the seminal paper [8] and continuing with [29, 2, 6, 10, 3, 23, 11, 16, 18], various schemes have been proposed and proven to asymptotically or *practically* stabilize linear time-invariant (LTI) systems at sufficiently high data rates. The first rigorous results on minimum data rates were in [29, 2], where it was shown that a discretized scalar plant with parameter a was stabilizable iff the data rate was not less than $\log_2 |a|$ bits per sampling interval. Similar tight bounds were subsequently obtained for noiseless autoregressive moving average [20] and linear state-space systems [27, 22, 14], using different formulations and techniques. With regard to stochastic plants, separation principles, causal *rate-distortion* theorems [28, 5], and the notion of feedback capacity [24] have been introduced.

This paper focuses on finite-dimensional, stochastic linear plants, under very mild assumptions on the noise and initial state probability distributions. In particular, the objective is to construct a coding and control scheme which achieves mean square state stability *while consuming as little data rate as possible*. The problem is formulated precisely in the next section, and the main result, which specifies the infimum stabilizing data rate, is stated. Somewhat counter-intuitively, the infimum rate depends only on the unstable dynamics of the plant and not on the noise statistics.

The remainder of the paper essentially constitutes the proof. As the presence of noise makes it difficult to extend the asymptotic quantization approach of [20, 22], a completely different method is developed here. In section 4, the well-known *entropy power inequality* of information theory [7, 9] is used to derive a strict lower bound on the data rate of any stabilizing, causal coding and control scheme, regardless of structure. It is shown that as the feedback data rate approaches this bound from above, the mean square state norms become arbitrarily large. In section 5, a specific, finite-dimensional scheme is then proposed. By applying a new, finite-level quantizer error inequality, it is proven to achieve mean square stability at any data rate exceeding the critical bound.

2. Formulation. First, certain conventions are defined. Vectors are written in bold-face type, matrices in bold-face upper-case, random variables in upper-case, and their realizations in corresponding lower-case letters. All random variables are assumed to exist on a common probability space with measure P. The probability

density of random vector $\mathbf{X}$ in Euclidean space with respect to (w.r.t.) Lebesgue measure λ on the space is denoted by $p_{\mathbf{X}}$, the probability density conditioned on the σ-field generated by an event $A = a$ by $p_{\mathbf{X}|a}$, expectation by E, and expectation conditioned on $A = a$ by E_a. The *(differential) entropy* of $\mathbf{X}$ is written $\mathrm{H}\{\mathbf{X}\} \triangleq -\mathrm{E}\{\ln p_{\mathbf{X}}(\mathbf{X})\}$; the *conditional entropy* of $\mathbf{X}$, given $A = a$, as $\mathrm{H}_a\{\mathbf{X}\} \triangleq -\mathrm{E}_a\{\ln p_{\mathbf{X}|A}(\mathbf{X})\}$; and the *average conditional entropy* of $\mathbf{X}$, given $A = a$, $B = b$, and averaged over B, as $\mathrm{H}_a\{\mathbf{X}|B\} \triangleq -\mathrm{E}_a\{\mathrm{H}_{A,B}\{\mathbf{X}\}\}$. Sequences $\{a_j\}_{j=0}^{k}$ are denoted $\tilde{a}_k$ (defined as the empty sequence when $k < 0$), and $\|\cdot\|$ represents either the Euclidean norm on a real vector space or the matrix norm induced by it. The $d \times d$ identity matrix is written $\mathbf{I}_d$, the $m \times n$ null matrix $\mathbf{0}_{m\times n}$, real numbers $\mathbb{R}$, positive reals $\mathbb{R}_+$, complex numbers $\mathbb{C}$, integers $\mathbb{Z}$, positive integers $\mathbb{Z}_+$, and nonnegative integers $\mathbb{W}$.

Consider the partially observed, discrete-time, stochastic linear system

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{v}_k, \quad \mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{w}_k \qquad \forall k \in \mathbb{W}, \tag{2.1}$$

with state $\mathbf{x}_k$ and process noise $\mathbf{v}_k \in \mathbb{R}^n$, control signal $\mathbf{u}_k \in \mathbb{R}^m$, and measurement $\mathbf{y}_k$ and measurement noise $\mathbf{w}_k \in \mathbb{R}^p$. It is assumed that the following hold:

- A1. $(\mathbf{A}, \mathbf{B})$ is reachable and $(\mathbf{C}, \mathbf{A})$, observable;
- A2. $\mathbf{x}_0$, $\mathbf{v}_k$, and $\mathbf{w}_k$ are realizations of random variables $\mathbf{X}_0$, $\mathbf{V}_k$, and $\mathbf{W}_k$, respectively, where $\mathbf{X}_0, \mathbf{V}_k, \mathbf{W}_j$ are mutually independent $\forall k, j \in \mathbb{W}$;
- A3. $\exists \varepsilon > 0$ s.t. $\mathbf{X}_0, \mathbf{V}_k, \mathbf{W}_k$ have uniformly bounded $(2+\varepsilon)$th absolute moments over $k \in \mathbb{W}$;
- A4. the probability distribution of each random variable $\mathbf{V}_k$ is absolutely continuous with respect to Lebesgue measure λ on $\mathbb{R}^n$;
- A5. $\inf_{k\in\mathbb{W}} \mathrm{H}\{\mathbf{V}_k^{\mathrm{u}}\} > -\infty$, where $\mathbf{V}_k^{\mathrm{u}} \in \mathbb{R}^{f\times n}$ is the process noise seen by the $f \geq 1$ unstable eigenvectors of $\mathbf{A}$; i.e., the process noise injects a minimum amount of uncertainty into the unstable dynamics.

Suppose that the sensor producing the measurements is connected to the controller via a digital channel, onto which one symbol s_k from a finite alphabet $\mathcal{S}_k$, of possibly time-varying size $\mu_k \geq 1$, is transmitted during the $(k+1)$th sampling interval. It is assumed that each transmitted symbol is received without error, as in Shannon source coding, after a delay of d intervals. The *(asymptotic average) data rate* of the channel may then be defined as

$$R \triangleq \liminf_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \log_2 \mu_k. \tag{2.2}$$

This is a more general definition than in [21, 22], in which the alphabet size μ_k is constant. In particular, it permits the alphabet $\mathcal{S}_k$ to vary periodically. For technical reasons, it is also assumed that $\mu_k/k \to 0$ as $k \to \infty$.

As the symbols in the channel are discrete-valued but the plant measurements are continuous-valued, analog-to-digital conversion, or coding, is required. In practice, constraints such as complexity and finite memory may be important but, in the spirit of source coding, such limitations will be largely ignored here in order to concentrate on the communication aspect of the problem. Each transmitted symbol may thus depend on all past and present measurements and past symbols,

$$s_k = \gamma_k(\tilde{\mathbf{y}}_k, \tilde{s}_{k-1}) \qquad \forall k \in \mathbb{W}, \tag{2.3}$$

where $\gamma_k : \mathbb{R}^{p\times(k+1)} \times \tilde{\mathcal{S}}_{k-1} \to \mathcal{S}_k$ is the coder mapping at time k. Note in particular that s_k does not necessarily correspond to a quantized version of the latest

measurement alone. At time k the controller has the symbols $s_0, \ldots, s_{k-d}$ available to it and can then generate a control signal of the general form

$$\mathbf{u}_k = \delta_k(\tilde{s}_{k-d}) \quad \forall k \in \mathbb{W}, \tag{2.4}$$

where $\delta_k : \tilde{\mathcal{S}}_{k-d} \to \mathbb{R}^m$ is the controller mapping at time k. As $\tilde{s}_k$ is the empty sequence $\{\}$ when $k < 0$, the first d control signals $\mathbf{u}_0, \ldots, \mathbf{u}_{d-1}$ are just preset inputs. Similarly, in the coder equation (2.3) at time $k = 0$, s_0 is a function only of $\mathbf{y}_0$.

Now, define the *coder-controller* as the triple of alphabet, coder, and controller mapping sequences $(\tilde{\mathcal{S}}_\infty, \tilde{\gamma}_\infty, \tilde{\delta}_\infty)$. The objective here is to construct a coder-controller which stabilizes the plant in the mean square sense,

$$\sup_{k \in \mathbb{W}} \mathrm{E}\|\mathbf{X}_k\|^2 < \infty, \tag{2.5}$$

while using as small a data rate as possible. The main result of this paper is now stated as follows.

THEOREM 2.1. *Given assumptions* A1–A5*, any coder-controller which stabilizes the plant* (2.1) *in the mean square sense* (2.5) *must have a data rate* R *(see* (2.2)*) strictly satisfying*

$$R > \sum_{|\eta_j| \geq 1} \log_2 |\eta_j| =: H, \tag{2.6}$$

where $\eta_1, \ldots, \eta_n$ *are the open-loop eigenvalues. As* R *approaches this bound from above, the supremum mean square state norm* (2.5) *approaches* ∞*.*

This inequality is also tight. In other words, for any number $R' > H$*, a mean-square-stabilizing coder-controller at data rate* $R \leq R'$ *can be constructed, which furthermore is finite-dimensional with periodic alphabet.*

This result assumes nothing about the coding and control laws except causality, and imposes only mild requirements on the noise distributions. It thus draws a fundamental line of demarcation between what is and is not achievable with stochastic linear systems when communication rates are limited. In this sense, H plays a role similar to source entropy in errorless Shannon source coding, and can be taken as a measure of the rate at which information is generated by an unstable, stochastic linear plant. Hence (2.6) states that, to achieve stability, the channel must transport data as fast as it is produced.

A more physical insight can be gained by rewriting the inequality above as $2^R > \prod_{|\eta_j| \geq 1} |\eta_j|$. The right-hand side (RHS) is simply the factor by which a volume in the unstable subspace increases at each time step due to the plant dynamics, while the left-hand side (LHS) is the asymptotic average number of disjoint regions into which the coder can partition the volume. In other words, the system is stabilizable iff the dynamical increase in "uncertainty volume" due to unstable dynamics is outweighed by the partitioning induced by the coder.

Note also that the data rate bound above is completely independent of the noise distributions and link delay, a consequence of the weak notion of stability used here. Increasing the noise variances (or delay) would obviously increase the mean square state norms, but as long as (2.6) is satisfied it remains possible to keep the state uniformly bounded in a mean square sense. The reason for this is that the noise increases state uncertainty volumes in an additive rather than multiplicative fashion but is averaged out exponentially, in effect, by the coder. However, as the data rate

approaches the critical limit, this exponential averaging becomes increasingly weak, leading to an unbounded increase in uncertainty volumes and hence mean square states.

Finally, it is remarked that in the problem formulation above there is no explicit communication constraint between the controller and actuator. This is reasonable if they are colocated, but even otherwise the formulation above is applicable, since from a plant output-to-input perspective the location of the controller is purely nominal. The symbols that would be transmitted by it over an additional digital link to the actuator would have to be converted once again into inputs, making intermediate calculations redundant. In other words the "bottle-neck" link determines the effective data rate, as expected, and Theorem 2.1 still applies. This is stated below more precisely.

PROPOSITION 2.2. *Suppose that two cascaded digital links connect the sensor to the actuator, with associated mappings*

$$s_k^1 = \gamma_k^1(\tilde{\mathbf{y}}_k, \tilde{s}_{k-1}^1) \in \mathcal{S}_k^1, \qquad \textit{(link-1 coder)} \tag{2.7}$$

$$s_k^2 = \gamma_k^2(\tilde{s}_{k-d_1}^1) \in \mathcal{S}_k^2, \qquad \textit{(link-2 coder)} \tag{2.8}$$

$$\mathbf{u}_k = \delta_k^2(\tilde{s}_{k-d_2}^2) \in \mathbb{R}^m. \qquad \textit{(actuator)} \tag{2.9}$$

Let R_1 be the data rate (2.2) *of the first link and R_2 that of the second. Then this coding and control scheme can be expressed as a single-link coder-controller of the form* (2.3)–(2.4)*, with delay $d = d_1 + d_2$ and data rate $R = \min\{R_1, R_2\}$.*

Conversely, any single-link coder-controller with periodic alphabet, data rate R, and delay d can always be expressed as a two-link coding and control scheme of the form above, with periodic alphabet sizes, link data rates both equal to R, and arbitrary delays $d_1, d_2 \in \mathbb{W}$ s.t. $d_1 + d_2 = d$.

Proof. See Appendix A for the proof.

The remainder of this paper is devoted to proving Theorem 2.1, in three stages. In the next section, the system dynamics are transformed into a simpler form. In section 4, the strict necessity of (2.6) is established via an inductive argument using the *entropy power inequality* of information theory [9]. Finally, the sufficiency of (2.6) is demonstrated in section 5 by constructing a coder-controller and using a new quantizer error result to recursively bound the mean square state norms.

3. Real Jordan form. Before proceeding, it is convenient to transform the system so as to decouple its dynamical modes. The obvious approach of putting the matrix $\mathbf{A}$ into *Jordan canonical form* generally requires a transformation matrix with complex elements. As this would complicate the analysis somewhat, the *real Jordan canonical form* [15] is used here.

Let $\lambda_1, \ldots, \lambda_b$ be the *distinct* eigenvalues of $\mathbf{A} \in \mathbb{R}^{n\times n}$, ordered by nonincreasing magnitude with conjugates excluded, and let the algebraic multiplicity of each λ_i be m_i. The real Jordan canonical form $\mathbf{J}$ then has the block diagonal structure

$$\mathbf{J} \equiv \operatorname{diag}(\mathbf{J}_1, \ldots, \mathbf{J}_b) \in \mathbb{R}^{n\times n}, \tag{3.1}$$

where the block $\mathbf{J}_i \in \mathbb{R}^{n_i \times n_i}$ with

$$n_i \triangleq \begin{cases} m_i & \text{if } \lambda_i \in \mathbb{R}, \\ 2m_i & \text{otherwise.} \end{cases} \tag{3.2}$$

More detail regarding the structure of each block can be found in, e.g., [15, pp. 150–153] or [26]. For the purposes of this paper, the most important fact is that

each block $\mathbf{J}_i$ is *similar* to the block diagonal matrix of all *standard* Jordan blocks corresponding to λ_i, λ_i^*. Hence $\mathbf{J}_i$ has either exactly one distinct eigenvalue λ_i or a pair of complex conjugate eigenvalues λ_i, λ_i^*, each with multiplicity m_i.

The real Jordan canonical form is related to the matrix $\mathbf{A}$ via a real similarity matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$ s.t. $\mathbf{T}^{-1}\mathbf{J}\mathbf{T} = \mathbf{A}$. Defining the transformed state

$$\acute{\mathbf{x}}_k \triangleq \mathbf{T}\mathbf{x}_k \quad \forall k \in \mathbb{W}, \tag{3.3}$$

the system equations (2.1) can then be written

$$\acute{\mathbf{x}}_{k+1} = \mathbf{J}\acute{\mathbf{x}}_k + \mathbf{T}\mathbf{B}\mathbf{u}_k + \mathbf{T}\mathbf{v}_k \in \mathbb{R}^n, \quad \mathbf{y}_k \triangleq \mathbf{C}\mathbf{T}^{-1}\acute{\mathbf{x}}_k + \mathbf{w}_k \in \mathbb{R}^p \qquad \forall k \in \mathbb{W}. \tag{3.4}$$

By partitioning the transformed state vector into the vectors $\acute{\mathbf{x}}_k^{(1)}, \ldots, \acute{\mathbf{x}}_k^{(b)}$ corresponding to each subsystem, the dynamical equation above can be rewritten more explicitly as

$$\acute{\mathbf{x}}_{k+1}^{(i)} = \mathbf{J}_i\acute{\mathbf{x}}_k^{(i)} + (\mathbf{T}\mathbf{B}\mathbf{u}_k)^{(i)} + (\mathbf{T}\mathbf{v}_k)^{(i)} \in \mathbb{R}^{d_i} \quad \forall k \in \mathbb{W},\ i \in [1, \ldots, b], \tag{3.5}$$

where $(\cdot)^{(i)}$ denotes that portion of the vector argument that feeds into the ith subsystem.

The original system has thus been decomposed into b real subsystems, with dynamics characterized by either a single eigenvalue or a pair of complex conjugate eigenvalues, possibly repeated. As $\mathbf{T}$ is invertible, it follows that the problems of stabilizing (3.4) and (2.1) are equivalent.

4. Proof of necessity. The first step towards proving Theorem 2.1 is to establish the necessity of (2.6) for mean square stability. In order to do so, a recursive lower bound for $\mathrm{E}\|\acute{\mathbf{X}}_k\|^2$ shall be sought, which is independent of the coder-controller and easier to analyze in terms of dynamics and data rate. More precisely, this bound will be sought for the state vector corresponding to subsystems with eigenvalue $|\lambda_i| \geq 1$.

If a strict inequality in (2.6) was not desired, a lower bound could quickly be obtained by observing that since the noise terms are independent, the mean square state norm cannot increase if they are all suppressed. In other words, the mean square state norm is bounded below by the mean square state norm of the plant with a random initial state but no noise. This is precisely the situation explored in [22] and, by a slight modification of the quantization argument used there, the nonstrict version of (2.6) is easily seen to be necessary for mean square stability.

However, this reduction to a noiseless system does not reveal that stability is in fact impossible at a data rate equal to the critical bound H. More importantly, it states nothing about behavior *near* H, in particular the fact that, as the data rate approaches H, the supremum mean square state norm (2.5) becomes arbitrarily large, drastically degrading performance, regardless of the coder-controllers used. This is made apparent by the entropy-based analysis below.

Denote the index set of unstable subsystems and their total dimension, respectively, by

$$\mathcal{U} \triangleq \{i : |\lambda_i| \geq 1\}, \quad f \triangleq \sum_{i\in\mathcal{U}} n_i, \tag{4.1}$$

and stack the unstable subsystem states $\acute{\mathbf{x}}_k^{(i)}$, $i \in \mathcal{U}$, to construct

$$\mathbf{x}_k^{\mathrm{u}} := \left[\acute{\mathbf{x}}_k^{(1)\mathrm{T}} \cdots \acute{\mathbf{x}}_k^{(|\mathcal{U}|)\mathrm{T}}\right]^{\mathrm{T}} \equiv \mathbf{R}\acute{\mathbf{x}}_k \in \mathbb{R}^f \qquad \forall k \in \mathbb{W},$$

$$\text{where } \mathbf{R} \triangleq \begin{bmatrix}\mathbf{I}_f & \mathbf{0}_{f\times(n-f)}\end{bmatrix} \in \mathbb{R}^{f\times n}. \tag{4.2}$$

Now, suppose that the coder-controller $(\tilde{\mathcal{S}}_\infty, \tilde{\gamma}_\infty, \tilde{\delta}_\infty)$ stabilizes (2.1), and hence (3.4), in mean square state norm. As $\|\mathbf{x}_k^{\mathrm{u}}\| \leq \|\hat{\mathbf{x}}_k\|$, it follows that $\{\mathbf{X}_k^{\mathrm{u}}\}_{k\in\mathbb{W}}$ is also bounded in mean square norm. Following the usual definition of *entropy power* (see, e.g., [9]), let the *conditional entropy power* of a random variable $\mathbf{X} \in \mathbb{R}^f$, given an event $A = a$, be

$$\mathrm{N}_a\{\mathbf{X}\} \triangleq (2\pi\mathrm{e})^{-1}\mathrm{e}^{2\mathrm{H}_a\{\mathbf{X}\}/f}. \tag{4.3}$$

In connection with the uncertainty volume interpretation of Theorem 2.1, $(\mathrm{N}_a\{\mathbf{X}\})^{f/2}$ can be regarded as the volume of the effective support set of $p_{\mathbf{X}|a}$. Furthermore,

$$\mathrm{N}_a\{\mathbf{X}\} \leq \mathrm{e}^{1/f-1}\mathrm{E}_a\|\mathbf{X}\|^2, \tag{4.4}$$

with equality iff $\mathbf{X}$ is symmetric Gaussian with zero mean when conditioned on $A = a$ (see Appendix B). This is essentially a statement of the well-known entropy-maximizing property of Gaussian distributions.

By analogy with the notation for average conditional entropy, denote the *average conditional entropy power* of $\mathbf{X}$, given $A = a$, averaged over A, by

$$\mathrm{N}\{\mathbf{X}|A\} \triangleq \mathrm{E}\left\{\mathrm{N}_A\{\mathbf{X}\}\right\}. \tag{4.5}$$

Setting $\mathbf{X} = \mathbf{X}_k^{\mathrm{u}}$ and $A = \tilde{S}_{k-d-1}$, the random variable associated with the sequence $\tilde{s}_{k-d-1}$, it then follows that

$$\begin{aligned} n_k &:= \mathrm{N}\{\mathbf{X}_k^{\mathrm{u}}|\tilde{S}_{k-d-1}\} = \mathrm{E}\big\{\mathrm{N}_{\tilde{S}_{k-d-1}}\{\mathbf{X}_k^{\mathrm{u}}\}\big\}, \\ &\leq \mathrm{e}^{1/f-1}\mathrm{E}\big\{\mathrm{E}_{\tilde{S}_{k-d-1}}\|\mathbf{X}_k^{\mathrm{u}}\|^2\big\} = \mathrm{e}^{1/f-1}\mathrm{E}\|\mathbf{X}_k^{\mathrm{u}}\|^2 \quad \forall k \in \mathbb{W}, \end{aligned} \tag{4.6}$$

so that $\{n_k\}_{k\in\mathbb{W}}$ must also be bounded. Note here that n_k can be interpreted as the average unstable subspace uncertainty volume, given the symbol sequence $\tilde{s}_{k-d-1}$.

Another important property of conditional entropy power is its superadditivity for summed independent random variables, i.e.,

$$\mathrm{N}_a\{\mathbf{X}+\mathbf{Y}\} \geq \mathrm{N}_a\{\mathbf{X}\} + \mathrm{N}_a\{\mathbf{Y}\}, \tag{4.7}$$

where $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^f$ are mutually independent when conditioned on an event $A = a$ (see, e.g., [7, 9]). By means of a recursive argument that employs this so-called *entropy power inequality*, it shall be shown that any stabilizing data rate must satisfy (2.6) strictly. First, observe from (4.2) that

$$\mathbf{RJ} = \mathbf{J}^{\mathrm{u}}\mathbf{R}, \quad \text{where } \mathbf{J}^{\mathrm{u}} \triangleq \mathrm{diag}(\mathbf{J}_1, \ldots, \mathbf{J}_{|\mathcal{U}|}) \in \mathbb{R}^{f\times f}. \tag{4.8}$$

Left-multiplying (3.4) by $\mathbf{R}$ and using (2.4), the dynamical equation for $\mathbf{x}_k^{\mathrm{u}}$ is then

$$\mathbf{x}_{k+1}^{\mathrm{u}} = \mathbf{J}^{\mathrm{u}}\mathbf{x}_k^{\mathrm{u}} + \mathbf{RTv}_k + \mathbf{RTB}\delta_k(\tilde{s}_{k-d}), \tag{4.9}$$

$$\begin{aligned} \Rightarrow \mathrm{N}_{\tilde{S}_{k-d}}\{\mathbf{X}_{k+1}^{\mathrm{u}}\} &= \mathrm{N}_{\tilde{S}_{k-d}}\{\mathbf{J}^{\mathrm{u}}\mathbf{X}_k^{\mathrm{u}} + \mathbf{RTV}_k + \mathbf{RTB}\delta_k(\tilde{S}_{k-d})\} \\ &= \mathrm{N}_{\tilde{S}_{k-d}}\{\mathbf{J}^{\mathrm{u}}\mathbf{X}_k^{\mathrm{u}} + \mathbf{RTV}_k\} \qquad (4.10) \\ &\geq \mathrm{N}_{\tilde{S}_{k-d}}\{\mathbf{J}^{\mathrm{u}}\mathbf{X}_k^{\mathrm{u}}\} + \mathrm{N}_{\tilde{S}_{k-d}}\{\mathbf{RTV}_k\} = \mathrm{N}_{\tilde{S}_{k-d}}\{\mathbf{J}^{\mathrm{u}}\mathbf{X}_k^{\mathrm{u}}\} + \mathrm{N}\{\mathbf{RTV}_k\} \qquad (4.11) \\ &= |\det\mathbf{J}^{\mathrm{u}}|^{2/f}\mathrm{N}_{\tilde{S}_{k-d}}\{\mathbf{X}_k^{\mathrm{u}}\} + \mathrm{N}\{\mathbf{RTV}_k\} \quad \forall k \in \mathbb{W}. \qquad (4.12) \end{aligned}$$

The equality in (4.10) is due to the property that $\mathrm{H}_a\{\mathbf{X}+g(A)\} = \mathrm{H}_a\{\mathbf{X}\}$ for any function g (translation invariance). The inequality (4.11) uses the mutual independence of $\mathbf{V}_k$ and $\mathbf{X}_k, \tilde{S}_{k-d}$, the latter both being determined by $\mathbf{X}_0, \mathbf{V}_j, \mathbf{W}_j$, $j \le k-1$, to apply the entropy power inequality (4.7), and finally (4.12) expresses the effect of an invertible linear transformation on differential entropy; see Corollary 9.6.4 in [7].

Now, denote the conditional entropy power of $\mathbf{X}$, given $A = a$, $S = s$ and averaged only over S, by $\mathrm{N}_a\{\mathbf{X}|S\} \triangleq \mathrm{E}_a\{\mathrm{N}_{A,S}\{\mathbf{X}\}\}$. The next step utilizes the result below.

LEMMA 4.1. *Let $\mathbf{X} \in \mathbb{R}^f$ and $S \in \mathcal{S}$, a finite alphabet, be random variables conditioned on an event $A = a$. Then*

$$\mathrm{N}_a\{\mathbf{X}|S\} \ge |\mathcal{S}|^{-2/f}\mathrm{N}_a\{\mathbf{X}\}. \tag{4.13}$$

Proof. See Appendix C for the proof.

As $(\mathrm{N}_a\{\mathbf{X}|S\})^{f/2}$ can be viewed as the average uncertainty volume of $\mathbf{X}$, given S conditioned on $A = a$, this inequality states that knowledge of a correlated random variable $S \in \mathcal{S}$ with $|\mathcal{S}|$ distinct values reduces the average uncertainty volume of $\mathbf{X}$ by at most a factor of $|\mathcal{S}|$. In a sense, this is an extension of deterministic volume partitioning to a stochastic setting.

Setting $\mathbf{X} = \mathbf{X}^{\mathrm{u}}_k$, $A = \tilde{S}_{k-d-1}$, $S = S_{k-d}$, and averaging (4.12),

$$\begin{aligned}
n_{k+1} &\equiv \mathrm{E}\left\{\mathrm{N}_{\tilde{S}_{k-d}}\{\mathbf{X}^{\mathrm{u}}_{k+1}\}\right\} \\
&\ge |\det\mathbf{J}^{\mathrm{u}}|^{2/f}\mathrm{E}\left\{\mathrm{E}_{\tilde{S}_{k-d-1}}\left\{\mathrm{N}_{\tilde{S}_{k-d}}\{\mathbf{X}^{\mathrm{u}}_k\}\right\}\right\} + \mathrm{E}\left\{\mathrm{N}\{\mathbf{RTV}_k\}\right\} \\
&= |\det\mathbf{J}^{\mathrm{u}}|^{2/f}\mathrm{E}\left\{\mathrm{N}_{\tilde{S}_{k-d-1}}\{\mathbf{X}^{\mathrm{u}}_k|S_{k-d}\}\right\} + \mathrm{N}\{\mathbf{RTV}_k\} \\
&\ge \left|\frac{\det\mathbf{J}^{\mathrm{u}}}{|\mathcal{S}_{k-d}|}\right|^{2/f}\mathrm{E}\left\{\mathrm{N}_{\tilde{S}_{k-d-1}}\{\mathbf{X}^{\mathrm{u}}_k\}\right\} + \mathrm{N}\{\mathbf{RTV}_k\} \\
&\equiv \left|\frac{\det\mathbf{J}^{\mathrm{u}}}{\mu_{k-d}}\right|^{2/f} n_k + \mathrm{N}\{\mathbf{RTV}_k\} \quad \forall k \in \mathbb{W}.
\end{aligned} \tag{4.14}$$

By assumption A5 in section 2, $\exists\theta \in \mathbb{R}$ s.t. $\mathrm{H}\{\mathbf{RTV}_j\} \ge \theta\ \forall j \in \mathbb{W}$. Hence

$$\mathrm{N}\{\mathbf{RTV}_j\} \ge (2\pi)^{-1}\mathrm{e}^{2\theta/f-1} =: \beta > 0 \quad \forall j \in \mathbb{W}, \tag{4.15}$$

by definition (4.3). Substituting this into (4.14) and for convenience setting $\mu_i = 1$ when $i < 0$,

$$\begin{aligned}
n_{k+1} &\ge \left|\frac{\det\mathbf{J}^{\mathrm{u}}}{\mu_{k-d}}\right|^{2/f} n_k + \beta \ge \beta\sum_{j=0}^{k}\prod_{i=j+1}^{k}\left|\frac{\det\mathbf{J}^{\mathrm{u}}}{\mu_{i-d}}\right|^{2/f} \\
&\equiv \beta\sum_{j=0}^{k}\frac{h_k}{h_j}, \qquad \text{where } h_k := \prod_{i=0}^{k}\left|\frac{\det\mathbf{J}^{\mathrm{u}}}{\mu_{i-d}}\right|^{2/f} \\
\Rightarrow\ \infty > \alpha &\triangleq \frac{1}{\beta}\sup_{k\in\mathbb{Z}_+} n_k \ge \sum_{j=0}^{k}\frac{h_k}{h_j} \ \Leftrightarrow\ \frac{\alpha-1}{h_k} \ge \sum_{j=0}^{k-1}\frac{1}{h_j} \quad \forall k \in \mathbb{Z}_+.
\end{aligned} \tag{4.16}$$

As $h_k > 0$, $\alpha > 1$. By upward induction on k it can be verified that

$$
\begin{aligned}
& h_k \le \alpha(1-1/\alpha)^k h_0 \quad \forall k \in \mathbb{Z}_+ \\
\Rightarrow\ & 0 > \log_2(1-1/\alpha) \ge \frac{1}{k}\log_2\left(\frac{h_k}{\alpha}\right) \\
\Rightarrow \log_2\left(1-\frac{1}{\alpha}\right) &= \frac{2(k+1)}{fk}\log_2|\det \mathbf{J}^{\mathrm{u}}| - \frac{2}{fk}\sum_{i=0}^{k}\log_2\mu_{i-d} - \frac{\log_2\alpha}{k} \quad \forall k \in \mathbb{Z}_+ \\
& \ge \frac{2}{f}\log_2|\det \mathbf{J}^{\mathrm{u}}| - \liminf_{k\to\infty}\left[\frac{2}{fk}\sum_{i=0}^{k}\log_2\mu_{i-d} - \frac{2\log_2|\det \mathbf{J}^{\mathrm{u}}|}{fk} + \frac{\log_2\alpha}{k}\right] \\
& \equiv (2/f)\left(\log_2|\det \mathbf{J}^{\mathrm{u}}| - R\right) \equiv (2/f)(H-R). \qquad (4.17)
\end{aligned}
$$

This proves the strict necessity of (2.6). Furthermore, after rearranging the inequalities above and using (4.6),

$$
\sup_{k\in\mathbb{Z}_+} \mathrm{E}\|\mathbf{X}_k^{\mathrm{u}}\|^2 \ge \mathrm{e}^{1-1/f}\sup_{k\in\mathbb{Z}_+} n_k = \mathrm{e}^{1-1/f}\beta\alpha \ge \frac{\mathrm{e}^{1-1/f}\beta}{1-2^{-2(R-H)/f}}. \qquad (4.18)
$$

As β in (4.15) depends only on process noise statistics, this is a universal lower bound on the supremum mean square state of *all* coder-controllers with data rate R. Hence as $R \searrow H$, the supremum mean square state becomes arbitrarily large. □

Note that the argument above can be adapted to deal with stability in the sense of uniform boundedness with bounded disturbances. The idea is to replace the average conditional entropy power n_k with the maximum state uncertainty volume, given past symbols, and then use the deterministic analogue of the entropy power inequality, the *Brunn–Minkowski inequality* (see, e.g., [7]). This states that, given two λ-measurable regions $X, Y \subset \mathbb{R}^f$, the volume of the set sum $X+Y \triangleq \{x+y | x \in X, y \in Y\}$ satisfies $\lambda(X+Y)^{1/f} \ge \lambda(X)^{1/f} + \lambda(Y)^{1/f}$. It then follows that (2.6) is also strictly necessary for uniformly bounded stabilizability.

5. Achievability of data rate bound. The final step in proving Theorem 2.1 is to establish that (2.6) is attainable, i.e., that the system (2.1) can in principle be stabilized at any data rate arbitrarily close to but greater than the critical bound H. The general, entropy-based argument of the preceding section does not offer many clues as to how to prove this, so in this section a completely different approach is taken. Based on a semiheuristic line of reasoning, a finite-dimensional coder-controller with periodically varying coding alphabet is constructed in section 5.2. By means of a new quantizer bound, it is then demonstrated in section 5.3 that it achieves means square stability at data rates arbitrarily close to H.

The chief complications in the design and analysis of this scheme arise from the unbounded support of the noise terms. With uniformly bounded noise, any coding and control law which achieves asymptotic contraction *without* disturbances, in the sense that $\exists \gamma \in (0,1)$ s.t. $\forall$ sufficiently large k, $\|\mathbf{x}_k\| < \gamma r$, $\forall r > 0$, $\|\mathbf{x}_0\| \le r$, can easily be modified to achieve uniformly bounded stability. The idea is to recast such a law as an equivalent open-loop scheme which generates symbols according to the initial state alone. Assuming no noise, $\exists$ a sufficiently large $\tau \in \mathbb{Z}_+$ s.t. any state with norm $\le r$ will after τ steps have norm $\le \gamma r$. The effect of bounded disturbances then boosts the radius of this worst-case region by an additive constant $c \cdot \tau$, so the open-loop scheme can then be reapplied with $r^{\mathrm{new}} = \gamma r + c\tau$. As the recursion $r \mapsto \gamma r + c\tau$ is stable, a uniform bound on the state at times $0, \tau, 2\tau, \ldots$ is guaranteed and trivially leads to a bound over all integer times.

The situation is quite different when dealing with unbounded stochastic disturbances, because of the impossibility of a coder-controller which uniformly contracts mean square norms in an analogous sense. Briefly, the reason for this is that even though a distribution may have finite second absolute moment, the tail integral $\int_{\|\mathbf{x}\|\geq t} \|\mathbf{x}\|^2 \mathrm{d}P_{\mathbf{X}}(\mathbf{x})$ can approach 0 arbitrarily slowly with large t. In section 5.3 this difficulty is overcome by dealing with a functional M_ε (see (5.9)) instead of the mean square state norm. Before proceeding to the construction and analysis of the stabilizing coder-controller, several structural issues are first discussed below.

5.1. Structural issues. In order to make the analysis tractable, a certain amount of structure will need to be imposed on the general coder-controller equations (2.3)–(2.4). It is known (see, e.g., [28]) that for a linear Gauss–Markov system under a mean quadratic cost there exist optimal coding and control schemes with the following form:

1. Prior to coding, a Kalman filter is applied to recursively calculate the linear minimum variance prediction $\bar{\mathbf{x}}_{k+d|k}$ of $\acute{\mathbf{x}}_{k+d}$, given the measurement and control sequences $\tilde{\mathbf{y}}_k, \tilde{\mathbf{u}}_{k+d-1}$. Note that the control signals are not observed directly by the coder, but inferred from knowledge of the symbol sequence $\tilde{s}_{k-1}$ and the controller mappings.
2. Based on the past symbol sequence $\tilde{s}_{k-1}$, the latest prediction $\bar{\mathbf{x}}_{k+d}$ is recursively (and possibly nonuniformly) quantized to yield a coded estimate

$$\hat{\mathbf{x}}_{k+d} \equiv Q_k(\bar{\mathbf{x}}_{k+d|k}, \tilde{s}_{k-1}) \equiv \varpi_k(s_k)$$

   with μ_k possible values. The index s_k of the selected quantizer point $\varpi_k(s_k)$ is transmitted.
3. Upon receiving s_k at time $k+d$, the controller uses it and the previous symbols to regenerate $\hat{\mathbf{x}}_{k+d}$ and applies a certainty-equivalent linear control law $\mathbf{u}_{k+d} \equiv \mathbf{L}\hat{\mathbf{x}}_{k+d}$.

Although no Gaussian assumptions are made in this paper, it is convenient to use a modified version of this tristage structure as a basis for constructing a stabilizing scheme.

Considering the first stage, recall that the linear minimum variance predictions satisfy the separation principle

$$\mathrm{E}\|\acute{\mathbf{X}}_k\|^2 = \mathrm{E}\|\acute{\mathbf{X}}_k - \bar{\mathbf{X}}_{k|k-d}\|^2 + \mathrm{E}\|\bar{\mathbf{X}}_{k|k-d}\|^2 \quad \forall k \in \mathbb{W},$$

even with non-Gaussian noise. The first term on the RHS is uniformly bounded, by the observability and $(2+\varepsilon)$th moment assumptions A1 and A3 in section 2, and independent of the control law (see, e.g., [1]). Hence the mean square stability of the partially observed system (3.4) is equivalent to that of the fully observed filter process. Furthermore, this process satisfies a recursive equation of the same form as (3.4), i.e.,

$$\bar{\mathbf{x}}_{k+1+d|k+1} = \mathbf{J}\bar{\mathbf{x}}_{k+d|k} + \mathbf{TBu}_k + \mathbf{z}_{k+1} \quad \forall k \in \mathbb{W}, \tag{5.1}$$

where, by assumption A3, the $(2+\varepsilon)$th absolute moments of the *innovation* $\mathbf{Z}_{k+1}$ can be shown to be uniformly bounded over $k \in \mathbb{W}$.

The second stage above is not quite so straightforward, since the optimal quantizer $Q_k(\cdot,\cdot)$ is generally time-varying and stores all past symbols. As the objective here

is not optimality but stability, it is natural to investigate whether simpler quantizer structures will suffice. One possibility is a static, memoryless coder,

$$\hat{\mathbf{x}}_{k+d} \equiv Q(\bar{\mathbf{x}}_{k+d|k}) \equiv \varpi(s_k) \in \mathbb{R}^n, \tag{5.2}$$

where Q is a fixed quantizer with points $\varpi(0), \varpi(1), \ldots, \varpi(\mu-1)$. Another option is a *finite-state, predictive quantizer* (see, e.g., [12]), in which the latest coded state estimate is stored and the *prediction error* is recursively coded according to a finite-valued internal variable $\iota_k \in \mathcal{I}$,

$$Q\left(\bar{\mathbf{x}}_{k+d|k} - (\mathbf{J}+\mathbf{TBL})\hat{\mathbf{x}}_{k-1+d|k-1}, \iota_k\right) \equiv \varpi(s_k), \tag{5.3}$$

$$\begin{aligned} \hat{\mathbf{x}}_{k+d} &\equiv (\mathbf{J}+\mathbf{TBL})\hat{\mathbf{x}}_{k-1+d} + \varpi(s_k), \\ \iota_{k+1} &\equiv g(\iota_k, s_k), \end{aligned} \tag{5.4}$$

where $\mathbf{L}$ is a certainty-equivalent control gain such that $\mathbf{J}+\mathbf{TBL}$ is stable. The symbol s_k, corresponding to the index of the selected quantizer point, is then used to update the finite state, $(\iota_k, s_k) \mapsto \iota_{k+1}$. Examples are *differential pulse code* and *delta modulation* in speech processing.

For noise distributions with compact support, it can be shown that either type of coder can achieve stability. It may seem as if this should also hold in the case of infinite support, since if stability has been achieved, then the states and prediction errors remain with high probability in some bounded region, which could then be quantized without memory. However, this somewhat circular argument fails drastically if the plant is strictly unstable and either the initial state or a process noise term has infinite support in all directions.

PROPOSITION 5.1. *Suppose that the plant* (2.1) *has at least one open-loop eigenvalue with magnitude strictly greater than* 1*, and that, for any nonzero* $\mathbf{h} \in \mathbb{R}^n$*, either*

$$\mathrm{P}\{\mathbf{h}^{\mathrm{T}}\mathbf{X}_0 > \theta\} > 0 \quad \forall \theta \in \mathbb{R} \quad \textit{or} \quad \exists t \in \mathbb{W} \textit{ s.t. } \mathrm{P}\{\mathbf{h}^{\mathrm{T}}\mathbf{V}_t > \theta\} > 0 \quad \forall \theta \in \mathbb{R}. \tag{5.5}$$

Then for any static memoryless coder (5.2) *or finite-state predictive quantizer* (5.3)–(5.4)*, the* r*th absolute state moments are unbounded with time,* $\forall r > 0$*, regardless of the number of quantization points.*

Proof. See Appendix D for the proof.

This distinguishes the stochastic, communication-limited stabilization problem from the deterministic, bounded disturbance version, for which either memoryless or finite-state quantization suffices. The reason for the difference is basically that the finite range of the quantizer causes controller saturation. If the initial state or process noise has infinite support, there is consequently a finite chance that at some time k, the propagated state $\mathbf{A}\mathbf{x}_k$ is beyond reach of the control signal. The unstable plant dynamics then amplify this shortfall, causing the same phenomenon to occur with increasing probability at subsequent times, and inevitably leading to instability.

An obvious solution is to use an adaptive quantizer with possibly unbounded range, thereby allowing the control signal to "catch up" with the state. One simple approach is to use a predictive scheme with a scaling factor $l_k > 0$ which is recursively adjusted according to the symbols transmitted:

$$\begin{aligned} Q\left(\frac{\bar{\mathbf{x}}_{k+d|k} - (\mathbf{J}+\mathbf{TBL})\hat{\mathbf{x}}_{k-1+d|k-1}}{l_k}\right) &\equiv \varpi(s_k), \\ \hat{\mathbf{x}}_{k+d} &\equiv (\mathbf{J}+\mathbf{TBL})\hat{\mathbf{x}}_{k-1+d} + l_k\varpi(s_k), \\ l_{k+1} &\equiv g(l_k, s_k) \in \mathbb{R}_+. \end{aligned}$$

This approach, similar to [6, 19], is adopted in section 5.2.

Another characteristic of the coder-controller constructed here is that its symbol alphabet varies periodically with time, a point of departure from the time-invariant, constant data rate schemes in [3] and elsewhere. If the symbol alphabet were fixed, then the data rate in bits/interval could only take the discrete values $\log_2 \mu$, $\mu = 1, 2, \ldots$, making it impossible in general to attain data rates arbitrarily close to H. In [3] this is not an issue, since the underlying plant is in continuous time and the corresponding data rate bound, in bits/*second*, can be approached by increasing the sampling period. In the scenario considered in this paper, it is assumed that the sampling interval is not adjustable. However, by using a symbol alphabet which varies periodically, *average* data rates as close as desired to H can be achieved with sufficiently large cycle lengths, similar to the way that irrationals are approximated by rationals.

A periodic alphabet is also suggested by a consideration of the plant dynamics. By the block structure of the real Jordan form $\mathbf{J}$, the filter process (5.1) consists of b subsystems with decoupled dynamical matrices $\mathbf{J}_i \in \mathbb{R}^{n_i \times n_i}$, $i = 1, \ldots, b$. The speed at which the ith subsystem grows in any direction is determined by the eigenvalue λ_i and, intuitively, a larger $|\lambda_i|$ necessitates using a higher data rate. A natural approach is to cycle through the subsystems and encode each at a rate determined by the corresponding level of instability.

Notwithstanding the discussion above, the implementation of time-varying alphabets can be difficult. In the coder-controller presented below, the alphabet size is in fact kept constant for the initial part of each cycle, and no data is transmitted for the remainder. The subsystems are then allocated different effective data rates by means of a time-sharing protocol. More explicitly, time is divided into cycles of sufficiently large duration $\tau \in \mathbb{Z}_+$ and, within each cycle, the components of the unstable subsystem states $\bar{\mathbf{x}}_k^{(i)} \in \mathbb{R}^{n_i}$ are allocated transmission slots of fixed length τ_i, roughly proportional to $\log_2 |\lambda_i|$. During each slot a fixed alphabet of size $\mu_k \equiv \mu \geq 2$ is used to quantize the corresponding subsystem state component with a total of μ^{τ_i} levels. Towards the end of each cycle, there is then a quiet slot during which no information is transmitted, i.e., $\mu_k = 1$.

5.2. Stabilizing coder-controller. The coder-controller to be applied is defined below and analyzed in subsection 5.3. First, however, the static quantizer which is its basis is constructed, and a key lemma which stochastically bounds the quantizer errors is presented.

5.2.1. Quantizer. Recall that the floating point representation of a number $x \geq 1$ can be generated recursively by means of the following algorithm:

1. At iteration i, let $x \in J_i$, where $J_0 = [1, \infty)$.
2. At iteration $i+1$, if J_i was the semi-infinite interval $[10^i, \infty)$, then partition it into nine contiguous, disjoint subintervals of length 10^i and one semi-infinite interval $[10^{i+1}, \infty)$. If J_i was bounded, however, then partition it into ten equally long subintervals. In both cases set J_{i+1} to the subinterval containing x.
3. Repeat step 2 until $i =$ some predefined $\nu \geq 1$.
4. Approximate x by the lower limit of J_ν.

At termination, each interval $[10^{i-1}, 10^i)$, $i = 1, \ldots, \nu$, has been partitioned into $9 \times 10^{\nu - i}$ subintervals of equal length, and each x in it approximated as a floating point number with $\nu - i$ significant figures. Including the semi-infinite "overload" subinterval $[10^\nu, \infty)$, there are 10^ν subintervals in total, and so this algorithm can be

viewed as a nonuniform quantization of $x \geq 1$ with 10^ν points.

The scalar quantizer which underpins the coder-controller constructed here is basically an extension of this floating point scheme to $x \in \mathbb{R}$, with a base which may be nondecimal. First, select $\varrho > 1$ and let

$$r_i \stackrel{\Delta}{=} \varrho^{i-1}, \quad i \in \mathbb{Z}_+. \tag{5.6}$$

Then select a base $\mu \geq 2$ and for any integer $\nu \geq 2$ generate μ^ν disjoint, exhaustive intervals symmetrically about the origin by

(i) partitioning $[-r_1, r_1] = [-1, 1]$ into $(\mu^2 - 2)\mu^{\nu-2}$ intervals of length $2/[(\mu^2 - 2)\mu^{\nu-2}]$,

(ii) partitioning $(r_{i-1}, r_i]$ and $[-r_i, -r_{i-1})$ each into $(\mu - 1)\mu^{\nu-i}$ intervals of length $(\varrho^{i-1} - \varrho^{i-2})/[(\mu - 1)\mu^{\nu-i}]$ $\forall i \in [2, 3, \ldots, \nu]$,

(iii) leaving (r_ν, ∞) and $(-\infty, -r_\nu)$ as the right- and left-most intervals.

In general, there are precisely $(\mu^2 - 2)\mu^{\nu-2} + 2 + \sum_{2 \leq i \leq \nu} 2(\mu - 1)\mu^{\nu-i} = \mu^\nu$ intervals. Label them $I(0), I(1), \ldots, I(\mu^\nu - 1)$, from left- to right-most, and $\forall x \in \mathbb{R}$ let

$$\kappa_\nu(\omega) \stackrel{\Delta}{=} \begin{cases} \text{half-length of } I(\omega) & \text{if } 1 \leq \omega \leq \mu^\nu - 2, \\ 0.5(1 - 1/\mu)^{-1}(r_{\nu+1} - r_\nu) & \text{if } \omega = \mu^\nu - 1, \\ -0.5(1 - 1/\mu)^{-1}(r_{\nu+1} - r_\nu) & \text{if } \omega = 0, \end{cases} \tag{5.7}$$

$$q_\nu(x) := \varpi_\nu(\omega) := \left\{ \begin{array}{ll} \text{midpoint of } I(\omega) & \text{if } 1 \leq \omega \leq \mu^\nu - 2, \\ r_\nu + \kappa_\nu(\omega) & \text{if } \omega = \mu^\nu - 1, \\ -r_\nu - \kappa_\nu(\omega) & \text{if } \omega = 0, \end{array} \right\} \text{ if } x \in I(\omega). \tag{5.8}$$

The precise form of the equations above is immaterial, some of the constants being selected solely to simplify subsequent analysis. Observe that, like the floating point quantizer, the intervals of $q_{\nu+1}$ can be generated recursively by partitioning each interval of q_ν into μ subintervals. Furthermore, as $\nu \to \infty$, the range $[-r_\nu, r_\nu]$ covered by finite I's becomes unbounded, and at the same time any number x is eventually captured in an interval with length $\to 0$. Both these competing properties are necessary for the quantization error of a random variable with infinite support to approach zero pointwise as ν increases.

However, to show the attainability of (2.6), convergence in a stronger sense will be required; in particular, the mean square quantization error should diminish like the inverse square of the number of levels, $\mu^{-2\nu}$. Guaranteeing this for fat-tailed distributions is the real motivation for the exponential form of (5.6). If the distribution being quantized had exponentially decaying tails, then it can be shown that (5.6) leads to a waste of quantizer levels on regions of very low probability and, consequently, a shortage of levels in high-probability regions. However, as the tails may decay according to a power law, a sufficiently large ϱ ensures that both the low- and high-probability mean square error contributions die off like $\mu^{-2\nu}$.

In fact, a slightly stronger result can be proven. Before stating it formally, for any random variable $L \in \mathbb{R}_+$ and random vector $\mathbf{X}$ define the functional

$$\mathrm{M}_\varepsilon\{\mathbf{X}|L\} := \mathrm{E}\{L^2\} + \mathrm{E}\{\|\mathbf{X}\|^{2+\varepsilon} L^{-\varepsilon}\}. \tag{5.9}$$

This cannot be smaller than the mean square norm of $\mathbf{X}$, since

$$\begin{aligned} \mathrm{E}\|\mathbf{X}\|^2 &= \mathrm{E}\left\{\|\mathbf{X}\|^2 \left[\chi(\|\mathbf{X}\| \leq L) + \chi(\|\mathbf{X}\| > L)\right]\right\} \\ &\leq \mathrm{E}\{L^2\} + \mathrm{E}\left\{\|\mathbf{X}\|^2 (\|\mathbf{X}\|/L)^\varepsilon \chi(\|\mathbf{X}\| > L)\right\} \\ &\leq \mathrm{E}\{L^2\} + \mathrm{E}\left\{\|\mathbf{X}\|^{2+\varepsilon} L^{-\varepsilon}\right\} \equiv \mathrm{M}_\varepsilon\{\mathbf{X}|L\}, \end{aligned} \tag{5.10}$$

where χ is the usual indicator function. The mean square quantizer errors generated by q_ν can then be bounded as follows.

LEMMA 5.2. *Let $X \in \mathbb{R}$, $L > 0$ be random variables with $\mathrm{E}|X|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$. If the quantizer parameter ϱ of (5.6) and the base $\mu \in [2, 3, \ldots)$ are selected so that $\varrho > \mu^{2/\varepsilon}$, then the quantizer errors $X - Lq_\nu(X/L)$ satisfy*

$$\mathrm{M}_\varepsilon \left\{ X - Lq_\nu\left(\frac{X}{L}\right) \middle| L\kappa_\nu(\Omega) \right\} \leq \frac{\zeta}{\mu^{2\nu}} \mathrm{M}_\varepsilon\{X|L\} \quad \forall \nu \in [2, 3, \ldots), \tag{5.11}$$

where q_ν, κ_ν are defined in (5.8)–(5.7), $\Omega \in [0, 1, \ldots \mu^\nu - 1]$ is the index of the quantizer level $q_\nu(X/L)$, and $\zeta > 0$ is determined only by ε, μ, and ϱ.

Proof. See Appendix E for the proof.

By (5.10), the LHS exceeds the mean square quantizer error $\mathrm{E}|X - Lq_\nu(X/L)|^2$, so this result upper-bounds the latter and guarantees that it decreases as fast as the square of the number of quantizer levels. The condition $\varrho > \mu^{2/\varepsilon}$ is crucial here, as it relates the speed at which the quantizer range increases with ν to the fatness of the distribution tails, and thereby ensures that the contribution of the overload regions decays faster than $\mu^{-2\nu}$ for any fixed integer $\mu \geq 2$.

However, the real utility of this lemma lies in the appearance of M_ε on both sides, together with the independence of the constant ζ of the distribution of X. These two facts permit (5.11) to be used recursively when proving coder-controller stability in subsection 5.3. In contrast, a similar inequality relating the mean square norms of the error and X, via a density-independent factor decaying like the number of levels squared, is impossible.[1] What can be done instead is to upper-bound the mean square error by some higher moment of X as in Lemma 6.6 of [13], an approach which does not permit recursive application.

5.2.2. Time-sharing protocol and coder. The quantizer above will now be used to construct a coder-controller. The measurements are first passed through a Kalman filter to generate a fully observed process of the form (5.1). In order to simplify the analysis, and reduce subscript clutter, a nonpredictive filter with output $\bar{\mathbf{x}}_k := \bar{\mathbf{x}}_{k|k}$ is used. As discussed in section 5.1, its mean square stability is equivalent to that of the original system (2.1), and its innovations $\mathbf{z}_k$, $k \in \mathbb{Z}_+$, are uniformly bounded in the $(2+\varepsilon)$th absolute moment.

Divide times $k \in \mathbb{W}$ into *cycles* $[j\tau, \ldots, (j+1)\tau - 1]$, $j \in \mathbb{W}$, of uniform integer duration $\tau \in \mathbb{Z}_+$. Let R' be any given number greater than H from (2.6), and select any integer $\mu \geq 2^{R'}$. With $\mathcal{U}$ denoting the index set of unstable subsystems (4.1), subdivide each cycle into f *transmission slots* of duration τ_i, for each scalar component of $\bar{\mathbf{x}}_{j\tau}^{(i)} \in \mathbb{R}^{n_i}$, followed by a *quiet slot* of duration $\tau - \sum_{i\in\mathcal{U}} n_i\tau_i$, where

$$\tau_i \triangleq \lfloor \tau \log_\mu(\xi|\lambda_i|) \rfloor + 1 \quad \forall i \in \mathcal{U}, \tag{5.12}$$

with $\lfloor\cdot\rfloor$ denoting rounding down. If the parameter ξ is chosen to satisfy

$$0 < f \log_2 \xi < R' - \sum_{i\in\mathcal{U}} n_i \log_2 |\lambda_i| \equiv R' - \sum_{|\eta_l|\geq 1} \log_2 |\eta_l|, \tag{5.13}$$

[1] The reason for this is essentially that even if p_X has a finite second moment, $|x|^2 p_X(x)$ may decay so slowly with large x that the overload regions dominate the mean square error, making it decrease more slowly than the inverse square number of levels.

then the choice of transmission slot durations (5.12) is feasible, since

$$\sum_{i\in\mathcal{U}} n_i\tau_i \le \sum_{i\in\mathcal{U}} n_i\left(\tau\frac{\log_2(\xi|\lambda_i|)}{\log_2\mu}+1\right) \le \tau\frac{\sum_{i\in\mathcal{U}} n_i\log_2|\lambda_i| + f\log_2\xi}{R'} + f. \tag{5.14}$$

As the coefficient of τ is less than unity by (5.13), the sum of transmission slot durations is less than any sufficiently large cycle length τ.

Let the symbol alphabet $\mathcal{S}_k = \mathbb{Z}_\mu$ during transmission slots, and $= \{0\}$ during the quiet slot. Then by reasoning similar to the above, the asymptotic average data rate R of this periodic alphabet, equal to the average data rate over one cycle, satisfies

$$\begin{aligned} R &= \frac{1}{\tau}\sum_{i\in\mathcal{U}} n_i\tau_i\log_2\mu \le \frac{\log_2\mu}{\tau}\sum_{i\in\mathcal{U}} n_i\left(\tau\frac{\log_2(\xi|\lambda_i|)}{\log_2\mu}+1\right)\\ &= \sum_{i\in\mathcal{U}} n_i\log_2|\lambda_i| + f\log_2\xi + \frac{f\log_2\mu}{\tau} < R' \end{aligned}$$

for all sufficiently large τ. As R' is any number exceeding H, this confirms that the data rate of this protocol can be made arbitrarily close to H, leaving aside for now the question of stability.

Just before the start of a cycle at time $k = j\tau$, let $l_j \in \mathbb{R}_+$ be the adaptive quantizer scaling factor discussed in section 5.1, and $\hat{\mathbf{x}}_{j\tau} \in \mathbb{R}^n$ be an estimate of $\bar{\mathbf{x}}_{j\tau}$ internal to the coder. The *coder state* is then defined as $\psi_j \triangleq (\hat{\mathbf{x}}_{j\tau}, l_j)$. Indexing the scalar components of vectors $\in \mathbb{R}^{n_i}$ by an additional superscript $h \in [1,\dots,n_i]$, at the start of the transmission slot for $\bar{x}_{j\tau}^{(i,h)}$ let it be scaled and quantized via

$$\varpi_{\tau_i}(\omega_j^{(i,h)}) \equiv q_{\tau_i}([\bar{x}_{j\tau}^{(i,h)} - \hat{x}_{j\tau}^{(i,h)}]/l_j) \quad \forall h \in [1,\dots,n_i], \tag{5.15}$$

where ϖ_ν and q_ν are defined in (5.8). The index $\omega_j^{(i,h)} \in [0,\dots,\mu^\tau - 1]$ of the selected quantizer level is then expanded as τ_i base-μ digits and transmitted. After this has been done for all unstable state components, the coder state is updated via

$$\hat{\mathbf{x}}_{(j+1)\tau} = \mathbf{J}^\tau\left[\hat{\mathbf{x}}_{j\tau} + l_j\varpi(\omega_j)\right] + \sum_{k=j\tau}^{(j+1)\tau-1} \mathbf{J}^{(j+1)\tau-1-k}\mathbf{TBL}\hat{\mathbf{x}}_k, \tag{5.16}$$

$$\text{where } \hat{\mathbf{x}}_{k+1} = (\mathbf{J}+\mathbf{TBL})\hat{\mathbf{x}}_k \ \ \forall k \in [j\tau,\dots,j\tau+\tau-2], \ \ \hat{\mathbf{x}}_0 = \mathbf{0}, \tag{5.17}$$

$$l_{j+1} = \max_{i\in\mathcal{U}, h\in[1,\dots,n_i]}\left\{\sigma,\, l_j|\lambda_i|^\tau\kappa_{\tau_i}\left(\omega_j^{(i,h)}\right)\right\} \quad \forall j \in \mathbb{W}, \ \ l_0 = \sigma. \tag{5.18}$$

In the above, $\varpi(\omega_j) \in \mathbb{R}^n$ is the vector with (i,h)th component $\varpi_{\tau_i}(\omega_j^{(i,h)})$ for $i \in \mathcal{U}$ and 0 for $i \notin \mathcal{U}$, $\mathbf{L} \in \mathbb{R}^{p\times n}$ is the certainty-equivalent controller gain matrix, and $\sigma^{2+\varepsilon}$ is a uniform upper bound on the $(2+\varepsilon)$th absolute moment of

$$\mathbf{G}_j := \sum_{i=1}^{\tau} \mathbf{J}^{\tau-i}\mathbf{Z}_{j\tau+i} \ \ \forall j \in \mathbb{W}. \tag{5.19}$$

5.2.3. Controller. Similar to the coder, define a controller internal state $\psi_j^{\text{con}} \triangleq (\hat{\mathbf{x}}_{j\tau}^{\text{con}}, l_j^{\text{con}}) \in \mathbb{R}^n \times \mathbb{R}_+$, initialized when $j = 0$ to $(\mathbf{0}, \sigma)$. At any time $k \in [j\tau,\dots, j\tau+\tau-1]$ in the cycle, a certainty-equivalent control signal

$$\mathbf{u}_k = \mathbf{L}\hat{\mathbf{x}}_k^{\text{con}} \tag{5.20}$$

is applied, where $\hat{\mathbf{x}}_k^{\rm con}$ is given by (5.17) (with superscripts "con"), and $\mathbf{L}$ is the given gain matrix s.t. $\mathbf{J}+\mathbf{TBL}$ is strictly stable.

By the time-sharing protocol, the last transmission slot during this cycle ends at time $j\tau + c(\tau) - 1$, where $c(\tau) \triangleq \sum_{i\in\mathcal{U}} n_i\tau_i$. Recalling that the channel has a delay d, at time $k = j\tau + c(\tau) - 1 + d$ the controller then has available all the symbols $s_{j\tau}, \ldots, s_{j\tau+c-1}$, comprising the base-$\mu$ expansions of the quantizer level indices $\omega_j^{(i,h)}$, $h \in [1, \ldots, n_i]$, $i \in \mathcal{U}$. By reasoning similar to (5.14), $c(\tau) + d \leq \tau$ sufficiently large, so that these indices are guaranteed to be received before the beginning of the next cycle at time $(j+1)\tau$. The controller then updates its internal state via the same recursive equations (5.16)–(5.18) as the coder.

5.3. Analysis. A uniform bound on the mean square norms of the filter process (5.1) using the coder-controller above will now be derived, for a data rate arbitrarily close to the lower bound (2.6). First, it is shown that the coder error $\mathbf{F}_k \triangleq \bar{\mathbf{X}}_k - \hat{\mathbf{X}}_k$ is uniformly bounded in mean square norm over times $k = j\tau$, $j \in \mathbb{W}$, by using the functional M_ε defined in (5.9) and Lemma 5.2. The mean square stability of $\mathbf{F}_k$ over all integer times is then deduced, which in turn will be shown to imply that of $\bar{\mathbf{X}}_k$, $k \in \mathbb{W}$.

Observe that since the initial controller and coder internal states $\psi_0^{\rm con}, \psi_0$ are equal and, furthermore, the same update equations (5.16)–(5.18) are used for each, it follows that $\hat{\mathbf{x}}_k^{\rm con} = \hat{\mathbf{x}}_k$ and $l_j^{\rm con} = l_j$ $\forall j, k \in \mathbb{W}$. The superscript "con" is thus dropped in the analysis. Substituting (5.20) into the filter recursion (5.1) and iterating over a cycle,

$$\begin{aligned}\bar{\mathbf{x}}_{(j+1)\tau} &= \mathbf{J}^\tau \bar{\mathbf{x}}_{j\tau} + \sum_{k=j\tau}^{(j+1)\tau-1} \mathbf{J}^{(j+1)\tau-1-k}\left(\mathbf{TBL}\hat{\mathbf{x}}_k + \mathbf{z}_{k+1}\right)\\ &= \mathbf{J}^\tau \bar{\mathbf{x}}_{j\tau} + \mathbf{g}_j + \sum_{k=j\tau}^{(j+1)\tau-1} \mathbf{J}^{(j+1)\tau-1-k}\mathbf{TBL}\hat{\mathbf{x}}_k,\end{aligned}$$

where $\mathbf{g}_j \triangleq \sum_{k=1}^{\tau} \mathbf{J}^{\tau-k}\mathbf{z}_{j\tau+k}$ (see (5.19)). Subtracting this from (5.16), and then exploiting the block diagonal structure $\mathbf{J} \equiv \mathrm{diag}(\mathbf{J}_1, \ldots \mathbf{J}_b)$ of the real Jordan form, where $\mathbf{J}_i \in \mathbb{R}^{n_i\times n_i}$,

$$\begin{aligned}&\mathbf{f}_{(j+1)\tau} = \mathbf{J}^\tau[\mathbf{f}_{j\tau} - l_j\varpi(\omega_j)] + \mathbf{g}_j,\\ \Leftrightarrow\ &\mathbf{f}^{(i)}_{(j+1)\tau} = \mathbf{J}_i^\tau\left[\mathbf{f}^{(i)}_{j\tau} - l_j\varpi(\omega_j)^{(i)}\right] + \mathbf{g}_j^{(i)} \in \mathbb{R}^{n_i} \quad \forall i \in [1, \ldots, b], j \in \mathbb{W}.\end{aligned} \tag{5.21}$$

By (5.15) and the definition of $\varpi(\omega_j)$, $\varpi(\omega_j)^{(i)} \triangleq \mathbf{0}$ $\forall i \notin \mathcal{U}$, in which case the RHS above simply becomes the recursion $\mathbf{f}^{(i)}_{(j+1)\tau} = \mathbf{J}_i^\tau \mathbf{f}^{(i)}_{j\tau} + \mathbf{g}_j^{(i)}$. Recall that each block $\mathbf{J}_i$ has exactly either one real eigenvalue λ_i or two complex conjugate eigenvalues λ_i, λ_i^*. As $|\lambda_i| < 1$ $\forall i \notin \mathcal{U}$, and furthermore the noise term has a uniform moment bound $\mathrm{E}\|\mathbf{G}_j^{(i)}\|^2 \leq \sigma^2$, it immediately follows that $\mathrm{E}\|\mathbf{F}^{(i)}_{j\tau}\|^2$ must be uniformly bounded for all strictly stable subsystems.

Hence, only the unstable subsystems $i \in \mathcal{U}$ need be considered. For each such i, $\varpi(\omega_j)^{(i)} \in \mathbb{R}^{n_i}$ is defined to be the quantizer point vector with hth component $q_{\tau_i}(f^{(i,h)}_{j\tau}/l_j)$. Applying square norms and the triangle inequality to (5.21), $\forall i \in \mathcal{U}$, $j \in \mathbb{W}$,

$$\begin{aligned}\|\mathbf{f}^{(i)}_{(j+1)\tau}\|^2 &\le 2^2\left[\|\mathbf{J}_i^\tau\|^2\|\mathbf{f}^{(i)}_{j\tau} - l_j\varpi(\omega_j)^{(i)}\|^2 + \|\mathbf{g}_j^{(i)}\|^2\right]\\ &= 4\left[\|\mathbf{g}_j^{(i)}\|^2 + \|\mathbf{J}_i^\tau\|^2\sum_{h=1}^{n_i}\left|f^{(i,h)}_{j\tau} - l_j q_{\tau_i}\left(f^{(i,h)}_{j\tau}/l_j\right)\right|^2\right],\end{aligned} \tag{5.22}$$

using (5.15) and the definition of $\varpi(\omega_j)$. As each $\mathbf{J}_i$ is *similar* to the block diagonal matrix of all Jordan blocks associated with λ_i, a trivial adaptation of a result in [15, p. 138] states that $\exists\zeta_0 > 0$ s.t.

$$\|\mathbf{J}_i^\tau\| \le \zeta_0\tau^{n_i-1}|\lambda_i|^\tau \quad \forall i \in [1,\dots,b], \tau \in \mathbb{Z}_+. \tag{5.23}$$

Let the stacked vector of unstable subsystem errors be $\mathbf{f}^{\mathrm{u}}_k \triangleq [\mathbf{f}^{(1)\mathrm{T}}_k, \dots, \mathbf{f}^{(|\mathcal{U}|)\mathrm{T}}_k]^{\mathrm{T}}$, and define $\mathbf{g}^{\mathrm{u}}_j$ in a similar way. By summing (5.22) over $i \in \mathcal{U}$, applying the growth rate bound above, and twice using the trivial inequality $(\sum_{1\le l\le r}|y_l|)^\alpha \le r^\alpha\sum_{1\le l\le r}|y_l|^\alpha$ $\forall r \in \mathbb{Z}_+$, $\alpha > 0$, it then follows that $\exists\phi \ge 1$ s.t.

$$\begin{aligned}\|\mathbf{f}^{\mathrm{u}}_{(j+1)\tau}\|^{2+\varepsilon} &= \left(\sum_{i\in\mathcal{U}}\|\mathbf{f}^{(i)}_{(j+1)\tau}\|^2\right)^{1+\varepsilon/2}\\ &\le \phi\left(\|\mathbf{g}^{\mathrm{u}}_j\|^{2+\varepsilon} + \sum_{i\in\mathcal{U}}|\tau^{n_i-1}\lambda_i^\tau|^{2+\varepsilon}\sum_{h=1}^{n_i}\left|f^{(i,h)}_{j\tau} - l_j q_{\tau_i}\left(f^{(i,h)}_{j\tau}/l_j\right)\right|^{2+\varepsilon}\right).\end{aligned}$$

Dividing by l^ε_{j+1} and taking expectations,

$$\begin{aligned}&\mathrm{E}\left\{\|\mathbf{F}^{\mathrm{u}}_{(j+1)\tau}\|^{2+\varepsilon}L^{-\varepsilon}_{j+1}\right\}\\ &\le \phi\left(\mathrm{E}\left\{\frac{\|\mathbf{G}^{\mathrm{u}}_j\|^{2+\varepsilon}}{L^\varepsilon_{j+1}}\right\} + \sum_{i\in\mathcal{U}}\left|\tau^{n_i-1}\lambda_i^\tau\right|^{2+\varepsilon}\sum_{h=1}^{n_i}\mathrm{E}\left\{\frac{|F^{(i,h)}_{j\tau} - L_j q_{\tau_i}(F^{(i,h)}_{j\tau}/L_j)|^{2+\varepsilon}}{L^\varepsilon_{j+1}}\right\}\right)\\ &\le \phi\left(\mathrm{E}\left\{\frac{\|\mathbf{G}^{\mathrm{u}}_j\|^{2+\varepsilon}}{\sigma^\varepsilon}\right\} + \sum_{i\in\mathcal{U}}\left|\tau^{n_i-1}\lambda_i^\tau\right|^{2+\varepsilon}\sum_{h=1}^{n_i}\mathrm{E}\left\{\frac{|F^{(i,h)}_{j\tau} - L_j q_{\tau_i}(F^{(i,h)}_{j\tau}/L_j)|^{2+\varepsilon}}{[L_j|\lambda_i|^\tau\kappa_{\tau_i}(\Omega^{(i,h)}_j)]^\varepsilon}\right\}\right)\\ &= \phi\left(\mathrm{E}\left\{\frac{\|\mathbf{G}^{\mathrm{u}}_j\|^{2+\varepsilon}}{\sigma^\varepsilon}\right\} + \sum_{i\in\mathcal{U}}\tau^{(n_i-1)(2+\varepsilon)}|\lambda_i|^{2\tau}\sum_{h=1}^{n_i}\mathrm{E}\left\{\frac{|F^{(i,h)}_{j\tau} - L_j q_{\tau_i}(F^{(i,h)}_{j\tau}/L_j)|^{2+\varepsilon}}{[L_j\kappa_{\tau_i}(\Omega^{(i,h)}_j)]^\varepsilon}\right\}\right),\end{aligned} \tag{5.24}$$

where the second inequality is a consequence of the definition of l_{j+1} (see (5.18)).

Now, let

$$\vartheta_j := \mathrm{M}_\varepsilon\{\mathbf{F}^{\mathrm{u}}_{(j+1)\tau}|L_j\} \equiv \mathrm{E}\{L_j^2\} + \mathrm{E}\{\|\mathbf{F}^{\mathrm{u}}_{(j+1)\tau}\|^{2+\varepsilon}L_j^{-\varepsilon}\} \quad \forall j \in \mathbb{W}. \tag{5.25}$$

By (5.10), the RHS is never less than $\mathrm{E}\|\mathbf{F}^{\mathrm{u}}_{(j+1)\tau}\|^2$, so to establish the mean square boundedness of the errors it is sufficient to show that $\sup_{j\in\mathbb{W}}\vartheta_j < \infty$. Observe that

$$\begin{aligned}\mathrm{E}\{L^2_{j+1}\} &\equiv \mathrm{E}\left\{\max_{i\in\mathcal{U},h\in[1,\dots,n_i]}\left\{\sigma^2,\, L_j^2|\lambda_i|^{2\tau}\kappa_{\tau_i}\left(\Omega^{(i,h)}_j\right)^2\right\}\right\}\\ &\le \sigma^2 + \sum_{i\in\mathcal{U}}|\lambda_i|^{2\tau}\sum_{h=1}^{n_i}\mathrm{E}\left|L_j\kappa_{\tau_i}\left(\Omega^{(i,h)}_j\right)\right|^2.\end{aligned}$$

Adding this to (5.24), noting that $\mathrm{E}\|\mathbf{G}_j^{\mathrm{u}}\|^{2+\varepsilon} \leq \sigma^{2+\varepsilon}$, $\tau \geq 1$, and using definition (5.25),

$$\vartheta_{j+1} \leq \phi\left(\sigma^2 + \sum_{i\in\mathcal{U}} \tau^{(n_i-1)(2+\varepsilon)}|\lambda_i|^{2\tau} \sum_{h=1}^{n_i} \mathrm{M}_\varepsilon\left\{F_{j\tau}^{(i,h)} - L_j q_{\tau_i}\left(\frac{F_{j\tau}^{(i,h)}}{L_j}\right)\middle| L_j\kappa_{\tau_i}\left(\Omega_j^{(i,h)}\right)\right\}\right). \tag{5.26}$$

Applying Lemma 5.2 to each term in the inner sum, with $X = F_{j\tau}^{(i,h)}$, $L = L_j$, $\nu = \tau_i$, and $\Omega = \Omega_j^{(i,h)}$ $\forall j \in \mathbb{W}$,

$$\begin{aligned}\vartheta_{j+1} &\leq \phi\left(\sigma^2 + \sum_{i\in\mathcal{U}} \tau^{(n_i-1)(2+\varepsilon)}|\lambda_i|^{2\tau} \sum_{h=1}^{n_i} \frac{\zeta}{\mu^{2\tau_i}}\mathrm{M}_\varepsilon\left\{F_{j\tau}^{(i,h)}|L_j\right\}\right)\\ &\leq \phi\left(\sigma^2 + \sum_{i\in\mathcal{U}} \tau^{(n_i-1)(2+\varepsilon)}|\lambda_i|^{2\tau} \sum_{h=1}^{n_i} \frac{\zeta}{\mu^{2\tau_i}}\mathrm{M}_\varepsilon\left\{\mathbf{F}_{j\tau}^{\mathrm{u}}|L_j\right\}\right)\\ &\equiv \phi\sigma^2 + \phi\zeta\left(\sum_{i\in\mathcal{U}} n_i\tau^{(n_i-1)(2+\varepsilon)}\frac{|\lambda_i|^{2\tau}}{\mu^{2\tau_i}}\right)\vartheta_j, \end{aligned} \tag{5.27}$$

where the second inequality is obtained from the definition of M_ε in (5.9) and the trivial fact that the magnitude of a vector is never less than the magnitude of any of its components.

The inequality above is a first order, sublinear recursion for ϑ_j with a forcing term. By (5.12) and the fact that $x - \lfloor x\rfloor < 1$ $\forall x \in \mathbb{R}$, it follows that $\tau_i > \tau\log_\mu(\xi|\lambda_i|)$ $\forall i \in \mathcal{U}$, $\tau \in \mathbb{Z}_+$. This is equivalent to $\xi^\tau|\lambda_i|^\tau < \mu^{\tau_i}$, which when substituted into the above yields

$$\vartheta_{j+1} \leq \phi\sigma^2 + \phi\zeta\left(\sum_{i\in\mathcal{U}} n_i\tau^{(n_i-1)(2+\varepsilon)}\frac{1}{\xi^{2\tau}}\right)\vartheta_j \qquad \forall j \in \mathbb{W}.$$

As $\xi > 1$ by the left inequality of (5.13), $\tau^{(n_i-1)(2+\varepsilon)}\xi^{-2\tau} \to 0$ as $\tau \to \infty$ $\forall i$. Hence, by choosing a sufficiently large, finite cycle length τ, the coefficient of ϑ_j above can be made strictly less than 1. As the τ-dependent noise term σ^2 is finite for any fixed τ, the recursion above is then stable and yields uniformly bounded ϑ_j. By definition (5.25) and the inequality (5.10), $\mathrm{E}\|\mathbf{F}_{j\tau}^{\mathrm{u}}\|^2$ is then also uniformly bounded over $j \in \mathbb{W}$. Recalling the discussion after (5.21), the overall error vector $\mathbf{F}_{j\tau}$ must be as well.

The rest of the proof is straightforward. Subtracting (5.17) from (5.1), iterating forward r steps from time $j\tau$, and taking norms, at any time $k \equiv j\tau + r$ with $r \in [0, \ldots, \tau-1]$,

$$\|\mathbf{f}_k\| \leq \|\mathbf{J}^r\|\|\mathbf{f}_{j\tau}\| + \sum_{l=0}^{r-1}\|\mathbf{J}^{r-l}\mathbf{TBL}\|\|\mathbf{Z}_{j\tau+l+1}\|.$$

As $\mathbf{F}_{j\tau}$ and $\mathbf{Z}_{j\tau+l+1}$ are uniformly bounded in mean square norm and r can only take a finite number of values, the RHS and hence LHS are also uniformly mean square bounded over $k \in \mathbb{W}$. Rewriting (5.1) as

$$\bar{\mathbf{x}}_{k+1} = (\mathbf{J}+\mathbf{TBL})\bar{\mathbf{x}}_k + \mathbf{TBL}(\hat{\mathbf{x}}_k - \mathbf{x}_k) + \mathbf{z}_{k+1} = (\mathbf{J}+\mathbf{TBL})\bar{\mathbf{x}}_k - \mathbf{TBLf}_k + \mathbf{z}_{k+1} \ \ \forall k \in \mathbb{W},$$

the strict stability of $\mathbf{J} + \mathbf{TBL}$ then ensures the uniform boundedness of the mean square filter outputs $\mathrm{E}\|\bar{\mathbf{X}}_{k+1}\|^2$ over $k \in \mathbb{W}$. This completes the proof that the coder-controller constructed in subsection 5.2 stabilizes the system (2.1) at data rates arbitrarily close to, but exceeding, the critical bound (2.6). □

In the foregoing analysis, the assumption that the coder and controller internal states have the same initial value is crucial. Even if true, real digital channels invariably introduce data errors, causing the coder and controller states to eventually differ. It is thus important to emphasize that the scheme presented here is not intended in the present form to be a practical solution to communication-limited stabilization problems, but is primarily a theoretical construct for demonstrating stabilizability in the limited sense of (2.5). Nonetheless it does possess some attributes, such as finite dimensionality, which make implementation easy and may serve as a foundation for a more practical scheme. In this respect, an important and as yet open extension of this research is the *internal* stability of finite-dimensional, data-rate-limited control loops, i.e., ensuring that the plant and coder-controller internal states remain mean square stable with a random *overall* initial condition, channel errors, and process and measurement noise. It is easy to see that redundancy must be incorporated in the transmitted symbols to counteract channel noise, but it is not evident if an analogue of the well-known *source-channel separation theorem* of information theory [25] applies.

6. Conclusion. In this paper, the problem of stabilizing a general stochastic linear system in mean square state norm under a feedback data rate constraint was investigated. By employing information-theoretic techniques and a new quantizer error bound, an expression was derived for the smallest data rate above which such a system is stabilizable by a coding and control law, without imposing any structural or computational constraints and with very mild conditions on the system noise. This infimum rate is determined only by the unstable eigenvalues of the dynamical matrix, and it was demonstrated that, as the data rate approaches it from above, the mean square states become unbounded for any coder-controller. To establish the attainability of this bound, a finite-dimensional scheme was constructed and shown to achieve stability at data rates arbitrarily close to the bound. Extensions of these results to nonlinear systems, linear systems with Markov parameters, and decentralized control are being investigated.

Appendix A. Proof of Proposition 2.2. Suppose that $R_1 \leq R_2$. By direct substitution of (2.8) into (2.9), each input $\mathbf{u}_k$ depends (in a time-varying way) on the link-1 symbol sequence $\tilde{s}^1_{k-d_1-d_2}$,

$$\mathbf{u}_k \equiv \phi_k(\tilde{s}^1_{k-d_1-d_2}) \quad \forall k \in \mathbb{W}.$$

This mapping, (2.7), and the alphabet sequence $\tilde{\mathcal{S}}^1_\infty$ then constitute a coder-controller with data rate R_1 (see (2.2)) and link delay $d = d_1 + d_2$.

Now suppose that $R_2 < R_1$. By (2.7), the link-1 symbol sequence $\tilde{s}^1_{k-d_1}$ is also a time-varying function of the measurement sequence $\tilde{\mathbf{y}}_{k-d_1}$. Hence (2.8) can be rewritten in the form

$$s^2_k \equiv \theta_k(\tilde{\mathbf{y}}_{k-d_1}).$$

Defining the d_1-step-ahead link-2 symbol $c_k \triangleq s^2_{k+d_1}$, $\forall k \in \mathbb{W}$, the expression above and the actuator mapping (2.9) become

$$c_k = \theta_k(\tilde{\mathbf{y}}_k) \in \mathcal{S}^2_{k+d_1}, \quad \mathbf{u}_k = \delta^2_k(\tilde{c}_{k-d_1-d_2}).$$

This is a single-link coder-controller with delay $d = d_1 + d_2$. As the asymptotic average data rate is independent of constant time shifts of the alphabet, its value remains R_2.

The proof of the second part is straightforward. Let the coder (2.7) for link-1, with delay d_1, be given by (2.3), and set the coder and actuator for link-2, with delay $d_2 = d - d_1$, to be

$$s_k^2 = s_{k-d_1}^1 \equiv s_{k-d_1}, \quad \mathbf{u}_k = \delta_k(\tilde{s}_{k-d_2}^2) \equiv \delta_k(\tilde{s}_{k-d_1-d_2}).$$

Evidently, with regard to the plant this is equivalent to the single-link coder-controller (2.3)–(2.4). Furthermore, the link-2 alphabet is obviously periodic if that of link-1 is, with the same average data rate. □

Appendix B. Proof of inequality (4.4). The argument is essentially that of Lemma 5 in [9]. Denote the mean square norm of $\mathbf{X}$, given $A = a$, by σ^2, and let ϕ be the symmetric, f-dimensional Gaussian distribution with zero mean and variance σ^2. By the nonnegativity of the *Kullback–Leibler information distance* D,

$$\begin{aligned} 0 \le D(p_{\mathbf{X}|a} \| \phi) &\triangleq \int_{\mathbb{R}^f} p_{\mathbf{X}|a}(\mathbf{x}) \ln \frac{p_{\mathbf{X}|a}(\mathbf{x})}{\phi(\mathbf{x})} \mathrm{d}\lambda(\mathbf{x}) \\ &= \int_{\mathbb{R}^f} p_{\mathbf{X}|a}(\mathbf{x}) \ln p_{\mathbf{X}|a}(\mathbf{x}) \mathrm{d}\lambda(\mathbf{x}) - \int_{\mathbb{R}^f} p_{\mathbf{X}|a}(\mathbf{x}) \ln \phi(\mathbf{x}) \mathrm{d}\lambda(\mathbf{x}) \\ &= -\mathrm{H}_a\{\mathbf{X}\} - \int_{\mathbb{R}^f} p_{\mathbf{X}|a}(\mathbf{x}) \left(-\frac{f}{2} \ln(2\pi\sigma^2) - \frac{\|\mathbf{x}\|^2}{2\sigma^2} \right) \mathrm{d}\lambda(\mathbf{x}) \\ &= -\mathrm{H}_a\{\mathbf{X}\} + 0.5 f \ln(2\pi\sigma^2) + 0.5 \\ \Rightarrow \mathrm{N}_a\{\mathbf{X}\} &\equiv (2\pi\mathrm{e})^{-1} \mathrm{e}^{2\mathrm{H}_a\{\mathbf{X}\}/f} \le \mathrm{e}^{1/f-1} \sigma^2. \end{aligned}$$ □

Appendix C. Proof of Lemma 4.1. By standard properties of joint and average differential entropy [7, 9],

$$\mathrm{H}_a\{\mathbf{X}|S\} = \mathrm{H}_a\{\mathbf{X}, S\} - \mathrm{H}_a S \ge \mathrm{H}_a\{\mathbf{X}\} - \mathrm{H}_a S \ge \mathrm{H}_a\{\mathbf{X}\} - \ln|\mathcal{S}|.$$

Differential entropy is undefined for discrete-valued random variables, but the joint entropy above may be taken to denote $-\mathrm{E}_a \ln\big(p_{\mathbf{X}|S,A}(\mathbf{X})\mathrm{P}\{S|A\}\big)$, while $\mathrm{H}_a\{S\}$ represents the base-e discrete entropy of S, given $A = a$. The first inequality arises from the fact that the entropy of joint random variables can never be smaller than the individual entropies, while the second inequality is a consequence of the fact that the base-e entropy of a random variable in a finite alphabet $\mathcal{S}$ is at most $\ln|\mathcal{S}|$. Using Jensen's inequality for convex functions [7] and the lower bound above,

$$\begin{aligned} 2\pi\mathrm{e}\mathrm{N}_a\{\mathbf{X}|S\} &= \mathrm{E}_a\{\mathrm{e}^{2\mathrm{H}_{S,A}\{\mathbf{X}\}/f}\} \ge \mathrm{e}^{\mathrm{E}_a\{2\mathrm{H}_{S,A}\{\mathbf{X}\}/f\}} \\ &\equiv \mathrm{e}^{2\mathrm{H}_a\{\mathbf{X}|S\}/f} \ge \mathrm{e}^{2(\mathrm{H}_a\{\mathbf{X}\} - \ln|\mathcal{S}|)/f} = |\mathcal{S}|^{-2/f} 2\pi\mathrm{e}\mathrm{N}_a\mathbf{X}. \end{aligned}$$ □

Appendix D. Proof of Proposition 5.1. Consider the finite-state predictive quantizer (5.4). As there are a finite number $|\mathcal{I}|$ of possible internal variables ι_k, $Q(\cdot)$ is bounded. By the strict stability of $\mathbf{J} + \mathbf{TBL}$, it then follows from (5.4) that $\hat{\mathbf{x}}_k$ is bounded over k, and hence $\exists \rho > 0$ s.t. $\|\mathbf{u}_k\| = \|\mathbf{L}\hat{\mathbf{x}}_k\| \le \rho$.

Now, convert (2.1) into *standard* Jordan form via a complex similarity matrix $\mathbf{S}$. There is then at least one scalar component $x_k \in \mathbb{C}$ of the transformed state vector that satisfies the scalar, decoupled recursion

$$x_{k+1} = \eta x_k + v_k + u_k = \sum_{j=-1}^{k} \eta^{k-j}(v_j + u_j) \in \mathbb{C},$$

where $|\eta| > 1$ and v_k, u_k are the corresponding scalar components of $\mathbf{S}\mathbf{v}_k, \mathbf{S}\mathbf{u}_k$, respectively. For convenience, $u_{-1} \triangleq 0$ and $\mathbf{v}_{-1} \triangleq \mathbf{S}\mathbf{x}_0$. Evidently $V_{-1}, V_0, \ldots$ are still independent, and $|u_k| \le \rho$. Defining

$$\beta_k := \rho \sum_{j=0}^{k} |\eta|^{k-j} \ge \left| \sum_{j=0}^{k} \eta^{k-j} u_j \right|,$$

$$g_k := \sum_{j=-1}^{k} \eta^{k-j} v_j, \quad \bar{v}_k \triangleq \sum_{-1 \le j \le k, j \ne t} \eta^{-j} v_j \quad \forall k \in \mathbb{W},$$

where the time $t \ge -1$ is specified in (5.5), it follows that

$$\begin{aligned}
&\mathrm{P}\{\|\mathbf{S}\mathbf{X}_{k+1}\| \ge \beta_k\} \ge \mathrm{P}\{|X_{k+1}| \ge \beta_k\} = \mathrm{P}\left\{ \left| G_k - \sum_{j=0}^{k} \eta^{k-j} U_j \right| \ge \beta_k \right\} \\
&\ge \mathrm{P}\{|G_k| - \beta_k \ge \beta_k\} = \mathrm{P}\{|G_k| \ge 2\beta_k\} \\
&= \mathrm{P}\left\{ \left| \sum_{j=-1}^{k} \eta^{k-j} V_j \right| \ge 2\rho \sum_{j=0}^{k} |\eta|^{k-j} \right\} = \mathrm{P}\left\{ \left| \sum_{j=-1}^{k} \eta^{-j} V_j \right| \ge 2\rho \sum_{j=0}^{k} |\eta|^{-j} \right\} \\
&\ge \mathrm{P}\left\{ \left| \sum_{j=-1}^{k} \eta^{-j} V_j \right| \ge \theta \right\} \equiv \mathrm{P}\{|\eta^{-t} V_t + \bar{V}_k| \ge \theta\} \ge \mathrm{P}\left\{ \Re(\eta^{-t} V_t) + \Re(\bar{V}_k) \ge \theta \right\} \\
&\ge \mathrm{P}\left\{ \Re(\eta^{-t} V_t) \ge \alpha\theta,\ \Re(\bar{V}_k) \ge (1-\alpha)\theta \right\} \\
&= \mathrm{P}\left\{ \Re(\eta^{-t} V_t) \ge \alpha\theta \right\} \mathrm{P}\left\{ \Re(\bar{V}_k) \ge (1-\alpha)\theta \right\} \quad \forall \alpha \in \mathbb{R}, k \ge t,
\end{aligned} \tag{D.1}$$

where $\theta \triangleq 2\rho \sum_{j \in \mathbb{W}} |\eta|^{-j} = 2\rho/(1 - |\eta|^{-1})$ and the last step follows from the mutual independence of V_j, $j \ge -1$.

Furthermore, as $\mathrm{E}|V_j|^2$ is uniformly bounded, it follows from Holder's inequality that

$$\mathrm{E}|\Re(\bar{V}_k)|^2 \le \mathrm{E}\left| \sum_{j \ge -1, j \ne t} \eta^{-j} V_j \right|^2 \le \sum_{j \ge -1} |\eta|^{-j} \mathrm{E}\left\{ \sum_{j \ge -1} |\eta|^{-j} |V_j|^2 \right\} < \infty \quad \forall k \in \mathbb{W}.$$

By Theorem 22.6 in [4], $\Re(\bar{V}_k)$ then converges with probability 1, and thus in distribution, to a random variable $\bar{V}$. Hence $\exists \alpha_* \in \mathbb{R}$, $\epsilon > 0$, $k_* \in \mathbb{W}$ s.t. $\forall k \ge k_*$,

$$\mathrm{P}\left\{ \Re(\bar{V}_k) \ge (1-\alpha_*)\theta \right\} \ge \mathrm{P}\left\{ \bar{V} \ge (1-\alpha_*)\theta \right\} - \epsilon > 0.$$

In addition, since $\Re(\eta^{-t} V_t)$ is just a scalar linear function of $\mathbf{V}_t$, (5.5) implies that $\mathrm{P}\{\Re(\eta^{-t} V_t) > \vartheta\} > 0$ $\forall \vartheta \in \mathbb{R}$. Applying this and the inequality above to (D.1),

$$\mathrm{P}\{\|\mathbf{X}_{k+1}\| \ge \beta_k\} \ge \mathrm{P}\left\{ \Re(\eta^{-t} V_t) \ge \alpha_* \theta \right\} \left[\mathrm{P}\left\{ \bar{V} \ge (1-\alpha_*)\theta \right\} - \epsilon \right] \equiv \nu > 0 \quad \forall k \ge k_*.$$

It then follows that, $\forall r > 0$,

$$\mathrm{E}\|\mathbf{X}_{k+1}\|^r \ge \mathrm{E}\left\{ \|\mathbf{X}_{k+1}\|^r \chi(\|\mathbf{X}_{k+1}\| \ge \beta_k) \right\} \ge \beta_k^r \mathrm{P}\{\|\mathbf{X}_{k+1}\| \ge \beta_k\} \ge \beta_k^r \nu \to \infty,$$

since $\beta_k \to \infty$. The same reasoning applies to static memoryless coding (5.2). □

Appendix E. Proof of Lemma 5.2. Let $\phi = l\kappa_\nu(\omega)$, where $\omega \in \mathbb{Z}_{\mu^\nu}$ is the index of the selected quantizer point $q_\nu(x/l) \equiv \varpi_\nu(\omega)$ and κ_ν, ϖ_ν are defined by (5.8), (5.7), respectively. If $1 \le \omega \le \mu^\nu - 2$, then the interval $I(\omega)$ which contains x/l is bounded with length $2\kappa_\nu(\omega) \equiv 2\phi/l$ and midpoint $q_\nu(x/l) \triangleq \varpi_\nu(\omega)$. In this case, $|x - lq_\nu(x/l)| < \phi$, and $\forall \omega \in [1, \dots \mu^\nu - 2]$,

$$\mathrm{E}_{\omega,l}\left\{|X - Lq_\nu(X/L)|^{2+\varepsilon}\Phi^{-\varepsilon}\right\} \le \mathrm{E}_{\omega,l}\{\Phi^{2+\varepsilon}\Phi^{-\varepsilon}\} = \phi^2. \tag{E.1}$$

If $\omega = \mu^\nu - 1$, then x/l lies inside the semi-infinite interval $I(\mu^\nu - 1)$, defined as $(\varpi_\nu(\omega) - \phi/l, \infty)$. Hence

$$\begin{aligned}
&\mathrm{E}_{\omega,l}\left\{|X - Lq_\nu(X/L)|^{2+\varepsilon}\Phi^{-\varepsilon}\right\} \\
&= \mathrm{E}_{\omega,l}\left\{|X - L\varpi_\nu(\Omega)|^{2+\varepsilon}\Phi^{-\varepsilon}\chi(|X - L\varpi_\nu(\Omega)| \le \Phi)\right\} \\
&\quad + \mathrm{E}_{\omega,l}\left\{[X - L\varpi_\nu(\Omega)]^{2+\varepsilon}\,\Phi^{-\varepsilon}\chi(X - L\varpi_\nu(\Omega) > \Phi)\right\} \\
&\le \mathrm{E}_{\omega,l}\{\Phi^{2+\varepsilon}\Phi^{-\varepsilon}\} + \mathrm{E}_{\omega,l}\left\{X^{2+\varepsilon}\Phi^{-\varepsilon}\chi(X - L\varpi_\nu(\Omega) > \Phi)\right\} \\
&= \phi^2 + \mathrm{E}_{\omega,l}\left\{|X|^{2+\varepsilon}\,[L\kappa_\nu(\omega)]^{-\varepsilon}\,\chi\left(X - L\varpi_\nu(\omega) > L\kappa_\nu(\omega)\right)\right\} \\
&\le \phi^2 + \kappa_\nu(\omega)^{-\varepsilon}\mathrm{E}_{\omega,l}\{|X|^{2+\varepsilon}L^{-\varepsilon}\} \\
&= \phi^2 + \left[0.5(1 - 1/\mu)^{-1}(1 - \varrho^{-1})\varrho^\nu\right]^{-\varepsilon}\mathrm{E}_{\omega,l}\{|X|^{2+\varepsilon}L^{-\varepsilon}\} \\
&\le \phi^2 + \left[0.5(1 - 4^{-1/\varepsilon})\varrho^\nu\right]^{-\varepsilon}\mathrm{E}_{\omega,l}\{|X|^{2+\varepsilon}L^{-\varepsilon}\},
\end{aligned} \tag{E.2}$$

since $1 - \varrho^{-1} > 1 - \mu^{-2/\varepsilon} > 1 - 4^{-1/\varepsilon}$ for $\mu \ge 2$. By the symmetry of the quantizer about the origin, the same bound applies if $\omega = 0$, corresponding to the other semi-infinite interval $I(0)$. Averaging this, (E.1) and (E.2) over Ω, L,

$$\begin{aligned}
&\mathrm{E}\left\{\frac{|X - Lq_\nu(X/L)|^{2+\varepsilon}}{\Phi^\varepsilon}\right\} \\
&\le \mathrm{E}\{\Phi^2\} + \left[0.5(1 - 4^{-1/\varepsilon})\right]^{-\varepsilon}\varrho^{-\varepsilon\nu}\mathrm{E}\left\{\frac{|X|^{2+\varepsilon}}{L^\varepsilon}\right\} =: \beta_\nu.
\end{aligned} \tag{E.3}$$

By the definitions of $\phi \equiv l\kappa_\nu(\omega)$ from (5.7), and q_ν from (5.8),

$$\begin{aligned}
&\mathrm{E}_l\{\Phi^2\} \\
&= \left[\frac{l}{(\mu^2 - 2)\mu^{\nu-2}}\right]^2 \mathrm{P}_l\left\{\frac{|X|}{L} \le r_1\right\} + \sum_{i=2}^{\nu}\left[\frac{(r_i - r_{i-1})l}{2(\mu - 1)\mu^{\nu-i}}\right]^2 \mathrm{P}_l\left\{r_{i-1} < \frac{|X|}{L} \le r_i\right\} \\
&\quad + \left[\frac{(r_{\nu+1} - r_\nu)l}{2(1 - 2/\mu)}\right]^2 \mathrm{P}_l\left\{\frac{|X|}{L} > r_\nu\right\} \\
&= \left[\frac{l}{(\mu^2 - 2)\mu^{\nu-2}}\right]^2 \mathrm{P}_l\left\{\frac{|X|}{L} \le r_1\right\} \\
&\quad + \left[\frac{\varrho^{-1} - \varrho^{-2}}{2(\mu - 1)\mu^\nu}\right]^2 \left[\sum_{i=2}^{\nu}(\varrho\mu)^{2i}l^2\mathrm{P}_l\left\{r_{i-1} < \frac{|X|}{L} \le r_i\right\} + (\varrho\mu)^{2(\nu+1)}l^2\mathrm{P}_l\left\{\frac{|X|}{L} > r_\nu\right\}\right] \\
&\le \left[\frac{\mu^2 l}{(\mu^2 - 2)\mu^\nu}\right]^2 + \left[\frac{\varrho^{-1} - \varrho^{-2}}{2(\mu - 1)\mu^\nu}\right]^2 \sum_{i=2}^{\nu+1}(\varrho\mu)^{2i}l^2\mathrm{P}_l\left\{\frac{|X|}{L} \ge r_{i-1}\right\} \\
&\le \left[\frac{2l}{\mu^\nu}\right]^2 + \left[\frac{1}{2(\mu - 1)\varrho\mu^\nu}\right]^2 \sum_{i=2}^{\nu+1}(\varrho\mu)^{2i}l^2\mathrm{P}_l\left\{\frac{|X|}{L} \ge r_{i-1}\right\},
\end{aligned} \tag{E.4}$$

since $\mu^2/(\mu^2-2) \le 2$ for $\mu \ge 2$ and $\varrho > 1$. By a Chebyshev inequality type of argument,

$$\begin{aligned} l^2 \mathrm{P}_l\{|X| > r_{i-1}L\} &\le l^2 \mathrm{E}_l \left\{ [|X|/(r_{i-1}L)]^{2+\varepsilon}\, \chi(|X| > r_{i-1}L) \right\} \\ &= r_{i-1}^{-2-\varepsilon} \mathrm{E}_l \left\{ |X|^{2+\varepsilon} L^{-\varepsilon} \chi(|X| > r_{i-1}L) \right\} \\ &\le r_{i-1}^{-2-\varepsilon} \mathrm{E}_l \{ |X|^{2+\varepsilon} L^{-\varepsilon} \} = \varrho^{-(i-2)(2+\varepsilon)} \mathrm{E}_l \{ |X|^{2+\varepsilon} L^{-\varepsilon} \} \quad \forall i \ge 2. \end{aligned}$$

Substituting this into (E.4), averaging over L, and letting $b \triangleq \mathrm{E}\{|X|^{2+\varepsilon} L^{-\varepsilon}\}$,

$$\mathrm{E}\Phi^2 \le \frac{4\mathrm{E}L^2}{\mu^{2\nu}} + \left[\frac{1}{2(\mu-1)\varrho\mu^\nu} \right]^2 \sum_{i=2}^{\nu+1} \frac{(\varrho\mu)^{2i} b}{\varrho^{(i-2)(2+\varepsilon)}} = \frac{4\mathrm{E}L^2}{\mu^{2\nu}} + \left[\frac{\varrho^{1+\varepsilon}}{2(\mu-1)\mu^\nu} \right]^2 \sum_{i=2}^{\nu+1} \left(\frac{\mu^2}{\varrho^\varepsilon} \right)^i b.$$

As $\varrho > \mu^{2/\varepsilon}$, the geometric sum on the RHS is bounded with limit $\mu^4 \varrho^{-2\varepsilon}/(1-\mu^2\varrho^{-\varepsilon})$. Hence

$$\mathrm{E}\Phi^2 \le \frac{4\mathrm{E}L^2}{\mu^{2\nu}} + \left[\frac{\varrho^{1+\varepsilon}}{2(\mu-1)\mu^\nu} \right]^2 \frac{\mu^4 \varrho^{-2\varepsilon}}{1-\mu^2\varrho^{-\varepsilon}} b \equiv \frac{4\mathrm{E}L^m + \zeta_0 b}{\mu^{2\nu}}.$$

Adding this to (E.3),

$$\begin{aligned} \mathrm{M}_\varepsilon\{X - Lq_\nu(X/L)|\Phi\} &\le \mathrm{E}\Phi^2 + \beta_\nu \le 2\frac{4\mathrm{E}L^2 + \zeta_0 b}{\mu^{2\nu}} + \frac{[0.5(1-4^{-1/\varepsilon})]^{-\varepsilon} b}{\varrho^{\varepsilon\nu}} \\ &\le 2\frac{4\mathrm{E}L^2 + \zeta_0 b}{\mu^{2\nu}} + \frac{[0.5(1-4^{-1/\varepsilon})]^{-\varepsilon} b}{\mu^{2\nu}} \\ &\le \max\left\{ 8,\, 2\zeta_0 + [0.5(1-4^{-1/\varepsilon})]^{-\varepsilon} \right\} \frac{\mathrm{E}L^2 + b}{\mu^{2\nu}}. \qquad \square \end{aligned}$$

Note that virtually the same argument holds for the mean mth power quantization error, $m > 0$, by defining $\mathrm{M}_{\varepsilon,m}\{X|L\} \triangleq \mathrm{E}L^m + \mathrm{E}\{|X|^{m+\varepsilon} L^{-\varepsilon}\}$ and setting $\varrho > \mu^{m/\varepsilon}$.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice–Hall, Englewood Cliffs, NJ, 1979.

[2] J. Baillieul, *Feedback designs for controlling device arrays with communication channel bandwidth constraints*, in ARO Workshop on Smart Structures, Pennsylvania State University, 1999.

[3] J. Baillieul, *Feedback designs in information-based control*, in Stochastic Theory and Control, Proceedings of a Workshop held in Lawrence, Kansas, University of Kansas, 2001, B. Pasik-Duncan, ed., Springer, New York, 2001, pp. 35–57.

[4] P. Billingsley, *Probability and Measure*, Wiley, New York, 1995.

[5] V. S. Borkar and S. K. Mitter, *LQG control with communication constraints*, in Communications, Computation, Control and Signal Processing, Dordrecht, Boston, 1997, pp. 365–373.

[6] R. W. Brockett and D. Liberzon, *Quantized feedback stabilization of linear systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1279–1289.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[8] D. F. Delchamps, *Stabilizing a linear system with quantized state feedback*, IEEE Trans. Automat. Control, 35 (1990), pp. 916–924.

[9] A. Dembo, T. M. Cover, and J. A. Thomas, *Information theoretic inequalities*, IEEE Trans. Inform. Theory, 37 (1991), pp. 1501–1518.

[10] N. Elia and S. K. Mitter, *Stabilization of linear systems with limited information*, IEEE Trans. Automat. Control, 46 (2001), pp. 1384–1400.

[11] F. Fagnani and S. Zampieri, *Stability analysis and synthesis for scalar linear systems with a quantized feedback*, IEEE Trans. Automat. Control, 48 (2003), pp. 1569–1584.
[12] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Norwell, MA, 1993.
[13] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Springer, New York, 2000.
[14] J. Hespanha, A. Ortega, and L. Vasudevan, *Towards the control of linear systems with minimum bit-rate*, in Proceedings of the 15th International Symposium on Mathematical Theory of Networks and Systems, University of Notre Dame, 2002; available online at www.nd.edu/∼mtns/main.html.
[15] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[16] H. Ishii and B. A. Francis, *Quadratic stabilization of sampled-data systems with quantization*, Automatica, 39 (2003), pp. 1793–1800.
[17] J. C. Kieffer and J. G. Dunham, *On a type of stochastic stability for a class of encoding schemes*, IEEE Trans. Inform. Theory, 29 (1983), pp. 703–717.
[18] D. Liberzon, *On stabilization of linear systems with limited information*, IEEE Trans. Automat. Control, 48 (2003), pp. 304–307.
[19] G. Nair and R. Evans, *A finite-dimensional coder-estimator for rate-constrained state estimation*, in Proceedings of the 14th IFAC World Congress, Vol. I, Beijing, 1999, Elsevier, Amsterdam, pp. 19–24.
[20] G. Nair and R. Evans, *Stabilization with data-rate-limited feedback: Tightest attainable bounds*, Systems Control Lett., 41 (2000), pp. 49–56.
[21] G. Nair and R. Evans, *Mean square stabilisability of stochastic linear systems with data rate constraints*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, 2002, IEEE Press, Piscataway, NJ, pp. 1632–1637.
[22] G. Nair and R. Evans, *Exponential stabilisability of finite-dimensional linear systems with limited data rates*, Automatica, 39 (2003), pp. 585–593.
[23] I. R. Petersen and A. V. Savkin, *Multi-rate stabilization of multivariable discrete-time linear systems via a limited capacity communication channel*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, IEEE Press, Piscataway, NJ, pp. 304–309.
[24] A. Sahai, *Evaluating channels for control: Capacity reconsidered*, in Proceedings of the American Control Conference, Chicago, 2000, IEEE, pp. 2358–2362.
[25] C. E. Shannon, *A mathematical theory of communication*, Bell Syst. Tech. J., (1948); Reprinted in Claude Elwood Shannon Collected Papers, IEEE Press, Piscataway, NJ, 1993.
[26] G. E. Shilov, *Linear Algebra*, Prentice–Hall, Englewood Cliffs, NJ, 1971.
[27] S. Tatikonda and S. Mitter, *Control under communication constraints*, in Proceedings of the 38th Annual Allerton Conference on Communications, Control, and Computing, Monticello, IL, 2000, pp. 182–190.
[28] S. Tatikonda, A. Sahai, and S. K. Mitter, *Control of LQG systems under communication constraints*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 1165–1170.
[29] W. S. Wong and R. W. Brockett, *Systems with finite communication bandwidth constraints* II: *Stabilization with limited information feedback*, IEEE Trans. Automat. Control, 44 (1999), pp. 1049–1053.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 437–465

FINITE ELEMENT METHODS IN LOCAL ACTIVE CONTROL OF SOUND*

ALFREDO BERMÚDEZ†, PABLO GAMALLO†, AND RODOLFO RODRÍGUEZ‡

Abstract. The active control of sound is analyzed in the framework of the mathematical theory of optimal control. After setting the problem in the frequency domain, we deal with the state equation, which is a Helmholtz partial differential equation. We show the existence of a unique solution and analyze a finite element approximation when the source term is a Dirac delta measure. Two optimization problems are successively considered. The first one concerns the choice of phases and amplitudes of the actuators to minimize the noise at the sensors' location. The second one consists of determining the optimal actuators' placement. Both problems are then numerically solved. Error estimates are settled and numerical results for some tests are reported.

Key words. dissipative acoustics, noise reduction, active control, optimal control problem, finite element approximation

AMS subject classifications. 49J20, 49K20, 65N15

DOI. 10.1137/S0363012903431785

1. Introduction.

Noise reduction is an important problem in acoustical and environmental engineering. While passive methods are good for middle and high frequencies, they are not efficient for low ones. However, the latter can be significantly reduced by active control techniques. This is an old concept that has generated increasing interest in recent years due to the development of fast digital signal processors (DSP). It is based on the principle of destructive interference of waves: an opposite pressure is generated by a secondary source to cancel an undesired noise. In order to achieve a significant reduction, this source must produce, with great precision, an equal amplitude but inverted replica of the noise to be canceled. Applications of these techniques can be used, for instance, to reduce noise in aircrafts or cars.

Reference books on this subject are [3] and [11]. The general principles of active control of noise were described in an early patent by Leug in 1936. A microphone detects the undesired noise and provides an input signal to an electronic control system. The transfer from the microphone to the loudspeaker is adjusted so that the sound wave generated will destructively interfere with the noise to be canceled.

In this paper we state the problem of active control of noise in the framework of the optimal control theory of distributed systems and present its mathematical and numerical analysis. For the sake of simplicity we consider that the noise to be canceled has one single frequency, although it is also possible to control broad-band or even nonperiodic noises. Two problems are successively considered. In a first step, complex amplitudes are taken as control variables with the objective of minimizing the pressure at some particular points in the domain. In a second step, the loudspeakers' location is optimized with respect to the same objective function. A third step that is

*Received by the editors July 21, 2003; accepted for publication (in revised form) November 18, 2003; published electronically July 2, 2004.

http://www.siam.org/journals/sicon/43-2/43178.html

†Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain (mabermud@usc.es, pgamallo@usc.es). Partially funded by MCYT, Spain, under grant DPI2001-1613-C02-02 and Xunta de Galicia research project PGIDT02PXIC20701PN.

‡GI²MA, Departamento de Ingeniería Matemática, Universidad de Concepción, Casilla 160-C, Concepción, Chile (rodolfo@ing-mat.udec.cl). Partially funded by FONDAP in Applied Mathematics, Chile.

not included here would consist of determining the microphones' location in view of minimizing the global noise, i.e., the norm of the pressure in the whole domain under consideration rather than at some finite number of points.

The outline of the paper is as follows. In section 2 we introduce the physical problem and pose it in the framework of the optimal control theory. In section 3 we analyze the state equation. Although our main concern is when the inner source terms are Dirac delta measures, to tackle this problem we analyze first the same equation with data in L^2. We prove the existence and uniqueness of the solution and analyze its regularity, including some local $W^{2,\infty}$ a priori estimates which are used in the following section. In section 4 we introduce a finite element method to approximate the state equation. Once more we study first the case with L^2 data and then with Dirac delta measures. In both cases we prove L^2 and pointwise error estimates. In section 5 we state an optimal control problem to determine the optimal amplitudes of the actuators and show that it is well-posed. Then we approximate it by using the finite element approximation of the state equation introduced in the previous section. Next, we prove an error estimate for the approximate optimal control. In section 6 we report some numerical results which confirm our theoretical assertions. In section 7 we study how to determine the optimal location of the actuators, again in the framework of the optimal control theory. We prove the existence of an optimal control in this case and settle the optimality conditions. Finally, we report the results of some numerical experiments.

2. Mathematical model. The optimal control problem. Let $\Omega \subset \mathbb{R}^n$ ($n = 2$ or 3) be a bounded, convex, two-dimensional polygonal or three-dimensional polyhedral domain enclosing a nondissipative acoustic fluid (i.e., inviscid, compressible, and barotropic). The propagation of acoustic waves in this domain is modeled by the well-known equation

$$\frac{1}{c^2}\frac{\partial^2 P(x,t)}{\partial t^2} - \Delta P(x,t) = F(x,t) \quad \text{in } \Omega,$$

where P is the pressure fluctuation, c the sound speed, and F an inner source term. In our case, F will correspond to the secondary source of noise produced by loudspeakers, which will be the control variable. Moreover, there is a primary noise source acting on a part Γ_{N} of the boundary of the domain, $\partial\Omega$, which is modeled by

$$\frac{\partial P(x,t)}{\partial \boldsymbol{n}} = G(x,t) \quad \text{on } \Gamma_{\text{N}},$$

where $\boldsymbol{n}$ is an outward unit normal vector to $\partial\Omega$. This means that normal displacements are imposed on $\Gamma_{\text{N}} = \bigcup_{j=1}^{J} \Gamma_{\text{N}}^j$, where $\Gamma_{\text{N}}^1, \ldots, \Gamma_{\text{N}}^J$ denote the plane faces of Γ_{N}. In practice, it corresponds to the effect of an external vibration source transmitted to the enclosure Ω by the vibrations of some of the walls. Finally, we assume that the rest of the boundary $\Gamma_{\text{Z}} := \partial\Omega \setminus \bar{\Gamma}_{\text{N}} = \bigcup_{k=1}^{K} \Gamma_{\text{Z}}^k$ is formed by damping plane walls $\Gamma_{\text{Z}}^1, \ldots, \Gamma_{\text{Z}}^K$ characterized by a frequency-dependent *wall impedance* $Z(\omega)$. We assume that $|\Gamma_{\text{Z}}| > 0$ and $|\Gamma_{\text{N}}| > 0$, too.

In this paper we consider that G is a harmonic source with angular frequency $\omega \in \mathbb{R}$, $\omega \neq 0$. Hence, the secondary source F must be chosen also harmonic with the same frequency, i.e.,

$$G(x,t) = \text{Re}[g(x)\,\text{e}^{-i\omega t}], \qquad F(x,t) = \text{Re}[f(x)\,\text{e}^{-i\omega t}],$$

where $g(x)$ and $f(x)$ are complex functions which correspond to the respective complex amplitudes. Actually, their modulus are the physical amplitudes while their arguments are the phase angles. Since the model is linear, the stationary solution of the wave equation is also harmonic with the same frequency:

$$P(x,t) = \mathrm{Re}[p(x)\,\mathrm{e}^{-i\omega t}].$$

In such a case, we are led to the following Helmholtz problem, whose solution is the complex pressure amplitude:

$$\left\{\begin{array}{ll} -\Delta p - \left(\dfrac{\omega}{c}\right)^2 p = f & \text{in } \Omega, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = \dfrac{i\omega\rho}{Z(\omega)}\, p & \text{on } \Gamma_{\mathrm{Z}}, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = g & \text{on } \Gamma_{\mathrm{N}}, \end{array}\right. \tag{2.1}$$

where ρ is the density of the fluid and $Z(\omega) \in \mathbb{C}$ is the wall impedance given by

$$Z(\omega) := \beta(\omega) + \frac{\alpha(\omega)}{\omega}\, i.$$

The boundary condition on Γ_{Z} allows modeling the behavior of absorbing viscoelastic materials covering the enclosure walls which are typically used as passive systems to reduce low-frequency noise. The frequency-dependent coefficients $\alpha(\omega)$ and $\beta(\omega)$ are related to the viscous and elastic responses of the isolating material, respectively. Both are strictly positive functions of the angular frequency ω. However, in what follows, we will assume only that $\beta(\omega) \neq 0$.

In our case, the secondary source f will be a linear combination of N Dirac delta measures supported at some given points, $y_1, \ldots, y_N \in \Omega$ with complex amplitudes $u_1, \ldots, u_N$ to be determined:

$$f = \sum_{i=1}^{N} u_i \delta_{y_i} \quad \text{with } u_i \in \mathbb{C},\ i = 1, \ldots, N. \tag{2.2}$$

This amounts to considering loudspeakers as acoustic monopoles (see, for instance, [11]).

In order to state the noise active control as an optimal control problem, we make the following choices:

- the *state of the system* is given by the pressure $p(x)$ in the domain Ω;
- the *control variable* $\boldsymbol{u}$ is the vector of complex amplitudes of the loudspeakers (actuators),

$$\boldsymbol{u} := (u_1, \ldots, u_N) \in \mathbb{C}^N,$$

which define the source term f in (2.1) by means of (2.2);

- the *set of admissible controls* is a convex closed set $U_{\mathrm{ad}} \subseteq \mathbb{C}^N$;
- the *model of the system* relating the control variable to the state is the Helmholtz problem (2.1);
- the *observation* $\boldsymbol{z}$ is the set of pressure values at M microphones (sensors) located at given points $w_1, \ldots, w_M \in \Omega$,

$$\boldsymbol{z}(\boldsymbol{u}) := (p(w_1), \ldots, p(w_M)) \in \mathbb{C}^M,$$

where, for $\boldsymbol{u} \in \mathbb{C}^N$, p denotes the solution of problem (2.1) with f given by (2.2); in the next section it will be shown that evaluating pressure at points $w_i \in \Omega$ makes sense as long as they do not coincide with the locations of the actuators;

- the *cost function* to be minimized depends on the observation and eventually on the cost of the control itself, namely,

$$J(\boldsymbol{u}) := \frac{1}{2}\|\boldsymbol{z}(\boldsymbol{u})\|^2 + \frac{\nu}{2}\|\boldsymbol{u}\|^2, \tag{2.3}$$

where $\nu \geq 0$ is a weighting factor, and $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{C}^N$ or $\mathbb{C}^M$. Thus we are led to the following *optimal control problem*:

Find $\boldsymbol{u}^{\rm op} \in U_{\rm ad}$ *such that*

$$J(\boldsymbol{u}^{\rm op}) = \inf_{\boldsymbol{u}\in U_{\rm ad}} J(\boldsymbol{u}). \tag{2.4}$$

Any solution $\boldsymbol{u}^{\rm op}$ of this minimization problem will be called an *optimal control*.

3. State equation. In this section we prove the existence and uniqueness of solution of the state equation and analyze its regularity, which will be used to study the optimal control problem. Our goal is the Helmholtz equation with singular data because, in our case, f is a linear combination of Dirac delta measures. However, we tackle this problem by first analyzing the same equation with data in $L^2(\Omega)$.

3.1. Data in $L^2(\Omega)$. We consider the Helmholtz problem (2.1) with $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_{\rm N})$. Multiplying the first equation by a test function $q \in H^1(\Omega)$, taking into account the boundary conditions, and using a Green's formula, we obtain the following weak formulation of (2.1):

Find $p \in H^1(\Omega)$ *such that*

$$\begin{aligned} &\int_\Omega \nabla p \cdot \nabla \bar{q}\, dx - \frac{i\omega\rho}{Z(\omega)} \int_{\Gamma_{\rm Z}} p\bar{q}\, d\Gamma - \Big(\frac{\omega}{c}\Big)^2 \int_\Omega p\bar{q}\, dx \\ &\qquad = \int_\Omega f\bar{q}\, dx + \int_{\Gamma_{\rm N}} g\bar{q}\, d\Gamma \quad \forall q \in H^1(\Omega). \end{aligned} \tag{3.1}$$

We denote by a_ω the sesquilinear continuous form in $H^1(\Omega)$ appearing in the left-hand side of this problem:

$$a_\omega(p,q) := \int_\Omega \nabla p \cdot \nabla \bar{q}\, dx - \frac{i\omega\rho}{Z(\omega)} \int_{\Gamma_{\rm Z}} p\bar{q}\, d\Gamma - \Big(\frac{\omega}{c}\Big)^2 \int_\Omega p\bar{q}\, dx, \quad p, q \in H^1(\Omega).$$

It is clear that a_ω is not positive definite and therefore the *Lax–Milgram lemma* can not be applied to show the existence and uniqueness of the solution. Instead, we show below that a_ω satisfies a Gårding's inequality. Then, according to Fredholm's alternative, uniqueness implies the existence of solution.

To prove uniqueness we consider the homogeneous problem

$$\widetilde{p} \in H^1(\Omega): \quad a_\omega(\widetilde{p}, q) = 0 \quad \forall q \in H^1(\Omega). \tag{3.2}$$

Due to the damping viscous term $\beta \neq 0$ of the wall impedance Z, from the viewpoint of physics, no solution $\widetilde{p} \neq 0$ of problem (3.2) should be expected when $\omega \in \mathbb{R}$, $\omega \neq 0$. Indeed, we have the following result.

LEMMA 3.1. *If* $|\Gamma_{\rm Z}| > 0$, $\omega \in \mathbb{R}$, $\omega \neq 0$, *and* $\beta \neq 0$, *then* $\widetilde{p} = 0$ *is the unique solution of problem* (3.2).

Proof. Let $\widetilde{p}$ be a solution of (3.2). By choosing $q = \widetilde{p}$ in (3.2) we obtain

$$A - \frac{i\omega\rho}{\beta + \frac{\alpha}{\omega}i}B - \left(\frac{\omega}{c}\right)^2 C = A - \frac{\alpha\rho}{\beta^2 + \frac{\alpha^2}{\omega^2}}B - \left(\frac{\omega}{c}\right)^2 C - i\frac{\omega\rho\beta}{\beta^2 + \frac{\alpha^2}{\omega^2}}B = 0,$$

where $A := \int_\Omega |\nabla\widetilde{p}|^2\, dx$, $B := \int_{\Gamma_z} |\widetilde{p}|^2\, d\Gamma$, and $C = \int_\Omega |\widetilde{p}|^2\, dx$ are real numbers. Then the imaginary part must vanish too. Hence, for $\beta \neq 0$ and $\omega \neq 0$, we have $B = 0$ and, consequently, $\widetilde{p} = 0$ on Γ_z. Then, by taking test functions $q \in \mathcal{C}^\infty(\bar{\Omega})$ in (3.2), we obtain that $\widetilde{p}$ satisfies

$$\begin{cases} -\Delta\widetilde{p} = \left(\dfrac{\omega}{c}\right)^2 \widetilde{p} & \text{in } \Omega, \\ \dfrac{\partial\widetilde{p}}{\partial \boldsymbol{n}} = 0 & \text{on } \partial\Omega, \\ \widetilde{p} = 0 & \text{on } \Gamma_z. \end{cases}$$

According to these equations, if $\widetilde{p} \neq 0$, then it would be an eigenfunction of the Laplace operator satisfying simultaneously Neumann and Dirichlet homogeneous conditions on Γ_z, which is not possible because of the *unique prolongation theorem*. Thus $\widetilde{p} = 0$, and we conclude the proof. ∎

On the other hand, the following lemma shows that the sesquilinear form a_ω satisfies a Gårding's inequality.

LEMMA 3.2. *There exist strictly positive constants γ and C_ω, the latter depending on $\omega \in \mathbb{R}$, such that*

$$|a_\omega(q,q) + C_\omega \|q\|^2_{L^2(\Omega)}| \geq \gamma \|q\|^2_{H^1(\Omega)} \quad \forall q \in H^1(\Omega). \tag{3.3}$$

Proof. For all $q \in H^1(\Omega)$ and $\omega \in \mathbb{R}$, we have

$$\mathrm{Re}[a_\omega(q,q)] = \|q\|^2_{H^1(\Omega)} - \frac{\alpha\rho}{\beta^2 + \frac{\alpha^2}{\omega^2}}\|q\|^2_{L^2(\Gamma_z)} - \left(1 + \frac{\omega^2}{c^2}\right)\|q\|^2_{L^2(\Omega)}.$$

Thus, if $\alpha \leq 0$, then (3.3) holds with $\gamma = 1$ and $C_\omega = 1 + \omega^2/c^2$. Otherwise, from the trace theorem (see, for instance, [2]), $\exists C > 0$ such that

$$\|q\|_{L^2(\Gamma_z)} \leq C\|q\|^{1/2}_{L^2(\Omega)}\|q\|^{1/2}_{H^1(\Omega)} \quad \forall q \in H^1(\Omega).$$

Hence, $\forall \varepsilon > 0$ we have that

$$\|q\|^2_{L^2(\Gamma_z)} \leq \varepsilon\|q\|^2_{H^1(\Omega)} + \frac{C^2}{4\varepsilon}\|q\|^2_{L^2(\Omega)} \quad \forall q \in H^1(\Omega).$$

Then, the choice $\varepsilon = (\beta^2 + \alpha^2/\omega^2)/(2\alpha\rho) > 0$ leads to

$$\mathrm{Re}[a_\omega(q,q) + C_\omega\|q\|^2_{L^2(\Omega)}] \geq \gamma\|q\|^2_{H^1(\Omega)}$$

with $\gamma := 1/2$ and $C_\omega := 1 + \omega^2/c^2 + C^2/(8\varepsilon^2)$. Therefore, since $|\cdot| \geq \mathrm{Re}(\cdot)$, we end the proof. ∎

Now we are able to conclude the following existence, uniqueness, and regularity of solution results for the Helmholtz problem given above. From now on C denotes a strictly positive constant not necessarily the same at each occurrence.

THEOREM 3.3. *Let $\omega \in \mathbb{R}$, $\omega \neq 0$, $\beta \neq 0$, $f \in L^2(\Omega)$, and $g \in L^2(\Gamma_{\rm N})$. Then problem* (3.1) *has a unique solution $p \in H^1(\Omega)$. Moreover, if $g|_{\Gamma_{\rm N}^j} \in H^{1/2}(\Gamma_{\rm N}^j)$, $j = 1, \ldots, J$, then $p \in H^2(\Omega)$ and the following estimate holds*:

$$\|p\|_{H^2(\Omega)} \leq C \left[\|f\|_{L^2(\Omega)} + \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\rm N}^j)} \right]. \tag{3.4}$$

Proof. Uniqueness has been proved in Lemma 3.1. Because of Lemma 3.2 we know that the problem satisfies Fredholm's alternative (see, for instance, Theorem 6.5.15 of [10]). Then existence is a consequence of uniqueness. Moreover, because of the *open mapping theorem*,

$$\|p\|_{H^1(\Omega)} \leq C[\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma_{\rm N})}]. \tag{3.5}$$

Next, by testing (3.1) with functions $q \in \mathcal{C}^\infty(\bar{\Omega})$ we obtain that p satisfies

$$\begin{cases} -\Delta p + p = \left(1 + \dfrac{\omega^2}{c^2}\right) p + f & \text{in } \Omega, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = \dfrac{i\omega\rho}{Z(\omega)} p & \text{on } \Gamma_{\rm Z}, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = g & \text{on } \Gamma_{\rm N}. \end{cases}$$

Since the domain Ω is a convex two-dimensional polygon or three-dimensional polyhedron, the standard a priori estimate for this Neumann problem (see [5] and [9]) yields

$$\|p\|_{H^2(\Omega)} \leq C \left[\|p\|_{L^2(\Omega)} + \|f\|_{L^2(\Omega)} + \sum_{k=1}^{K} \|p\|_{H^{1/2}(\Gamma_{\rm Z}^k)} + \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\rm N}^j)} \right],$$

where C is a constant independent of f, g, and p. Then (3.4) is a direct consequence of this inequality and (3.5). ∎

Remark 3.4. The convexity assumption on Ω is used only to obtain the estimate (3.4) and the analogous one for the Green's function (3.11) below. Similar results are valid for smooth nonconvex domains, too.

3.2. Dirac delta measures. Let us now consider the Helmholtz problem (2.1) with homogeneous Neumann boundary data and inner source $f = \delta_y$ being the Dirac delta measure supported at an inner point $y \in \Omega$. Its solution is the Green's function $G^y \in L^2(\Omega)$ of problem (2.1):

$$\begin{cases} -\Delta G^y - \left(\dfrac{\omega}{c}\right)^2 G^y = \delta_y & \text{in } \Omega, \\ \dfrac{\partial G^y}{\partial \boldsymbol{n}} = \dfrac{i\omega\rho}{Z(\omega)} G^y & \text{on } \Gamma_{\rm Z}, \\ \dfrac{\partial G^y}{\partial \boldsymbol{n}} = 0 & \text{on } \Gamma_{\rm N}, \end{cases} \tag{3.6}$$

where the first equation must be understood in the sense of distributions.

It is simple to show that this problem has a unique solution. Indeed, let Φ^y be the fundamental solution of the Helmholtz equation in the whole space, namely, the solution of

$$-\Delta\Phi^y - \left(\frac{\omega}{c}\right)^2 \Phi^y = \delta_y \quad \text{in } \mathbb{R}^n.$$

This fundamental solution is explicitly known (see, for instance, [6]):

$$\Phi^y(x) := \begin{cases} \dfrac{1}{4} Y_0\left(\dfrac{\omega}{c}|x-y|\right) & \text{if } n = 2, \\ \dfrac{\cos\left(\dfrac{\omega}{c}|x-y|\right)}{4\pi|x-y|} & \text{if } n = 3, \end{cases} \tag{3.7}$$

where Y_0 denotes the zero-order second-kind Bessel function. It clearly satisfies $\Phi^y \in \mathcal{C}^\infty(\mathbb{R}\setminus\{y\})$ and $\Phi^y|_\Omega \in L^2(\Omega)$. Then G^y is a solution of (3.6) if and only if $G^y = \Phi^y|_\Omega + p^y$, with p^y satisfying

$$\begin{cases} -\Delta p^y - \left(\dfrac{\omega}{c}\right)^2 p^y = 0 & \text{in } \Omega, \\ \dfrac{\partial p^y}{\partial \boldsymbol{n}} = \dfrac{i\omega\rho}{Z(\omega)} p^y + \dfrac{i\omega\rho}{Z(\omega)} \Phi^y - \dfrac{\partial \Phi^y}{\partial \boldsymbol{n}} & \text{on } \Gamma_{\mathrm{Z}}, \\ \dfrac{\partial p^y}{\partial \boldsymbol{n}} = -\dfrac{\partial \Phi^y}{\partial \boldsymbol{n}} & \text{on } \Gamma_{\mathrm{N}}. \end{cases} \tag{3.8}$$

The variational formulation of this problem consists of finding $p^y \in H^1(\Omega)$ such that

$$a_\omega(p^y, q) = \int_{\Gamma_{\mathrm{Z}}} \left[\frac{i\omega\rho}{Z(\omega)}\Phi^y - \frac{\partial \Phi^y}{\partial \boldsymbol{n}}\right] \bar{q}\, d\Gamma - \int_{\Gamma_{\mathrm{N}}} \frac{\partial \Phi^y}{\partial \boldsymbol{n}} \bar{q}\, d\Gamma \quad \forall q \in H^1(\Omega), \tag{3.9}$$

with a_ω as defined above. Then the arguments in the proof of Theorem 3.3 allow us to conclude that this problem has a unique solution and, hence, problem (3.6) does too. Moreover, these arguments also show that $p^y \in H^2(\Omega)$ and, furthermore,

$$\|p^y\|_{H^2(\Omega)} \le C \left[\sum_{k=1}^{K} \left\| \frac{i\omega\rho}{Z(\omega)}\Phi^y - \frac{\partial \Phi^y}{\partial \boldsymbol{n}} \right\|_{H^{1/2}(\Gamma_{\mathrm{Z}}^k)} + \sum_{j=1}^{J} \left\| \frac{\partial \Phi^y}{\partial \boldsymbol{n}} \right\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)}\right]. \tag{3.10}$$

Consequently p^y is a continuous function and, hence, $G^y = \Phi^y + p^y$ is continuous in $\Omega\setminus\{y\}$. This shows that evaluating the pressure at a point $w \in \Omega$ where a microphone is located makes sense as long as $w \neq y$. Therefore, the control problem (2.4) is well-posed whenever the sets of sensors' and actuators' locations do not intersect.

Furthermore, if $d > 0$ is such that $\mathcal{B}_d(y) := \{x \in \mathbb{R}^n : |x-y| < d\} \subset\subset \Omega$, then $\|\Phi^y\|_{H^2(\Omega\setminus\bar{\mathcal{B}}_d(y))}$ is bounded by a constant which depends on d. Thus, from the estimate (3.10) we have that

$$\|G^y\|_{H^2(\Omega\setminus\bar{\mathcal{B}}_d(y))} \le \|\Phi^y\|_{H^2(\Omega\setminus\bar{\mathcal{B}}_d(y))} + \|p^y\|_{H^2(\Omega)} \le C, \tag{3.11}$$

with C depending on d, too.

To end this subsection we present an alternative characterization of the solution of problem (3.6) obtained by *transposition techniques*. This will be used to prove convergence in $L^2(\Omega)$ of the numerical scheme introduced in the following section.

To this goal, note that given $q \in L^2(\Omega)$, the adjoint problem to (3.6) reads

$$\left\{\begin{array}{ll} -\Delta r - \left(\dfrac{\omega}{c}\right)^2 r = q & \text{in } \Omega, \\ \dfrac{\partial r}{\partial \boldsymbol{n}} = -\dfrac{i\omega\rho}{\bar{Z}(\omega)} r & \text{on } \Gamma_{\mathrm{Z}}, \\ \dfrac{\partial r}{\partial \boldsymbol{n}} = 0 & \text{on } \Gamma_{\mathrm{N}}. \end{array}\right. \tag{3.12}$$

In spite of the conjugate in the coefficient of the middle equation, Theorem 3.3 applies to this problem since it has been proved with the only assumptions that $\omega \neq 0$ and $\beta \neq 0$. Hence (3.12) has a unique solution which satisfies $r \in H^2(\Omega)$ and

$$\|r\|_{H^2(\Omega)} \leq C \|q\|_{L^2(\Omega)}. \tag{3.13}$$

In what follows we prove that the solution G^y of (3.6) is the unique function in $L^2(\Omega)$ which satisfies

$$\int_\Omega G^y \bar{q}\, dx = \langle \delta_y, r \rangle \quad \forall q \in L^2(\Omega). \tag{3.14}$$

Indeed, standard computations (see, for instance, Chapter 2.2.4 of [7]) show that

$$\langle \delta_y, r \rangle = \int_\Omega \Phi^y \left(-\Delta \bar{r} - \frac{\omega^2}{c^2} \bar{r} \right) dx + \int_{\partial\Omega} \left(\Phi^y \frac{\partial \bar{r}}{\partial \boldsymbol{n}} - \frac{\partial \Phi^y}{\partial \boldsymbol{n}} \bar{r} \right) d\Gamma.$$

On the other hand, integration by parts in (3.9) with r as a test function yields

$$\int_\Omega p^y \left(-\Delta \bar{r} - \frac{\omega^2}{c^2} \bar{r} \right) dx - \frac{i\omega\rho}{Z(\omega)} \int_{\Gamma_{\mathrm{Z}}} (p^y + \Phi^y) \bar{r}\, d\Gamma + \int_{\partial\Omega} \left(p^y \frac{\partial \bar{r}}{\partial \boldsymbol{n}} + \frac{\partial \Phi^y}{\partial \boldsymbol{n}} \bar{r} \right) d\Gamma = 0.$$

Then, by adding these two equations, we obtain

$$\langle \delta_y, r \rangle = \int_\Omega G^y \left(-\Delta \bar{r} - \frac{\omega^2}{c^2} \bar{r} \right) dx - \frac{i\omega\rho}{Z(\omega)} \int_{\Gamma_{\mathrm{Z}}} G^y \bar{r}\, d\Gamma + \int_{\partial\Omega} G^y \frac{\partial \bar{r}}{\partial \boldsymbol{n}} = \int_\Omega G^y \bar{q}\, dx,$$

where we have used the three equations of (3.12) for the last equality. Thus we conclude (3.14).

3.3. Local $W^{2,\infty}$ a priori estimates. The following lemma yields L^∞ local a priori estimates for the second derivatives of the solutions of problems (3.1) and (3.6). These bounds will be used in the following section to obtain pointwise error estimates for the finite element method proposed therein to solve these problems.

LEMMA 3.5. *Let D_0 and D_1 be disjoint open subsets of Ω satisfying $D_i \subset\subset \Omega$, $i = 0, 1$. Let $d > 0$ be such that* $\operatorname{dist}(D_i, \partial\Omega) \geq d$, *$i = 0, 1$, and* $\operatorname{dist}(D_0, D_1) \geq d$. *Then the following hold*:

1. *For each $y \in D_0$, the solution of problem* (3.6) *satisfies $G^y|_{D_1} \in W^{2,\infty}(D_1)$ and there exists a constant $C > 0$ depending on d such that*

$$\|G^y\|_{W^{2,\infty}(D_1)} \leq C \quad \forall y \in D_0.$$

2. *Let $f \in L^2(\Omega)$ be such that* $\operatorname{supp}(f) \subset D_0$. *Let $g \in L^2(\Gamma_{\mathrm{N}})$ be such that $g|_{\Gamma_{\mathrm{N}}^j} \in H^{1/2}(\Gamma_{\mathrm{N}}^j)$, $j = 1, \ldots, J$. Let p be the solution of problem* (3.1). *Then there exists a constant $C > 0$ depending on d such that*

$$\|p\|_{W^{2,\infty}(D_1)} \leq C \left[\|f\|_{L^2(\Omega)} + \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)} \right].$$

Proof. Consider the following subsets of Ω: $D_2 := \{x \in \mathbb{R}^n : \operatorname{dist}(x, D_1) < \frac{d}{4}\}$ and $D_3 := \{x \in \mathbb{R}^n : \operatorname{dist}(x, D_1) < \frac{d}{2}\}$. Then $\operatorname{dist}(D_3, \partial\Omega) \geq \frac{d}{2}$ and $\operatorname{dist}(D_0, D_3) \geq \frac{d}{2}$. Let χ be a $\mathcal{C}^\infty$ cut-off real function supported in D_3 such that $\chi|_{D_2} = 1$ and $\|\chi\|_{W^{2,\infty}(\mathbb{R}^n)}$ is bounded by a constant depending on d.

Given $y \in D_0$, we write the solution of problem (3.6) in the form $G^y = \Phi^y|_\Omega + p^y$, with Φ^y given by (3.7) and p^y being the solution of (3.8). Explicit differentiation of (3.7) shows that, for all $y \in D_0$, since $\operatorname{dist}(y, D_3) \geq \frac{d}{2}$, $\|\Phi^y\|_{W^{2,\infty}(D_3)}$ is bounded by a constant depending on d but not on y. So, to prove the first part of the theorem we only need to estimate $\|p^y\|_{W^{2,\infty}(D_1)}$.

Given $z \in D_1$, let G^z be the solution of problem (3.6) with y substituted by z. Then we have

$$p^y(z) = \langle \delta_z, \chi p^y \rangle = \left\langle -\Delta G^z - \frac{\omega^2}{c^2} G^z, \chi p^y \right\rangle = \left\langle G^z, -\Delta(\chi p^y) - \frac{\omega^2}{c^2} \chi p^y \right\rangle,$$

and

$$\begin{aligned} -\Delta(\chi p^y) - \frac{\omega^2}{c^2} \chi p^y &= \chi \left(-\Delta p^y - \frac{\omega^2}{c^2} p^y \right) - 2\nabla\chi \cdot \nabla p^y - (\Delta\chi) p^y \\ &= -[2\nabla\chi \cdot \nabla p^y + (\Delta\chi) p^y]. \end{aligned}$$

Hence, $\operatorname{supp}(-\Delta(\chi p^y) - \frac{\omega^2}{c^2} \chi p^y) \subset \operatorname{supp}(\nabla\chi) \subset D_3 \setminus \bar{D}_2$, and thus

$$p^y(z) = -\int_{D_3 \setminus \bar{D}_2} G^z(x) [2\nabla\chi(x) \cdot \nabla \bar{p}^y(x) + \Delta\chi(x)\, \bar{p}^y(x)]\, dx.$$

Because of the symmetry of the operator involved, the Green's function is symmetric, i.e., $G^z(x) = G^x(z)\ \forall x, z \in \Omega :\ x \neq z$. Then by differentiating the expression above we obtain $\forall \alpha \in \mathbb{N}^n$

$$D_z^\alpha p^y(z) = -\int_{D_3 \setminus \bar{D}_2} D_x^\alpha G^z(x) [2\nabla\chi(x) \cdot \nabla \bar{p}^y(x) + \Delta\chi(x)\, \bar{p}^y(x)]\, dx.$$

Consequently, by (3.11) and (3.10), we have $\forall \alpha \in \mathbb{N}^n$ such that $|\alpha| := \sum_{l=1}^n \alpha_l \leq 2$

$$\begin{aligned} |D_z^\alpha p^y(z)| &\leq \|G^z\|_{H^2(D_3 \setminus \bar{D}_2)} \|\chi\|_{W^{2,\infty}(D_3 \setminus \bar{D}_2)} \|p^y\|_{H^1(D_3 \setminus \bar{D}_2)} \\ &\leq C \left[\sum_{k=1}^{K} \left\| \frac{i\omega\rho}{Z(\omega)} \Phi^y - \frac{\partial \Phi^y}{\partial \boldsymbol{n}} \right\|_{H^{1/2}(\Gamma_{\mathrm{Z}}^k)} + \sum_{j=1}^{J} \left\| \frac{\partial \Phi^y}{\partial \boldsymbol{n}} \right\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)} \right] \leq C, \end{aligned}$$

with C depending on d but not on the particular point $z \in D_1$. Thus we conclude the proof of the first part of the lemma.

For the second part, let $f \in L^2(\Omega)$ with $\mathrm{supp}(f) \subset D_0$. We proceed exactly as above and use that

$$-\Delta(\chi p) - \frac{\omega^2}{c^2}(\chi p) = \chi\left(-\Delta p - \frac{\omega^2}{c^2} p\right) - 2\nabla\chi\cdot\nabla p - \Delta\chi\, p = -(2\nabla\chi\cdot\nabla p + \Delta\chi\, p),$$

because of $-\Delta p - \frac{\omega^2}{c^2} p = f$ and $\mathrm{supp}(f) \cap \mathrm{supp}\,\chi = \emptyset$. Then we obtain from (3.11) and Theorem 3.3

$$\begin{aligned} |D_z^\alpha p(z)| &\le \left| \int_{D_3\setminus\bar{D}_2} D_x^\alpha G^z(x)[2\nabla\chi(x)\cdot\nabla\bar{p}^y(x) + \Delta\chi(x)\,\bar{p}^y(x)]\,dx \right| \\ &\le \|G^z\|_{H^2(D_3\setminus\bar{D}_2)} \|\chi\|_{W^{2,\infty}(D_3\setminus\bar{D}_2)} \|p\|_{H^1(D_3\setminus\bar{D}_2)} \\ &\le C\left[\|f\|_{L^2(\Omega)} + \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)} \right] \quad \forall \alpha \in \mathbb{N}^n :\ |\alpha| \le 2, \end{aligned}$$

with C again depending on d but not on the particular point $z \in D_1$. Thus we conclude the proof. ∎

4. Numerical approximation of the state equation. For the construction of the finite element spaces we consider a quasi-uniform family of shape-regular triangulations $\{\mathcal{T}_h\}_{h>0}$ of Ω. More precisely, for each element $T \in \mathcal{T}_h$ (T being a two-dimensional triangle or a three-dimensional tetrahedron) we associate two parameters: h_T and ρ_T. The first one denotes the diameter of T and the second one the diameter of the largest ball contained in T. We denote $h := \max_{T\in\mathcal{T}_h} h_T$ and make the following hypothesis of regularity of the triangulation: there exist positive constants σ_1 and σ_2 such that

$$\frac{h_T}{\rho_T} \le \sigma_1, \quad \frac{h}{h_T} \le \sigma_2 \quad \forall\, T \in \mathcal{T}_h,\ \forall h > 0.$$

We associate with each triangulation $\mathcal{T}_h$ a finite element space $\mathcal{V}_h$ which consists of functions globally continuous in Ω and linear on each element $T \in \mathcal{T}_h$. Then, the discrete problem associated with problem (3.1) is the following:

Find $p_h \in \mathcal{V}_h$ such that

$$a_\omega(p_h, q_h) = \langle f, q_h\rangle + \int_{\Gamma_{\mathrm{N}}} g\bar{q}_h\, d\Gamma \quad \forall q_h \in \mathcal{V}_h. \tag{4.1}$$

Notice that this problem is well defined for $f \in L^2(\Omega)$ as well as for f given by (2.2), because the functions in $\mathcal{V}_h$ are continuous.

4.1. Data in $L^2(\Omega)$. Again, to tackle the numerical approximation when the source term is a Dirac delta measure, we consider first the problem with data in $L^2(\Omega)$.

Since a_ω is continuous and satisfies the Gårding's inequality (3.3), and since the continuous problem (3.1) has a unique solution, the following existence and approximation result is readily obtained from [12].

THEOREM 4.1. *Given $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_{\mathrm{N}})$, let p be the solution of problem (3.1). Then there exists $h_0 > 0$ such that, $\forall\, h \in (0, h_0]$, problem (4.1) has a unique solution p_h. Moreover, if $g|_{\Gamma_{\mathrm{N}}^j} \in H^{1/2}(\Gamma_{\mathrm{N}}^j)$, $j = 1, \dots, J$, then the following estimate holds*:

$$\|p - p_h\|_{L^2(\Omega)} \le Ch^2 \left[\|f\|_{L^2(\Omega)} + \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)} \right].$$

Since the observation $\boldsymbol{z}(\boldsymbol{u})$ consists of point values of the solution of problem (3.1), a pointwise error estimate will be used in the following section to obtain an error bound for the approximate control. To this aim, we have the following result, which is a consequence of the interior maximum norm estimates proved in [13].

THEOREM 4.2. *Given $f \in L^2(\Omega)$ such that $\mathrm{supp}(f) \subset\subset \Omega$ and $g \in L^2(\Gamma_{\mathrm{N}})$ such that $g|_{\Gamma_{\mathrm{N}}^j} \in H^{1/2}(\Gamma_{\mathrm{N}}^j)$, $j = 1, \ldots, J$, let p be the solution of problem (3.1). Given $w \in \Omega \setminus \mathrm{supp}(f)$, let $d > 0$ be such that $\mathrm{dist}(w, \partial\Omega) \geq d$, $\mathrm{dist}(w, \mathrm{supp}(f)) \geq d$, and $\mathrm{dist}(\mathrm{supp}(f), \partial\Omega) \geq d$. Let $h_0 > 0$ be such that, $\forall\, h \in (0, h_0]$, problem (4.1) has a unique solution p_h. Then, there exist strictly positive constants $h_1 < h_0$ and C, both depending on d, such that the following pointwise error estimate holds $\forall h \in (0, h_1]$:*

$$|p(w) - p_h(w)| \leq Ch^2 \ln\left(\frac{1}{h}\right) \left[\|f\|_{L^2(\Omega)} + \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)} \right].$$

Proof. Let $D := \mathcal{B}_{d/8}(w)$, $D_1 := \mathcal{B}_{d/4}(w)$, and $D_2 := \mathcal{B}_{d/2}(w)$. Then $D_2 \subset\subset \Omega$ and $D_2 \cap \mathrm{supp}(f) = \emptyset$, with $\mathrm{dist}(D_2, \partial\Omega) \geq \frac{d}{2}$ and $\mathrm{dist}(D_2, \mathrm{supp}(f)) \geq \frac{d}{2}$.

Now, for $h < h_0$, since p and p_h are solutions of (3.1) and (4.1), respectively, $\forall q_h \in \mathcal{V}_h$ with $\mathrm{supp}(q_h) \subset D_1$ there holds

$$\int_\Omega \nabla(p - p_h) \cdot \nabla \bar{q}_h \, dx - \left(\frac{\omega}{c}\right)^2 \int_\Omega (p - p_h)\bar{q}_h \, dx = 0.$$

Then, according to Theorem 5.1 in [13], there exist $C > 0$ and $h_1 > 0$ such that, $\forall h \in (0, h_1]$ and $\forall q_h \in \mathcal{V}_h$, the following inequality holds:

$$\|p - p_h\|_{L^\infty(D)} \leq C \ln\left(\frac{1}{h}\right) [\|p - q_h\|_{L^\infty(D_1)} + \|p - p_h\|_{L^2(D_1)}].$$

Because of Lemma 3.5 (part 2) and the standard error estimate for the Lagrange interpolation (see, for instance, [2]), if $h < d/4$, then

$$\inf_{q_h \in \mathcal{V}_h} \|p - q_h\|_{L^\infty(D_1)} \leq Ch^2 \|p\|_{W^{2,\infty}(D_2)} \leq C \left[\|f\|_{L^2(\Omega)} + \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)} \right].$$

Thus, the theorem follows from the last two inequalities and Theorem 4.1. ∎

4.2. Dirac delta measures. We consider now problem (4.1) with $g = 0$ and f being a Dirac delta measure:

$$G_h^y \in \mathcal{V}_h : \quad a_\omega(G_h^y, q_h) = \langle \delta_y, q_h \rangle \quad \forall q_h \in \mathcal{V}_h. \tag{4.2}$$

This is a discretization of problem (3.6). Let us recall that it is well defined because the functions in $\mathcal{V}_h$ are continuous. Then, by proceeding as in the previous subsection, it can be shown that $\exists h_0$ such that $\forall h \in (0, h_0]$, this problem has a unique solution. Here and thereafter, h_0 denotes a maximum mesh size that is not necessarily the same at each occurrence.

To show convergence in $L^2(\Omega)$, we are going to use the scheme proposed for elliptic problems in [14] (see also [4]). This scheme allows splitting the approximation error into two parts: the first one due to the error in the approximation of the Dirac delta measure by $L^2(\Omega)$ functions and the second one due to the approximation error of the Helmholtz equation with data in $L^2(\Omega)$.

Consider the following auxiliary variational problem:

$$\widetilde{G} \in H^1(\Omega): \quad a_\omega(\widetilde{G}, q) = \langle \delta_h, q \rangle \quad \forall q \in H^1(\Omega),$$

where $\delta_h \in L^2(\Omega)$ is an approximation of δ_y satisfying the following properties:

1. $\langle \delta_h, q_h \rangle = \int_\Omega \delta_h \bar{q}_h \, dx = \bar{q}_h(y) \quad \forall q_h \in \mathcal{V}_h$;
2. $\|\delta_h\|_{L^2(\Omega)} \leq Ch^{-n/2}$;
3. $\|\delta_y - \delta_h\|_{H^{-2}(\Omega)} \leq Ch^{2-n/2}$;
4. $\delta_h = 0$ outside the elements in which y lies.

A construction of such δ_h is given in [14].

We use the triangular inequality to separate the error in two terms:

$$\|G^y - G_h^y\|_{L^2(\Omega)} \leq \|G^y - \widetilde{G}\|_{L^2(\Omega)} + \|\widetilde{G} - G_h^y\|_{L^2(\Omega)}.$$

From property 1 above we conclude that δ_y and δ_h are identical functionals on $\mathcal{V}_h$. Consequently G_h^y can be seen as a finite element approximation of $\widetilde{G}$, too. Therefore, from Theorem 4.1 and property 2 of δ_h, we obtain the following approximation result for h sufficiently small:

$$\|\widetilde{G} - G_h^y\|_{L^2(\Omega)} \leq Ch^2 \|\delta_h\|_{L^2(\Omega)} \leq Ch^{2-n/2}.$$

To estimate the remaining term, note that by property 4 above δ_h has its support included in a certain set $D \subset\subset \Omega$, for h sufficiently small. Then we can apply to $\widetilde{G}$ the arguments of the transposition technique that allow us to prove (3.14). By so doing, we obtain that $\forall q \in L^2(\Omega)$, if $r \in H^2(\Omega)$ is the solution of (3.12), then

$$\int_\Omega (G^y - \widetilde{G}) \bar{q} = \langle \delta_y - \delta_h, r \rangle.$$

Hence, by taking $q = G^y - \widetilde{G}$ and using the a priori estimate (3.13), we have

$$\|G^y - \widetilde{G}\|^2_{L^2(\Omega)} \leq \|\delta_y - \delta_h\|_{H^{-2}(\Omega)} \|r\|_{H^2(\Omega)} \leq C \|\delta_y - \delta_h\|_{H^{-2}(\Omega)} \|G^y - \widetilde{G}\|_{L^2(\Omega)}.$$

Consequently, from property 3 of δ_h we have

$$\|G^y - \widetilde{G}\|_{L^2(\Omega)} \leq C \|\delta_y - \delta_h\|_{H^{-2}(\Omega)} \leq Ch^{2-n/2}.$$

Thus we have proved the following result.

THEOREM 4.3. *For $y \in \Omega$, let $G^y \in L^2(\Omega)$ be the solution of problem* (3.6). *There exists $h_0 > 0$ such that, $\forall\, h \in (0, h_0]$, problem* (4.2) *has a unique solution G_h^y. Moreover, there exists a strictly positive constant C such that the following error bound holds*:

$$\|G^y - G_h^y\|_{L^2(\Omega)} \leq Ch^{2-n/2}.$$

Next, we prove an error estimate for $|G^y(w) - G_h^y(w)|$, for $w \neq y$. This will be used in the following section to obtain an error bound for the approximate control. A similar result with the explicit dependence of the constant of the estimate on $|w - y|$ has been proved in Theorem 6.1(i) of [13], but for a coercive second-order linear operator with homogeneous Neumann boundary conditions on a smooth domain. The additional hypotheses of this theorem were used only to obtain an explicit estimate for the corresponding Green's function (see (4.9) of [13]). However, to the best of

the authors' knowledge, such an estimate is not available for our Helmholtz problem with mixed Neumann–Robin boundary conditions on a convex polyhedron. Instead, we will use the local maximum norm a priori estimate proved in Lemma 3.5 (part 1) to obtain a similar result, although without making explicit the dependence of the constant of the estimate on $|w-y|$.

THEOREM 4.4. *Let $y \in \Omega$ and $G^y \in L^2(\Omega)$ be the solution of problem (3.6). Let $h_0 > 0$ be such that, $\forall h \in (0, h_0]$, problem (4.2) has a unique solution G_h^y. Given $w \in \Omega$, $w \neq y$, let $d > 0$ be such that $|w-y| \geq d$, $\operatorname{dist}(w, \partial\Omega) \geq d$, and $\operatorname{dist}(y, \partial\Omega) \geq d$. Then there exist strictly positive constants $h_1 < h_0$ and C, both depending on d, such that, $\forall h \in (0, h_1]$,*

$$|G^y(w) - G_h^y(w)| \leq Ch^2 \ln\left(\frac{1}{h}\right).$$

Proof. The proof is essentially identical to that of Theorem 6.1(i) of [13]. Thus, we include here only its main steps and emphasize those which differ in our case.

Let $D_1 := \mathcal{B}_{d/4}(w)$, $D_2 := \mathcal{B}_{d/2}(w)$, and $D' := \mathcal{B}_{d/4}(y)$. Then $\operatorname{dist}(D_1, D') \geq \frac{d}{2}$. Applying Theorem 5.1 in [13] we know that there exist constants $C > 0$ and $h_1 > 0$, both depending on d, such that, $\forall h \in (0, h_1]$ and $\forall q_h \in \mathcal{V}_h$,

$$|G^y(w) - G_h^y(w)| \leq C \ln\left(\frac{1}{h}\right) [\|G^y - q_h\|_{L^\infty(D_1)} + \|G^y - G_h^y\|_{L^1(D_1)}].$$

The first term in the right-hand side can be bounded by using Lemma 3.5 (part 1) and the standard error estimate for the Lagrange interpolation (see, for instance, [2]); namely, if $h < d/4$, then

$$\inf_{q_h \in \mathcal{V}_h} \|G^y - q_h\|_{L^\infty(D_1)} \leq Ch^2 \|G^y\|_{W^{2,\infty}(D_2)} \leq Ch^2.$$

For the second term, we apply the same duality argument as in the proof of Theorem 6.1(i) of [13]. By doing so, we can repeat all the steps of this proof with the exception of the following one: Given $q \in \mathcal{C}_0^\infty(D_1)$, let

$$r \in W^{1,\infty}(\Omega): \quad a_\omega(s, r) = \int_\Omega s\bar{q}\, dx \quad \forall s \in W^{1,1}(\Omega);$$

it has to be proved that $\|r\|_{W^{2,\infty}(D')} \leq C\|q\|_{L^\infty(D_1)}$. In our case, we do it by repeating the arguments in the proof of Lemma 3.5 (part 2).

The rest of the proof runs essentially as that of Theorem 6.1(i) of [13]. ∎

5. Optimal amplitudes of actuators. Numerical methods. From now on we assume that the primary source g is such that $g|_{\Gamma_N^j} \in H^{1/2}(\Gamma_N^j)$, $j = 1, \ldots, J$, and the secondary source is given by (2.2) in terms of the control variable $\boldsymbol{u} = (u_1, \ldots, u_N) \in \mathbb{C}^N$: $f := \sum_{i=1}^N u_i \delta_{y_i}$, with $y_i \in \Omega$, $i = 1, \ldots, N$.

Let p be the solution of problem (2.1) with such f. Due to the linearity of the Helmholtz equation, p can be written in terms of the control variable as follows:

$$p = p_0 + \sum_{i=1}^N u_i G^{y_i}, \tag{5.1}$$

where p_0 is the pressure field arising from the primary source g without any control; more precisely, p_0 is the solution of problem (2.1) for $\boldsymbol{u} = \boldsymbol{0}$ (i.e., $f = 0$). In its

turn, G^{y_i} is the solution of problem (3.6) with $y = y_i$, $i = 1, \dots, N$. This corresponds to the pressure field when the system is only excited by the ith loudspeaker with unit amplitude, excluding the effect of the primary source. Note that according to Theorem 3.3 and (3.11), it makes sense to evaluate p at points $w \in \Omega$, $w \neq y_i$, $i = 1, \dots, N$.

Let $w_1, \dots, w_M \in \Omega$ be such that $\{y_1, \dots, y_N\}$ and $\{w_1, \dots, w_M\}$ are disjoint. Proving the existence of an optimal control is an easy task because the control space is finite-dimensional. It relies upon the fact that the mapping giving the observation from the control, namely,

$$\begin{aligned} \mathbb{C}^N &\longrightarrow \mathbb{C}^M, \\ \boldsymbol{u} &\longmapsto \boldsymbol{z}(\boldsymbol{u}) = (p(w_i), \dots, p(w_M)), \end{aligned}$$

is affine (and thus continuous). The mapping $\boldsymbol{z}(\boldsymbol{u})$ is the so-called *transfer function*, which establishes the relation between controls and observations.

Therefore, the cost function (2.3) is quadratic. The first term of the cost function is convex since the observation $\boldsymbol{z}(\boldsymbol{u})$ is affine and the second one is strictly convex when $\nu > 0$. Therefore, it is clear that the function J is strictly convex under either of the two following assumptions:

- $\nu > 0$,
- $\nu \geq 0$ and $\boldsymbol{z}(\boldsymbol{u})$ is one-to-one,

in which case there exists a unique optimal control.

We notice that $\boldsymbol{z}(\boldsymbol{u})$ is one-to-one if and only if the observations corresponding to each single actuator are linearly independent. Obviously this can happen only if the number of microphones is greater than or equal to the number of loudspeakers: $M \geq N$.

To write the control problem in matrix form, we introduce the vectors

$$\begin{aligned} \boldsymbol{z}_0 &:= (p_0(w_1), \dots, p_0(w_M)) \in \mathbb{C}^M, \\ \boldsymbol{z}_i &:= (G^{y_i}(w_1), \dots, G^{y_i}(w_M)) \in \mathbb{C}^M, \quad i = 1, \dots, N. \end{aligned}$$

Note that according to Theorem 3.3 and (3.11), respectively, there hold

$$\|\boldsymbol{z}_0\| \leq C \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)} \quad \text{and} \quad \|\boldsymbol{z}_i\| \leq C, \quad i = 1, \dots, N. \tag{5.2}$$

The observation $\boldsymbol{z}$ can be written in terms of the control variable $\boldsymbol{u} \in \mathbb{C}^N$ and the observations $\boldsymbol{z}_0, \boldsymbol{z}_1, \dots, \boldsymbol{z}_N$ in the following way:

$$\boldsymbol{z}(\boldsymbol{u}) = \boldsymbol{z}_0 + \sum_{i=1}^{N} u_i \boldsymbol{z}_i.$$

Then the cost function becomes

$$\begin{aligned} J(\boldsymbol{u}) &= \frac{1}{2} \left\| \boldsymbol{z}_0 + \sum_{i=1}^{N} u_i \boldsymbol{z}_i \right\|^2 + \frac{\nu}{2} \|\boldsymbol{u}\|^2 \\ &= \frac{1}{2} \left[\bar{\boldsymbol{z}}_0^{\mathrm{t}} \boldsymbol{z}_0 + 2 \operatorname{Re} \left(\sum_{i=1}^{N} \bar{u}_i \bar{\boldsymbol{z}}_i^{\mathrm{t}} \boldsymbol{z}_0 \right) + \sum_{i,j=1}^{N} u_j \bar{u}_i (\bar{\boldsymbol{z}}_i^{\mathrm{t}} \boldsymbol{z}_j + \nu \delta_{ij}) \right]. \end{aligned}$$

Let us define the matrix $\boldsymbol{Z} \in \mathbb{C}^{N\times N}$ and the vector $\boldsymbol{d} \in \mathbb{C}^N$ by

$$(\boldsymbol{Z})_{ij} := \bar{\boldsymbol{z}}_i^{\mathrm{t}} \boldsymbol{z}_j, \quad i,j = 1,\dots,N,$$
$$(\boldsymbol{d})_i := \bar{\boldsymbol{z}}_i^{\mathrm{t}} \boldsymbol{z}_0, \quad i = 1,\dots,N.$$

Then the optimal control problem (2.4) is equivalent to the following quadratic programming problem:

Find $\boldsymbol{u}^{\mathrm{op}} \in U_{\mathrm{ad}}$ *such that*

$$J(\boldsymbol{u}^{\mathrm{op}}) = \inf_{\boldsymbol{u}\in U_{\mathrm{ad}}} \frac{1}{2}[\bar{\boldsymbol{u}}^{\mathrm{t}}(\boldsymbol{Z} + \nu \boldsymbol{I})\boldsymbol{u} + 2\,\mathrm{Re}(\bar{\boldsymbol{u}}^{\mathrm{t}}\boldsymbol{d}) + \|\boldsymbol{z}_0\|^2].$$

Although the cost function is defined in a finite-dimensional space, it involves the solution of a partial differential equation which has to be approximated by means of some discretization process as, for instance, the finite element method described in section 4. This leads to approximate observations and thereby to an approximate cost function.

Similar definitions hold for the approximate observations. Given $\boldsymbol{u} \in \mathbb{C}^N$, let

$$\boldsymbol{z}_h(\boldsymbol{u}) := (p_h(w_i),\dots,p_h(w_M)) \in \mathbb{C}^M,$$

where p_h is the solution of the discrete problem (4.1) with f defined by (2.2) as above. Let $\boldsymbol{z}_{ih} \in \mathbb{C}^M$ be defined by

$$\boldsymbol{z}_{0h} := (p_{0h}(w_1),\dots,p_{0h}(w_M)) \in \mathbb{C}^M,$$
$$\boldsymbol{z}_{ih} := (G_h^{y_i}(w_1),\dots,G_h^{y_i}(w_M)) \in \mathbb{C}^M, \quad i = 1,\dots,N,$$

where p_{0h} is the solution of problem (4.1) for $\boldsymbol{u} = \boldsymbol{0}$ (i.e., $f = 0$), and $G_h^{y_i}$ is the solution of the discrete problem (4.2) for $y = y_i$, $i = 1,\dots,N$. Then $\boldsymbol{z}_h(\boldsymbol{u}) \in \mathbb{C}^M$ is given by

$$\boldsymbol{z}_h(\boldsymbol{u}) := \boldsymbol{z}_{0h} + \sum_{i=1}^{N} u_i \boldsymbol{z}_{ih}.$$

Let $\boldsymbol{Z}_h \in \mathbb{C}^{N\times N}$ and $\boldsymbol{d}_h \in \mathbb{C}^N$ be defined by

$$(\boldsymbol{Z}_h)_{ij} := \bar{\boldsymbol{z}}_{ih}^{\mathrm{t}} \boldsymbol{z}_{jh}, \quad i,j = 1,\dots,N,$$
$$(\boldsymbol{d}_h)_i := \bar{\boldsymbol{z}}_{ih}^{\mathrm{t}} \boldsymbol{z}_{0h}, \quad i = 1,\dots,N.$$

Then the approximate cost function can be written as

$$J_h(\boldsymbol{u}) := \frac{1}{2}\|\boldsymbol{z}_h(\boldsymbol{u})\|^2 + \frac{\nu}{2}\|\boldsymbol{u}\|^2 = \frac{1}{2}[\bar{\boldsymbol{u}}^{\mathrm{t}}(\boldsymbol{Z}_h + \nu \boldsymbol{I})\boldsymbol{u} + 2\,\mathrm{Re}(\bar{\boldsymbol{u}}^{\mathrm{t}}\boldsymbol{d}_h) + \|\boldsymbol{z}_{0h}\|^2].$$

These definitions lead us to the following discrete optimal control problem:

Find $\boldsymbol{u}_h^{\mathrm{op}} \in U_{\mathrm{ad}}$ *such that*

$$J_h(\boldsymbol{u}_h^{\mathrm{op}}) = \inf_{\boldsymbol{u}\in U_{\mathrm{ad}}} \frac{1}{2}[\bar{\boldsymbol{u}}^{\mathrm{t}}(\boldsymbol{Z}_h + \nu \boldsymbol{I})\boldsymbol{u} + 2\,\mathrm{Re}(\bar{\boldsymbol{u}}^{\mathrm{t}}\boldsymbol{d}_h) + \|\boldsymbol{z}_{0h}\|^2]. \tag{5.3}$$

The argument $\boldsymbol{u}_h^{\mathrm{op}}$, where the minimum is attained, is expected to be an approximation of the optimal control $\boldsymbol{u}^{\mathrm{op}}$. Our next goal is to obtain an estimate for this approximation. To this aim, we denote

$$\delta \boldsymbol{d} := \boldsymbol{d} - \boldsymbol{d}_h \quad \text{and} \quad \delta \boldsymbol{Z} := \boldsymbol{Z} - \boldsymbol{Z}_h.$$

LEMMA 5.1. *There exist strictly positive constants C and h_0 such that, $\forall\, h \in (0, h_0]$, the following inequalities hold:*

$$\|\delta \boldsymbol{d}\| \le Ch^2 \ln\left(\frac{1}{h}\right) \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)}, \qquad \|\delta \boldsymbol{Z}\| \le Ch^2 \ln\left(\frac{1}{h}\right).$$

Moreover, if $\boldsymbol{Z}$ is positive definite, then so is $\boldsymbol{Z}_h$ for h small enough.

Proof. First we settle an error estimate for the observations $\boldsymbol{z}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$. We denote the corresponding errors by

$$\delta \boldsymbol{z}_i := \boldsymbol{z}_{ih} - \boldsymbol{z}_i, \quad i = 0, \ldots, N.$$

From Theorem 4.2, for h small enough we have

$$\|\delta \boldsymbol{z}_0\| = \left[\sum_{k=1}^{M} |p_0(w_k) - p_{0h}(w_k)|^2\right]^{1/2} \le Ch^2 \ln\left(\frac{1}{h}\right) \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)},$$

whereas from Theorem 4.4,

$$\|\delta \boldsymbol{z}_i\| = \left[\sum_{k=1}^{M} |G^{y_i}(w_k) - G_h^{y_i}(w_k)|^2\right]^{1/2} \le Ch^2 \ln\left(\frac{1}{h}\right).$$

Therefore, if h is small enough,

$$\begin{aligned}
\|\delta \boldsymbol{d}\| &= \left[\sum_{i=1}^{N} |\bar{\boldsymbol{z}}_0^{\mathrm{t}} \boldsymbol{z}_i - (\bar{\boldsymbol{z}}_0 + \delta \bar{\boldsymbol{z}}_0)^{\mathrm{t}} (\boldsymbol{z}_i + \delta \boldsymbol{z}_i)|^2\right]^{1/2} \\
&\le \left[\sum_{i=1}^{N} (\|\boldsymbol{z}_0\| \|\delta \boldsymbol{z}_i\| + \|\delta \boldsymbol{z}_0\| \|\boldsymbol{z}_i\| + \|\delta \boldsymbol{z}_0\| \|\delta \boldsymbol{z}_i\|)^2\right]^{1/2} \\
&\le Ch^2 \ln\left(\frac{1}{h}\right) \sum_{j=1}^{J} \|g\|_{H^{1/2}(\Gamma_{\mathrm{N}}^j)},
\end{aligned}$$

the last inequality because of (5.2).

The error bound for $\|\delta \boldsymbol{Z}\|$ is proved essentially in the same way. Furthermore, since $\boldsymbol{Z}_h$ converges to $\boldsymbol{Z}$, if $\boldsymbol{Z}$ is positive definite, then for h small enough $\boldsymbol{Z}_h$ is positive definite too. ∎

As an immediate consequence of the above lemma we have the existence and uniqueness of solution of the discrete optimal control problem, for h sufficiently small.

COROLLARY 5.2. *Let us assume that $\nu > 0$ or $\nu \ge 0$ and $\boldsymbol{z}(\boldsymbol{u})$ is one-to-one. Then, there exists $h_0 > 0$ such that, $\forall\, h \in (0, h_0]$, problem (5.3) has a unique solution.*

To obtain a bound for $\|\boldsymbol{u}^{\mathrm{op}} - \boldsymbol{u}_h^{\mathrm{op}}\|$, we prove first the following a priori error estimate for the solution of a variational inequality subject to data perturbations.

LEMMA 5.3. *Let U_{ad} be a convex subset of $\mathbb{C}^N$, $\boldsymbol{b} \in \mathbb{C}^N$, and $\boldsymbol{A} \in \mathbb{C}^{N \times N}$ a positive definite Hermitian matrix. Let $\alpha > 0$ be such that*

$$\bar{\boldsymbol{v}}^{\mathrm{t}} \boldsymbol{A} \boldsymbol{v} \ge \alpha \|\boldsymbol{v}\|^2 \quad \forall \boldsymbol{v} \in \mathbb{C}^N.$$

Let $\delta\boldsymbol{b} \in \mathbb{C}^N$ and $\delta\boldsymbol{A} \in \mathbb{C}^{N\times N}$ be such that $\|\delta\boldsymbol{A}\| < \alpha$. Let $\boldsymbol{u} \in U_{\rm ad}$ and $(\boldsymbol{u}+\delta\boldsymbol{u}) \in U_{\rm ad}$ be the solutions of the following variational inequalities:

$$\mathrm{Re}[(\bar{\boldsymbol{v}} - \bar{\boldsymbol{u}})^{\rm t}(\boldsymbol{A}\boldsymbol{u} + \boldsymbol{b})] \geq 0 \quad \forall \boldsymbol{v} \in U_{\rm ad}, \tag{5.4}$$

$$\mathrm{Re}\{[\bar{\boldsymbol{v}} - (\bar{\boldsymbol{u}} + \bar{\delta\boldsymbol{u}})]^{\rm t}[(\boldsymbol{A} + \delta\boldsymbol{A})(\boldsymbol{u} + \delta\boldsymbol{u}) + (\boldsymbol{b} + \delta\boldsymbol{b})]\} \geq 0 \quad \forall \boldsymbol{v} \in U_{\rm ad}. \tag{5.5}$$

Then

$$\|\delta\boldsymbol{u}\| \leq \frac{1}{\alpha - \|\delta\boldsymbol{A}\|}(\|\delta\boldsymbol{A}\|\|\boldsymbol{u}\| + \|\delta\boldsymbol{b}\|). \tag{5.6}$$

Furthermore, if $U_{\rm ad} \ni \boldsymbol{0}$ and $\|\delta\boldsymbol{A}\| < \theta\alpha$ with $0 < \theta < 1$, then

$$\|\delta\boldsymbol{u}\| \leq \frac{1}{(1-\theta)\alpha}\left(\frac{\|\boldsymbol{b}\|}{\alpha}\|\delta\boldsymbol{A}\| + \|\delta\boldsymbol{b}\|\right). \tag{5.7}$$

Proof. By taking $\boldsymbol{v} = \boldsymbol{u} + \delta\boldsymbol{u}$ in (5.4) and $\boldsymbol{v} = \boldsymbol{u}$ in (5.5) we obtain

$$\mathrm{Re}[\bar{\delta\boldsymbol{u}}^{\rm t}(\boldsymbol{A}\boldsymbol{u} + \boldsymbol{b})] \geq 0,$$
$$\mathrm{Re}\{-\bar{\delta\boldsymbol{u}}^{\rm t}[(\boldsymbol{A} + \delta\boldsymbol{A})(\boldsymbol{u} + \delta\boldsymbol{u}) + (\boldsymbol{b} + \delta\boldsymbol{b})]\} \geq 0.$$

By adding these inequalities we obtain

$$\mathrm{Re}(\bar{\delta\boldsymbol{u}}^{\rm t}\boldsymbol{A}\delta\boldsymbol{u}) \leq \mathrm{Re}\{-\bar{\delta\boldsymbol{u}}^{\rm t}[\delta\boldsymbol{A}(\boldsymbol{u} + \delta\boldsymbol{u}) + \delta\boldsymbol{b}]\}.$$

Then, since $\boldsymbol{A}$ is Hermitian and positive definite, we have

$$\alpha\|\delta\boldsymbol{u}\|^2 \leq \bar{\delta\boldsymbol{u}}^{\rm t}\boldsymbol{A}\delta\boldsymbol{u} = \mathrm{Re}(\bar{\delta\boldsymbol{u}}^{\rm t}\boldsymbol{A}\delta\boldsymbol{u}) \leq \|\delta\boldsymbol{u}\|(\|\delta\boldsymbol{A}\|\|\delta\boldsymbol{u}\| + \|\delta\boldsymbol{A}\|\|\boldsymbol{u}\| + \|\delta\boldsymbol{b}\|),$$

and, therefore,

$$(\alpha - \|\delta\boldsymbol{A}\|)\|\delta\boldsymbol{u}\| \leq \|\delta\boldsymbol{A}\|\|\boldsymbol{u}\| + \|\delta\boldsymbol{b}\|.$$

Hence, for $\|\delta\boldsymbol{A}\| < \alpha$, we obtain (5.6). Moreover, if $\|\delta\boldsymbol{A}\| < \theta\alpha$ with $0 < \theta < 1$, then

$$\|\delta\boldsymbol{u}\| \leq \frac{1}{(1-\theta)\alpha}(\|\delta\boldsymbol{A}\|\|\boldsymbol{u}\| + \|\delta\boldsymbol{b}\|). \tag{5.8}$$

On the other hand, if $U_{\rm ad} \ni \boldsymbol{0}$, then we can take $\boldsymbol{v} = \boldsymbol{0}$ in (5.4) and we obtain $\mathrm{Re}[-\bar{\boldsymbol{u}}^{\rm t}(\boldsymbol{A}\boldsymbol{u} + \boldsymbol{b})] \geq 0$. Then

$$\alpha\|\delta\boldsymbol{u}\|^2 \leq \bar{\boldsymbol{u}}^{\rm t}\boldsymbol{A}\boldsymbol{u} = \mathrm{Re}(\bar{\boldsymbol{u}}^{\rm t}\boldsymbol{A}\boldsymbol{u}) \leq \mathrm{Re}(-\bar{\boldsymbol{u}}^{\rm t}\boldsymbol{b}) \leq \|\delta\boldsymbol{u}\|\|\delta\boldsymbol{b}\|.$$

Hence, $\|\delta\boldsymbol{u}\| \leq \|\delta\boldsymbol{b}\|/\alpha$, and (5.7) follows from this inequality and (5.8). Thus we conclude the proof. □

From the above lemma and the error estimates of Lemma 5.1 it is easy to prove the following result.

THEOREM 5.4. *Let us assume that $\nu > 0$ or $\nu \geq 0$ and $\boldsymbol{z}(\boldsymbol{u})$ is one-to-one. If $U_{\rm ad} \ni \boldsymbol{0}$, then there exist $C > 0$ and $h_0 > 0$ such that, $\forall\, h \in (0, h_0]$,*

$$\|\boldsymbol{u}^{\rm op} - \boldsymbol{u}_h^{\rm op}\| \leq Ch^2 \ln\left(\frac{1}{h}\right)\sum_{j=1}^{J}\|g\|_{H^{1/2}(\Gamma_{\rm N}^j)}.$$

Proof. Let $\delta\boldsymbol{u} := \boldsymbol{u}_h^{\rm op} - \boldsymbol{u}^{\rm op}$. The exact and the approximate optimal controls satisfy the variational inequalities

$$\begin{aligned}
&\mathrm{Re}\{(\bar{\boldsymbol{v}} - \bar{\boldsymbol{u}}^{\rm op})^{\rm t}[(\boldsymbol{Z} + \nu\boldsymbol{I})\boldsymbol{u}^{\rm op} + \boldsymbol{d}]\} \geq 0 \quad \forall \boldsymbol{v} \in U_{\rm ad},\\
&\mathrm{Re}\{(\boldsymbol{v} - \boldsymbol{u}_h^{\rm op})^{\rm t}[(\boldsymbol{Z} + \nu\boldsymbol{I} + \delta\boldsymbol{Z})(\boldsymbol{u}^{\rm op} + \delta\boldsymbol{u}) + (\boldsymbol{d} + \delta\boldsymbol{d})]\} \geq 0 \quad \forall \boldsymbol{v} \in U_{\rm ad}.
\end{aligned}$$

Since $\boldsymbol{Z} + \nu\boldsymbol{I}$ is a Hermitian positive definite matrix, let $\alpha > 0$ be such that

$$\bar{\boldsymbol{v}}^{\rm t}(\boldsymbol{Z} + \nu\boldsymbol{I})\boldsymbol{v} \geq \alpha\|\boldsymbol{v}\|^2 \quad \forall \boldsymbol{v} \in U_{\rm ad}.$$

According to Lemma 5.1, for h sufficiently small $\|\delta\boldsymbol{Z}\| < \alpha/2$. Then we can apply Lemma 5.3 to the variational inequalities above and obtain

$$\|\boldsymbol{u}^{\rm op} - \boldsymbol{u}_h^{\rm op}\| \leq \frac{2}{\alpha}\left(\frac{\|\boldsymbol{d}\|}{\alpha}\|\delta\boldsymbol{Z}\| + \|\delta\boldsymbol{d}\|\right).$$

Thus, we conclude the proof from this inequality, (5.2), and Lemma 5.1. ∎

Remark 5.5. The assumption made on the admissible set, $U_{\rm ad} \ni \boldsymbol{0}$, to prove the error estimate of this theorem is not restrictive at all in practice. It just means that a vanishing control is also admissible.

6. Numerical results. In this section we present some numerical results for a three-dimensional test. In order to assess the effect of the control we use the following measure of attenuation:

$$\text{Attenuation (dB)} = -10\log_{10}\left[\frac{J(\boldsymbol{u}^{\rm op})}{J(\boldsymbol{0})}\right].$$

The data of the test are the following:

- the domain is $\Omega = [0,1]\,\mathrm{m} \times [0,1]\,\mathrm{m} \times [0,1]\,\mathrm{m}$;
- the physical parameters are $\omega = 680\,\mathrm{s}^{-1}$, $c = 340\,\mathrm{m\,s}^{-1}$, and $\rho = 1\,\mathrm{kg\,m}^{-3}$;
- the amplitude of the primary source of noise is $g(x,y,0) = \mathrm{e}^{iy}\,\mathrm{kg\,m}^{-2}\,\mathrm{s}^{-2}$ on the wall $z = 0$;
- there is one loudspeaker located at $y_1 = (\frac{4}{6}, \frac{4}{6}, \frac{4}{6})\,\mathrm{m}$;
- there are two microphones at $w_1 = (\frac{1}{6}, \frac{2}{6}, \frac{1}{6})\,\mathrm{m}$ and $w_2 = (\frac{5}{6}, \frac{1}{6}, \frac{3}{6})\,\mathrm{m}$;
- on the wall $z = 1$, the acoustic impedance is $Z = (102 + 340\,i) \times 10^3\,\mathrm{kg\,m}^{-2}\,\mathrm{s}^{-1}$;
- the rest of the walls are perfectly rigid;
- the admissible set of controls is $U_{\rm ad} = \mathbb{C}$ and the weighting factor is $\nu = 0$.

The optimal control has been computed for several meshes which have been obtained by uniformly refining the coarse mesh shown in Figure 6.1.

Then a more accurate value of the optimal control has been determined by extrapolating the controls computed on these meshes. This more accurate value has been used to compute the relative errors of the real and imaginary parts of the computed controls. These errors are shown in Figure 6.2, where it can be clearly observed that the order of convergence is essentially $\mathcal{O}(h^2)$ as predicted by the theoretical results.

For $h = 1/24$, the computed attenuation is 0.75 dB. The modulus of the complex pressure field on the plane containing the actuator and the two sensors is shown in Figures 6.3 and 6.4. The first one corresponds to the system without control, whereas the second one shows the pressure with the optimal control.

Finally Figure 6.5 shows the local attenuation field computed on the same plane: $\mathrm{Att}(w) = -10\log_{10}[|p_h(w)|^2/|p_{0h}(w)|^2]$.

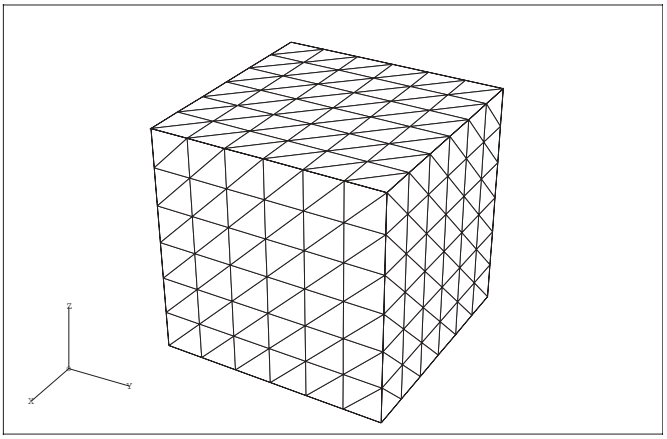


FIG. 6.1. *Coarsest mesh* ($h = 1/6\,\mathrm{m}$).

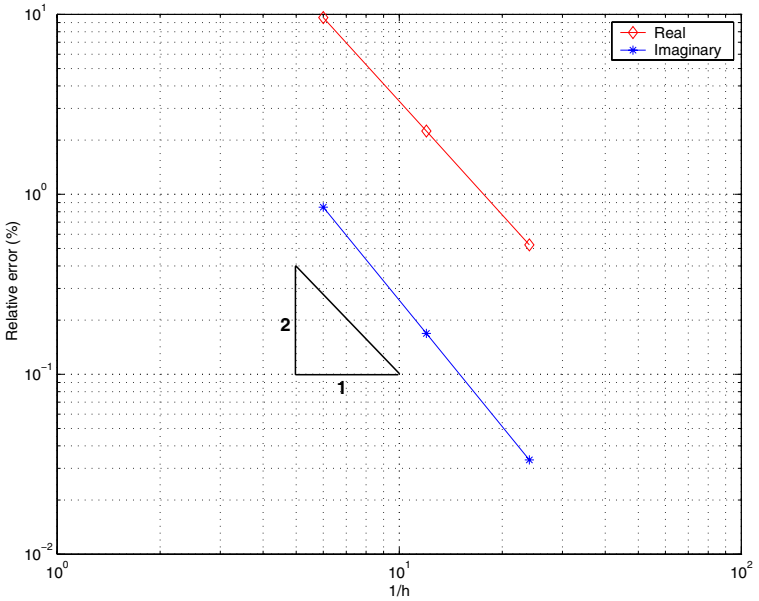


FIG. 6.2. *Relative error (%) as a function of* $1/h$ *in log-log scale.*

In this case, it can be observed that there exist zones where the noise is reinforced, that is, where primary and secondary sources interfere in a constructive way. This happens, for instance, in the location of the first sensor. Indeed, the noise level without control is low around this sensor and too high around the other one (see Figure 6.3). Thus, to obtain a minimum of the cost functional, the optimal amplitude of the actuator is such that it produces a noise reinforcement in the first sensor. Anyway, the comparison of Figures 6.3 and 6.4 shows that the global attenuation has been significant in the whole domain, except for the vicinity of the actuator.

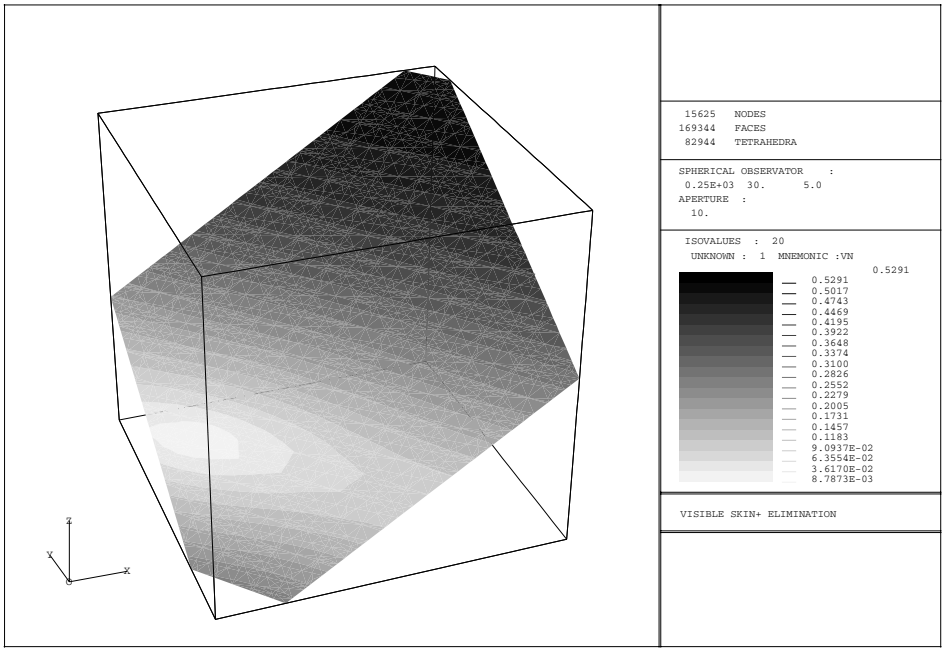


FIG. 6.3. *Modulus of the pressure field without control* ($h = 1/24\,\mathrm{m}$).

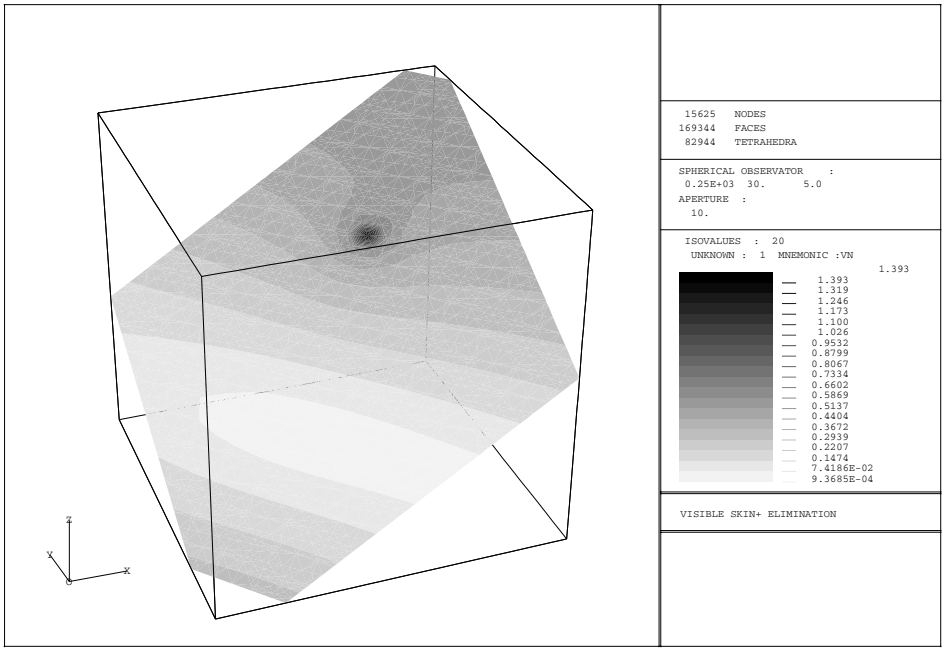


FIG. 6.4. *Modulus of the pressure field with control* ($h = 1/24\,\mathrm{m}$).

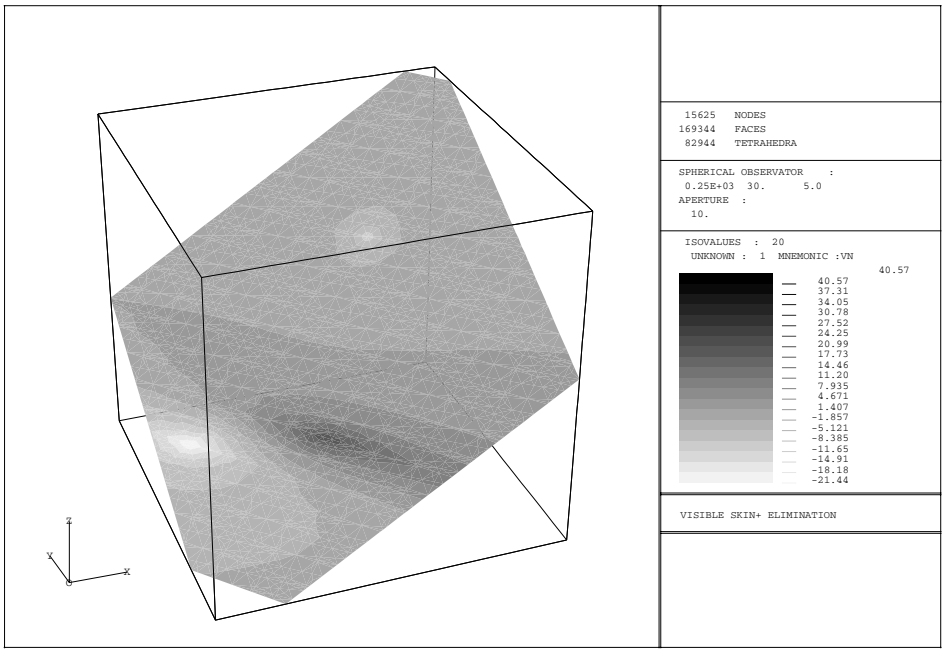


FIG. 6.5. *Attenuation field* ($h = 1/24\,\mathrm{m}$).

7. Optimal location of actuators. In the previous sections, the positions of the sensors (microphones) were given and the positions of the actuators (loudspeakers) were given, too. Then we have used the complex amplitudes of the actuators as the unique control variable and have determined their optimal values with the objective of minimizing the pressure level at those points where the sensors were located. Now we assume that the positions of the actuators can also be chosen in certain subsets of the domain and we will try to determine those that minimize the same objective function as above, when the complex amplitudes are optimal with respect to these positions. This is the most important problem when a system of active control of sound has to be implemented to reduce noise in an enclosure. It can also be formulated as an optimal control problem.

In this case the *control variables* are the complex amplitudes (modulus and phases) and the positions of the actuators,

$$\boldsymbol{u} = (u_1, \ldots, u_N) \in \mathbb{C}^N \quad \text{and} \quad \boldsymbol{y} = (y_1, \ldots, y_N) \in \Omega^N,$$

respectively, which define the secondary source by means of (2.2).

We consider the *set of admissible controls* $U_{\mathrm{ad}} \times Y_{\mathrm{ad}} \subset \mathbb{C}^N \times \Omega^N$, where $U_{\mathrm{ad}} \subset \mathbb{C}^N$ and $Y_{\mathrm{ad}} \subset\subset (\Omega \setminus \{w_1, \ldots, w_M\})^N$ are closed convex subsets.

The *observation* $\boldsymbol{z}(\boldsymbol{u}, \boldsymbol{y})$ is again the set of pressure values at the microphones' locations $w_1, \ldots, w_M \in \Omega$. The *transfer function* is now

$$\begin{aligned} U_{\mathrm{ad}} \times Y_{\mathrm{ad}} &\longrightarrow \mathbb{C}^M, \\ (\boldsymbol{u}, \boldsymbol{y}) &\longmapsto \boldsymbol{z}(\boldsymbol{u}, \boldsymbol{y}) = (p(w_i), \ldots, p(w_M)), \end{aligned}$$

where, for each admissible set of values of the control variables, $\boldsymbol{u} \in U_{\mathrm{ad}}$ and $\boldsymbol{y} \in Y_{\mathrm{ad}}$, p denotes again the solution of the *state equation* (2.1) with f given by (2.2). Notice

that the sensors' locations are excluded from the domain of admissible locations for the actuators, to ensure the continuity of the transfer function.

The *cost function* is given again by

$$J(\boldsymbol{u},\boldsymbol{y}) := \frac{1}{2}\|\boldsymbol{z}(\boldsymbol{u},\boldsymbol{y})\|^2 + \frac{\nu}{2}\|\boldsymbol{u}\|^2,$$

and then the *optimal control problem* is as follows:

Find $(\boldsymbol{u}^{\rm op},\boldsymbol{y}^{\rm op}) \in U_{\rm ad}\times Y_{\rm ad}$ *such that*

$$J(\boldsymbol{u}^{\rm op},\boldsymbol{y}^{\rm op}) = \inf_{(\boldsymbol{u},\boldsymbol{y})\in U_{\rm ad}\times Y_{\rm ad}} J(\boldsymbol{u},\boldsymbol{y}). \tag{7.1}$$

The difficulty now is that the dependency of the state with respect to the additional control variables (the positions of actuators) is no longer affine. Thus the cost function may have many local minima, and therefore gradient-like methods are not suitable to solve the problem. In practice, the number of feasible locations is typically finite, and hence the optimization problem becomes an integer programming problem. In these cases one can use, for instance, genetic or simulated annealing algorithms. Once the optimal locations have been determined, it will be possible to improve them in some given neighborhoods by performing a local minimization using the gradient of the cost function with respect to the location of the actuators. As we will see below, the first order optimality conditions allow computing the gradient of the cost function through an adjoint state.

7.1. Existence of an optimal control. Optimality conditions. We recall that the pressure field defining the observations can be written in terms of the control variables by means of (5.1): $p = p_0 + \sum_{i=1}^N u_i G^{y_i}$. Then the transfer function is affine with respect to $\boldsymbol{u}$, although it is nonlinear with respect to $\boldsymbol{y}$.

Therefore, the optimal control problem (7.1) has a solution as a direct consequence of the following facts:

- the function $J(\boldsymbol{u},\boldsymbol{y})$ is continuous in $U_{\rm ad}\times Y_{\rm ad}$;
- the set $Y_{\rm ad}$ is compact;
- $U_{\rm ad}$ is a closed set; and
- $J(\boldsymbol{u},\boldsymbol{y}) \to \infty$ when $\|\boldsymbol{u}\| \to \infty$ $\forall \boldsymbol{y} \in Y_{\rm ad}$.

In what follows we deduce the optimality condition for a local minimum of the cost function $J(\boldsymbol{u},\boldsymbol{y})$ given by (7.1). We notice that in the present case J is convex with respect to the amplitudes $\boldsymbol{u}$ but not with respect to the actuators' positions $\boldsymbol{y}$. Therefore, this optimality condition will be necessary but not sufficient.

The cost function can be written explicitly in terms of the control variables as follows:

$$J(\boldsymbol{u},\boldsymbol{y}) = \frac{1}{2}\sum_{k=1}^M |p(w_k)|^2 + \frac{\nu}{2}\|\boldsymbol{u}\|^2 = \frac{1}{2}\sum_{k=1}^M \left| p_0(w_k) + \sum_{i=1}^N u_i G^{y_i}(w_k)\right|^2 + \frac{\nu}{2}\sum_{i=1}^N |u_i|^2.$$

This function is differentiable in $\mathbb{C}^N \times (\Omega\setminus\{w_1,\ldots,w_M\})^N$. Hence, it is well known that if it attains a local minimum in the convex subset $U_{\rm ad}\times Y_{\rm ad}$ at $(\boldsymbol{u}^{\rm op},\boldsymbol{y}^{\rm op})$, then the following inequality holds:

$$DJ(\boldsymbol{u}^{\rm op},\boldsymbol{y}^{\rm op})(\boldsymbol{u}-\boldsymbol{u}^{\rm op},\boldsymbol{y}-\boldsymbol{y}^{\rm op}) \geq 0 \quad \forall (\boldsymbol{u},\boldsymbol{y}) \in U_{\rm ad}\times Y_{\rm ad}.$$

Since J depends on the complex variables $u_1, \ldots, u_N$, some care must be taken to compute its differential. Indeed, by using that

$$D\left(\frac{1}{2}\|\boldsymbol{v}\|^2\right)(\delta\boldsymbol{v}) = \mathrm{Re}(\bar{\boldsymbol{v}}^{\mathrm{t}}\delta\boldsymbol{v}) = \mathrm{Re}(\bar{\delta\boldsymbol{v}}^{\mathrm{t}}\boldsymbol{v}) \quad \forall \boldsymbol{v}, \delta\boldsymbol{v} \in \mathbb{C}^M \text{ or } \mathbb{C}^N,$$

straightforward computations lead to

$$D_{\boldsymbol{u}}J(\boldsymbol{u},\boldsymbol{y})(\delta\boldsymbol{u}) = \mathrm{Re}\left\{\sum_{i=1}^{N}\sum_{k=1}^{M}\left[p(w_k)\frac{\partial\bar{p}(w_k)}{\partial u_i}\,\bar{\delta u_i}\right] + \nu\sum_{i=1}^{N}u_i\bar{\delta u_i}\right\} \tag{7.2}$$

and

$$D_{\boldsymbol{y}}J(\boldsymbol{u},\boldsymbol{y})(\delta\boldsymbol{y}) = \mathrm{Re}\left[\sum_{i=1}^{N}\sum_{k=1}^{M}p(w_k)\nabla_{y_i}\bar{p}(w_k)\cdot\delta y_i\right] \tag{7.3}$$

$\forall(\boldsymbol{u},\boldsymbol{y}) \in U_{\mathrm{ad}} \times Y_{\mathrm{ad}}$, $\forall\delta\boldsymbol{u} \in \mathbb{C}^N$, and $\forall\delta\boldsymbol{y} \in (\mathbb{R}^n)^N$. As a consequence of all this we obtain the optimality condition of the following theorem.

THEOREM 7.1. *If* $(\boldsymbol{u}^{\mathrm{op}}, \boldsymbol{y}^{\mathrm{op}}) \in U_{\mathrm{ad}} \times Y_{\mathrm{ad}}$ *is a solution of the control problem* (7.1)*, then it satisfies*

$$\mathrm{Re}\left\{\sum_{i=1}^{N}\sum_{k=1}^{M}\left[p(w_k)\frac{\partial\bar{p}(w_k)}{\partial u_i}\,(\bar{u}_i - \bar{u}_i^{\mathrm{op}}) + p(w_k)\nabla_{y_i}\bar{p}(w_k)\cdot(y_i - y_i^{\mathrm{op}})\right] + \nu\sum_{i=1}^{N}u_i^{\mathrm{op}}\,(\bar{u}_i - \bar{u}_i^{\mathrm{op}})\right\} \geq 0 \quad \forall(\boldsymbol{u},\boldsymbol{y}) \in U_{\mathrm{ad}} \times Y_{\mathrm{ad}}.$$

Standard duality arguments can be used to compute the gradient of the cost function. To this aim, given $(\boldsymbol{u},\boldsymbol{y}) \in U_{\mathrm{ad}} \times Y_{\mathrm{ad}}$, we recall the *state equation*,

$$\begin{cases} -\Delta p - \left(\dfrac{\omega}{c}\right)^2 p = \displaystyle\sum_{i=1}^{N} u_i\delta_{y_i} & \text{in } \Omega, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = \dfrac{i\omega\rho}{Z(\omega)}\,p & \text{on } \Gamma_{\mathrm{Z}}, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = g & \text{on } \Gamma_{\mathrm{N}}, \end{cases} \tag{7.4}$$

and introduce the *adjoint state equation*,

$$\begin{cases} -\Delta r - \left(\dfrac{\omega}{c}\right)^2 r = \displaystyle\sum_{k=1}^{M} p(w_k)\delta_{w_k} & \text{in } \Omega, \\ \dfrac{\partial r}{\partial \boldsymbol{n}} = -\dfrac{i\omega\rho}{\bar{Z}(\omega)}\,r & \text{on } \Gamma_{\mathrm{Z}}, \\ \dfrac{\partial r}{\partial \boldsymbol{n}} = 0 & \text{on } \Gamma_{\mathrm{N}}, \end{cases} \tag{7.5}$$

with p being the solution of the state equation (7.4). Then we have the following result.

THEOREM 7.2. *Let* $(\boldsymbol{u},\boldsymbol{y}) \in U_{\rm ad} \times Y_{\rm ad}$ *and* $(\delta\boldsymbol{u},\delta\boldsymbol{y}) \in \mathbb{C}^N \times (\mathbb{R}^n)^N$. *Let* p *be the solution of the state equation* (7.4) *and let* r *be the solution of the adjoint state equation* (7.5). *Then*

$$D_{\boldsymbol{u}}J(\boldsymbol{u},\boldsymbol{y})(\delta\boldsymbol{u}) = \mathrm{Re}\left\{\sum_{i=1}^{N}[r(y_i)+\nu u_i]\bar{\delta u}_i\right\}$$

and

$$D_{\boldsymbol{y}}J(\boldsymbol{u},\boldsymbol{y})(\delta\boldsymbol{u}) = \mathrm{Re}\left[\sum_{i=1}^{N}\bar{u}_i\nabla r(y_i)\cdot\delta y_i\right].$$

Proof. Let G^{w_k} be the solution of problem (3.6) with $y = w_k$, $k = 1,\dots,M$. Then

$$r = \sum_{k=1}^{M} p(w_k)\bar{G}^{w_k}.$$

On the other hand, from (5.1), for $w_k \neq y_i$, $i = 1,\dots,N$, $k = 1,\dots,M$,

$$\frac{\partial\bar{p}(w_k)}{\partial u_i} = \frac{\partial}{\partial u_i}\left[\bar{p}_0(w_k) + \sum_{i=1}^{N} u_i\bar{G}^{y_i}(w_k)\right] = \bar{G}^{y_i}(w_k) = \bar{G}^{w_k}(y_i)$$

and

$$\nabla_{y_i}\bar{p}(w_k) = \bar{u}_i\nabla_{y_i}\bar{G}^{y_i}(w_k) = \bar{u}_i\nabla\bar{G}^{w_k}(y_i),$$

where we have used the symmetry of $G^y(w)$ with respect to y and w in Ω.

Consequently, from (7.2) and (7.3) we have

$$\begin{aligned} D_{\boldsymbol{u}}J(\boldsymbol{u},\boldsymbol{y})(\delta\boldsymbol{u}) &= \mathrm{Re}\left\{\sum_{i=1}^{N}\sum_{k=1}^{M}[p(w_k)\bar{G}^{w_k}(y_i)\,\bar{\delta u}_i] + \nu\sum_{i=1}^{N}u_i\bar{\delta u}_i\right\} \\ &= \mathrm{Re}\left\{\sum_{i=1}^{N}[r(y_i)+\nu u_i]\bar{\delta u}_i\right\} \end{aligned}$$

and

$$D_{\boldsymbol{y}}J(\boldsymbol{u},\boldsymbol{y})(\delta\boldsymbol{y}) = \mathrm{Re}\left[\sum_{i=1}^{N}\sum_{k=1}^{M}p(w_k)\bar{u}_i\nabla\bar{G}^{w_k}(y_i)\cdot\delta y_i\right] = \mathrm{Re}\left[\sum_{i=1}^{N}\bar{u}_i\nabla r(y_i)\cdot\delta y_i\right].$$

Thus we conclude the proof. ∎

As a consequence of this theorem, we can write the optimality condition of problem (7.1) in terms of the solution p of the state equation and the solution r of the adjoint state equation. Thus we obtain the following *Euler inequality*:

$$\mathrm{Re}\left(\sum_{i=1}^{N}\{[r(y_i^{\rm op})+\nu u_i^{\rm op}](\bar{u}_i-\bar{u}_i^{\rm op}) + \bar{u}_i^{\rm op}\,\nabla r(y_i^{\rm op})\cdot(y_i-y_i^{\rm op})\}\right) \geq 0$$
$$\forall(\boldsymbol{u},\boldsymbol{y}) \in U_{\rm ad}\times Y_{\rm ad}.$$

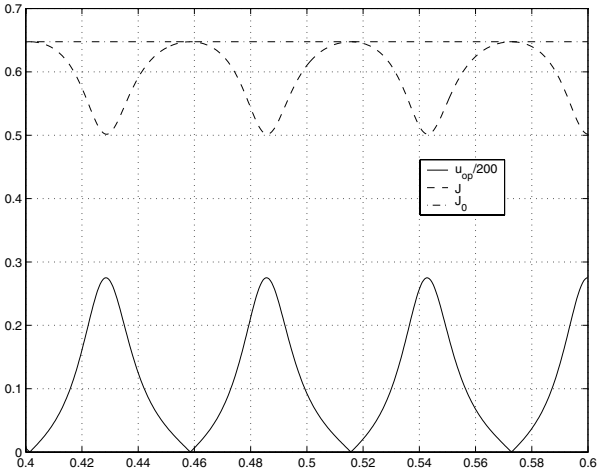


FIG. 7.1. *J and $u_{\rm op}$ as functions of the loudspeaker position* ($\omega = 18700\,\mathrm{s}^{-1}$).

7.2. Numerical experiments. In this section we present two numerical tests. The goal of the first one is to show that many local minima can actually arise. It is a one-dimensional problem, and then the Helmholtz equation becomes

$$\begin{cases} -\dfrac{d^2p}{dx^2} - \left(\dfrac{\omega}{c}\right)^2 p = f, & x \in (a,b), \\ \dfrac{dp}{dx} = \dfrac{i\omega\rho}{Z(\omega)}\, p & \text{at } x = a, \\ \dfrac{dp}{dx} = g & \text{at } x = b. \end{cases}$$

This one-dimensional equation can be easily solved when the secondary source is a linear combination of Dirac delta measures:

$$f = \sum_{i=1}^{N} u_i \delta_{y_i}, \quad u_i \in \mathbb{C},\ y_i \in (a,b),\ i = 1, \ldots, N.$$

Thus, we can determine the corresponding optimal amplitudes and then the optimal value of the cost function.

First, we consider the following data:

- the domain is the segment $[0,1]\,\mathrm{m}$;
- the physical parameters are $\rho = 1\,\mathrm{kg\,m^{-3}}$, $c = 340\,\mathrm{m\,s^{-1}}$, and $\omega = 18700\,\mathrm{s^{-1}}$;
- the amplitude of the primary source is $g = 1\,\mathrm{kg\,m^{-2}\,s^{-2}}$ at $x = 1\,\mathrm{m}$;
- the wall impedance at $x = 0$ is $Z = 34 \times 10^7 + 34 \times 10^3\, i\,\mathrm{kg\,m^{-2}\,s^{-1}}$;
- there is one actuator located at any point in the segment $Y_{\rm ad} = [0.4, 0.6]\,\mathrm{m}$;
- there are 4 sensors at the points $w_1 = 0.15\,\mathrm{m}$, $w_2 = 0.25\,\mathrm{m}$, $w_3 = 0.65\,\mathrm{m}$, and $w_4 = 0.85\,\mathrm{m}$;
- the admissible set of amplitudes is $U_{\rm ad} = \mathbb{C}$ and the weighting factor is $\nu = 0$.

For each position of the actuator in $Y_{\rm ad}$, we compute the optimal amplitude and represent the corresponding value of the cost function. Figure 7.1 shows this function. We observe several local minima. Furthermore, in this case, the values of J at all these minima are the same and correspond to maximum values of the optimal amplitude.

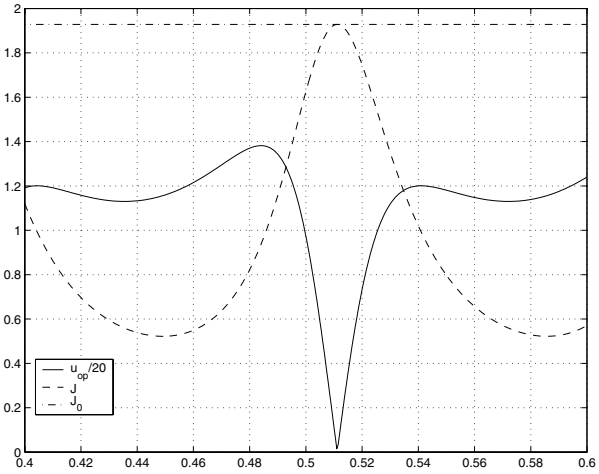


FIG. 7.2. *J and $u_{\rm op}$ as functions of the loudspeaker position* ($\omega = 7820\,\mathrm{s}^{-1}$).

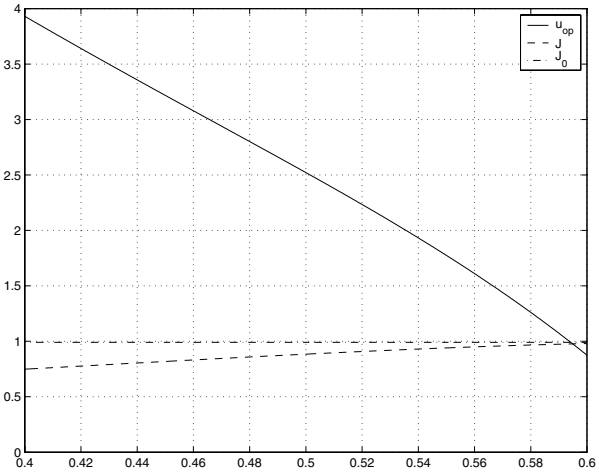


FIG. 7.3. *J and $u_{\rm op}$ as functions of the loudspeaker position* ($\omega = 1700\,\mathrm{s}^{-1}$).

Second, we analyze the system as the angular frequency decreases. Figures 7.2 and 7.3 show similar graphs for $\omega = 7820\,\mathrm{s}^{-1}$ and $\omega = 1700\,\mathrm{s}^{-1}$, respectively.

We observe in Figure 7.2 that the local minima of the cost functional do not necessarily coincide with local maxima of the optimal amplitude.

We also notice that the number of local minima diminishes with the angular frequency. For instance, in the case of Figure 7.3, J has only one minimum in the interval $Y_{\rm ad} = [0.4, 0.6]\,\mathrm{m}$.

Figure 7.4 shows the corresponding graph for the same frequency $\omega = 1700\,\mathrm{s}^{-1}$, when the admissible set of locations for the actuator is the whole domain of the problem: $Y_{\rm ad} = [0, 1]\,\mathrm{m}$. In this case it can be seen that, as expected, complete attenuation is attained as the actuator gets close to the primary source at $x = 1$.

The second test corresponds to a three-dimensional enclosure. We use the simu-

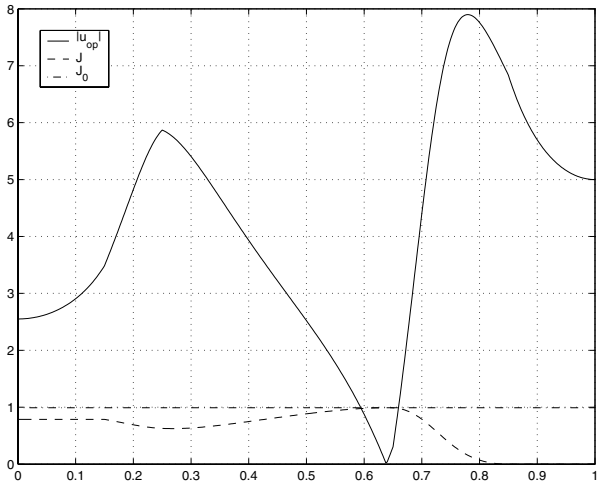


FIG. 7.4. *J and $u_{\rm op}$ as functions of the loudspeaker position* ($\omega = 1700\,\mathrm{s}^{-1}$, $Y_{\rm ad} = [0,1]\,\mathrm{m}$).

TABLE 7.1
Possible locations of the actuators.

Coordinates (m)	Coordinates (m)	Coordinates (m)	Coordinates (m)
(0.1,0.1,0.1)	(0.5,0.4,0.1)	(0.5,0.1,0.3)	(0.8,0.8,0.3)
(0.9,0.9,0.1)	(0.5,0.6,0.1)	(0.5,0.9,0.3)	(0.6,0.4,0.4)
(0.1,0.9,0.1)	(0.1,0.5,0.3)	(0.5,0.5,0.3)	(0.4,0.6,0.4)
(0.9,0.1,0.1)	(0.9,0.5,0.3)	(0.2,0.2,0.3)	(0.6,0.6,0.4)

lated annealing algorithm (see, for instance, [1]) to determine the optimal location of loudspeakers among a given finite number of feasible ones.

The data of the test are the following:

- domain $\Omega = [0,1]\,\mathrm{m} \times [0,1]\,\mathrm{m} \times [0,1]\,\mathrm{m}$;
- physical parameters $\rho = 1\,\mathrm{kg\,m^{-3}}$, $c = 340\,\mathrm{m\,s^{-1}}$, and $\omega = 1360\,\mathrm{s^{-1}}$;
- wall impedance at $z = 1\,\mathrm{m}$, $Z = (102 + 340\,i) \times 10^3\,\mathrm{kg\,m^{-2}\,s^{-1}}$;
- primary source at wall $z = 0$ with amplitude $g(x,y,0) = \mathrm{e}^{iy}\,\mathrm{kg\,m^{-2}\,s^{-2}}$;
- the rest of the walls are perfectly rigid;
- we have to locate 8 actuators and consider 16 possible locations which are shown in Table 7.1;
- there are 10 sensors and their positions are shown in Table 7.2;
- admissible set of amplitudes $U_{\rm ad} = \mathbb{C}^8$ and weighting factor $\nu = 0$.

We have used a mesh like that of Figure 6.1 for $h = 1/24\,\mathrm{m}$.

The attenuation is computed again by

$$\text{Attenuation (dB)} = -10\log_{10}\left(\frac{J(\boldsymbol{u}^{\rm op},\boldsymbol{y}^{\rm op})}{J(\boldsymbol{0},\boldsymbol{y})}\right).$$

Notice that $J(\boldsymbol{0},\boldsymbol{y})$ is the value of the cost function with no control, and hence it does not depend on $\boldsymbol{y}$. In Table 7.3 we show the attenuation obtained for different executions of the simulated annealing algorithm. We also include the value obtained with the exhaustive search, i.e., by computing the cost function for all of the possible $\binom{16}{8} = 12870$ configurations.

TABLE 7.2
Positions of the sensors.

Coordinates (m)	Coordinates (m)
(0.2,0.2,0.6)	(0.5,0.1,0.8)
(0.2,0.8,0.6)	(0.1,0.5,0.8)
(0.8,0.2,0.6)	(0.5,0.9,0.8)
(0.8,0.8,0.6)	(0.7,0.5,0.6)
(0.9,0.5,0.8)	(0.3,0.5,0.8)

TABLE 7.3
Simulated annealing: number of iterations and optimal attenuation.

No. of Iterations	Attenuation (dB)
569	73.7
599	73.7
710	73.7
785	73.7
800	68.6
1201	73.7
1498	73.7
12870 (exhaustive)	73.7

Remark 7.3. Some experiments show that the use of a basis consisting of rigid cavity vibration modes can be more efficient in terms of computer effort (CPU time and memory) than the purely finite element technique introduced in this paper. This is due to the fact that the number of vibration modes needed to obtain accurate results for low frequencies (which is the typical case in active control of sound) is not large. Then the time to calculate these first modes by solving the corresponding eigenvalue problem by finite element methods, together with that to solve the Helmholtz problems in this small vibration modes basis, can be significantly less than the time needed to solve the same number of Helmholtz problem in the large finite element basis. This approach will be reported elsewhere.

Acknowledgments. The authors wish to thank Dr. Pedro Cobo from IAC/CSIC, Madrid, Spain, for very useful discussions.

This work was partially done while R. Rodríguez participated in the program *Computational Challenges in PDEs* at the Newton Institute of the University of Cambridge in June 2003. He is grateful for the support and stimulating atmosphere at the institute.

Last but not least, the authors wish to thank Ricardo Durán and Ricardo Nochetto for very helpful discussions.

REFERENCES

[1] K. H. BAEK AND S. J. ELLIOT, *Natural algorithms for choosing source locations in active control systems*, J. Sound Vibration, 186 (1995), pp. 245–267.

[2] S. BRENNER AND L. R. SCOTT, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, Berlin, Heidelberg, 1996.

[3] L. L. BERANEK AND I. L. VER, *Noise and Vibration Control Engineering. Principles and Applications*, John Wiley, New York, 1992.

[4] E. CASAS, *L^2 estimates for the finite element method for the Dirichlet problem with singular data*, Numer. Math., 47 (1985), pp. 627–632.

[5] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains: Smoothness and Asymptotics of Solutions*, Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.
[6] R. Dautray and J. L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Springer-Verlag, Berlin, 1984–1985.
[7] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, RI, 1998.
[8] P. Gamallo, *Contribución al estudio matemático de problemas de simulación elastoacústica y control activo del ruido*, Ph.D. thesis, Universidade de Santiago de Compostela, Spain, 2002.
[9] P. Grisvard, *Elliptic Problems for Non-smooth Domains*, Pitman, Boston, 1985.
[10] W. Hackbusch, *Elliptic Partial Differential Equations, Theory and Numerical Treatment*, Springer-Verlag, Berlin, Heidelberg, New York, 1992.
[11] P. A. Nelson and S. J. Elliot, *Active Control of Sound*, Academic Press, London, 1999.
[12] A. Schatz, *An observation concerning Ritz–Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.
[13] A. Schatz and L. Wahlbin, *Interior maximum norm estimates for finite element methods*, Math. Comp., 31 (1977), pp. 414–442.
[14] R. Scott, *Finite element convergence for singular data*, Numer. Math., 21 (1973), pp. 317–327.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 466–476

A BOGOLYUBOV-TYPE THEOREM WITH A NONCONVEX CONSTRAINT IN BANACH SPACES*

F. S. DE BLASI†, G. PIANIGIANI‡, AND A. A. TOLSTONOGOV§

Abstract. We prove an analogue of the classical Bogolyubov theorem, with a nonconvex constraint. In the case we consider, the constraint is the solution set of a Cauchy problem for a differential inclusion with a nonconvex right-hand side satisfying a Lipschitz condition. Our approach is based on a relaxation argument, as in the Filippov–Wazewski theorem.

Key words. differential inclusion, Banach space, minimization, relaxation

AMS subject classifications. Primary, 49J24; Secondary, 34A60

DOI. 10.1137/S0363012903423156

1. Introduction and main result.

In relaxation theory for variational problems, the classical Bogolyubov theorem [3] has been extended in several directions by many authors including Young [16], MacShane [10], Warga [15], Ioffe [7], Ioffe and Tikhomirov [8, 9], and Ekeland and Temam [5]. More recently an analogue of Bogolyubov's theorem with a convex constraint has been obtained by Suslov [12]. The aim of the present paper is to present a Bogolyubov-type theorem in a case in which the constraint is not necessarily convex. In our approach we use a relaxation argument in the spirit of the classical Filippov–Wazewski theorem (see [1, Chap. 2]).

Throughout the paper, I is the unit interval $[0,1]$ with Lebesgue measure μ and σ-algebra Σ of the μ-measurable subsets of I, $\mathbb{E}$ is a separable Banach space with norm $||\cdot||$, and $\mathcal{K}(\mathbb{E})$ is the space of all nonempty compact subsets of $\mathbb{E}$ endowed with the Hausdorff metric h. Further, $\Sigma \otimes \mathcal{B}(\mathbb{E})$ is the σ-algebra of $I \times \mathbb{E}$ generated by the sets $A \times B$, with $A \in \Sigma$ and $B \in \mathcal{B}(\mathbb{E})$, where $\mathcal{B}(\mathbb{E})$ is the σ-algebra of the Borel subsets of $\mathbb{E}$. We set $B(a,r) = \{x \in \mathbb{E} \mid \|x - a\| \le r\}$ and denote by $\overline{\mathbb{R}}$ the set of the extended real numbers.

Let M be a metric space. A multifunction $F : M \to K(\mathbb{E})$ is said to be *upper semicontinuous* (resp., *lower semicontinuous*) if for each closed $C \subset \mathbb{E}$ the set $\{x \in M \mid F(x) \cap C \neq \emptyset\}$ (resp., $\{x \in M \mid F(x) \subset C\}$) is closed in $\mathbb{E}$. F is said to be *continuous* when it is both lower and upper semicontinuous. In our setting, the above definitions are equivalent to the corresponding definitions of upper semicontinuity, lower semicontinuity, and continuity in the sense of Hausdorff. A multifunction $F : C \to K(\mathbb{E})$ with closed values is said to be *measurable* if for each closed set $C \subset \mathbb{E}$ the set $\{t \in I \mid F(t) \cap C \neq \emptyset\} \in \Sigma$.

Given $F : I \times \mathbb{E} \to \mathcal{K}(\mathbb{E})$, consider the Cauchy problem

$$(\mathrm{C}_F) \qquad \dot{x} \in F(t,x), \quad x(0) = x_0,$$

*Received by the editors February 22, 2003; accepted for publication (in revised form) January 13, 2004; published electronically July 23, 2004. This research was supported in part by the Russian Foundation for Basic Research (grant 03-01-00203).

http://www.siam.org/journals/sicon/43-2/42315.html

†Centro Vito Volterra, Dipartimento di Matematica, Università di Roma II, Via della Ricerca Scientifica, 00133 Roma, Italy (deblasi@mat.uniromaz.it).

‡Dipartimento di Matematica per le Decisioni, Università di Firenze, Via Lombroso, 6/17, 50134 Firenze, Italy (giulio.pianigiani@dmd.unifi.it).

§Institute for System Dynamics and Control Theory, Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia (aatol@icc.ru).

and the convexified Cauchy problem

$$(\mathrm{C}_{\overline{\mathrm{co}}\,F}) \qquad \dot{x} \in \overline{\mathrm{co}}\,F(t,x), \quad x(0) = x_0,$$

where $x_0 \in \mathbb{E}$ and $\overline{\mathrm{co}}\,F(t,x)$ denotes the closed convex hull of $F(t,x)$.

By a *solution* of the Cauchy problem (C_F) (resp., $(\mathrm{C}_{\overline{\mathrm{co}}\,F})$) we mean an absolutely continuous function $x : I \to \mathbb{E}$ given by $x(t) = x_0 + \int_0^t u(s)ds$, where $u : I \to \mathbb{E}$ is Bochner integrable, satisfying the differential inclusion (C_F) (resp., $(\mathrm{C}_{\overline{\mathrm{co}}\,F})$) almost everywhere (a.e.) in I. We denote by $\mathcal{M}_F$ (resp., $\mathcal{M}_{\overline{\mathrm{co}}\,F}$) the solution set of the Cauchy problem (C_F) (resp., $(\mathrm{C}_{\overline{\mathrm{co}}\,F})$). $\mathcal{M}_F$ and $\mathcal{M}_{\overline{\mathrm{co}}\,F}$ are equipped with the induced metric of $C(I,\mathbb{E})$, the Banach space of the continuous functions from I to $\mathbb{E}$.

Let $F : I \times \mathbb{E} \to \mathcal{K}(\mathbb{E})$ and $g : I \times \mathbb{E} \times \mathbb{E} \to \mathbb{R}$ be given. We say that F satisfies condition (H_F) if

$$(\mathrm{H}_F) \qquad \begin{cases} \text{(i) } t \to F(t,x) \text{ is measurable for every } x \in \mathbb{E}, \\ \text{(ii) } h(F(t,x),F(t,y)) \le k(t)\|x-y\| \quad \text{for every } x,y \in \mathbb{E},\ t \in I \text{ a.e.}, \\ \text{(iii) } h(F(t,0),0) \le m(t) \quad \text{for } t \in I \text{ a.e.}, \end{cases}$$

where $m, k \in L^1(I,\mathbb{R})$ and $m(t), k(t) > 0$.

We say that g satisfies condition (H_g) if

$$(\mathrm{H}_g) \qquad \begin{cases} \text{(i) } t \to g(t,x,u) \text{ is measurable for every } x,u \in \mathbb{E}, \\ \text{(ii) } (x,u) \to g(t,x,u) \text{ is continuous for } t \in I \text{ a.e.}, \\ \text{(iii) } |g(t,x,u)| \le a(t) + b(t)||x|| + c||u|| \quad \text{for every } x,y \in \mathbb{E},\ t \in I \text{ a.e.}, \end{cases}$$

where $a, b \in L^1(I,\mathbb{R})$, and $a(t), b(t), c > 0$.

In what follows we will use the following Banach space version, proved in [13], of the classical relaxation theorem of Filippov and Wazewski.

Relaxation theorem. Let F satisfy (H_F). Then $\mathcal{M}_F$ is nonempty, $\mathcal{M}_{\overline{\mathrm{co}}\,F}$ is compact, and

$$\overline{\mathcal{M}}_F = \mathcal{M}_{\overline{\mathrm{co}}\,F}, \tag{1.1}$$

where $\overline{\mathcal{M}}_F$ denotes the closure of $\mathcal{M}_F$ in $C(I,\mathbb{E})$.

For any $f : \mathbb{E} \to \overline{\mathbb{R}}$, we denote by $f^{**} : \mathbb{E} \to \overline{\mathbb{R}}$ the bipolar of f (see [5, Chap. 1]). It is known that if f has an affine continuous minorant, then f^{**} coincides with the upper envelope of all affine continuous minorants of f. Hence f^{**} is lower semicontinuous and convex.

Given $F : I \times \mathbb{E} \to \mathcal{K}(\mathbb{E})$ and $g : I \times \mathbb{E} \times \mathbb{E} \to \mathbb{R}$, we define $g_F : I \times \mathbb{E} \times \mathbb{E} \to \overline{\mathbb{R}}$ by

$$g_F(t,x,u) = \begin{cases} g(t,x,u) & \text{if} \quad u \in F(t,x), \\ +\infty & \text{if} \quad u \notin F(t,x). \end{cases} \tag{1.2}$$

Further, we denote by $u \to g_F^{**}(t,x,u)$ the bipolar of the function $u \to g_F(t,x,u)$.

The following theorem is an analogue of Bogolyubov's theorem [3] with a nonconvex constraint.

THEOREM 1.1. *Let F, g satisfy (H_F) and (H_g). Then, for every $x \in \mathcal{M}_{\overline{\mathrm{co}}F}$, there exists a sequence $\{x_n\} \subset \mathcal{M}_F$ such that*

(i) *x_n converges to x in $C(I,\mathbb{E})$,*

(ii) $\lim_{n\to\infty} \sup_{t\in I} |\int_0^t (g_F^{**}(s,x(s),\dot{x}(s)) - g(s,x_n(s),\dot{x}_n(s)))ds| = 0$.

Observe that when F is merely continuous, both statements of Theorem 1.1 can fail, as is shown by the following example, which is patterned after Plis [11].

Example 1.2. Let $F : \mathbb{R}^2 \to \mathcal{K}(\mathbb{R}^2)$ and $g : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ be given by

$$F(x) = \{(-1, |x_1| + 2\sqrt{|x_2|}), (1, |x_1| + 2\sqrt{|x_2|})\}\ , \qquad g(x,u) = \|x\| + \|u\|,$$

where $x = (x_1, x_2)$, $u = (u_1, u_2)$, and $\|.\|$ is the Euclidean norm. Clearly F satisfies $(\mathrm{H}_F)_{\mathrm{i}}$ and $(\mathrm{H}_F)_{\mathrm{iii}}$ but not $(\mathrm{H}_F)_{\mathrm{ii}}$, while g satisfies (H_g).

As $F(0) = \{p^+, p^-\}$, where $p^+ = (1,0)$, $p^- = (-1,0)$, one has $g_F(0,u) = 1$ if $u \in \{p^+, p^-\}$, while $g_F(0,u) = +\infty$ if $u \notin \{p^+, p^-\}$. Further, $g_F^{**}(0,0) = 1$. In fact, $g_F^{**}(0,0) \geq 1$, for 1 is a minorant of $u \to g_F(0,u)$. On the other hand, $u \to g_F^{**}(0,u)$ is convex, and hence

$$g_F^{**}(0,0) = g_F^{**}\left(0, \frac{p^+ + p^-}{2}\right) \leq \frac{g_F^{**}(0,p^+) + g_F^{**}(0,p^-)}{2} \leq \frac{g_F(0,p^+) + g_F(0,p^-)}{2} = 1,$$

whence $g_F^{**}(0,0) = 1$.

Observe now that the function $x_0 = (0,0)$ is a solution of the convexified Cauchy problem $\dot{x} \in \overline{\mathrm{co}}\ F(x), x(0) = 0$, i.e., $x_0 \in \mathcal{M}_{\overline{\mathrm{co}}F}$. For an arbitrary $x \in \mathcal{M}_F$, $x(t) = (x_1(t), x_2(t))$, $t \in I$, one has $|\dot{x}_1(t)| = 1$ and $\dot{x}_2(t) = |x_1(t)| + 2\sqrt{|x_2(t)|}$ for $t \in I$ a.e., $x_1(0) = x_2(0) = 0$. Since $|x_1(t)| \neq 0$, $t \in I$ a.e., it follows that $x_2(t) \geq t^2$, and thus $||x(t)|| \geq t^2$. Moreover, $\|\dot{x}(t)\| \geq 1$, $t \in I$ a.e., and hence

$$\left|\int_0^t (g(x(s),\dot{x}(s)) - g_F^{**}(0,0))ds\right| = \int_0^t (\|x(s)\| + \|\dot{x}(s)\| - 1)ds \geq \int_0^t \|x(s)\|ds \geq \frac{t^3}{3}\ , \ t \in I.$$

Therefore, for the solution $x_0 \in \mathcal{M}_{\overline{\mathrm{co}}F}$ neither statement of Theorem 1.1 is satisfied.

It is known [14] that the equality (1.1) remains valid if $(\mathrm{H}_F)_{\mathrm{ii}}$ is replaced by

$$h(\overline{\mathrm{co}}\, F(t,x), \overline{\mathrm{co}}\, F(t,y)) \leq k(t)||x - y|| \ \text{ for every } x, y \in \mathbb{E},\ t \in I \text{ a.e.}, \tag{1.3}$$

where $k \in L^1(I, \mathbb{R})$, $k(t) > 0$.

The following example shows that property (ii) in Theorem 1.1 can fail if $H(F)_{\mathrm{ii}}$ is replaced by (1.3).

Example 1.3. Let $F : \mathbb{R} \to \mathcal{K}(\mathbb{R})$ and $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be given by

$$F(x) = \begin{cases} \{-1, 1\} & \text{if} \quad x \neq 0, \\ \{-1, -1/2, 1/2, 1\} & \text{if} \quad x = 0, \end{cases} \qquad g(x,u) = |x| + |u|.$$

Here F satisfies $(\mathrm{H}_F)_{\mathrm{i}}$, $(\mathrm{H}_F)_{\mathrm{iii}}$, and (1.3), while g satisfies (H_g), yet (ii) does not hold.

In fact, $x_0 = 0$ is a solution of the convexified Cauchy problem $\dot{x} \in \overline{\mathrm{co}}\ F(x)$, $x(0) = 0$, i.e., $x_0 \in \mathcal{M}_{\overline{\mathrm{co}}F}$. Moreover for every $x \in \mathcal{M}_F$ one has $|\dot{x}(t)| = 1$, $t \in I$ a.e., which implies $g(x(t), \dot{x}(t)) \geq 1$, $t \in I$ a.e. Since in addition $g_F^{**}(0,0) = 1/2$, it follows that

$$\left|\int_0^t (g(x(s),\dot{x}(s)) - g_F^{**}(0,0))ds\right| \geq \frac{t}{2}, \quad t \in I,$$

showing that (ii) cannot be satisfied.

Remark 1.4. Observe that the above Bogolyubov-type Theorem 1.1 is also an extension of the theorem of Filippov and Wazewski.

2. Lemmas. For $f : I \times \mathbb{E} \times \mathbb{E} \to \overline{\mathbb{R}}$ we denote by $\operatorname{epi} f(t,x)$ the *epigraph* of the function $u \to f(t,x,u)$, i.e.,

$$\operatorname{epi} f(t,x) = \{(u,\lambda) \in \mathbb{E} \times \mathbb{R} \mid f(t,x,u) \le \lambda\}.$$

In what follows we consider the Banach space $\mathbb{Y} = \mathbb{E} \times \mathbb{R}$ equipped with the norm $||(x,\lambda)|| = \max(||x||, |\lambda|)$ and denote by $h_{\mathbb{Y}}$ the Hausdorff metric of $\mathcal{K}(\mathbb{Y})$.

For F, g satisfying (H_F), (H_g) define $G : I \times \mathbb{E} \to \mathcal{K}(\mathbb{Y})$ by

$$G(t,x) = \{(u, g(t,x,u)) \in \mathbb{E} \times \mathbb{R} \mid u \in F(t,x)\}. \tag{2.1}$$

LEMMA 2.1. *Let F, g satisfy (H_F), (H_g). Then, for each $x \in \mathbb{E}$ and $t \in I$ a.e. we have*

(i) $(u,\lambda) \in \overline{\operatorname{co}}\, G(t,x)$ *implies* $u \in \overline{\operatorname{co}}\, F(t,x)$,

(ii) $u \in \overline{\operatorname{co}}\, F(t,x)$ *implies* $(u,\lambda) \in \overline{\operatorname{co}}\, G(t,x)$ *for some* $\lambda \in \mathbb{R}$.

Proof. (i) Let $(u,\lambda) \in \overline{\operatorname{co}}\, G(t,x)$, and let $\{\sum_{i=1}^{p_n} \alpha_n^i (u_n^i, \lambda_n^i)\}$, where $\sum_{i=1}^{p_n} \alpha_n^i = 1$, $\alpha_n^i > 0$, be a sequence of convex combinations of points $(u_n^i, \lambda_n^i) \in G(t,x)$, converging to (u,λ). As $u_n^i \in F(t,x)$, one has $\sum_{i=1}^{p_n} \alpha_n^i u_n^i \in \operatorname{co} F(t,x)$, and thus $u \in \overline{\operatorname{co}}\, F(t,x)$.

(ii) Let $u \in \overline{\operatorname{co}}\, F(t,x)$, and let $\{\sum_{i=1}^{p_n} \alpha_n^i u_n^i\}$ be a sequence of convex combinations of points $u_n^i \in F(t,x)$, converging to u. Clearly, $\sum_{i=1}^{p_n} \alpha_n^i (u_n^i, g(t,x,u_n^i)) \in \operatorname{co} G(t,x)$. In view of $(\mathrm{H}_g)_{\mathrm{iii}}$, the sequence $\{\sum_{i=1}^{p_n} \alpha_n^i g(t,x,u_n^i)\}$ is bounded, and thus, passing to a subsequence, one can assume that it converges to some $\lambda \in \mathbb{R}$. Hence $(u,\lambda) \in \overline{\operatorname{co}}\, G(t,x)$, completing the proof.

Property (ii) of the next lemma could be derived from known results on normal integrands (see Castaing and Valadier [4, Chap. 7]), yet a direct proof is provided to make the exposition self-contained.

LEMMA 2.2. *Let F, g satisfy (H_F), (H_g). Then, for each $x \in \mathbb{E}$ and $t \in I$ a.e. we have*

(i) $$g_F^{**}(t,x,u) = \begin{cases} \min\{\lambda \in \mathbb{R} \mid (u,\lambda) \in \overline{\operatorname{co}}\, G(t,x)\} & \text{if} \quad u \in \overline{\operatorname{co}}\, F(t,x), \\ +\infty \quad \text{if} \quad u \notin \overline{\operatorname{co}}\, F(t,x). \end{cases}$$

In particular, $(u, g_F^{**}(t,x,u)) \in \overline{\operatorname{co}}\, G(t,x)$ for every $u \in \overline{\operatorname{co}}\, F(t,x)$. Moreover,

(ii) *for every $\varepsilon > 0$ there is a closed set $I_\varepsilon \subset I$, with $\mu(I \setminus I_\varepsilon) < \varepsilon$, such that $g_F^{**}(t,x,u)$ restricted to $I_\varepsilon \times \mathbb{E} \times \mathbb{E}$ is lower semicontinuous.*

Proof. (i) Let $x \in \mathbb{E}$ and $t \in I$ a.e. be arbitrary. By $(\mathrm{H}_g)_{\mathrm{iii}}$ the function $u \to g_F(t,x,u)$ has an affine continuous minorant and hence by [5, Chap. 1],

$$\operatorname{epi} g_F^{**}(t,x) = \overline{\operatorname{co}}\, \operatorname{epi} g_F(t,x). \tag{2.2}$$

Let $u \in \mathbb{E}$ and suppose $g_F^{**}(t,x,u) < +\infty$. As $(u, g_F^{**}(t,x,u)) \in \operatorname{epi} g_F^{**}(t,x)$, then, by (2.2), there exists a sequence of convex combinations $\sum_{i=1}^{p_n} \alpha_n^i (u_n^i, \lambda_n^i)$ of points $(u_n^i, \lambda_n^i) \in \operatorname{epi} g_F(t,x)$, converging to $(u, g_F^{**}(t,x,u))$. Clearly $u_n^i \in F(t,x)$, for $g_F(t,x,u_n^i) \le \lambda_n^i < +\infty$, and thus $u \in \overline{\operatorname{co}}\, F(t,x)$. Consequently, $g_F^{**}(t,x,u) = +\infty$ for every $u \notin \overline{\operatorname{co}}\, F(t,x)$.

Let $u \in \overline{\operatorname{co}}\, F(t,x)$, and set

$$\mu = \min\{\lambda \in \mathbb{R} \mid (u,\lambda) \in \overline{\operatorname{co}}\, G(t,x)\}. \tag{2.3}$$

By virtue of Lemma 2.1(ii) and the compactness of $\overline{\operatorname{co}}\, G(t,x)$, the definition of μ is meaningful. We have $g_F^{**}(t,x,u) = \mu$. To show this, take any $\lambda \in \mathbb{R}$ so that $(u,\lambda) \in \overline{\operatorname{co}}\, G(t,x)$, and consider a sequence of convex combinations $\sum_{i=1}^{p_n} \alpha_n^i (u_n^i, \lambda_n^i)$ of points $(u_n^i, \lambda_n^i) \in G(t,x)$, where $\lambda_n^i = g(t,x,u_n^i)$, converging to (u,λ). As $u_n^i \in F(t,x)$,

one has $\lambda_n^i = g_F(t,x,u_n^i) \geq g_F^{**}(t,x,u_n^i)$. Thus, $\sum_{i=1}^{p_n} \alpha_n^i \lambda_n^i \geq \sum_{i=1}^{p_n} \alpha_n^i g_F^{**}(t,x,u_n^i)$, which implies $\lambda \geq g_F^{**}(t,x,u)$, by the convexity of $g_F^{**}(t,x,u)$ with respect to u. As $(u,\lambda) \in \overline{\mathrm{co}}\, G(t,x)$ and λ is arbitrary, it follows that $\mu \geq g_F^{**}(t,x,u)$.

On the other hand, $(u, g_F^{**}(t,x,u)) \in \operatorname{epi} g_F^{**}(t,x)$ and hence, by (2.2), there exists a sequence of convex combinations $\sum_{i=1}^{p_n} \alpha_n^i (u_n^i, \lambda_n^i)$ of points $(u_n^i, \lambda_n^i) \in \operatorname{epi} g_F(t,x)$, converging to $(u, g_F^{**}(t,x,u))$. Clearly, $+\infty > \lambda_n^i \geq g_F(t,x,u_n^i) = g(t,x,u_n^i)$ and $u_n^i \in F(t,x)$, whence $\sum_{i=1}^{p_n} \alpha_n^i \lambda_n^i \geq \sum_{i=1}^{p_n} \alpha_n^i g(t,x,u_n^i)$. Thus, setting $\gamma = \liminf_{n\to\infty} \sum_{i=1}^{p_n} \alpha_n^i g(t,x,u_n^i)$, one obtains $g_F^{**}(t,x,u) \geq \gamma$. Furthermore, $(u_n^i, g(t,x,u_n^i)) \in G(t,x)$, which implies that $(\sum_{i=1}^{p_n} \alpha_n^i u_n^i, \sum_{i=1}^{p_n} \alpha_n^i g(t,x,u_n^i)) \in \mathrm{co} G(t,x)$, and hence $(u,\gamma) \in \overline{\mathrm{co}}\, G(t,x)$. Then, by (2.3), $\gamma \geq \mu$ and, a fortiori, $g_F^{**}(t,x,u) \geq \mu$. Therefore $g_F^{**}(t,x,u) = \mu$, and (i) is proved.

(ii). Let $\varepsilon > 0$. F, g satisfy (H_F) and (H_g), and thus, by [6], there is a closed set $I_\varepsilon \subset I$, with $\mu(I \setminus I_\varepsilon) < \varepsilon$, such that the restrictions of $F(t,x)$ and $g(t,x,u)$ to $I_\varepsilon \times \mathbb{E}$ and $I_\varepsilon \times \mathbb{E} \times \mathbb{E}$ are both continuous. Without loss of generality, the restrictions to I_ε of the functions $k(t), m(t), a(t), b(t)$ can also be supposed to be continuous. For every closed $V \subset I \times \mathbb{R}$ the set $Z = \{(t,x) \in I_\epsilon \times \mathbb{E} \mid G(t,x) \cap V \neq \emptyset\}$ is closed. To show this, consider any sequence of points $(t_n, x_n) \in Z$ converging to $(t,x) \in I_\epsilon \times \mathbb{E}$, and take, for each $n \in \mathbb{N}$, a point $(u_n, g(t_n,x_n,u_n)) \in G(t_n,x_n) \cap V$. As $u_n \in F(t_n,x_n)$ and $F(t_n,x_n)$ converges to $F(t,x)$, a compact set, there exists a subsequence, say $\{u_n\}$, which converges to some $u \in F(t,x)$. Thus $(u, g(t,x,u)) \in G(t,x) \cap V$, and thus the set Z is closed, which proves that $G(t,x)$, and so also $\overline{\mathrm{co}}\, G(t,x)$, restricted to $I_\epsilon \times \mathbb{E}$, are upper semicontinuous.

The function $g_F^{**}(t,x,u)$ restricted to $I_\varepsilon \times \mathbb{E} \times \mathbb{E}$ is lower semicontinuous. It suffices to show that, for every $(t,x,u) \in I_\varepsilon \times \mathbb{E} \times \mathbb{E}$ and any sequence $\{(t_n,x_n,u_n)\} \subset I_\varepsilon \times \mathbb{E} \times \mathbb{E}$ converging to (t,x,u), one has

$$g_F^{**}(t,x,u) \leq \lambda, \quad \text{where} \quad \lambda = \liminf_{n\to\infty} g_F^{**}(t_n,x_n,u_n). \tag{2.4}$$

Suppose $\lambda < +\infty$ (the case $\lambda = +\infty$ is trivial). By virtue of $(\mathrm{H}_g)_{\mathrm{iii}}$, there exists $M \in \mathbb{R}$ such that for all $n \in \mathbb{N}$ large enough one has $g(t_n,x_n,u_n) \geq M$, which implies $g_F^{**}(t_n,x_n,u_n) \geq M$, and so $\lambda > -\infty$. As λ is finite, passing to a subsequence (without changing notation) one can assume that each $g_F^{**}(t_n,x_n,u_n)$ is finite, whence by (i), $u_n \in \overline{\mathrm{co}}\, F(t_n,x_n)$, and thus $(u_n, g_F^{**}(t,x_n,u_n)) \in \overline{\mathrm{co}}\, G(t_n,x_n)$ for all $n \in \mathbb{N}$. Since the map $\overline{\mathrm{co}}\, G(t,x)$ restricted to $I_\varepsilon \times \mathbb{E}$ is upper semicontinuous, it follows that $(u,\lambda) \in \overline{\mathrm{co}}\, G(t,x)$. Then, by the definition of $g_F^{**}(t,x,u)$, (2.4) holds; that is, the function $g_F^{**}(t,x,u)$ restricted to $I_\varepsilon \times \mathbb{E} \times \mathbb{E}$ is lower semicontinuous, proving (ii). This completes the proof.

Remark 2.3. By Lemma 2.2(ii) it follows that, for each $x \in \mathcal{M}_{\overline{\mathrm{co}}\, F}$ and any measurable selector $u(t) \in \overline{\mathrm{co}}\, F(t,x(t))$, the function $t \to g_F^{**}(t,x(t),u(t))$ is measurable on I.

LEMMA 2.4. *Let F satisfy (H_F), and let g satisfy $(\mathrm{H}_g)_{\mathrm{i}}$, $(\mathrm{H}_g)_{\mathrm{iii}}$, and*

$$|g(t,x,u) - g(t,y,v)| \leq l(||x-y|| + ||u-v||) \quad \textit{for every } x,y,u,v \in \mathbb{E},\ \ t \in I \textit{ a.e.}, \tag{2.5}$$

where $l > 0$. Then the multifunction $G : I \times \mathbb{E} \to \mathcal{K}(\mathbb{Y})$ given by (2.1) *has the following properties:*

(i) *the map $t \to G(t,x)$ is measurable for every $x \in \mathbb{E}$,*
(ii) *$h_{\mathbb{Y}}(G(t,x),G(t,y)) \leq L(t)||x-y||$ for every $x,y, \in \mathbb{E},\ t \in I$ a.e.,*
(iii) *$h_{\mathbb{Y}}(G(t,0),0) \leq M(t),\ t \in I$ a.e.,*

where $L, M \in L^1(I, \mathbb{R})$ *and* $L(t), M(t) > 0$.

Proof. As in the proof of Lemma 2.2, for each $\varepsilon > 0$ there is a closed set $I_\varepsilon \subset I$, with $\mu(I \setminus I_\varepsilon) < \varepsilon$, such that the map $G(t,x)$ restricted to $I_\varepsilon \times \mathbb{E}$ is continuous. Hence the map $t \to G(t,x)$ is measurable for every $x \in \mathbb{E}$, proving (i). To show (ii), take $x, y \in \mathbb{E}$, $t \in I$ a.e. By virtue of $(\mathrm{H}_F)_{\mathrm{ii}}$, for each $u \in F(t,x)$ there exists $v \in F(t,y)$ such that $||u-v|| \le k(t)||x-y||$. Moreover, by (2.1), $(u, g(t,x,u)) \in G(t,x)$ and $(v, g(t,y,v)) \in G(t,y)$, and thus, in view of (2.5), $|g(t,x,u) - g(t,y,v)| \le l(1 + k(t))||x-y||$. Consequently,

$$G(t,x) \subset G(t,y) + L(t)||x-y||B,$$

where B is the closed unit ball in $\mathbb{Y}$ and $L(t) = \max\{k(t), l(1+k(t))\}$. By changing the role of x and y, (ii) follows. In view of $(\mathrm{H}_F)_{\mathrm{iii}}$, $||u|| \le m(t)$ for every $u \in F(t,0)$. Further, $(\mathrm{H}_g)_{\mathrm{iii}}$ implies $|g(t,0,u)| \le a(t) + cm(t)$. Hence, setting $M(t) = \max\{m(t), a(t) + cm(t)\}$, (iii) is also satisfied, completing the proof.

3. Proof of Theorem 1.1. We will first construct a convenient Lipschitzean ε-approximation φ of g. For $t \in I$ set $U(t) = \{x(t) \mid x \in \mathcal{M}_{\overline{\mathrm{co}}\, F}\}$. Since $\mathcal{M}_{\overline{\mathrm{co}}\, F}$ is compact in $C(I, \mathbb{E})$, using the Ascoli–Arzelà theorem, one can show that the function $t \to U(t)$ is continuous from I to $\mathcal{K}(\mathbb{E})$, and the set $A = \cup_{t \in I} U(t)$ is compact in $\mathbb{E}$. By virtue of [6], for every $\varepsilon > 0$ there exists a closed set $I_\varepsilon \subset I$, with $\mu(I \setminus I_\varepsilon) < \varepsilon$, such that the restriction of $F(t,x)$ to $I_\varepsilon \times \mathbb{E}$ is continuous. A is compact, and thus the maps $F(t, \overline{\mathrm{co}}\, A)$ and $\overline{\mathrm{co}}\, F(t, \overline{\mathrm{co}}\, A)$, restricted to I_ε, are both compact valued and continuous; moreover, $\overline{\mathrm{co}}\, F(t, \overline{\mathrm{co}}\, A)$ is measurable on I.

Let $x \in \mathcal{M}_{\overline{\mathrm{co}}\, F}$ be arbitrary. Take $R > 0$ so that $A \subset B(0,R)$. In view of $(\mathrm{H}_F)_{\mathrm{ii}}$ and $(\mathrm{H}_F)_{\mathrm{iii}}$, for $u \in \overline{\mathrm{co}}\, F(t, x(t))$ and $t \in I$ a.e. one has $F(t, x(t)) \subset B(0, m(t) + Rk(t))$. From the latter and $(\mathrm{H}_g)_{\mathrm{iii}}$ it follows that

$$|g(t, x(t), u)| \le M(t) \quad \text{for every } u \in \overline{\mathrm{co}}\, F(t, x(t)),\ t \in I \text{ a.e.}, \tag{3.1}$$

where the function $M(t) = a(t) + Rb(t) + c(m(t) + Rk(t))$ is in $L^1(I, \mathbb{R})$.

Moreover, $-M(t) \le g_F^{**}(t, x(t), u)$ for all $u \in \mathbb{E}$, $t \in I$ a.e., since $-M(t)$ is a minorant of $u \to g_F(t, x(t), u)$. On the other hand, as $u \to g_F^{**}(t, x(t), u)$ is convex, lower semicontinuous, and $g_F^{**}(t, x(t), u) \le g_F(t, x(t), u) = g(t, x(t), u) \le M(t)$ for all $u \in F(t, x(t))$, $t \in I$ a.e., it follows that $g_F^{**}(t, x(t), u) \le M(t)$ for all $u \in \overline{\mathrm{co}}\, F(t, x(t))$, $t \in I$ a.e. Therefore

$$|g_F^{**}(t, x(t), u)| \le M(t) \ \text{ for all } u \in \overline{\mathrm{co}}\, F(t, x(t)),\ t \in I \ \text{ a.e.} \tag{3.2}$$

Let $\epsilon > 0$. By virtue of [6], (3.1), and (3.2), there exists a closed set $I_\varepsilon \subset I$ such that the map $t \to \overline{\mathrm{co}}\, F(t, \overline{\mathrm{co}}\, A)$ restricted to I_ε is continuous, g restricted to $I_\varepsilon \times \mathbb{E} \times \mathbb{E}$ is continuous, and furthermore

$$\int_{I \setminus I_\varepsilon} |g(t, x(t), \dot{x}(t))| dt \le \frac{\varepsilon}{5}, \quad \int_{I \setminus I_\varepsilon} |g_F^{**}(t, x(t), \dot{x}(t))| dt \le \frac{\varepsilon}{5} \quad \text{for all } x \in \mathcal{M}_{\overline{\mathrm{co}}\, F}. \tag{3.3}$$

By Tietze's extension theorem the function g restricted to $I_\varepsilon \times \mathbb{E} \times \mathbb{E}$ admits a continuous extension, say $\overline{g}$, to all of $I \times \mathbb{E} \times \mathbb{E}$. Set $B = \cup_{t \in I_\varepsilon} \overline{\mathrm{co}}\, F(t, \overline{\mathrm{co}}\, A)$ and observe that B is compact. Let $\overline{g}_\varepsilon : I \times \overline{\mathrm{co}}\, A \times \overline{\mathrm{co}}\, B \to \mathbb{R}$ be a locally Lipschitzean function satisfying

$$|\overline{g}_\varepsilon(t,x,u) - \overline{g}(t,x,u)| \le \frac{\varepsilon}{5} \quad \text{for all } (t,x,u) \in I \times \overline{\mathrm{co}}\, A \times \overline{\mathrm{co}}\, B.$$

Since $\overline{g}_\varepsilon$ is defined on a compact convex set, $\overline{g}_\varepsilon$ is actually Lipschitzean. Hence, by the MacShane lemma [5, p. 276], $\overline{g}_\varepsilon$ admits a bounded Lipschitzean extension g_ε (with the same Lipschitz constant) to the whole space $I \times \mathbb{E} \times \mathbb{E}$.

Define $\varphi : I \times \mathbb{E} \times \mathbb{E} \to \mathbb{R}$ by

$$\varphi(t,x,u) = \begin{cases} g_\varepsilon(t,x,u) & \text{if } t \in I_\epsilon, \\ 0 & \text{if } t \in I \backslash I_\varepsilon. \end{cases} \tag{3.4}$$

It is easily seen that φ satisfies $(\mathrm{H}_g)_{\mathrm{i}}$, $(\mathrm{H}_g)_{\mathrm{iii}}$ and (2.5) (with different $a(t), b(t), c, l$); moreover,

$$|\varphi(t,x,u) - g(t,x,u)| \le \frac{\varepsilon}{5} \qquad \text{for all} \quad (t,x,u) \in I_\varepsilon \times \overline{\mathrm{co}}\, A \times \overline{\mathrm{co}}\, B. \tag{3.5}$$

Consider now the function φ_F given by (1.2), with φ in place of g. Then,

$$\varphi_F(t,x,u) - \frac{\varepsilon}{5} \le g_F(t,x,u) \le \varphi_F(t,x,u) + \frac{\varepsilon}{5} \qquad \text{for all } (t,x,u) \in I_\varepsilon \times \overline{\mathrm{co}}\, A \times \overline{\mathrm{co}}\, B. \tag{3.6}$$

In fact, let $(t,x) \in I_\epsilon \times \overline{\mathrm{co}}\, A$. If $u \in F(t,x)$, then $u \in \overline{\mathrm{co}}\, B$, and thus (3.6) follows from (3.5) as $g_F(t,x,u) = g(t,x,u)$ and $\varphi_F(t,x,u) = \varphi(t,x,u)$. If $u \notin F(t,x)$, then $\varphi_F(t,x,u) = g_F(t,x,u) = +\infty$, and hence (3.6) holds trivially. From (3.6) one has

$$\varphi_F^{**}(t,x,u) - \frac{\varepsilon}{5} \le g_F^{**}(t,x,u) \le \varphi_F^{**}(t,x,u) + \frac{\varepsilon}{5} \qquad \text{for all } (t,x,u) \in I_\varepsilon \times \overline{\mathrm{co}}\, A \times \mathbb{E}.$$

Observe that if $u \in \overline{\mathrm{co}}\, F(t,x)$, then $\varphi_F^{**}(t,x,u)$ and $g_F^{**}(t,x,u)$ are both finite by Lemma 2.2(i), and hence

$$|\varphi_F^{**}(t,x,u) - g_F^{**}(t,x,u)| \le \frac{\varepsilon}{5} \qquad \text{for all } t \in I_\varepsilon,\ x \in \overline{\mathrm{co}}\, A,\ u \in \overline{\mathrm{co}}\, F(t,x). \tag{3.7}$$

Furthermore,

$$\varphi_F^{**}(t,x,u) = 0 \quad \text{for } t \in I \setminus I_\varepsilon,\ x \in E,\ u \in \overline{\mathrm{co}}\, F(t,x). \tag{3.8}$$

In fact, $\varphi_F^{**}(t,x,u) \ge 0$ for all $u \in E$, as 0 is a minorant of $u \to \varphi_F(t,x,u)$. On the other hand, since $u \to \varphi_F^{**}(t,x,u)$ is convex, lower semicontinuous, and $\varphi_F^{**}(t,x,u) \le \varphi_F(t,x,u) = \varphi(t,x,u) = 0$ for all $u \in F(t,x)$, it follows that $\varphi_F^{**}(t,x,u) \le 0$, and thus (3.8) holds.

In view of (3.5), (3.4), (3.3) and (3.7), (3.8), (3.3), for each $x \in \mathcal{M}_{\overline{\mathrm{co}}\, F}$ one has

$$\int_I |\varphi(t,x(t),\dot{x}(t)) - g(t,x(t),\dot{x}(t))|dt \tag{3.9}$$

$$\le \int_{I_\epsilon} |\varphi(t,x(t),\dot{x}(t)) - g(t,x(t),\dot{x}(t))|dt + \int_{I \setminus I_\epsilon} |g(t,x(t),\dot{x}(t))|dt \le \frac{2\varepsilon}{5}$$

and

$$\int_I |\varphi_F^{**}(t,x(t),\dot{x}(t)) - g_F^{**}(t,x(t),\dot{x}(t))|dt \tag{3.10}$$

$$\le \int_{I_\epsilon} |\varphi_F^{**}(t,x(t),\dot{x}(t)) - g_F^{**}(t,x(t),\dot{x}(t))|dt + \int_{I \setminus I_\epsilon} |g_F^{**}(t,x(t),\dot{x}(t))|dt \le \frac{2\varepsilon}{5}.$$

Consider now the Cauchy problem

$$\begin{cases} \dot{x}(t) \in F(t, x(t)), \ \ x(0) = x_0, \\ \dot{\lambda}(t) = \varphi(t, x(t), \dot{x}(t)), \ \ \lambda(0) = 0. \end{cases}$$

Let $G : I \times \mathbb{E} \to \mathcal{K}(\mathbb{Y})$ be the multifunction given by (2.1), with φ in place of g, and set $z = (x, \lambda) \in \mathbb{Y}$. Then the previous Cauchy problem and the corresponding convexified problem can be written in the form

$$\begin{aligned} \dot{z}(t) &\in G(t, x(t)), \ \ z(0) = (x_0, 0), \\ \dot{z}(t) &\in \overline{\text{co}}\, G(t, x(t)), \ \ z(0) = (x_0, 0). \end{aligned}$$

Denote by $\mathcal{M}_G$, $\mathcal{M}_{\overline{\text{co}}\, G}$ the solution sets of the above Cauchy problems. In view of Lemma 2.4, the map $G(t, x)$ satisfies (i)–(iii) and thus, by the relaxation theorem, $\mathcal{M}_{\overline{\text{co}}\, G}$ is nonempty and compact and

$$\overline{\mathcal{M}}_G = \mathcal{M}_{\overline{\text{co}}\, G}. \tag{3.11}$$

Let $x \in \mathcal{M}_{\overline{\text{co}}\, F}$ be arbitrary, and set $v(t) = \int_0^t \varphi_F^{**}(s, x(s), \dot{x}(s))ds, \ t \in I$. Since $\dot{x}(t) \in \overline{\text{co}}\, F(t, x(t)), \ t \in I$ a.e., Lemma 2.2(i) implies that $(\dot{x}(t), \dot{v}(t)) \in \overline{\text{co}}\, G(t, x(t)), \ t \in I$ a.e., and hence $(x, v) \in \mathcal{M}_{\overline{\text{co}}\, G}$. By (3.11) there exists $(y, w) \in \mathcal{M}_G$ such that $\|(x, v) - (y, w)\| < \varepsilon/5$. Thus $y \in \mathcal{M}_F$, by Lemma 2.1(i); moreover, $w(t) = \int_0^t \varphi(s, y(s), \dot{y}(s))ds$ and

$$\sup_{t \in I} ||x(t) - y(t)|| \le \frac{\varepsilon}{5}, \qquad \sup_{t \in I} \left| \int_0^t (\varphi_F^{**}(s, x(s), \dot{x}(s)) - \varphi(s, y(s), \dot{y}(s)))ds \right| \le \frac{\varepsilon}{5}. \tag{3.12}$$

Since, for each $t \in I$,

$$\left| \int_0^t (g_F^{**}(s, x(s), \dot{x}(s)) - g(s, y(s), \dot{y}(s)))ds \right| \le \int_I |g_F^{**}(s, x(s), \dot{x}(s)) - \varphi_F^{**}(s, x(s), \dot{x}(s))|ds$$

$$+ \sup_{t \in I} \left| \int_0^t (\varphi_F^{**}(s, x(s), \dot{x}(s)) - \varphi(s, y(s), \dot{y}(s)))ds \right| + \int_I |\varphi(s, y(s), \dot{y}(s)) - g(s, y(s), \dot{y}(s))|ds,$$

then, by virtue of (3.10), (3.12), and (3.9), as $t \in I$ is arbitrary, it follows that

$$\sup_{t \in I} \left| \int_0^t (g_F^{**}(s, x(s), \dot{x}(s)) - g(s, y(s), \dot{y}(s)))ds \right| \le \varepsilon.$$

Thus, setting $\epsilon = 1/n$, $\varphi_n = \varphi$, and $y = x_n$, one obtains a sequence $\{x_n\} \subset \mathcal{M}_F$ for which (i) and (ii) are satisfied. This completes the proof.

4. An application. We present an application of the previous result. In what follows, $\mathbb{E}$ is a separable reflexive Banach space. Consider the minimization problem

$$\text{(P)} \qquad \text{minimize} \int_I g(t, x(t), \dot{x}(t))dt \quad \text{for } x \in \mathcal{M}_F,$$

and associate with (P) the relaxed minimization problem

$$\text{(P}^{**}\text{)} \qquad \text{minimize} \int_I g_F^{**}(t, x(t), \dot{x}(t))dt \quad \text{for } \ x \in \mathcal{M}_{\overline{\text{co}}\, F}.$$

THEOREM 4.1. *Let F, g satisfy* (H_F), (H_g). *Then,*

$$\min(\mathrm{P}^{**}) = \inf(\mathrm{P}). \tag{4.1}$$

Furthermore, for each solution x of problem (P^{**}), *there exists a minimizing sequence* $\{x_n\}$ *of problem* (P) *such that*

(i) x_n *converges to x in* $C(I,E)$,

(ii) $\lim_{n\to\infty} \sup_{t\in I} | \int_0^t (g_F^{**}(s,x(s),\dot{x}(s)) - g(s,x_n(s),\dot{x}_n(s)))ds| = 0$.

Conversely, if $\{x_n\}$ *is a minimizing sequence of problem* (P), *then there exists a subsequence* $\{x_{n_k}\}$ *and a solution x of problem* (P^{**}) *so that* (i) *and* (ii) *are satisfied.*

Proof. Set $A = \{x(t) \mid t \in I, x \in \mathcal{M}_{\overline{\mathrm{co}}\, F}\}$. As was shown in the proof of Theorem 1.1, the set A is compact in $\mathbb{E}$. Define $\varphi : I \times \mathbb{E} \times \mathbb{E} \to \overline{\mathbb{R}}$ by

$$\varphi(t,x,u) = \begin{cases} g_F^{**}(t,x,u), & t \in I,\ x \in A,\ u \in \mathbb{E}, \\ +\infty & \text{elsewhere.} \end{cases} \tag{4.2}$$

By Lemma 2.2(ii), there exists a sequence of closed sets $I_n \subset I_{n+1} \subset I$ with $\mu(I \setminus I_n) < 1/n$, $n \in \mathbb{N}$, such that the function $g_F^{**}(t,x,u)$ restricted to $I_n \times \mathbb{E} \times \mathbb{E}$ is lower semicontinuous. As $\varphi(t,x,u)$ equals $g_F^{**}(t,x,u)$ on the closed set $I_n \times A \times \mathbb{E}$, while $\varphi(t,x,u) = +\infty$ on $I_n \times (\mathbb{E} \setminus A) \times \mathbb{E}$, the function $\varphi(t,x,u)$ restricted to $I_n \times \mathbb{E} \times \mathbb{E}$ is lower semicontinuous. Moreover, $J = I \setminus \cup_{n=1}^{\infty} I_n$ is a Borel set, whence the function $\tilde{\varphi} : I \times \mathbb{E} \times \mathbb{E} \to \overline{\mathbb{R}}$ given by

$$\tilde{\varphi}(t,x,u) = \begin{cases} \varphi(t,x,u), & t \in J,\ x \in \mathbb{E},\ u \in \mathbb{E}, \\ 0 & \text{elsewhere} \end{cases}$$

is Borel and thus $\Sigma \otimes \mathcal{B}(\mathbb{E} \times \mathbb{E})$ measurable. Therefore, $\varphi(t,x,u) = \tilde{\varphi}(t,x,u)$ for all $x, u \in \mathbb{E}$ and $t \in I$ a.e., for $\mu(I \setminus J) = 0$. Thus, without loss of generality, we can assume that $\varphi(t,x,u)$ is $\Sigma \otimes \mathcal{B}(\mathbb{E} \times \mathbb{E})$ measurable. Moreover, the function $u \to \varphi(t,x,u)$ is convex. Thanks to $(\mathrm{H}_F)_{\mathrm{ii}}$, $(\mathrm{H}_F)_{\mathrm{iii}}$, $(\mathrm{H}_g)_{\mathrm{iii}}$, the inequality (2.3) in [2] is satisfied, and so $\varphi(t,x,u)$ meets all assumptions of Theorem 2.1 in [2].

Consider now the integral functional $J_\varphi : C(I,\mathbb{E}) \times L^1(I,\mathbb{E}) \to \overline{\mathbb{R}}$ defined by

$$J_\varphi(x,u) = \int_I \varphi(t,x(t),u(t))dt,$$

where the space $L^1(I,\mathbb{E})$ is endowed with the weak topology $\sigma(L^1, L^\infty)$. By Theorem 2.1 in [2], the functional $(x,u) \to J_\varphi(x,u)$ is sequentially lower semicontinuous on $C(I,\mathbb{E}) \times L^1(I,\mathbb{E})$. As $\mathcal{R}_{\overline{\mathrm{co}}\, F} = \{(x,\dot{x}) \mid x \in \mathcal{M}_{\overline{\mathrm{co}}\, F}\}$ is a metrizable compact subset of $C(I,\mathbb{E}) \times L^1(I,\mathbb{E})$, the functional $J_\varphi(x,u)$ has a minimum on $\mathcal{R}_{\overline{\mathrm{co}}\, F}$; i.e., there exists $(x,\dot{x}) \in \mathcal{R}_{\overline{\mathrm{co}}\, F}$ such that

$$\int_I \varphi(t,x(t),\dot{x}(t))dt = \min\{J_\varphi(y,\dot{y}) \mid (y,\dot{y}) \in \mathcal{R}_{\overline{\mathrm{co}}\, F}\}.$$

Then, by (4.2),

$$\int_I g_F^{**}(t,x(t),\dot{x}(t))dt = \min(\mathrm{P}^{**}),$$

and hence problem (P^{**}) has a solution.

As $x \in \mathcal{M}_{\overline{\text{co}}\, F}$, then, by Theorem 1.1, there exists a sequence $\{x_n\} \subset \mathcal{M}_F$ satisfying (i), (ii). By (ii),

$$\lim_{n\to\infty} \int_I g(t, x_n(t), \dot{x}_n(t))dt = \int_I g_F^{**}(t, x(t), \dot{x}(t))dt,$$

and thus $\inf(\mathrm{P}) \leq \min(\mathrm{P}^{**})$. Furthermore, for each $x \in \mathcal{M}_F$, one has $g(t, x(t), \dot{x}(t)) = g_F(t, x(t), \dot{x}(t)) \geq g_F^{**}(t, x(t), \dot{x}(t))$ for $t \in I$ a.e., which implies $\inf(\mathrm{P}) \geq \min(\mathrm{P}^{**})$. Therefore (4.1) holds, and $\{x_n\}$ is actually a minimizing sequence of problem (P).

Let $\{x_n\} \subset \mathcal{M}_F$ be a minimizing sequence of problem (P). Then $(x_n, v_n) \in \mathcal{M}_G$, $n \in \mathbb{N}$, where $v_n(t) = \int_0^t g(s, x_n(s), \dot{x}_n(s))ds$ and G is given by (2.1). As $\mathcal{M}_{\overline{\text{co}}\, G}$ is compact in $C(I, \mathbb{Y})$, there exists a subsequence $\{(x_{n_k}, v_{n_k})\}$ which converges to some $(x, v) \in \mathcal{M}_{\overline{\text{co}}\, G}$, where $x \in \mathcal{M}_{\overline{\text{co}}\, F}$ by Lemma 2.1(i). Since $\{x_{n_k}\}$ is a minimizing sequence of problem (P), $\min(\mathrm{P}^{**}) = \inf(\mathrm{P})$ and, by Lemma 2.2(i), $g_F^{**}(t, x(t), \dot{x}(t)) \leq \dot{v}(t),\ t \in I$ a.e., then one has

$$\lim_{k\to\infty} \int_I g(t, x_{n_k}(t), \dot{x}_{n_k}(t))dt = \min(\mathrm{P}^{**}) \leq \int_I g_F^{**}(t, x(t), \dot{x}(t))dt \leq \int_I \dot{v}(t)dt.$$

However, v_{n_k} converges to v in $C(I, \mathbb{R})$, and thus the first and the last terms in the above expression are equal. Consequently,

$$\min(\mathrm{P}^{**}) = \int_I g_F^{**}(t, x(t), \dot{x}(t))dt.$$

Moreover, $g_F^{**}(t, x(t), \dot{x}(t)) = \dot{v}(t),\ t \in I$ a.e., and thus $\int_0^t g(s, x_{n_k}(s), \dot{x}_{n_k}(s))ds$ converges to $\int_0^t g_F^{**}(s, x(s), \dot{x}(s))ds$ in $C(I, \mathbb{R})$. Therefore x is a solution of problem (P^{**}) and (ii) is satisfied (with x_{n_k} in place of x_n). As (i) also holds, the proof is complete.

Remark 4.2. A similar result was obtained in [14, Appendix] under the assumption that the function $(x, u) \to g(t, x, u)$ is Lipschitzean.

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[2] E. J. Balder, *Necessary and sufficient condition for L_1-strong-weak lower semicontinuity of integral functionals*, Nonlinear Anal., 11 (1987), pp. 1399–1404.

[3] N. N. Bogolyubov, *Sur quelques method nouvelles dans le calculus des variations*, Ann. Mat. Pura Appl. (4), 7 (1930), pp. 249–271.

[4] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, 1977.

[5] I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnel*, Dunod, Gauthier-Villars, Paris 1974.

[6] C. J. Himmelberg, *Precompact contraction of metric uniformities and the continuity of $F(t, x)$*, Rend. Sem. Mat. Univ. Padova, 50 (1973), pp. 185–188.

[7] A. D. Ioffe, *Transformations of correctly posed variational problems*, Soviet Math. Dokl., 7 (1968), pp. 623–627.

[8] A. D. Ioffe and V. M. Tikhomirov, *The duality of convex functions and extremal problems*, Soviet Math. Dokl., 9 (1968), pp. 685–687.

[9] A. D. Ioffe and V. M. Tikhomirov, *Theory of Extremal Problems*, North–Holland, Amsterdam, 1979.

[10] E. J. MacShane, *Existence theorem for Bolza problem in the calculus of variations*, Duke Math. J., 7 (1940), pp. 28–61.

[11] A. Plis, *Trajectories and quasitrajectories of an orientor field*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 11 (1963), pp. 369–370.

[12] S. I. Suslov, *A theorem of Bogolyubov with constraints in the form of differential inclusion*, Sibirsk. Mat. Zh., 35 (1994), pp. 902–914.

[13] A. A. Tolstonogov, *On properties of solutions of differential inclusions in a Banach spaces*, Soviet. Math. Dokl., 20 (1979), pp. 960–964.
[14] A. A. Tolstonogov, *Differential Inclusions in a Banach Space*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.
[15] J. Warga, *Functions of relaxed control*, SIAM J. Control, 5 (1967), pp. 628–641.
[16] L. C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Soc. Sci. Varsovie, 30 (1937), pp. 212–234.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 477–501

HESSIAN RIEMANNIAN GRADIENT FLOWS IN CONVEX PROGRAMMING*

FELIPE ALVAREZ†, JÉRÔME BOLTE‡, AND OLIVIER BRAHIC§

Abstract. In view of solving theoretically constrained minimization problems, we investigate the properties of the gradient flows with respect to Hessian Riemannian metrics induced by Legendre functions. The first result characterizes Hessian Riemannian structures on convex sets as metrics that have a specific integration property with respect to variational inequalities, giving a new motivation for the introduction of Bregman-type distances. Then, the general evolution problem is introduced, and global convergence is established under quasi-convexity conditions, with interesting refinements in the case of convex minimization. Some explicit examples of these gradient flows are discussed. Dual trajectories are identified, and sufficient conditions for dual convergence are examined for a convex program with positivity and equality constraints. Some convergence rate results are established. In the case of a linear objective function, several optimality characterizations of the orbits are given: optimal path of viscosity methods, continuous-time model of Bregman-type proximal algorithms, geodesics for some adequate metrics, and projections of $\dot{q}$-trajectories of some Lagrange equations and completely integrable Hamiltonian systems.

Key words. gradient flow, Hessian Riemannian metric, Legendre-type convex function, existence, global convergence, Bregman distance, Lyapunov functional, quasi-convex minimization, convex and linear programming, Legendre transform coordinates, Lagrange and Hamilton equations

AMS subject classifications. 34G20, 34A12, 34D05, 90C25

DOI. 10.1137/S0363012902419977

1. Introduction.

The aim of this paper is to study the existence, global convergence, and geometric properties of gradient flows with respect to a specific class of Hessian Riemannian metrics on convex sets. Our work is indeed deeply related to the constrained minimization problem

$$\text{(P)} \qquad \min\{f(x) \mid x \in \overline{C},\ Ax = b\},$$

where $\overline{C}$ is the closure of a nonempty, *open*, and convex subset C of $\mathbb{R}^n$; A is an $m \times n$ real matrix with $m \leq n$; $b \in \mathbb{R}^m$ and $f \in C^1(\mathbb{R}^n)$. A strategy for solving (P) consists of endowing C with a Riemannian metric g, restricting it to the relative interior of the feasible set $\mathcal{F} := C \cap \{x \mid Ax = b\}$, and then considering the trajectories generated by the steepest descent vector field. We focus on those metrics that are induced by the Hessian $H = \nabla^2 h$ of a *Legendre-type* convex function h defined on C (cf. Definition 3.3), that is, $g_{ij} = \frac{\partial^2 h}{\partial x_i \partial x_j}$. This leads to the initial value problem

$$\text{(H-SD)} \qquad \dot{x}(t) + \nabla_H f_{|\mathcal{F}}(x(t)) = 0, \qquad x(0) \in \mathcal{F},$$

*Received by the editors December 12, 2002; accepted for publication December 1, 2003; published electronically July 23, 2004.

http://www.siam.org/journals/sicon/43-2/41997.html

†Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático (CNRS UMR 2071), Universidad de Chile, Blanco Encalada 2120, Santiago, Chile (falvarez@dim.uchile.cl). This author was supported by Fondecyt 1020610, Fondap en Matemáticas Aplicadas and Programa Iniciativa Científica Milenio.

‡ACSIOM-CNRS FRE 2311, Département de Mathématiques, case 51, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier cedex 5, France (bolte@math.univ-montp2.fr). This author was partially supported by Ecos-Conicyt C00E05.

§GTA-CNRS UMR 5030, Département de Mathématiques, case 51, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier cedex 5, France (brahic@math.univ-montp2.fr).

where (H-SD) stands for H-steepest descent.

The use of Riemannian methods in optimization has increased recently. For interior point methods in linear programming, see Karmarkar [31], Bayer and Lagarias [7], and Nesterov and Todd [37]; for continuous-time models of proximal-type algorithms and related topics, see Iusem, Svaiter, and Da Cruz Neto [29], Bolte and Teboulle [8], and Attouch and Teboulle [3]. For a systematic dynamical system approach to constrained optimization based on double bracket flows, see Brockett [10, 11], Helmke and Moore [24], and the references therein. See Smith [41] and Udriste [43] for general optimization techniques on Riemannian manifolds. On the other hand, the structure of (H-SD) is also at the heart of some important problems in applied mathematics. For connections with population dynamics and game theory, see Akin [1] and Hofbauer and Sygmund [27]. We will see that (H-SD) can be reformulated as the differential inclusion $\frac{d}{dt}\nabla h(x(t)) + \nabla f(x(t)) \in \operatorname{Im} A^T$, $x(t) \in \mathcal{F}$, which is formally similar to some evolution problems in infinite dimensional spaces arising in thermodynamical systems; see Kenmochi and Pawlow [32] and references therein.

A classical approach in the asymptotic analysis of dynamical systems consists of exhibiting attractors of the orbits by using Lyapunov functionals. Our choice of Hessian Riemannian metrics is based on this idea. In fact, we consider first the important case where f is convex, a condition that permits us to reformulate (P) as a variational inequality problem: find $a \in \overline{\mathcal{F}}$ such that $(\nabla_{H} f_{|\mathcal{F}}(x), x-a)_x^H \geq 0$ for all x in $\mathcal{F}$. In order to identify a suitable Lyapunov functional, this variational problem is met through the following integration problem: *find the metrics $(\cdot,\cdot)^H$ for which the vector fields $V^a : \mathcal{F} \to \mathbb{R}^n$, $a \in \mathcal{F}$, defined by $V^a(x) = x - a$, are $(\cdot,\cdot)^H$-gradient vector fields.* Our first result (cf. Theorem 3.1) establishes that such metrics are given by the Hessian of strictly convex functions, and in that case the vector fields V^a appear as gradients with respect to the second variable of some distance-like functions that are called D-functions. Indeed, if $(\cdot,\cdot)^H$ is induced by the Hessian $H = \nabla^2 h$ of $h : \mathcal{F} \mapsto \mathbb{R}$, we have for all a, x in $\mathcal{F}$, $\nabla_{H} D_h(a,.)(x) = x - a$, where $D_h(a,x) = h(a) - h(x) - dh(x)(a-x)$. See Duistermaat [19] for a related characterization of Hessian metrics.

Motivated by the previous result and with the aim of solving (P), we are then naturally led to consider Hessian Riemannian metrics that cannot be smoothly extended out of $\mathcal{F}$. Such a requirement is fulfilled by the Hessian of a *Legendre (convex) function* h, whose definition is recalled in section 3. We give then a differential inclusion reformulation of (H-SD), which permits us to show that in the case of a linear objective function f, the flow of $-\nabla_{H} f_{|\mathcal{F}}$ stands at the crossroad of many optimization methods. In fact, following [29], we prove that viscosity methods and Bregman proximal algorithms produce their paths or iterates in the orbit of (H-SD). The D-function of h plays an essential role for this. In section 4.4 we give a systematic method for constructing Legendre functions based on barrier functions for convex inequality problems, which is illustrated with some examples; relations to other works are discussed.

Section 4 deals with global existence and convergence properties. After having given a nontrivial well-posedness result (cf. Theorem 4.1), we prove in section 4.2 that $f(x(t)) \to \inf_{\overline{\mathcal{F}}} f$ as $t \to +\infty$ whenever f is convex. A natural problem that arises is the trajectory convergence to a critical point. Since one expects the limit to be a (local) solution to (P), which may belong to the boundary of C, the notion of critical point must be understood in the sense of the optimality condition for a local minimizer a of f over $\overline{\mathcal{F}}$:

$$(\mathcal{O}) \qquad \nabla f(a) + N_{\overline{\mathcal{F}}}(a) \ni 0, \qquad a \in \overline{\mathcal{F}},$$

where $N_{\overline{\mathcal{F}}}(a)$ is the normal cone to $\overline{\mathcal{F}}$ at a, and ∇f is the Euclidean gradient of f. This involves an asymptotic singular behavior that is rather unusual in the classical theory of dynamical systems, where the critical points are typically supposed to be in the manifold. In section 4.3 we assume that the Legendre-type function h is a *Bregman function with zone C* (see [5] and [34] for comprehensive surveys) and prove that, under a quasi convexity assumption on f, the trajectory converges to some point a satisfying $(\mathcal{O})$. When f is convex, the preceding result amounts to the convergence of $x(t)$ toward a global minimizer of f over $\overline{\mathcal{F}}$. We also give a variational characterization of the limit and establish an abstract result on the rate of convergence under uniqueness of the solution. We consider in section 4.5 the case of linear programming, for which asymptotic convergence as well as a variational characterization are proved without the Bregman-type condition. Within this framework, we also give some estimates on the convergence rate that are valid for the specific Legendre functions commonly used in practice. In section 4.6, we consider the interesting case of positivity and equality constraints, introducing a *dual* trajectory $\lambda(t)$ that, under some appropriate conditions, converges to a solution to the dual problem of (P) whenever f is convex, even if primal convergence is not ensured.

Finally, for a linear objective function, and inspired by the seminal work [7], we define in section 5 a change of coordinates called *Legendre transform coordinates*, which permits us to show that the orbits of (H-SD) may be seen as straight lines in a positive cone. This leads to additional geometric interpretations of the flow of $-\nabla_{H} f_{|\mathcal{F}}$. On the one hand, the orbits are geodesics with respect to an appropriate metric and, on the other hand, they may be seen as $\dot{q}$-trajectories of some Lagrangian, with consequences in terms of completely integrable Hamiltonians.

Notation. $\operatorname{Ker} A = \{x \in \mathbb{R}^n \mid Ax = 0\}$. The orthogonal complement of $\mathcal{A}_0$ is denoted by $\mathcal{A}_0^{\perp}$, and $\langle \cdot, \cdot \rangle$ is the standard Euclidean scalar product of $\mathbb{R}^n$. Let us denote by $\mathbb{S}^n_{++}$ the cone of real symmetric definite positive matrices. Let $\Omega \subset \mathbb{R}^n$ be an open set. If $f : \Omega \to \mathbb{R}$ is differentiable, then ∇f stands for the Euclidean gradient of f. If $h : \Omega \mapsto \mathbb{R}$ is twice differentiable, then its Euclidean Hessian at $x \in \Omega$ is denoted by $\nabla^2 h(x)$ and is defined as the endomorphism of $\mathbb{R}^n$ whose matrix in canonical coordinates is given by $\left[\frac{\partial^2 h(x)}{\partial x_i \partial x_j}\right]_{i,j \in \{1,\dots,n\}}$. Thus, for all $x \in \Omega$, $d^2 h(x) = \langle \nabla^2 h(x) \cdot, \cdot \rangle$.

2. Preliminaries.

2.1. The minimization problem and optimality conditions.

Given a positive integer $m < n$, a full rank matrix $A \in \mathbb{R}^{m \times n}$, and $b \in \operatorname{Im} A$, let us define

$$\mathcal{A} = \{x \in \mathbb{R}^n \mid Ax = b\}. \tag{2.1}$$

Set $\mathcal{A}_0 = \mathcal{A} - \mathcal{A} = \operatorname{Ker} A$. Of course, $\mathcal{A}_0^{\perp} = \operatorname{Im} A^T$, where A^T is the transpose of A. Let C be a nonempty, open, and convex subset of $\mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R}$ a $\mathcal{C}^1$ function. Consider the constrained minimization problem

$$\inf\{f(x) \mid x \in \overline{C},\ Ax = b\}. \tag{P}$$

The set of optimal solutions of (P') is denoted by $S(P')$. We call f the *objective function* of (P'). The *feasible set* of (P') is given by $\overline{\mathcal{F}} = \{x \in \mathbb{R}^n \mid x \in \overline{C},\ Ax = b\} = \overline{C} \cap \mathcal{A}$, and $\mathcal{F}$ stands for the *relative interior* of $\overline{\mathcal{F}}$, that is,

$$\mathcal{F} = \operatorname{ri} \overline{\mathcal{F}} = \{x \in \mathbb{R}^n \mid x \in C,\ Ax = b\} = C \cap \mathcal{A}. \tag{2.2}$$

Throughout this article, we assume that

$$\mathcal{F} \neq \emptyset. \tag{2.3}$$

It is well known that a necessary condition for a to be locally minimal for f over $\overline{\mathcal{F}}$ is $(\mathcal{O})$: $-\nabla f(a) \in N_{\overline{\mathcal{F}}}(a)$, where $N_{\overline{\mathcal{F}}}(x) = \{\nu \in \mathbb{R}^n \mid \forall y \in \overline{\mathcal{F}},\ \langle y - x, \nu\rangle \leq 0\}$ is the *normal cone* to $\overline{\mathcal{F}}$ at $x \in \overline{\mathcal{F}}$ ($N_{\overline{\mathcal{F}}}(x) = \emptyset$ when $x \notin \overline{\mathcal{F}}$); see, for instance, [40, Theorem 6.12]. By [39, Corollary 23.8.1], $N_{\overline{\mathcal{F}}}(x) = N_{\overline{C}\cap\mathcal{A}}(x) = N_{\overline{C}}(x) + N_{\mathcal{A}}(x) = N_{\overline{C}}(x) + \mathcal{A}_0^{\perp}$ for all $x \in \overline{\mathcal{F}}$. Therefore, the necessary optimality condition for $a \in \overline{\mathcal{F}}$ is

$$-\nabla f(a) \in N_{\overline{C}}(a) + \mathcal{A}_0^{\perp}. \tag{2.4}$$

If f is convex, then this condition is also sufficient for $a \in \overline{\mathcal{F}}$ to be in $S(\mathrm{P})$.

2.2. Riemannian gradient flows on the relative interior of the feasible set. Let M be a smooth manifold. The tangent space to M at $x \in M$ is denoted by T_xM. If $f : M \mapsto \mathbb{R}$ is a $\mathcal{C}^1$ function, then $df(x)$ denotes its differential or tangent map $df(x) : T_xM \to \mathbb{R}$ at $x \in M$. A $\mathcal{C}^k$ metric on M, $k \geq 0$, is a family of scalar products $(\cdot,\cdot)_x$ on each T_xM, $x \in M$, such that $(\cdot,\cdot)_x$ depends in a $\mathcal{C}^k$ way on x. The couple $M, (\cdot,\cdot)_x$ is called a $\mathcal{C}^k$ Riemannian manifold. This structure permits us to identify T_xM with its dual, i.e., the cotangent space T_xM^*, and thus to define a notion of gradient vector. Indeed, given f in M, the gradient of f is denoted by $\nabla_{(\cdot,\cdot)} f$ and is uniquely determined by the following conditions:

(g_1) tangency condition: for all $x \in M$, $\nabla_{(\cdot,\cdot)} f(x) \in T_xM^* \simeq T_xM$;

(g_2) duality condition: for all $x \in M$, $v \in T_xM$, $df(x)(v) = (\nabla_{(\cdot,\cdot)} f(x), v)_x$.

We refer the reader to [18, 35] for further details.

Let us return to the minimization problem (P). Since C is open, we can take $M = C$ with the usual identification $T_xC \simeq \mathbb{R}^n$ for every $x \in C$. Given a continuous mapping $H : C \to \mathbb{S}^n_{++}$, the metric defined by

$$\forall x \in C,\ \forall u, v \in \mathbb{R}^n, \qquad (u,v)_x^H = \langle H(x)u, v\rangle \tag{2.5}$$

endows C with a $\mathcal{C}^0$ Riemannian structure. The corresponding Riemannian gradient vector field of the objective function f restricted to C, which we denote by $\nabla_{H} f_{|C}$, is given by

$$\nabla_{H} f_{|C}(x) = H(x)^{-1}\nabla f(x). \tag{2.6}$$

Next, take $N = \mathcal{F} = C \cap \mathcal{A}$, which is a smooth submanifold of C with $T_x\mathcal{F} \simeq \mathcal{A}_0$ for each $x \in \mathcal{F}$. Definition (2.5) induces a metric on $\mathcal{F}$ for which the gradient of the restriction $f_{|\mathcal{F}}$ is denoted by $\nabla_{H} f_{|\mathcal{F}}$. Conditions ($\mathrm{g}_1$) and ($\mathrm{g}_2$) imply that for all $x \in \mathcal{F}$

$$\nabla_{H} f_{|\mathcal{F}}(x) = P_x H(x)^{-1}\nabla f(x), \tag{2.7}$$

where, given $x \in C$, $P_x : \mathbb{R}^n \to \mathcal{A}_0$ is the $(\cdot,\cdot)_x^H$-orthogonal projection onto the linear subspace $\mathcal{A}_0$. Since A has full rank, it is easy to see that

$$P_x = I - H(x)^{-1}A^T(AH(x)^{-1}A^T)^{-1}A, \tag{2.8}$$

and we conclude that for all $x \in \mathcal{F}$

$$\nabla_{H} f_{|\mathcal{F}}(x) = H(x)^{-1}[I - A^T(AH(x)^{-1}A^T)^{-1}AH(x)^{-1}]\nabla f(x). \tag{2.9}$$

Given $x \in \mathcal{F}$, the vector $-\nabla_{H} f_{|\mathcal{F}}(x)$ can be interpreted as that direction in $\mathcal{A}_0$ such that f decreases the most steeply at x with respect to the metric $(\cdot,\cdot)_x^H$. The *steepest descent method* for the (local) minimization of f on the Riemannian manifold

$\mathcal{F}, (\cdot,\cdot)_x^H$ consists of finding the solution trajectory $x(t)$ of the vector field $-\nabla_H f_{|\mathcal{F}}$ with initial condition $x^0 \in \mathcal{F}$:

$$\left\{ \begin{array}{l} \dot{x} + \nabla_H f_{|\mathcal{F}}(x) = 0, \\ x(0) = x^0 \in \mathcal{F}. \end{array} \right. \tag{2.10}$$

3. Legendre gradient flows in constrained optimization.

3.1. Lyapunov functionals, variational inequalities, and Hessian metrics. This section is intended to motivate the particular class of Riemannian metrics that is studied in this paper in view of the asymptotic convergence of the solution to (2.10).

Let us consider the minimization problem (P) and assume that C is endowed with some Riemannian metric $(\cdot,\cdot)_x^H$ as defined in (2.5). Recall that $V : \mathcal{F} \mapsto \mathbb{R}$ is a *Lyapunov functional* for the vector field $-\nabla_H f_{|\mathcal{F}}$ if, for all $x \in \mathcal{F}$, $(-\nabla_H f_{|\mathcal{F}}(x), \nabla_H V(x))_x^H \leq 0$. If $x(t)$ is a solution to (2.10), this implies that $t \mapsto V(x(t))$ is nonincreasing. Although $f_{|\mathcal{F}}$ is indeed a Lyapunov functional for $-\nabla_H f_{|\mathcal{F}}$, this does not ensure the convergence of $x(t)$. (See, for instance, the counterexample of Palis and De Melo [38] in the Euclidean case.)

Suppose that the objective function f is convex. For simplicity, we also assume that $A = 0$ so that $\mathcal{F} = C$. In the framework of convex minimization, the set of minimizers of f over $\overline{C}$, denoted by $\operatorname{Argmin}_{\overline{C}} f$, is characterized in variational terms as follows:

$$a \in \operatorname{Argmin}_{\overline{C}} f \quad \Leftrightarrow \quad \forall x \in \overline{C}, \qquad \langle \nabla f(x), x - a \rangle \geq 0. \tag{3.1}$$

Setting $q_a(x) = \frac{1}{2}|x-a|^2$ for all $a \in \operatorname{Argmin}_{\overline{C}}$, one observes that $\nabla q_a(x) = x - a$ and thus, by (3.1), q_a is a Lyapunov functional for $-\nabla f$. This key property allows one to establish the asymptotic convergence as $t \to +\infty$ of the corresponding steepest descent trajectories; see [12] for more details in a very general nonsmooth setting. To use the same kind of arguments in a non-Euclidean context, observe that, by (2.6) together with the continuity of ∇f, the following variational Riemannian characterization holds:

$$a \in \operatorname{Argmin}_{\overline{C}} f \quad \Leftrightarrow \quad \forall x \in C, \qquad (\nabla_H f(x), x - a)_x^H \geq 0. \tag{3.2}$$

We are thus naturally led to the problem of *finding the Riemannian metrics on C for which the mappings $C \ni x \mapsto x - y \in \mathbb{R}^n$, $y \in C$, are gradient vector fields.* The next result gives a characterization of such metrics: they are induced by the Hessian of strictly convex functions.

THEOREM 3.1. *Assume that $H \in \mathcal{C}^1(C; \mathbb{S}^n_{++})$ or, in other words, that $(\cdot,\cdot)_x^H$ is a $\mathcal{C}^1$ metric. The family of vector fields $\{V^y : C \ni x \mapsto x - y \in \mathbb{R}^n\}$, $y \in C$ is a family of $(\cdot,\cdot)^H$-gradient vector fields iff there exists a strictly convex function $h \in \mathcal{C}^3(C)$ such that for all $x \in C$, $H(x) = \nabla^2 h(x)$. Additionally, defining $D_h : C \times C \mapsto \mathbb{R}$ by*

$$D_h(y,x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle, \tag{3.3}$$

we obtain $\nabla_H D_h(y,\cdot)(x) = x - y$.

Proof. The set of metrics complying with the "gradient" requirement is denoted by $\mathcal{M}$, that is, $(\cdot,\cdot)_x^H \in \mathcal{M} \Leftrightarrow H \in \mathcal{C}^1(C; \mathbb{S}^n_{++})$ and for all $y \in C$ there exists $\varphi_y \in \mathcal{C}^1(C; \mathbb{R})$, $\nabla_H \varphi_y(x) = x - y$. Let $(x_1, \ldots, x_n)$ denote the canonical coordinates

of $\mathbb{R}^n$, and write $\sum_{i,j} H_{ij}(x)dx_i dx_j$ for $(\cdot,\cdot)_x^H$. By (2.6), the mappings $x \mapsto x-y$, $y \in C$, define a family of $(\cdot,\cdot)_x^H$ gradients iff $k_y : x \mapsto H(x)(x-y)$, $y \in C$, is a family of Euclidean gradients. Setting $\alpha^y(x) = \langle k_y(x), \cdot\rangle$, $x, y \in C$, the problem amounts to finding necessary (and sufficient) conditions under which the 1-forms α^y are all exact. Let $y \in C$. Since C is convex, the Poincaré lemma [35, Theorem V.4.1] states that α^y is exact iff it is closed. In canonical coordinates we have $\alpha^y(x) = \sum_i \left(\sum_k H_{ik}(x)(x_k - y_k)\right) dx_i$, $x \in C$, and therefore α^y is exact iff for all $i, j \in \{1,\dots,n\}$ we have $\frac{\partial}{\partial x_j}\sum_k H_{ik}(x)(x_k-y_k) = \frac{\partial}{\partial x_i}\sum_k H_{jk}(x)(x_k-y_k)$, which is equivalent to $\sum_k \frac{\partial}{\partial x_j} H_{ik}(x)(x_k-y_k) + H_{ij}(x) = \sum_k \frac{\partial}{\partial x_i} H_{jk}(x)(x_k-y_k) + H_{ji}(x)$. Since $H_{ij}(x) = H_{ji}(x)$, this gives the following condition: $\sum_k \frac{\partial}{\partial x_j} H_{ik}(x)(x_k-y_k) = \sum_k \frac{\partial}{\partial x_i} H_{jk}(x)(x_k-y_k)$, for all $i, j \in \{1,\dots,n\}$. If we set $V_x = (\frac{\partial}{\partial x_j} H_{i1}(x), \dots, \frac{\partial}{\partial x_j} H_{in}(x))^T$ and $W_x = (\frac{\partial}{\partial x_i} H_{j1}(x), \dots, \frac{\partial}{\partial x_i} H_{jn}(x))^T$, the latter can be rewritten $\langle V_x - W_x, x-y\rangle = 0$, which must hold for all $(x,y) \in C \times C$. Fix $x \in C$. Let $\epsilon_x > 0$ be such that the open ball of center x with radius ϵ_x is contained in C. For every ν such that $|\nu| = 1$, take $y = x + \epsilon_x/2\nu$ to obtain that $\langle V_x - W_x, \nu\rangle = 0$. Consequently, $V_x = W_x$ for all $x \in C$. Therefore, $(\cdot,\cdot)_x^H \in \mathcal{M}$ iff

$$\forall x \in C,\ \forall i,j,k \in \{1,\dots,n\}, \qquad \frac{\partial}{\partial x_i} H_{jk}(x) = \frac{\partial}{\partial x_j} H_{ik}(x). \tag{3.4}$$

LEMMA 3.2. *If $H : C \mapsto \mathbb{S}^n_{++}$ is a differentiable mapping satisfying* (3.4)*, then there exists $h \in \mathcal{C}^3(C)$ such that, for all $x \in C$, $H(x) = \nabla^2 h(x)$. In particular, h is strictly convex.*

Proof of Lemma 3.2. For all $i \in \{1,\dots,n\}$, set $\beta^i = \sum_k H_{ik}dx_k$. By (3.4), β^i is closed and therefore exact. Let $\phi_i : C \mapsto \mathbb{R}$ be such that $d\phi_i = \beta^i$ on C, and set $\omega = \sum_k \phi_k dx_k$. We have that $\frac{\partial}{\partial x_j}\phi_i(x) = H_{ij}(x) = H_{ji}(x) = \frac{\partial}{\partial x_i}\phi_j(x)$ for all $x \in C$. This proves that ω is closed, and therefore there exists $h \in \mathcal{C}^2(C,\mathbb{R})$ such that $dh = \omega$. To conclude, we just have to notice that $\frac{\partial}{\partial x_i} h(x) = \phi_i$, and thus $\frac{\partial^2 h}{\partial x_j \partial x_i}(x) = H_{ji}(x)$ for all $x \in C$. ∎

To finish the proof of Theorem 3.1, we note that taking $\varphi_y = D_h(y,\cdot)$ with D_h being defined by (3.3), we obtain $\nabla\varphi_y(x) = \nabla^2 h(x)(x-y)$, and therefore $\nabla_H \varphi_y(x) = x - y$ in virtue of (2.6). ∎

Remark 3.1. (a) In the theory of Bregman proximal methods for convex optimization, the distance-like function D_h defined by (3.3) is called the *D-function* of h. Theorem 3.1 is a new and surprising motivation for the introduction of D_h in relation with variational inequality problems. (b) For a geometrical approach to Hessian Riemannian structures, the reader is referred to the recent work of Duistermaat [19].

Theorem 3.1 suggests that we endow C with a Riemannian structure associated with the Hessian $H = \nabla^2 h$ of a strictly convex function $h : C \mapsto \mathbb{R}$. As we will see under some additional conditions, the D-function of h is essential to establishing the asymptotic convergence of the trajectory. On the other hand, if it is possible to replace h by a sufficiently smooth strictly convex function $h' : C' \mapsto \mathbb{R}$ with $C' \supset\supset C$ and $h'_{|_C} = h$, then the gradient flows for h and h' are the same on C, but the steepest descent trajectories associated with the latter may leave the feasible set of (P) and in general they will not converge to a solution of (P). We shall see that to avoid this drawback it is sufficient to require that $|\nabla h(x^j)| \to +\infty$ for all sequences (x^j) in C converging to a boundary point of C. This may be interpreted as a sort of *barrier technique*, a classical strategy for enforcing feasibility in optimization theory.

3.2. Legendre-type functions and the (H-SD) dynamical system. In what follows, we adopt the standard notation of convex analysis theory; see [39]. Given a closed convex subset S of $\mathbb{R}^n$, we say that an extended-real-valued function $g : S \mapsto \mathbb{R} \cup \{+\infty\}$ belongs to the class $\Gamma_0(S)$ when g is lower semicontinuous, proper ($g \not\equiv +\infty$), and convex. For such a function $g \in \Gamma_0(S)$, its *effective domain* is defined by $\operatorname{dom} g = \{x \in S \mid g(x) < +\infty\}$. When $g \in \Gamma_0(\mathbb{R}^n)$, its *Legendre–Fenchel conjugate* is given by $g^*(y) = \sup\{\langle x, y\rangle - g(x) \mid x \in \mathbb{R}^n\}$, and its *subdifferential* is the set-valued mapping $\partial g : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ given by $\partial g(x) = \{y \in \mathbb{R}^n \mid \forall z \in \mathbb{R}^n,\ f(x) + \langle y, z - x\rangle \leq f(z)\}$. We set $\operatorname{dom} \partial g = \{x \in \mathbb{R}^n \mid \partial g(x) \neq \emptyset\}$.

DEFINITION 3.3 (see [39, Chapter 26]). *A function $h \in \Gamma_0(\mathbb{R}^n)$ is called:*

(i) essentially smooth *if h is differentiable on* $\operatorname{int}\operatorname{dom} h$*, with, moreover, $|\nabla h(x^j)| \to +\infty$ for every sequence $(x^j) \subset \operatorname{int} \operatorname{dom} h$ converging to a boundary point of $\operatorname{dom} h$ as $j \to +\infty$;*

(ii) of Legendre type *if h is essentially smooth and strictly convex on* $\operatorname{int} \operatorname{dom} h$.

We remark that by [39, Theorem 26.1], $h \in \Gamma_0(\mathbb{R}^n)$ is essentially smooth iff $\partial h(x) = \{\nabla h(x)\}$ if $x \in \operatorname{int} \operatorname{dom} h$, and $\partial h(x) = \emptyset$ otherwise; in particular, $\operatorname{dom} \partial h = \operatorname{int} \operatorname{dom} h$.

Motivated by the results of section 3.1, we define a Riemannian structure on C by introducing a function $h \in \Gamma_0(\mathbb{R}^n)$ such that

$$(\mathrm{H}_0) \qquad \begin{cases} \text{(i)} & h \text{ is of Legendre type with } \operatorname{int} \operatorname{dom} h = C. \\ \text{(ii)} & h_{|C} \in \mathcal{C}^2(C;\mathbb{R}) \text{ and } \forall x \in C, \nabla^2 h(x) \in \mathbb{S}^n_{++}. \\ \text{(iii)} & \text{The mapping } C \ni x \mapsto \nabla^2 h(x) \text{ is locally Lipschitz continuous.} \end{cases}$$

Here and subsequently, we take $H = \nabla^2 h$ with h satisfying (H_0). The Hessian mapping $C \ni x \mapsto H(x)$ endows C with the (locally Lipschitz continuous) Riemannian metric

$$\forall x \in C,\ \forall u, v \in \mathbb{R}^n, \qquad (u, v)^H_x = \langle H(x)u, v\rangle = \langle \nabla^2 h(x)u, v\rangle, \tag{3.5}$$

and we say that $(\cdot,\cdot)^H_x$ is the *Legendre metric* on C induced by the Legendre-type function h, which also defines a metric on $\mathcal{F} = C \cap \mathcal{A}$ by restriction. In addition to $f \in \mathcal{C}^1(\mathbb{R}^n)$, we suppose that the objective function satisfies

$$\nabla f \text{ is locally Lipschitz continuous on } \mathbb{R}^n. \tag{3.6}$$

The corresponding steepest descent method in the manifold $\mathcal{F}, (\cdot,\cdot)^H_x$, which we refer to as (H-SD) for short, is then the following continuous dynamical system:

$$\text{(H-SD)} \qquad \begin{cases} \dot{x}(t) + \nabla_{H} f_{|\mathcal{F}}(x(t)) = 0, & t \in (T_m, T_M), \\ x(0) = x^0 \in \mathcal{F}, \end{cases}$$

with $H = \nabla^2 h$ and where $-\infty \leq T_m < 0 < T_M \leq +\infty$ defines the interval corresponding to the unique maximal solution of (H-SD). Given an initial condition $x^0 \in \mathcal{F}$, we shall say that (H-SD) is *well posed* when its maximal solution satisfies $T_M = +\infty$. In section 4.1 we will give some sufficient conditions ensuring the well-posedness of (H-SD).

3.3. Differential inclusion formulation of (H-SD) and some consequences. It is easily seen that the solution $x(t)$ of (H-SD) satisfies

$$\begin{cases} \dfrac{d}{dt}\nabla h(x(t)) + \nabla f(x(t)) \ \in\ \mathcal{A}_0^{\perp} \text{ on } (T_m, T_M), \\ x(t) \ \in\ \mathcal{F} \ \text{ on } (T_m, T_M), \\ x(0) \ =\ x^0 \in \mathcal{F}. \end{cases} \tag{3.7}$$

This differential inclusion problem makes sense even when $x \in W^{1,1}_{loc}(T_m, T_M; \mathbb{R}^n)$, the inclusions being satisfied almost everywhere on (T_m, T_M). Actually, the following result establishes that (H-SD) and (3.7) describe the same trajectory.

PROPOSITION 3.4. *Let $x \in W^{1,1}_{loc}(T_m, T_M; \mathbb{R}^n)$. Then, x is a solution of* (3.7) *iff x is the solution of* (H-SD). *In particular,* (3.7) *admits a unique solution of class $\mathcal{C}^1$.*

Proof. Assume that x is a solution of (3.7), and let I' be the subset of (T_m, T_M) on which $t \mapsto (x(t), \nabla h(x(t))$ is derivable. We may assume that $x(t) \in \mathcal{F}$ and $\frac{d}{dt}\nabla h(x(t)) + \nabla f(x(t)) \in \mathcal{A}_0^\perp$ for all $t \in I'$. Since x is absolutely continuous, $\dot{x}(t) + H(x(t))^{-1}\nabla f((x(t)) \in H(x(t))^{-1}\mathcal{A}_0^\perp$ and $\dot{x}(t) \in \mathcal{A}_0$ for all $t \in I'$. However, the orthogonal complement of $\mathcal{A}_0$ with respect to the inner product $\langle H(x)\cdot, \cdot\rangle$ is exactly $H(x)^{-1}\mathcal{A}_0^\perp$ when $x \in \mathcal{F}$. It follows that $\dot{x} + P_x H(x)^{-1}\nabla f(x) = 0$ on I'. This implies that x is the $\mathcal{C}^1$ solution of (H-SD). ∎

Suppose that f is convex. On account of Proposition 3.4, (H-SD) can be interpreted as a continuous-time model for a well-known class of iterative minimization algorithms. In fact, an implicit discretization of (3.7) yields the following iterative scheme: $\nabla h(x^{k+1}) - \nabla h(x^k) + \mu_k \nabla f(x^{k+1}) \in \operatorname{Im} A^T,\ Ax^{k+1} = b$, where $\mu_k > 0$ is a step-size parameter and $x^0 \in \mathcal{F}$. This is the optimality condition for

$$x^{k+1} \in \operatorname{Argmin}\left\{ f(x) + 1/\mu_k D_h(x, x^k) \mid Ax = b \right\}, \tag{3.8}$$

where D_h is given by

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y\rangle, \qquad x \in \operatorname{dom} h,\ y \in \operatorname{dom} \partial h = C. \tag{3.9}$$

The above algorithm is accordingly called the *Bregman proximal minimization* method; for an insight into its importance in optimization see, for instance, [5, 14, 15, 28, 34].

Next, assume that $f(x) = \langle c, x\rangle$ for some $c \in \mathbb{R}^n$. As already noticed in [6, 23, 36] for the log-metric and in [29] for a fairly general h, in this case the (H-SD) gradient trajectory can be viewed as a *central optimal path*. Indeed, integrating (3.7) over $[0, t]$, we obtain $\nabla h(x(t)) - \nabla h(x^0) + tc \in \mathcal{A}_0^\perp$. Since $x(t) \in \mathcal{A}$, it follows that

$$x(t) \in \operatorname{Argmin}\left\{ \langle c, x\rangle + 1/t D_h(x, x^0) \mid Ax = b \right\}, \tag{3.10}$$

which corresponds to the so-called *viscosity method* relative to $g(x) = D_h(x, x^0)$; see [2, 4, 29] and Corollary 4.8. We note now that, for a linear objective function, (3.8) and (3.10) are essentially the same: the sequence generated by the former belongs to the optimal path defined by the latter. Indeed, setting $t_0 = 0$ and $t_{k+1} = t_k + \mu_k$ for all $k \geq 0$ ($\mu_0 = 0$) and integrating (3.7) over $[t_k, t_{k+1}]$, we obtain that $x(t_{k+1})$ satisfies the optimality condition for (3.8). The following result summarizes the previous discussion.

PROPOSITION 3.5. *Assume that f is linear and that the corresponding* (H-SD) *dynamical system is well posed. Then, the viscosity optimal path $\widetilde{x}(\varepsilon)$ relative to $g(x) = D_h(x, x^0)$ and the sequence (x^k) generated by* (3.8) *exist and are unique, with in addition $\widetilde{x}(\varepsilon) = x(1/\varepsilon)$ for all $\varepsilon > 0$, and $x^k = x(\sum_{l=0}^{k-1} \mu_l)$ for all $k \geq 1$, where $x(t)$ is the solution of* (H-SD).

Remark 3.2. In order to ensure asymptotic convergence for proximal-type algorithms, it is usually required that the step-size parameters satisfy $\sum \mu_k = +\infty$. By Proposition 3.5, this is necessary for the convergence of (3.8) in the sense that when (H-SD) is well posed, if x^k converges to some $x^* \in S(P)$, then either $x^0 = x^*$ or $\sum \mu_k = +\infty$.

4. Global existence, asymptotic analysis, and examples.

4.1. Well-posedness of (H-SD). In this section we establish the well-posedness of (H-SD) (i.e., $T_M = +\infty$) under three different conditions. In order to avoid any confusion, we say that a set $E \subset \mathbb{R}^n$ is *bounded* when it is so for the usual Euclidean norm $|y| = \sqrt{\langle y, y\rangle}$. First, we propose the following condition:

(WP_1) The lower level set $\{y \in \overline{\mathcal{F}} \mid f(y) \leq f(x^0)\}$ is bounded.

Notice that (WP_1) is weaker than the classical assumption requiring f to have bounded lower level sets in the H metric sense. Next, let D_h be the D-function of h that is defined by (3.9) and consider the following condition:

(WP_2)
$$\begin{cases} \text{(i) } \operatorname{dom} h = \overline{C} \text{ and } \forall a \in \overline{C},\ \forall \gamma \in \mathbb{R},\ \{y \in \mathcal{F} \,|D_h(a, y) \leq \gamma\} \text{ is bounded.} \\ \text{(ii) } S(P) \neq \emptyset \text{ and } f \text{ is quasi-convex (i.e., the lower level sets of } f \text{ are convex).} \end{cases}$$

When $\overline{\mathcal{F}}$ is unbounded (WP_1) and (WP_2) involve some a priori properties on f. This is actually not necessary for the well-posedness of (H-SD). Consider

(WP_3) $\exists\ K \geq 0,\ L \in \mathbb{R}$ such that $\forall x \in C,\ ||H(x)^{-1}|| \leq K|x| + L.$

This property is satisfied by relevant Legendre-type functions; take, for instance, (4.13).

THEOREM 4.1. *Assume that* (3.6) *and* (H_0) *hold and additionally that either* (WP_1)*,* (WP_2)*, or* (WP_3) *is satisfied. If* $\inf_{\mathcal{F}} f > -\infty$*, then the dynamical system* (H-SD) *is well posed. Consequently, the mapping* $t \mapsto f(x(t))$ *is nonincreasing and convergent as* $t \to +\infty$*.*

Proof. When no confusion may occur, we drop the dependence on the time variable t. By definition,

$$T_M = \sup\{T > 0 \,|\exists!\text{ solution } x \text{ of } (H\text{-}SD) \text{ on } [0, T) \text{ s.t. } x([0, T)) \subset \mathcal{F}\}.$$

We have that $T_M > 0$. The definition (2.8) of P_x implies that, for all $y \in \mathcal{A}_0$, $(H(x)^{-1}\nabla f(x) + \dot{x}, y + \dot{x})_x^H = 0$ on $[0, T_M)$ and therefore

$$\langle \nabla f(x) + H(x)\dot{x}, y + \dot{x}\rangle = 0 \text{ on } [0, T_M). \tag{4.1}$$

Letting $y = 0$ in (4.1) yields

$$\frac{d}{dt} f(x) + \langle H(x)\dot{x}, \dot{x}\rangle = 0. \tag{4.2}$$

By (H_0)(ii), $f(x(t))$ is convergent as $t \to T_M$. Moreover,

$$\langle H(x(\cdot))\dot{x}(\cdot), \dot{x}(\cdot)\rangle \in L^1(0, T_M; \mathbb{R}). \tag{4.3}$$

Suppose that $T_M < +\infty$. To obtain a contradiction, we begin by proving that x is bounded. If (WP_1) holds, then x is bounded because $f(x(t))$ is nonincreasing so that $x(t) \in \{y \in \overline{\mathcal{F}} \,|f(y) \leq f(x^0)\}$ for all $t \in [0, T_M)$. Assume now that f and h comply with (WP_2), and let $a \in \overline{\mathcal{F}}$. For each $t \in [0, T_M)$ take $y = x(t) - a$ in (4.1) to obtain $\langle \nabla f(x) + \frac{d}{dt}\nabla h(x), x - a + \dot{x}\rangle = 0$. By (4.2), this gives $\langle \frac{d}{dt}\nabla h(x), x - a\rangle + \langle \nabla f(x), x - a\rangle = 0$, which we rewrite as

$$\frac{d}{dt} D_h(a, x(t)) + \langle \nabla f(x(t)), x(t) - a\rangle = 0 \qquad \forall\ t \in [0, T_M). \tag{4.4}$$

Now let $a \in \overline{\mathcal{F}}$ be a minimizer of f on $\overline{\mathcal{F}}$. From the quasi-convexity property of f, it follows that, for all $t \in [0, T_M)$, $\langle \nabla f(x(t)), x(t) - a\rangle \geq 0$. Therefore, $D_h(a, x(t))$ is nonincreasing, and (WP_2)(ii) implies that x is bounded. Suppose that (WP_3) holds and fix $t \in [0, T_M)$; we have $|x(t)-x^0| \leq \int_0^t |\dot{x}(s)|ds \leq \int_0^t ||\sqrt{H(x(s))^{-1}}||\sqrt{H(x(s))}\dot{x}(s)|ds \leq (\int_0^t ||H(x(s))^{-1}||ds)^{1/2}(\int_0^t \langle H(x(s))\dot{x}(s), \dot{x}(s)\rangle ds)^{1/2}$. The latter follows from the Cauchy–Schwarz inequality, together with the fact that $\|H(x)\|^2$ is the biggest eigenvalue of $H(x)$. Thus $|x(t) - x^0| \leq 1/2[\int_0^t ||H(x(s))^{-1}||ds + \int_0^t \langle H(x(s))\dot{x}(s), \dot{x}(r)\rangle ds]$. Combining (WP_3) and (4.3), Gronwall's lemma yields the boundedness of x.

Let $\omega(x^0)$ be the set of limit points of x, and set $K = x([0, T_M)) \cup \omega(x^0)$. Since x is bounded, $\omega(x^0) \neq \emptyset$ and K is compact. If $K \subset C$, then the compactness of K implies that x can be extended beyond T_M, which contradicts the maximality of T_M. Let us prove $K \subset C$. We argue again by contradiction. Assume that $x(t_j) \to x^*$, with $t_j < T_M$, $t_j \to T_M$ as $j \to +\infty$, and $x^* \in \mathrm{bd} C = \overline{C} \backslash C$. Since h is of Legendre type, we have $|\nabla h(x(t_j))| \to +\infty$, and we may assume that $\nabla h(x(t_j))/|\nabla h(x(t_j))| \to \nu \in \mathbb{R}^n$ with $|\nu| = 1$.

LEMMA 4.2. *If $(x^j) \subset C$ is such that $x^j \to x^* \in \mathrm{bd}\, C$ and $\nabla h(x^j)/|\nabla h(x^j))| \to \nu \in \mathbb{R}^n$, h being a function of Legendre type with $C = \mathrm{int\,dom}\, h$, then $\nu \in N_{\overline{C}}(x^*)$.*

Proof of Lemma 4.2. By convexity of h, $\langle \nabla h(x^j) - \nabla h(y), x^j - y\rangle \geq 0$ for all $y \in C$. Dividing by $|\nabla h(x^j)|$ and letting $j \to +\infty$, we get $\langle \nu, y - x^*\rangle \leq 0$ for all $y \in C$, which holds also for $y \in \overline{C}$. Hence, $\nu \in N_{\overline{C}}(x^*)$. ∎

Therefore, $\nu \in N_{\overline{C}}(x^*)$. Let $\nu_0 = \Pi_{\mathcal{A}_0}\nu$ be the Euclidean orthogonal projection of ν onto $\mathcal{A}_0$, and take $y = \nu_0$ in (4.1). Using (4.2), integration gives

$$\langle \nabla h(x(t_j)), \nu_0\rangle = \left\langle \nabla h(x^0) - \int_0^{t_j} \nabla f(x(s))ds, \nu_0 \right\rangle. \tag{4.5}$$

By (H_0) and the boundedness property of x, the right-hand side of (4.5) is bounded under the assumption $T_M < +\infty$. Hence, to draw a contradiction from (4.5) it suffices to prove $\langle \nabla h(x(t_j)), \nu_0\rangle \to +\infty$. Since $\langle \nabla h(x(t_j))/|\nabla h(x(t_j))|, \nu_0\rangle \to |\nu_0|^2$, the proof of the result is complete if we check that $\nu_0 \neq 0$. This is a direct consequence of the following claim.

LEMMA 4.3. *Let C be a nonempty open convex subset of $\mathbb{R}^n$, and $\mathcal{A}$ an affine subspace of $\mathbb{R}^n$ such that $C \cap \mathcal{A} \neq \emptyset$. If $x^* \in (\mathrm{bd}\, C) \cap \mathcal{A}$, then $N_{\overline{C}}(x^*) \cap \mathcal{A}_0^{\perp} = \{0\}$ with $\mathcal{A}_0 = \mathcal{A} - \mathcal{A}$.*

Proof of Lemma 4.3. Let us argue by contradiction and suppose that we can pick some $v \neq 0$ in $\mathcal{A}_0^{\perp} \cap N_{\overline{C}}(x^*)$. For $y_0 \in C \cap \mathcal{A}$ we have $\langle v, x^* - y_0\rangle = 0$. For $r \geq 0$, $z \in \mathbb{R}^n$, let $B(z, r)$ denote the ball with center z and radius r. There exists $\epsilon > 0$ such that $B(y_0, \epsilon) \subset C$. Take w in $B(0, \epsilon)$ such that $\langle v, w\rangle < 0$; then $y_0 + w \in C$, and yet $\langle v, x^* - (y_0 + w)\rangle = \langle v, w\rangle < 0$. This contradicts the fact that v is in $N_{\overline{C}}(x^*)$. ∎

This completes the proof of the theorem. ∎

4.2. Value convergence for a convex objective function. As a first result concerning the asymptotic behavior of (H-SD), we have the following.

PROPOSITION 4.4. *If* (H-SD) *is well-posed and f is convex, then for all $a \in \mathcal{F}$, for all $t > 0$, $f(x(t)) \leq f(a) + \frac{1}{t} D_h(a, x^0)$, where D_h is defined by* (3.9)*; hence* $\lim_{t\to+\infty} f(x(t)) = \inf_{\overline{\mathcal{F}}} f$.

Proof. We begin by noticing that $f(x(t))$ converges as $t \to +\infty$ (see Theorem 4.1). Fix $a \in \mathcal{F}$. By (4.4), we have that the solution $x(t)$ of $(H\text{-}SD)$ satisfies $\frac{d}{dt} D_h(a, x(t)) + \langle \nabla f(x(t)), x(t) - a\rangle = 0$, $\forall t \geq 0$. The convex inequality$f(x) + \langle \nabla f(x), x - a\rangle \leq f(a)$ yields $D_h(a, x(t)) + \int_0^t [f(x(s)) - f(a)]ds \leq D_h(a, x^0)$. Using that

$D_h \geq 0$ and since $f(x(t))$ is non-increasing, we get the estimate. Letting $t \to +\infty$, it follows that $\lim_{t\to+\infty} f(x(t)) \leq f(a)$. Since $a \in \mathcal{F}$ was arbitrary chosen, the proof is complete. ∎

4.3. Bregman metrics and trajectory convergence. In this section we establish the convergence of $x(t)$ under some additional properties on the D-function of h. Let us begin with a definition.

DEFINITION 4.5. *A function $h \in \Gamma_0(\mathbb{R}^n)$ is called a* Bregman function with zone C *when the following conditions are satisfied:*

(i) $\operatorname{dom} h = \overline{C}$, *$h$ is continuous and strictly convex on $\overline{C}$ and $h_{|C} \in \mathcal{C}^1(C;\mathbb{R})$.*

(ii) *for all $a \in \overline{C}$, for all $\gamma \in \mathbb{R}$, $\{y \in C | D_h(a,y) \leq \gamma\}$ is bounded, where D_h is defined by* (3.9).

(iii) *for all $y \in \overline{C}$, for all $y^j \to y$ with $y^j \in C$, $D_h(y, y^j) \to 0$.*

Observe that this notion slightly weakens the usual definition of Bregman function that was proposed by Censor and Lent in [13]; see also [9]. Actually, a Bregman function in the sense of Definition 4.5 belongs to the class of B-functions introduced by Kiwiel (see [33, Definition 2.4]). Recall the following important asymptotic separation property.

LEMMA 4.6 (see [33, Lemma 2.16]). *If h is a Bregman function with zone C, then for all $y \in \overline{C}$, for all $(y^j) \subset C$ such that $D_h(y, y^j) \to 0$, we have $y^j \to y$.*

THEOREM 4.7. *Suppose that* (H_0) *holds, with h being a Bregman function with zone C. If f is quasi-convex satisfying* (3.6) *and $S(P) \neq \emptyset$, then* (H-SD) *is well-posed, and its solution $x(t)$ converges as $t \to +\infty$ to some $x^* \in \overline{\mathcal{F}}$ with $-\nabla f(x^*) \in N_{\overline{C}}(x^*) + \mathcal{A}_0^{\perp}$. If in addition f is convex then $x(t)$ converges to a solution of* (P).

Proof. Notice first that (WP_2) is satisfied. By Theorem 4.1, (H-SD) is well-posed, $x(t)$ is bounded, and for each $a \in S(P)$, $D_h(a, x(t))$ is nonincreasing and hence convergent. Set $f_\infty = \lim_{t\to+\infty} f(x(t))$ and define $L = \{y \in \overline{\mathcal{F}} \mid f(y) \leq f_\infty\}$. The set L is nonempty and closed. Since f is supposed to be quasi-convex, L is convex, and similar arguments as in the proof of Theorem 4.1 under (WP_2) show that $D_h(a, x(t))$ is convergent for all $a \in L$. Let $x^* \in L$ denote a cluster point of $x(t)$, and take $t_j \to +\infty$ such that $x(t_j) \to x^*$. Then, by Definition 4.5(iii), $\lim_t D_h(x^*, x(t)) = \lim_j D_h(x^*, x(t_j)) = 0$. Therefore, $x(t) \to x^*$, thanks to Lemma 4.6. Let us prove that x^* satisfies the optimality condition $-\nabla f(x^*) \in N_{\overline{C}}(x^*) + \mathcal{A}_0^{\perp}$. Fix $z \in \mathcal{A}_0$, and for each $t \geq 0$ take $y = -\dot{x}(t) + z$ in (4.1) to obtain $\langle \frac{d}{dt}\nabla h(x(t)) + \nabla f(x(t)), z\rangle = 0$. This gives

$$\frac{1}{t}\int_0^t \langle \nabla f(x(s)), z\rangle ds = \langle s(t), z\rangle, \tag{4.6}$$

where $s(t) = [\nabla h(x^0) - \nabla h(x(t))]/t$. If $x^* \in \mathcal{F}$, then $\nabla h(x(t)) \to \nabla h(x^*)$, and hence $\langle \nabla f(x^*), z\rangle = \lim_{t\to+\infty} \frac{1}{t}\int_0^t \langle \nabla f(x(s)), z\rangle ds = \lim_{t\to+\infty} \langle s(t), z\rangle = 0$. Thus $\Pi_{\mathcal{A}_0}\nabla f(x^*) = 0$. However, $N_{\overline{\mathcal{F}}}(x^*) = \mathcal{A}_0^{\perp}$ when $x^* \in \mathcal{F}$, which proves our claim in this case. Assume now that $x^* \notin \mathcal{F}$, which implies that $x^* \in \partial C \cap \mathcal{A}$. By (4.6), we have that $\langle s(t), z\rangle$ converges to $\langle \nabla f(x^*), z\rangle$ as $t \to +\infty$ for all $z \in \mathcal{A}_0$, and therefore $\Pi_{\mathcal{A}_0} s(t) \to \Pi_{\mathcal{A}_0}\nabla f(x^*)$ as $t \to +\infty$. On the other hand, by Lemma 4.2, we have that there exists $\nu \in -N_{\overline{C}}(x^*)$ with $|\nu| = 1$ such that $\nabla h(x(t_j))/|\nabla h(x(t_j))| \to \nu$ for some $t_j \to +\infty$. Since $N_{\overline{C}}(x^*)$ is positively homogeneous, we deduce that there exists a $\bar{\nu} \in -N_{\overline{C}}(x^*)$ such that $\Pi_{\mathcal{A}_0}\nabla f(x^*) = \Pi_{\mathcal{A}_0}\bar{\nu}$. Thus, $-\nabla f(x^*) \in -\Pi_{\mathcal{A}_0}\bar{\nu} + \mathcal{A}_0^{\perp} \subseteq N_{\overline{C}}(x^*) + \mathcal{A}_0^{\perp}$, which proves the theorem. ∎

Following [29], we remark that when f is linear, the limit point can be characterized as a sort of "D_h-projection" of the initial condition onto the optimal set $S(P)$.

In fact, we have the following result.

COROLLARY 4.8. *Under the assumptions of Theorem* 4.7*, if f is linear, then the solution $x(t)$ of* (H-SD) *converges as $t \to +\infty$ to the unique optimal solution x^* of*

$$\min_{x \in S(P)} D_h(x, x^0). \tag{4.7}$$

Proof. Let $x^* \in S(P)$ be such that $x(t) \to x^*$ as $t \to +\infty$. Let $\bar{x} \in S(P)$. Since $x(t) \in \mathcal{F}$, the optimality of $\bar{x}$ yields $f(x(t)) \geq f(\bar{x})$, and it follows from (3.10) that $D_h(x(t), x^0) \leq D_h(\bar{x}, x^0)$. Letting $t \to +\infty$ in the last inequality, we deduce that x^* solves (4.7). Noticing that $D_h(\cdot, x^0)$ is strictly convex due to Definition 4.5(i), we conclude the result. ∎

We finish this section with an abstract result concerning the rate of convergence under uniqueness of the optimal solution. We will apply this result in the next section. Suppose that f is convex and satisfies (2.3) and (3.6), with in addition $S(P) = \{a\}$. Given a Bregman function h complying with (H_0), consider the following growth condition:

$$f(x) - f(a) \geq \alpha D_h(a, x)^\beta \qquad \forall x \in U_a \cap \overline{C}, \tag{GC}$$

where U_a is a neighborhood of a and with $\alpha > 0$, $\beta \geq 1$. The next abstract result gives an estimation of the convergence rate with respect to the D-function of h.

PROPOSITION 4.9. *Assume that f and h satisfy the above conditions, and let $x : [0, +\infty) \to \mathcal{F}$ be the solution of* (H-SD)*. Then we have the following estimations:*

- *if $\beta = 1$, then there exists $K > 0$ such that $D_h(a, x(t)) \leq Ke^{-\alpha t}$ for all $t > 0$;*
- *if $\beta > 1$, then there exists $K' > 0$ such that $D_h(a, x(t)) \leq K'/t^{\frac{1}{\beta - 1}}$ for all $t > 0$.*

Proof. The assumptions of Theorem 4.7 are satisfied; this yields the well-posedness of (H-SD) and the convergence of $x(t)$ to a as $t \to +\infty$. Additionally, from (4.4) it follows that for all $t \geq 0$, $\frac{d}{dt} D_h(a, x(t)) + \langle \nabla f(x(t)), x(t) - a \rangle = 0$. By the convexity of f, we have $\frac{d}{dt} D_h(a, x(t)) + f(x(t)) - f(a) \leq 0$. Since $x(t) \to a$, there exists t_0 such that for all $t \geq t_0$, $x(t) \in U_a \cap \mathcal{F}$. Therefore by combining (GC) and the last inequality, it follows that

$$\frac{d}{dt} D_h(a, x(t)) + \alpha D_h(a, x(t))^\beta \leq 0 \qquad \forall t \geq t_0. \tag{4.8}$$

In order to integrate this differential inequality, let us first observe that we have the following equivalence: $D_h(a, x(t)) > 0$ for all $t \geq 0$ iff $x^0 \neq a$. Indeed, if $a \in \overline{\mathcal{F}} \setminus \mathcal{F}$, then the equivalence follows from $x(t) \in \mathcal{F}$ together with Lemma 4.6; if $a \in \mathcal{F}$, then the optimality condition that is satisfied by a is $\Pi_{\mathcal{A}_0} \nabla f(a) = 0$, and the equivalence is a consequence of the uniqueness of the solution $x(t)$ of (H-SD). Hence, we can assume that $x^0 \neq a$ and divide (4.8) by $D_h(a, x(t))^\beta$ for all $t \geq t_0$. A simple integration procedure then yields the result. ∎

4.4. Examples: Interior point flows in convex programming. This section gives a systematic method for constructing explicit Legendre metrics on a quite general class of convex sets. By so doing, we will also show that many systems studied earlier by various authors [6, 31, 20, 23, 36] appear as particular cases of (H-SD) systems.

Let $p \geq 1$ be an integer, and set $I = \{1, \ldots, p\}$. Let us assume that to each $i \in I$ there corresponds a $\mathcal{C}^3$ concave function $g_i : \mathbb{R}^n \to \mathbb{R}$ such that

$$\exists x^0 \in \mathbb{R}^n, \quad \text{s.t. } \forall i \in I,\ g_i(x^0) > 0. \tag{4.9}$$

Suppose that the open convex set C is given by

$$C = \{x \in \mathbb{R}^n \mid g_i(x) > 0, i \in I\}. \tag{4.10}$$

By (4.9) we have that $C \neq \emptyset$ and $\overline{C} = \{x \in \mathbb{R}^n \mid g_i(x) \geq 0, i \in I\}$. Let us introduce a class of convex functions of Legendre type $\theta \in \Gamma_0(\mathbb{R})$ satisfying

$$(\mathrm{H}_1) \quad \begin{cases} \text{(i) } (0,\infty) \subset \operatorname{dom}\theta \subset [0,\infty); \\ \text{(ii) } \theta \in \mathcal{C}^3(0,\infty) \text{ and } \lim_{s\to 0^+} \theta'(s) = -\infty; \\ \text{(iii) for all } s > 0,\ \theta''(s) > 0; \\ \text{(iv) either } \theta \text{ is nonincreasing or for all } i \in I, g_i \text{ is an affine function.} \end{cases}$$

Proposition 4.10. *Under* (4.9) *and* (H_1)*, the function* $h \in \Gamma_0(\mathbb{R}^n)$ *defined by*

$$h(x) = \sum_{i\in I} \theta(g_i(x)) \tag{4.11}$$

is essentially smooth, with $\operatorname{int}\operatorname{dom} h = C$ *and* $h \in \mathcal{C}^3(C)$*, where* C *is given by* (4.10)*. If we assume in addition the nondegeneracy condition*

$$\forall x \in C, \qquad \operatorname{span}\{\nabla g_i(x) \mid i \in I\} = \mathbb{R}^n, \tag{4.12}$$

then $H = \nabla^2 h$ *is positive definite on* C*, and consequently* h *satisfies* (H_0)*.*

Proof. Define $h_i \in \Gamma_0(\mathbb{R}^n)$ by $h_i(x) = \theta(g_i(x))$. We have that for all $i \in I$, $C \subset \operatorname{dom} h_i$. Hence $\operatorname{int}\operatorname{dom} h = \bigcap_{i\in I} \operatorname{int}\operatorname{dom} h_i \supseteq C \neq \emptyset$, and by [39, Theorem 23.8] we conclude that $\partial h(x) = \sum_{i\in I} \partial h_i(x)$ for all $x \in \mathbb{R}^n$. But $\partial h_i(x) = \theta'(g_i(x))\nabla g_i(x)$ if $g_i(x) > 0$, and $\partial h_i(x) = \emptyset$ if $g_i(x) \leq 0$; see [26, Theorem IX.3.6.1]. Therefore $\partial h(x) = \sum_{i\in I} \theta'(g_i(x))\nabla g_i(x)$ if $x \in C$, and $\partial h(x) = \emptyset$ otherwise. Since ∂h is a single-valued mapping, it follows from [39, Theorem 26.1] that h is essentially smooth and $\operatorname{int}\operatorname{dom} h = \operatorname{dom}\partial h = C$. Clearly, h is of class $\mathcal{C}^3$ on C. Assume now that (4.12) holds. For $x \in C$, we have $\nabla^2 h(x) = \sum_{i\in I} \theta''(g_i(x))\nabla g_i(x)\nabla g_i(x)^T + \sum_{i\in I} \theta'(g_i(x))\nabla^2 g_i(x)$. By (H_1)(iv), it follows that for any $v \in \mathbb{R}^n$, $\sum_{i\in I} \theta'(g_i(x))\langle \nabla^2 g_i(x)v, v\rangle \geq 0$. Let $v \in \mathbb{R}^n$ be such that $\langle \nabla^2 h(x)v, v\rangle = 0$, which yields $\sum_{i\in I} \theta''(g_i(x))\langle v, \nabla g_i(x)\rangle^2 = 0$. According to (H_1)(iii), the latter implies that $v \in \operatorname{span}\{\nabla g_i(x) | i \in I\}^\perp = \{0\}$. Hence $\nabla^2 h(x) \in \mathbb{S}^n_{++}$, and the proof is complete. ∎

If h is defined by (4.11) with $\theta \in \Gamma_0(\mathbb{R})$ satisfying (H_1), we say that θ is the *Legendre kernel* of h. Such kernels can be divided into two classes. The first class corresponds to those kernels θ for which $\operatorname{dom}\theta = (0,\infty)$ so that $\theta(0) = +\infty$, and these kernels are associated with *interior barrier* methods in optimization such as, for instance, the log-barrier $\theta_1(s) = -\ln(s)$, $s > 0$, and the inverse barrier $\theta_2(s) = 1/s$, $s > 0$. The kernels θ belonging to the second class satisfy $\theta(0) < +\infty$ and are connected with the notion of a *Bregman function* in proximal algorithm theory. Here are some examples: the Boltzmann–Shannon entropy $\theta_3(s) = s\ln(s) - s$, $s \geq 0$ (with $0\ln 0 = 0$); $\theta_4(s) = -\frac{1}{\gamma}s^\gamma$ with $\gamma \in (0,1)$, $s \geq 0$ (Kiwiel [33]); $\theta_5(s) = (\gamma s - s^\gamma)/(1-\gamma)$ with $\gamma \in (0,1)$, $s \geq 0$ (Teboulle [42]); the "$x\log x$" entropy $\theta_6(s) = s\ln s$, $s \geq 0$. In relation with Theorem 4.7 given in the previous section, note that the Legendre kernels θ_i, $i = 3,\ldots,6$, are all Bregman functions with zone $\mathbb{R}_+$. Moreover, it is easily seen that each corresponding Legendre function h defined by (4.11) is indeed a Bregman function with zone C.

In order to illustrate the type of dynamical systems given by (H-SD), consider the case of positivity constraints where $p = n$ and $g_i(x) = x_i$, $i \in I$. Thus $C = \mathbb{R}^n_{++}$

and $\overline{C} = \mathbb{R}^n_+$. Let us assume that there exists $x^0 \in \mathbb{R}^n_{++}$ such that $Ax^0 = b$. Recall that the corresponding minimization problem is (P), $\min\{f(x) \mid x \geq 0,\ Ax = b\}$, and take first the kernel θ_3 from above. The associated Legendre function (4.11) is given by

$$h(x) = \sum_{i=1}^{n} x_i \ln x_i - x_i, \quad x \in \mathbb{R}^n_+, \tag{4.13}$$

and the differential equation in (H-SD) is given by

$$\dot{x} + [I - XA^T(AXA^T)^{-1}A]X\nabla f(x) = 0, \tag{4.14}$$

where $X = \operatorname{diag}(x_1, \ldots, x_n)$. If $f(x) = \langle c, x\rangle$, for some $c \in \mathbb{R}^n$ and in absence of linear equality constraints, then (4.14) is $\dot{x} + Xc = 0$. The change of coordinates $y = \nabla h(x) = (\ln x_1, \ldots, \ln x_n)$ gives $\dot{y} + c = 0$. Hence, $x(t) = (x_1^0 e^{-c_1 t}, \ldots, x_n^0 e^{-c_n t})$, $t \in \mathbb{R}$, where $x^0 = (x_1^0, \ldots, x_n^0) \in \mathbb{R}^n_{++}$. If $c \in \mathbb{R}^n_+$, then $\inf_{x\in\mathbb{R}^n_+}\langle c, x\rangle = 0$ and $x(t)$ converges to a minimizer of $f = \langle c, \cdot\rangle$ on $\mathbb{R}^n_+$; if $c_{i_0} < 0$ for some i_0, then $\inf_{x\in\mathbb{R}^n_+}\langle c, x\rangle = -\infty$ and $x_{i_0}(t) \to +\infty$ as $t \to +\infty$. Next, take $A = (1, \ldots, 1) \in \mathbb{R}^{1\times n}$ and $b = 1$ so that the feasible set of (P) is given by $\overline{\mathcal{F}} = \Delta_{n-1} = \{x \in \mathbb{R}^n \mid x \geq 0,\ \sum_{i=1}^n x_i = 1\}$, that is, the $(n-1)$-dimensional simplex. In this case, (4.14) corresponds to $\dot{x} + [X - xx^T]\nabla f(x) = 0$, or componentwise

$$\dot{x}_i + x_i\left(\frac{\partial f}{\partial x_i} - \sum_{j=1}^{n} x_j \frac{\partial f}{\partial x_j}\right) = 0, \quad i = 1, \ldots, n. \tag{4.15}$$

For suitable choices of f, this is a *Lotka–Volterra*-type equation that naturally arises in population dynamics theory and, in that context, the structure $(\cdot,\cdot)^H$ with h as in (4.13) is usually referred to as the *Shahshahani* metric; see [1, 27] and the references therein.

Karmarkar studied (4.15) in [31] for a quadratic objective function as a continuous model of the interior point algorithm introduced by him in [30]. Equation (4.14) is studied by Faybusovich in [20, 21, 22] when (P) is a linear program, establishing connections with completely integrable Hamiltonian systems and exponential convergence rate, and by Herzel, Recchini, and Zirilli in [25], who prove quadratic convergence for an explicit discretization.

Take now the log barrier kernel θ_1 and $h(x) = -\sum_{i=1}^n \ln x_i$. Since $\nabla^2 h(x) = X^{-2}$ with X defined as above, the associated differential equation is

$$\dot{x} + [I - X^2A^T(AX^2A^T)^{-1}A]X^2\nabla f(x) = 0. \tag{4.16}$$

This equation was considered by Bayer and Lagarias in [6] for a linear program. In the particular case $f(x) = \langle c, x\rangle$ and without linear equality constraints, (4.16) amounts to $\dot{x} + X^2c = 0$, or $\dot{y} + c = 0$ for $y = \nabla h(x) = -X^{-1}e$, with $e = (1, \ldots, 1) \in \mathbb{R}^n$, which gives $x(t) = \big(1/(1/x_1^0 + c_1 t), \ldots, 1/(1/x_n^0 + c_n t)\big)$, $T_m \leq t \leq T_M$, with $T_m = \max\{-1/x_i^0 c_i \mid c_i > 0\}$ and $T_M = \min\{-1/x_i^0 c_i \mid c_i < 0\}$ (see [6, p. 515]). A similar system was considered in [23, 36] as a continuous log-barrier method for nonlinear inequality constraints and with $\mathcal{A}_0 = \mathbb{R}^n$.

New systems may be derived by choosing other kernels. For instance, taking $h(x) = -1/\gamma \sum_{i=1}^n x_i^\gamma$ with $\gamma \in (0, 1)$, $A = (1, \ldots, 1) \in \mathbb{R}^{1\times n}$, and $b = 1$, we obtain

$$\dot{x}_i + \frac{x_i^{2-\gamma}}{1-\gamma}\left(\frac{\partial f}{\partial x_i} - \sum_{j=1}^{n} \frac{x_j^{2-\gamma}}{\sum_{k=1}^n x_k^{2-\gamma}} \frac{\partial f}{\partial x_j}\right) = 0, \quad i = 1, \ldots, n. \tag{4.17}$$

4.5. Convergence results for linear programming. Let us consider the specific case of a linear program

$$\min_{x\in\mathbb{R}^n}\{\langle c,x\rangle \mid Bx\geq d,\ Ax=b\}, \tag{LP}$$

where A and b are as in section 2.1, $c\in\mathbb{R}^n$, B is a $p\times n$ full rank real matrix with $p\geq n$, and $d\in\mathbb{R}^p$. We assume that the optimal set satisfies

$$S(LP) \text{ is nonempty and bounded} \tag{4.18}$$

and that there exists a Slater point $x^0\in\mathbb{R}^n$, $Bx^0>d$, and $Ax^0=b$. Take the Legendre function

$$h(x)=\sum_{i=1}^{n}\theta(g_i(x)),\quad g_i(x)=\langle B_i,x\rangle-d_i, \tag{4.19}$$

where $B_i\in\mathbb{R}^n$ is the ith row of B and the Legendre kernel θ satisfies (H_1). By (4.18), (WP_1) holds, and therefore (H-SD) is well posed due to Theorem 4.1. Moreover, $x(t)$ is bounded and all its cluster points belong to $S(LP)$ by Proposition 4.4. The variational property (3.10) ensures the convergence of $x(t)$ and gives a variational characterization of the limit as well. Indeed, we have the following result.

PROPOSITION 4.11. *Let h be given by* (4.19) *with θ satisfying* (H_1). *Under* (4.18), *(H-SD) is well posed and $x(t)$ converges as $t\to+\infty$ to the unique solution x^* of*

$$\min_{x\in S(LP)}\sum_{i\notin I_0}D_\theta(g_i(x),g_i(x^0)), \tag{4.20}$$

where $I_0=\{i\in I \mid g_i(x)=0$ for all $x\in S(LP)\}$.

Proof. Assume that $S(LP)$ is not a singleton; otherwise there is nothing to prove. The relative interior ri $S(LP)$ is nonempty, and moreover ri $S(LP)=\{x\in\mathbb{R}^n \mid g_i(x)=0$ for $i\in I_0$, $g_i(x)>0$ for $i\notin I_0$, $Ax=b\}$. By compactness of $S(LP)$ and strict convexity of $\theta\circ g_i$, there exists a unique solution x^* of (4.20). Indeed, it is easy to see that $x^*\in$ ri (LP). Let $\bar{x}\in S(LP)$ and $t_j\to+\infty$ be such that $x(t_j)\to\bar{x}$. It suffices to prove that $\bar{x}=x^*$. When $\theta(0)<+\infty$, the latter follows by the same arguments as in Corollary 4.8. When $\theta(0)=+\infty$, the proof of [4, Theorem 3.1] can be adapted to our setting (see also [29, Theorem 2]). Set $x^*(t)=x(t)-\bar{x}+x^*$. Since $Ax^*(t)=b$ and $D_h(x,x^0)=\sum_{i=1}^m D_\theta(g_i(x),g_i(x^0))$, equation (3.10) gives

$$\langle c,x(t)\rangle+\frac{1}{t}\sum_{i=1}^{m}D_\theta(g_i(x(t)),g_i(x^0))\leq\langle c,x^*(t)\rangle+\frac{1}{t}\sum_{i=1}^{m}D_\theta(g_i(x^*(t)),g_i(x^0)). \tag{4.21}$$

However, $\langle c,x(t)\rangle=\langle c,x^*(t)\rangle$ and for all $i\in I_0$, $g_i(x^*(t))=g_i(x(t))>0$. Since $x^*\in$ ri $S(LP)$, for all $i\notin I_0$ and j large enough, $g_i(x^*(t_j))>0$. Thus, the right-hand side of (4.21) is finite at t_j, and it follows that $\sum_{i\notin I_0}D_\theta(g_i(\bar{x}),g_i(x^0))\leq\sum_{i\notin I_0}D_\theta(g_i(x^*),g_i(x^0))$. Hence, $\bar{x}=x^*$. ∎

Rate of convergence. We turn now to the case where there is no equality constraint so that the linear program is

$$\min_{x\in\mathbb{R}^n}\{\langle c,x\rangle \mid Bx\geq d\}. \tag{4.22}$$

We assume that (4.22) admits a unique solution a, and we study the rate of convergence when θ is a Bregman function with zone $\mathbb{R}_+$. To apply Proposition 4.9, we need the following result.

LEMMA 4.12. *Set* $C = \{x \in \mathbb{R}^n | Bx > d\}$. *If* (4.22) *admits a unique solution* $a \in \mathbb{R}^n$, *then there exists* $k_0 > 0$, *s.t. for all* $y \in \overline{C}$, $\langle c, y - a\rangle \geq k_0 \mathcal{N}(y - a)$, *where* $\mathcal{N}(x) = \sum_{i\in I} |\langle B_i, x\rangle|$ *is a norm on* $\mathbb{R}^n$.

Proof. Set $I_0 = \{i \in I \mid \langle B_i, a\rangle = d_i\}$. The optimality conditions for a imply the existence of a *multiplier* vector $\lambda \in \mathbb{R}^p_+$ such that $\lambda_i[d_i - \langle B_i, a\rangle] = 0$, for all $i \in I$, and $c = \sum_{i\in I} \lambda_i B_i$. Let $y \in \overline{C}$. We deduce that $\langle c, y - a\rangle = N(y - a)$, where $N(x) = \sum_{i\in I_0} \lambda_i |\langle B_i, x\rangle|$. By uniqueness of the optimal solution, it is easy to see that $\operatorname{span}\{B_i \mid i \in I_0\} = \mathbb{R}^n$; hence N is a norm on $\mathbb{R}^n$. Since $\mathcal{N}(x) = \sum_{i\in I} |\langle B_i, x\rangle|$ is also a norm on $\mathbb{R}^n$ (recall that B is a full rank matrix), we deduce that there exists k_0 such that $N(x) \geq k_0 \mathcal{N}(x)$. ∎

The following lemma is a sharper version of Proposition 4.9 in the linear context.

LEMMA 4.13. *Under the assumptions of Proposition* 4.11, *assume in addition that* θ *is a Bregman function with zone* $\mathbb{R}_-$ *and that there exist* $\alpha > 0$, $\beta \geq 1$, *and* $\varepsilon > 0$ *such that*

$$\forall s \in (0, \varepsilon), \qquad \alpha D_\theta(0, s)^\beta \leq s. \tag{4.23}$$

Then there exist positive constants K, L, M *such that for all* $t > 0$ *the trajectory of* (H-SD) *satisfies* $D_h(a, x(t)) \leq Ke^{-Lt}$ *if* $\beta = 1$, *and* $D_h(a, x(t)) \leq M/t^{\frac{1}{\beta-1}}$ *if* $\beta > 1$.

Proof. By Lemma 4.12, there exists k_0 such that for all $t > 0$,

$$\langle c, x(t) - a\rangle \geq \sum_{i\in I} k_0 |\langle B_i, x(t)\rangle - \langle B_i, a\rangle|. \tag{4.24}$$

Now, if we prove that there exists $\lambda > 0$ such that

$$|\langle B_i, x(t)\rangle - \langle B_i, a\rangle| \geq \lambda D_\theta(\langle B_i, a\rangle - d_i, \langle B_i, x(t)\rangle - d_i) \tag{4.25}$$

for all $i \in I$ and for t large enough, then from (4.24) it follows that $f(\cdot) = \langle c, \cdot\rangle$ satisfies the assumptions of Proposition 4.9, and the conclusion follows easily. Since $x(t) \to a$, to prove (4.25) it suffices to show that for all $r_0 \geq 0$ there exist $\eta, \mu > 0$ such that for all s, $|s - r_0| < \eta$, $\mu D_\theta(r_0, s)^\beta \leq |r_0 - s|$. The case where $r_0 = 0$ is a direct consequence of (4.23). Let $r_0 > 0$. An easy computation yields $\frac{d^2}{ds^2} D_\theta(r_0, s)_{|s=r_0} = \theta''(r_0)$, and by Taylor's expansion formula,

$$D_\theta(r_0, s) = \frac{\theta''(r_0)}{2}(s - r_0)^2 + o(s - r_0)^2 \tag{4.26}$$

with $\theta''(r_0) > 0$ due to (H_1)(iii). Let η be such that for all s, $|s - r_0| < \eta$, $s > 0$, $D_\theta(r_0, s) \leq \theta''(r_0)(s - r_0)^2$, and $D_\theta(r_0, s) \leq 1$; since $\beta \geq 1$, $D_\theta(r_0, s)^\beta \leq D_\theta(r_0, s) \leq \theta''(r_0)|s - r_0|$. ∎

To obtain Euclidean estimates, the functions $s \mapsto D_\theta(r_0, s)$, $r_0 \in \mathbb{R}_+$, have to be locally compared to $s \mapsto |r_0 - s|$. By (4.26) and the fact that $\theta'' > 0$, for each $r_0 > 0$ there exist $K, \eta > 0$ such that $|r_0 - s| \leq K\sqrt{D_\theta(r_0, s)}$, for all s, $|r_0 - s| < \eta$. This shows that, in practice, the Euclidean estimate depends only on a property of the type (4.23). Examples:

- The Boltzmann–Shannon entropies $\theta_3(s) = s\ln(s) - s$ and $\theta_6(s) = s\ln s$ satisfy $D_{\theta_i}(0, s) = s$, $s > 0$; hence for some $K, L > 0$, $|x(t) - a| \leq Ke^{-Lt}$, for all $t \geq 0$.
- With either $\theta_4(s) = -s^\gamma/\gamma$ or $\theta_5(s) = (\gamma s - s^\gamma)/(1 - \gamma)$, $\gamma \in (0, 1)$, we have $D_{\theta_i}(0, s) = (1 + 1/\gamma)s^\gamma$, $s > 0$; hence $|x(t) - a| \leq K/t^{\frac{\gamma}{2-2\gamma}}$, for all $t > 0$.

4.6. Dual convergence. In this section we focus on the case $C = \mathbb{R}^n_{++}$, so that the minimization problem is

$$\min\{f(x) \mid x \geq 0,\ Ax = b\}. \tag{P}$$

We assume

$$f \text{ is convex and } S(P) \neq \emptyset, \tag{4.27}$$

together with the Slater condition

$$\exists x^0 \in \mathbb{R}^n,\ x^0 > 0,\ Ax^0 = b. \tag{4.28}$$

In convex optimization theory, it is usual to associate with (P) the *dual* problem given by

$$\min\{p(\lambda) \mid \lambda \geq 0\}, \tag{D}$$

where $p(\lambda) = \sup\{\langle \lambda, x\rangle - f(x) \mid Ax = b\}$. For many applications, dual solutions are as important as primal ones. In the particular case of a linear program where $f(x) = \langle c, x\rangle$ for some $c \in \mathbb{R}^n$, writing $\lambda = c + A^T y$ with $y \in \mathbb{R}^m$, the linear dual problem may equivalently be expressed as $\min\{\langle b, y\rangle \mid A^T y + c \geq 0\}$. Thus, λ is interpreted as a vector of *slack* variables for the dual inequality constraints. In the general case, $S(\mathrm{D})$ is nonempty and bounded under (4.27) and (4.28), and moreover $S(\mathrm{D}) = \{\lambda \in \mathbb{R}^n \mid \lambda \geq 0,\ \lambda \in \nabla f(x^*) + \operatorname{Im} A^T,\ \langle \lambda, x^*\rangle = 0\}$, where x^* is any solution of (P); see, for instance, [26, Theorems VII.2.3.2 and VII.4.5.1].

Let us introduce a Legendre kernel θ satisfying (H_1) and define

$$h(x) = \sum_{i=1}^{n} \theta(x_i). \tag{4.29}$$

Suppose that (H-SD) is well posed. Integrating the differential inclusion (3.7), we obtain

$$\lambda(t) \in c(t) + \operatorname{Im} A^T, \tag{4.30}$$

where $c(t) = \frac{1}{t}\int_0^t \nabla f(x(\tau))d\tau$ and $\lambda(t)$ is the *dual trajectory* defined by

$$\lambda(t) = \frac{1}{t}[\nabla h(x^0) - \nabla h(x(t))]. \tag{4.31}$$

Assume that $x(t)$ is bounded. From (4.27), it follows that ∇f is constant on $S(\mathrm{P})$, and then it is easy to see that $\nabla f(x(t)) \to \nabla f(x^*)$ as $t \to +\infty$ for any $x^* \in S(\mathrm{P})$. Consequently, $c(t) \to \nabla f(x^*)$. By (4.31) together with [39, Theorem 26.5], we have $x(t) = \nabla h^*(\nabla h(x^0) - t\lambda(t))$, where the Fenchel conjugate h^* is given by $h^*(\lambda) = \sum_{i=1}^n \theta^*(\lambda_i)$. Take any solution $\widetilde{x}$ of $A\widetilde{x} = b$. Since $Ax(t) = b$, we have $\widetilde{x} - \nabla h^*(\nabla h(x^0) - t\lambda(t)) \in \operatorname{Ker} A$. On account of (4.30), $\lambda(t)$ is the unique optimal solution of

$$\lambda(t) \in \operatorname{Argmin}\left\{\langle \widetilde{x}, \lambda\rangle + \frac{1}{t}\sum_{i=1}^{n} \theta^*(\theta'(x_i^0) - t\lambda_i) \mid \lambda \in c(t) + \operatorname{Im} A^T\right\}. \tag{4.32}$$

By (H_1)(iii), θ' is increasing in $\mathbb{R}_{++}$. Set $\eta = \lim_{s\to+\infty} \theta'(s) \in (-\infty, +\infty]$. Since θ^* is a Legendre-type function, $\operatorname{int dom}\, \theta^* = \operatorname{dom} \partial\theta^* = \operatorname{Im} \partial\theta = (-\infty, \eta)$. From

$(\theta^*)' = (\theta')^{-1}$, it follows that $\lim_{u\to-\infty}(\theta^*)'(u) = 0$ and $\lim_{u\to\eta^-}(\theta^*)'(u) = +\infty$. Consequently, (4.32) can be interpreted as a *penalty approximation scheme* of the dual problem (D), where the dual positivity constraints are penalized by a separable strictly convex function. Similar schemes have been treated in [4, 16, 17, 28]. Consider the additional condition

(4.33) Either $\theta(0) < \infty$, or $S(P)$ is bounded, or f is linear.

As a direct consequence of [28, Propositions 10 and 11], we obtain that under (4.27), (4.28), (4.33) and (H_1), $\{\lambda(t) \mid t \to +\infty\}$ is bounded and its cluster points belong to $S(D)$. The convergence of $\lambda(t)$ is more difficult to establish. In fact, under some additional conditions on θ^* (see [16, Conditions (H_0)–(H_1)] or [28, Conditions (A7) and (A8)]) it is possible to show that $\lambda(t)$ converges to a particular element of the dual optimal set (the "θ^*-center" in the sense of [16, Definition 5.1] or the $D_h(\cdot, x^0)$-center as defined in [28, p. 616]), which is characterized as the unique solution of a *nested hierarchy* of optimization problems on the dual optimal set. We will not develop this point here. Let us only mention that for all the examples of section 4.4, θ_i^* satisfies such additional conditions and consequently we have the following result.

PROPOSITION 4.14. *Under* (4.27), (4.28), *and* (4.33), *for each of the explicit Legendre kernels given in section* 4.4, $\lambda(t)$ *given by* (4.31) *converges to a particular dual solution.*

5. Legendre transform coordinates.

5.1. Legendre functions on affine subspaces. The first objective of this section is to slightly generalize the notion of a Legendre-type function to the case of functions whose domains are contained in an affine subspace of $\mathbb{R}^n$. We begin by noticing that the Legendre-type property does not depend on canonical coordinates.

LEMMA 5.1. *Let $g \in \Gamma_0(\mathbb{R}^r)$, $r \geq 1$, and $T : \mathbb{R}^r \to \mathbb{R}^r$ an affine invertible mapping. Then g is of a Legendre type iff $g \circ T$ is of Legendre type.*

Proof. The proof is elementary and is left to the reader. □

From now on, $\mathcal{A}$ is the affine subspace defined by (2.1), whose dimension is $r = n - m$.

DEFINITION 5.2. *A function $g \in \Gamma_0(\mathcal{A})$ is said to be of* Legendre type *if there exists an affine invertible mapping $T : \mathcal{A} \to \mathbb{R}^r$ such that $g \circ T^{-1}$ is a Legendre-type function in $\Gamma_0(\mathbb{R}^r)$.*

By Lemma 5.1, the previous definition is consistent.

PROPOSITION 5.3. *Let $h \in \Gamma_0(\mathbb{R}^n)$ be a function of Legendre type with $C = \operatorname{int}\operatorname{dom} h$. If $\mathcal{F} = C \cap \mathcal{A} \neq \emptyset$, then the restriction $h_{|\mathcal{A}}$ of h to $\mathcal{A}$ is of Legendre type, and moreover $\operatorname{int}_{\mathcal{A}}\operatorname{dom} h_{|\mathcal{A}} = \mathcal{F}$ (where $\operatorname{int}_{\mathcal{A}} B$ stands for the interior of B in $\mathcal{A}$ as a topological subspace of $\mathbb{R}^n$).*

Proof. From the inclusions $\mathcal{F} \subset \operatorname{dom} h_{|\mathcal{A}} \subset \overline{\mathcal{F}} = \overline{C} \cap \mathcal{A}$ and since $\operatorname{ri}\overline{\mathcal{F}} = \mathcal{F}$, we conclude that $\operatorname{int}_{\mathcal{A}}\operatorname{dom} h_{|\mathcal{A}} = \mathcal{F} \neq \emptyset$. Let $T : \mathbb{R}^r \to \mathcal{A}$ be an invertible transformation with $Tz = Lz + x^0$ for all $z \in \mathbb{R}^r$, where $x^0 \in \mathcal{A}$ and $L : \mathbb{R}^r \to \mathcal{A}_0$ is a nonsingular linear mapping. Define $k = h_{|\mathcal{A}} \circ T$. Clearly, $k \in \Gamma_0(\mathbb{R}^r)$. Let us prove that k is essentially smooth. We have $\operatorname{dom} k = T^{-1}\operatorname{dom} h_{|\mathcal{A}}$ and therefore $\operatorname{int}\operatorname{dom} k = T^{-1}\mathcal{F}$. Since h is differentiable on C, we conclude that k is differentiable on $\operatorname{int}\operatorname{dom} k$. Now, let $(z^j) \in \operatorname{int}\operatorname{dom} k$ be a sequence that converges to a boundary point $z \in \operatorname{bd}\operatorname{dom} k$. Then, $Tz^j \in \operatorname{int}_{\mathcal{A}}\operatorname{dom} h_{|\mathcal{A}}$ and $Tz^j \to Tz \in \operatorname{bd}_{\mathcal{A}}\operatorname{dom} h_{|\mathcal{A}} \subset \operatorname{bd}\operatorname{dom} h$. Since h is essentially smooth, $|\nabla h(Tz^j)| \to +\infty$. Thus, to prove that $|\nabla k(z^j)| \to +\infty$, it suffices to show that there exists $\lambda > 0$ such that $|\nabla k(z^j)| \geq \lambda |\nabla h(Tz^j)|$ for all j

large enough. Note that $\nabla k(z^j) = \nabla[h_{|\mathcal{A}} \circ T](z^j) = L^* \nabla h_{|\mathcal{A}}(Tz^j) = L^* \Pi_{\mathcal{A}_0} \nabla h(Tz^j)$, where $L^* : \mathcal{A}_0 \to \mathbb{R}^r$ is defined by $\langle z, L^* x \rangle = \langle Lz, x \rangle$ for all $(z, x) \in \mathbb{R}^r \times \mathcal{A}_0$. Of course, L^* is linear with $\operatorname{Ker} L^* = \{0\}$. Therefore $\frac{\nabla k(z^j)}{|\nabla h(Tz^j)|} = L^* \Pi_{\mathcal{A}_0} \frac{\nabla h(Tz^j)}{|\nabla h(Tz^j)|}$. Let ω denote the nonempty and compact set of cluster points of the normalized sequence $\nabla h(Tz^j)/|\nabla h(Tz^j)|$, $j \in \mathbb{N}$. By Lemma 4.2, we have that $\omega \subset \{\nu \in N_{\overline{C}}(Tz) \,|\, |\nu| = 1\}$, and consequently Lemma 4.3 yields $\Pi_{\mathcal{A}_0} \omega \cap \{0\} = \emptyset$. By the compactness of ω, we obtain $\liminf_{j \to +\infty} |\Pi_{\mathcal{A}_0} \nabla h(Tz^j)|/|\nabla h(Tz^j)| > 0$, which proves our claim. Finally, the strict convexity of k on $\operatorname{dom} \partial k = \operatorname{int} \operatorname{dom} k = T^{-1}\mathcal{F}$ is a direct consequence of the strict convexity of h in $\mathcal{F}$. □

5.2. Legendre transform coordinates. The prominent fact of Legendre functions theory is that $h \in \Gamma_0(\mathbb{R}^n)$ is of Legendre type iff its Fenchel conjugate h^* is of Legendre type [39, Theorem 26.5], and $\nabla h : \operatorname{int} \operatorname{dom} h \to \operatorname{int} \operatorname{dom} h^*$ is onto with $(\nabla h)^{-1} = \nabla h^*$. In the case of Legendre functions on affine subspaces, we have the following generalization.

PROPOSITION 5.4. *If $g \in \Gamma_0(\mathcal{A})$ is of Legendre type in the sense of Definition 5.2, then $\nabla g(\operatorname{int}_{\mathcal{A}} \operatorname{dom} g)$ is a nonempty, open, and convex subset of $\mathcal{A}_0$. In addition, ∇g is a one-to-one continuous mapping from $\operatorname{int}_{\mathcal{A}} \operatorname{dom} g$ onto its image.*

Proof. Let $Tx = Lx + z_0$, with $L : \mathcal{A}_0 \to \mathbb{R}^r$ being a linear invertible mapping and $z_0 \in \mathbb{R}^p$. Set $k = g \circ T^{-1} \in \Gamma_0(\mathbb{R}^r)$, which is of Legendre type. We have $\operatorname{dom} k = T \operatorname{dom} g$. Define $L^* : \mathbb{R}^r \to \mathcal{A}_0$ by $\langle L^* z, x \rangle = \langle z, Lx \rangle$ for all $(z, x) \in \mathbb{R}^r \times \mathcal{A}_0$. We have that $\nabla g(x) = \nabla[k \circ T](x) = L^* \nabla k(Tx)$ for all $x \in \operatorname{int}_{\mathcal{A}} \operatorname{dom} g$. Therefore $\nabla g(\operatorname{int}_{\mathcal{A}} \operatorname{dom} g) = L^* \nabla k(T \operatorname{int}_{\mathcal{A}} \operatorname{dom} g) = L^* \nabla k(\operatorname{int}_{\mathbb{R}^r} \operatorname{dom} k) = L^* \operatorname{int}_{\mathbb{R}^r} \operatorname{dom} k^*$. Since $\operatorname{int}_{\mathbb{R}^r} \operatorname{dom} k^*$ is a nonempty, open, and convex subset of $\mathbb{R}^r$ and L^* is an invertible linear mapping, then $L^* \operatorname{int}_{\mathbb{R}^r} \operatorname{dom} k^*$ is an open and nonempty subset of $\mathcal{A}_0$. Moreover, by [39, Theorem 6.6], we have $L^* \operatorname{int}_{\mathbb{R}^r} \operatorname{dom} k^* = \operatorname{ri} L^* \operatorname{dom} k^*$. Consequently, $\nabla g(\operatorname{int}_{\mathcal{A}} \operatorname{dom} g) = \operatorname{ri} L^* \operatorname{dom} k^* = \operatorname{int}_{\mathcal{A}_0} L^* \operatorname{dom} k^* \neq \emptyset$. Finally, since $\nabla k : \operatorname{int}_{\mathbb{R}^r} \operatorname{dom} k \to \operatorname{int}_{\mathbb{R}^r} \operatorname{dom} k^*$ is one-to-one and continuous, the same result holds for $\nabla g = L^* \circ \nabla k \circ T$ on $\operatorname{int}_{\mathcal{A}} \operatorname{dom} g$. □

In what follows, we assume that h satisfies the basic condition (H_0) and $\mathcal{F} = C \cap \mathcal{A} \neq \emptyset$. The *Legendre transform coordinates mapping* on $\mathcal{F}$ associated with h is defined by

$$\begin{array}{rccl} \phi_h : & \mathcal{F} & \to & \mathcal{F}^* = \phi_h(\mathcal{F}), \\ & x & \mapsto & \phi_h(x) = \nabla(h_{|\mathcal{A}}) = \Pi_{\mathcal{A}_0} \nabla h(x). \end{array} \tag{5.1}$$

This definition retrieves the Legendre transform coordinates introduced by Bayer and Lagarias in [6] for the particular case of the log-barrier on a polyhedral set.

THEOREM 5.5. *Under the above definitions and assumptions, $\mathcal{F}^*$ is a convex, (relatively) open, and nonempty subset of $\mathcal{A}_0$; ϕ_h is a $\mathcal{C}^1$ diffeomorphism from $\mathcal{F}$ to $\mathcal{F}^*$; and for all $x \in \mathcal{F}$, $d\phi_h(x) = \Pi_{\mathcal{A}_0} H(x)$ and $d\phi_h(x)^{-1} = \sqrt{H(x)^{-1}} \Pi_{\sqrt{H(x)}\mathcal{A}_0} \sqrt{H(x)^{-1}}$, where $H(x) = \nabla^2 h(x)$.*

Proof. By Propositions 5.3 and 5.4, $\mathcal{F}^*$ is a convex, open, and nonempty subset of $\mathcal{A}_0$ and ϕ_h is a continuous bijection. By (H_0)(ii), ϕ_h is of class $\mathcal{C}^1$ on $\mathcal{F}$, and we have, for all $x \in \mathcal{F}$, $d\phi_h(x) = \Pi_{\mathcal{A}_0} \nabla^2 h(x) = \Pi_{\mathcal{A}_0} H(x)$. Let $v \in \mathcal{A}_0$ be such that $d\phi_h(x)v = 0$. It follows that $H(x)v \in \mathcal{A}_0^\perp$ and, in particular, $\langle H(x)v, v \rangle = 0$. Hence, $v = 0$, thanks to (H_0)(iii). The implicit function theorem implies then that ϕ_h is a $\mathcal{C}^1$ diffeomorphism. Finally, the formula concerning $d\phi_h(x)^{-1}$ is a direct consequence of the next lemma, which is analogous to [7, p. 545], and whose proof is omitted.

LEMMA 5.6. *Define the linear operators $L_i : \mathbb{R}^n \to \mathbb{R}^n$ by $L_1 = \Pi_{\mathcal{A}_0} H(x)$ and*

$L_2 = \sqrt{H(x)^{-1}}\Pi_{\sqrt{H(x)}\mathcal{A}_0}\sqrt{H(x)^{-1}}$. *Then* $L_2L_1v = v$ *for all* $v \in \mathcal{A}_0$. □

Similarly to the classical Legendre-type functions theory, the inverse of ϕ_h can be expressed in terms of Fenchel conjugates. For that purpose, we notice that inverting ϕ_h is a minimization problem. Indeed, given $y \in \mathcal{A}_0$, the problem of finding $x \in \mathcal{F}$ such that $y = \Pi_{\mathcal{A}_0}\nabla h(x)$ is equivalent to $x = \operatorname{Argmin}\{h(z) - \langle y, z\rangle | z \in \mathcal{A}\}$, or equivalently

$$x = \operatorname{Argmin}\{(h + \delta_{\mathcal{A}})(z) - \langle y, z\rangle\}, \tag{5.2}$$

where $\delta_{\mathcal{A}}$ is the *indicator* of $\mathcal{A}$, i.e., $\delta_{\mathcal{A}}(z) = 0$ if $z \in \mathcal{A}$ and $+\infty$ otherwise. Let us recall the definition of *epigraphical sum* of two functions $g_1, g_2 \in \Gamma_0(\mathbb{R}^n)$, which is given by $(g_1 \square g_2)(y) = \inf\{g_1(u) + g_2(v) | u + v = y\}$ for all $y \in \mathbb{R}^n$. We have $g_1 \square g_2 \in \Gamma_0(\mathbb{R}^n)$, and if g_1 and g_2 satisfy $\operatorname{ri}\operatorname{dom} g_1 \cap \operatorname{ri}\operatorname{dom} g_2 \neq \emptyset$, then $(g_1 + g_2)^* = g_1^* \square g_2^*$ (see [39]).

PROPOSITION 5.7. *We have that* $\phi_h^{-1} : \mathcal{F}^* \to \mathcal{F}$ *is given by* $\phi_h^{-1}(y) = \nabla[h^* \square (\delta_{\mathcal{A}_0^\perp} + \langle \cdot, \widetilde{x}\rangle)](y)$ *for any* $\widetilde{x} \in \mathcal{A}$, *and moreover* $\mathcal{F}^* = \Pi_{\mathcal{A}_0}\operatorname{int}\operatorname{dom} h^*$.

Proof. The optimality condition for (5.2) yields $y \in \partial(h + \delta_{\mathcal{A}})(x)$. Thus, $x \in \partial(h + \delta_{\mathcal{A}})^*(y)$. From $\mathcal{F} \neq \emptyset$, we conclude that the function $g \in \Gamma_0(\mathbb{R}^n)$ defined by $g = (h + \delta_{\mathcal{A}})^*$ satisfies $g = h^* \square \delta_{\mathcal{A}}^* = h^* \square (\delta_{\mathcal{A}_0^\perp} + \langle \cdot, \widetilde{x}\rangle)$ with $\widetilde{x} \in \mathcal{A}$. Moreover, by [39, Corollary 26.3.2], g is essentially smooth and we deduce that indeed $x = \nabla g(y)$. Since g is essentially smooth, $\operatorname{dom}\partial g = \operatorname{int}\operatorname{dom} g$. By the definition of an epigraphical sum, $g(y) = \inf\{h^*(u) + \delta_{\mathcal{A}_0^\perp}(v) + \langle v, \widetilde{x}\rangle | u + v = y\}$, and consequently we have that $y \in \operatorname{dom} g$ iff $y \in \operatorname{dom} h^* + \mathcal{A}_0^\perp$. Hence, $\operatorname{int}\operatorname{dom} g = \operatorname{int}\operatorname{dom} h^* + \mathcal{A}_0^\perp$ (see, for instance, [39, Corollary 6.6.2]). Recalling that $\mathcal{F}^*$ is a relatively open subset of $\mathcal{A}_0$, we deduce that $\mathcal{F}^* = \Pi_{\mathcal{A}_0}\operatorname{dom}\partial g = \Pi_{\mathcal{A}_0}\operatorname{int}\operatorname{dom} h^*$. □

5.3. Linear problems in Legendre transform coordinates.

5.3.1. Polyhedral sets in Legendre transform coordinates. One of the first applications of Legendre transform coordinates is to transform linear constraints into positive cones.

PROPOSITION 5.8. *Assume that* $C = \{x \in \mathbb{R}^n | Bx > d\}$, *where* B *is a* $p \times n$ *full rank matrix, with* $p \geq n$. *Suppose also that* h *is of the form* (4.19) *with* θ *satisfying* (H_1), *and let* $\eta = \lim_{s \to +\infty}\theta'(s) \in (-\infty, +\infty]$. *If* $\eta < +\infty$, *then* $\overline{\operatorname{dom} h^*} = \{y \in \mathbb{R}^n \mid y + B^T\lambda = 0,\ \lambda_i \geq -\eta\}$, *and* $\operatorname{dom} h^* = \mathbb{R}^n$ *when* $\eta = +\infty$.

Proof. By [40, Theorem 11.5], $\overline{\operatorname{dom} h^*} = \{y \in \mathbb{R}^n \mid \langle y, d\rangle \leq h^\infty(d)\, \forall d \in \mathbb{R}^n\}$, where h^∞ is the *recession* function, also known as *horizon* function, of h. The recession function is defined by $h^\infty(d) = \lim_{t \to +\infty}\frac{1}{t}[h(\bar{x} + td) - h(\bar{x})]$, $d \in \mathbb{R}^n$, where $\bar{x} \in \operatorname{dom} h$; this limit does not depend of $\bar{x}$ and eventually $h^\infty(d) = +\infty$ (see also [39]). In this case, it is easy to verify that $h^\infty(d) = \sum_{i=1}^p \theta^\infty(\langle B_i, d\rangle)$. Clearly, $\theta^\infty(-1) = +\infty$ and $\theta^\infty(1) = \lim_{s \to +\infty}\theta'(s) = \eta$. In particular, if $\eta = +\infty$, then $\operatorname{dom} h^* = \mathbb{R}^n$. If $\eta < +\infty$, then $y \in \overline{\operatorname{dom} h^*}$ iff for all $d \in \mathbb{R}^n$ such that $Bd \geq 0$, $\langle y, d\rangle \leq h^\infty(d) = \sum_{i=1}^p \eta\langle B_i, d\rangle$, that is $\langle y - \eta B^T e, d\rangle \leq 0$ with $e = (1, \ldots, 1)$. Thus, by the Farkas lemma, $y \in \overline{\operatorname{dom} h^*}$ iff there exists $\mu \geq 0$, $y - \eta B^T e + B^T\mu = 0$. □

As a direct consequence of Propositions 5.7 and 5.8, we have the following.

COROLLARY 5.9. *Under the assumptions of Proposition* 5.8, *if* $\eta = 0$, *then* $\mathcal{F}^*$ *is a positive convex cone, and if* $\eta = +\infty$, *then* $\mathcal{F}^* = \mathcal{A}_0$.

5.3.2. (H-SD)-trajectories as geodesic curves. In what follows, we assume that $f(x) = \langle c, x\rangle$ for some $c \in \mathbb{R}^n$. As another striking application of Legendre transform coordinates, we prove that the trajectories of (H-SD) may be seen as straight lines in $\mathcal{F}^* = \phi_h(\mathcal{F})$ and also as geodesic curves in $\mathcal{F}$ with respect to some appropriate metric, extending to the general case a result of [7] for the log-metric.

PROPOSITION 5.10. *For every $y \in \mathcal{F}^*$ we have $[(\phi_h)_* \nabla_{H} f_{|\mathcal{F}}]\,(y) = \Pi_{\mathcal{A}_0} c$, where $(\phi_h)_* \nabla_{H} f_{|\mathcal{F}}$ is the* push forward *vector field of $\nabla_{H} f_{|\mathcal{F}}$ by ϕ_h.*

Proof. Let $y \in \mathcal{F}^*$. By definition, $[(\phi_h)_* \nabla_{H} f_{|\mathcal{F}}](y) = d\phi_h(\phi_h^{-1}(y)) \nabla_{H} f_{|\mathcal{F}}(\phi_h^{-1}(y))$. Setting $x = \phi_h^{-1}(y)$, by Theorem 5.5 we get $[(\phi_h)_* \nabla_{H} f_{|\mathcal{F}}]\,(y) = d\phi_h(x) \nabla_{H} f_{|\mathcal{F}}(x) = \Pi_{\mathcal{A}_0} H(x) H(x)^{-1} [I - A^T (AH(x)^{-1} A^T)^{-1} AH(x)^{-1}] c = \Pi_{\mathcal{A}_0} c - \Pi_{\mathcal{A}_0} A^T z$, where $z = [(AH(x)^{-1} A^T)^{-1} AH(x)^{-1}] c$. Since $\operatorname{Im} A^T = \mathcal{A}_0^{\perp}$, the conclusion follows. ∎

It follows directly from Proposition 5.10 that $\Phi_h(x(t)) = \Phi_h(x^0) + t \Pi_{\mathcal{A}_0} c$ with $x(t)$ being a solution to (H-SD). Endow $\mathcal{F}^*$ with the Euclidean metric, which allows us to define on $\mathcal{F}$ the metric

$$(\cdot, \cdot)^{H^2} = (\phi_h)^* \langle \cdot, \cdot \rangle, \tag{5.3}$$

that is, $(u, v)_x^{H^2} = \langle d\phi_h(x) u, d\phi_h(x) v \rangle = \langle \Pi_{\mathcal{A}_0} H(x) u, \Pi_{\mathcal{A}_0} H(x) v \rangle$ for all $(x, u, v) \in \mathcal{F} \times \mathbb{R}^n \times \mathbb{R}^n$. For each initial condition $x^0 \in \mathcal{F}$, and for every $c \in \mathbb{R}^n$, we set

$$v = d\phi_h(x^0)^{-1} \Pi_{\mathcal{A}_0} c = \sqrt{H(x^0)^{-1}} \Pi_{\sqrt{H(x^0)}\mathcal{A}_0} \sqrt{H(x^0)^{-1}} \Pi_{\mathcal{A}_0} c. \tag{5.4}$$

THEOREM 5.11. *Let $(x^0, c) \in \mathcal{F} \times \mathbb{R}^n$, set $f(x) = \langle c, x \rangle$, for all $x \in C$, and define v as in* (5.4)*. If $\mathcal{F}$ is endowed with the metric $(\cdot, \cdot)^{H^2}$ given by* (5.3)*, then the solution $x(t)$ of* (H-SD) *is the unique geodesic passing through x^0 with velocity v.*

Proof. Since $\mathcal{F}$, $(\cdot, \cdot)^{H^2}$ is isometric to the Euclidean manifold $\mathcal{F}^*$, $\langle \cdot, \cdot \rangle$, the geodesic joining two points of $\mathcal{F}$ exists and is unique. Let us denote by $\gamma : J \subset \mathbb{R} \mapsto \mathcal{F}$ the geodesic passing through x^0 with velocity v. By definition of $(\cdot, \cdot)^{H^2}$, $\phi_h(\gamma)$ is a geodesic in $\mathcal{F}^*$, whence $\phi_h(\gamma(t)) = \phi_h(x^0) + t d\phi_h(x^0) v$, $t \in J$. In view of (5.4), this can be rewritten as $\phi_h(\gamma(t)) = \phi_h(x^0) + t \Pi_{\mathcal{A}_0} c$. By Proposition 5.10, $\gamma = \phi_h^{-1}(\phi_h(\gamma))$ solves (H-SD). ∎

5.3.3. Lagrange equations. Following the ideas of [7], we describe the orbits of (H-SD) as orthogonal projections on $\mathcal{A}$ of $\dot{q}$-trajectories of a specific *Lagrangian system*. Recall that, given a real-valued mapping $\mathcal{L}(q, \dot{q})$, called the Lagrangian, where $q = (q_1, \ldots, q_n)$ and $\dot{q} = (\dot{q}_1, \ldots, \dot{q}_n)$, the associated Lagrange equations of motion are the following:

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_i} = \frac{\partial \mathcal{L}}{\partial q_i}, \quad \frac{d}{dt} q_i = \dot{q}_i, \quad \forall i = 1, \ldots, n. \tag{5.5}$$

Their solutions are C^1-piecewise paths $\gamma : t \longmapsto (q(t), \dot{q}(t))$, defined for $t \in J \subset \mathbb{R}$, that satisfy (5.5) and appear as extremals of the functional $\widehat{\mathcal{L}}(\gamma) = \int_J \mathcal{L}(q(t), \dot{q}(t)) dt$. Notice that, in general, the solutions are not unique, in the sense that they do not only depend on the initial condition $\gamma(0)$. Let us introduce the Lagrangian $\mathcal{L} : \mathbb{R}^n \times C \to \mathbb{R}$ defined by

$$\mathcal{L}(q, \dot{q}) = \langle \Pi_{\mathcal{A}_0} c, q \rangle - h(\Pi_{\mathcal{A}} \dot{q}), \tag{5.6}$$

where $\Pi_{\mathcal{A}}$ is the orthogonal projection onto $\mathcal{A}$, i.e., $\Pi_{\mathcal{A}} x = \widetilde{x} + \Pi_{\mathcal{A}_0}(x - \widetilde{x})$ for any $\widetilde{x} \in \mathcal{A}$.

THEOREM 5.12. *For any solution $\gamma(t) = (q(t), \dot{q}(t))$ of the Lagrangian dynamical system* (5.5) *with Lagrangian given by* (5.6)*, the projection $x(t) = \Pi_{\mathcal{A}} \dot{q}(t)$ is the solution of* (H-SD) *with initial condition $x^0 = \Pi_{\mathcal{A}} \dot{q}(0)$.*

Proof. It is easy to verify that $\nabla(h \circ \Pi_{\mathcal{A}})(x) = \Pi_{\mathcal{A}_0} \nabla h(\Pi_{\mathcal{A}} x)$ for any $x \in \mathbb{R}^n$. Given a solution $\gamma(t) = (q(t), \dot{q}(t))$ of (5.6) defined on J, we set $p(t) = (p_1(t), \ldots, p_n(t)) =$

$\big(\frac{\partial \mathcal{L}}{\partial \dot{q}_1}(\gamma(t)), \dots, \frac{\partial \mathcal{L}}{\partial \dot{q}_n}(\gamma(t))\big)$. We have $p(t) = \nabla(h \circ \Pi_{\mathcal{A}})(\dot{q}(t)) = \Pi_{\mathcal{A}_0} \nabla h(\Pi_{\mathcal{A}} \dot{q}(t)) = \phi_h(\Pi_{\mathcal{A}} \dot{q}(t))$. Equations of motion become $\frac{d}{dt} p(t) = \Pi_{\mathcal{A}_0} c$, that is, $\frac{d}{dt} \phi_h(\Pi_{\mathcal{A}} \dot{q}(t)) = \Pi_{\mathcal{A}_0} c$. Since $\phi_h : \mathcal{F} \to \mathcal{F}^*$ is a diffeomorphism, the latter means, according to Proposition 5.10, that $\Pi_{\mathcal{A}} \dot{q}(t)$ is a trajectory for the vector field $\nabla_H f_{|_{\mathcal{F}}}$. Notice that, C being convex, as soon as $\dot{q}(0) \in C$, $\Pi_{\mathcal{A}} \dot{q}(0) \in C \cap \mathcal{A} = \mathcal{F}$, and what precedes forces $\Pi_{\mathcal{A}} \dot{q}(t)$ to stay in $\mathcal{F}$ for any $t \in J$. □

5.3.4. Completely integrable Hamiltonian systems. In the following, all mappings are supposed to be at least of class $\mathcal{C}^2$. Let us first recall the notion of a Hamiltonian system. Given an integer $r \geq 1$ and a real-valued mapping $\mathcal{H}(q,p)$ on $\mathbb{R}^{2r}$ with coordinates $(q,p) = (q_1, \dots, q_r, p_1, \dots, p_r)$, the *Hamiltonian vector field* $X_{\mathcal{H}}$ *associated with* $\mathcal{H}$ is defined by $X_{\mathcal{H}} = \sum_{i=1}^{r} \frac{\partial \mathcal{H}}{\partial p_i} \frac{\partial}{\partial q_i} - \frac{\partial \mathcal{H}}{\partial q_i} \frac{\partial}{\partial p_i}$. The trajectories of the dynamical system induced by $X_{\mathcal{H}}$ are the solutions to

$$\left\{ \begin{array}{ll} \dot{p}_i(t) = -\frac{\partial}{\partial q_i} \mathcal{H}(q(t), p(t)), & i = 1, \dots, r, \\ \dot{q}_i(t) = \frac{\partial}{\partial p_i} \mathcal{H}(q(t), p(t)), & i = 1, \dots, r. \end{array} \right. \tag{5.7}$$

Following a standard procedure, Lagrangian functions $\mathcal{L}(q, \dot{q})$ are associated with Hamiltonian systems by means of the so-called Legendre transform

$$\Phi : \left\{ \begin{array}{ccc} \mathbb{R}^{2r} & \longrightarrow & \mathbb{R}^{2r}, \\ (q, \dot{q}) & \longmapsto & (q, \frac{\partial \mathcal{L}}{\partial \dot{q}}(q, \dot{q})). \end{array} \right.$$

In fact, when Φ is a diffeomorphism, the Hamiltonian function $\mathcal{H}$ associated with the Lagrangian $\mathcal{L}$ is defined on $\Phi(\mathbb{R}^{2r})$ by $\mathcal{H}(p,q) = \sum_{i=1}^{r} p_i \dot{q}_i - \mathcal{L}(q, \dot{q}) = \langle p, \psi^{-1}(q,p) \rangle - \mathcal{L}(q, \psi^{-1}(q,p))$, where $(q, \psi^{-1}(q,p)) := \Phi^{-1}(q,p)$. With these definitions, Φ sends the trajectories of the corresponding Lagrangian system on the trajectories of the Hamiltonian system (5.7).

In general, the Lagrangian (5.6) does not lead to an invertible Φ on $\mathbb{R}^{2n}$. However, we are interested only in the projections $\Pi_{\mathcal{A}} \dot{q}$ of the trajectories, which, according to Theorem 5.12, take their values in $\mathcal{F}$. Moreover, notice that for any differentiable path $t \mapsto q^{\perp}(t)$ lying in $\mathcal{A}_0^{\perp}$, $t \mapsto (q(t), \dot{q}(t))$ is a solution of (5.5) iff $t \mapsto (q(t) + q^{\perp}(t), \dot{q}(t) + \dot{q}^{\perp}(t))$ is. This legitimates the idea of restricting $\mathcal{L}$ to $\mathcal{A}_0 \times \Pi_{\mathcal{A}_0} \mathcal{F}$. Hence and from now on, $\mathcal{L}$ denotes the function

$$\mathcal{L} : \left\{ \begin{array}{ccc} \mathcal{A}_0 \times \Pi_{\mathcal{A}_0} \mathcal{F} & \longrightarrow & \mathbb{R}, \\ (q, \dot{q}) & \longmapsto & \mathcal{L}(q, \dot{q}). \end{array} \right. \tag{5.8}$$

Taking $(q_1, \dots, q_r)$, with $r = n - m$, a linear system of coordinates induced by an Euclidean orthonormal basis for $\mathcal{A}_0$, we easily see that this "new" Lagrangian has trajectories $(q(t), \dot{q}(t))$ lying in $\mathcal{A}_0 \times \Pi_{\mathcal{A}_0} \mathcal{F}$, whose projections $\Pi_{\mathcal{A}} \dot{q}(t)$ are exactly the (H-SD) trajectories. Moreover, an easy computation yields $\frac{\partial \mathcal{L}}{\partial \dot{q}}(q, \dot{q}) = \Pi_{\mathcal{A}_0} \nabla h(\Pi_{\mathcal{A}_0} \dot{q}) = [\phi_h \circ \Pi_{\mathcal{A}}](\dot{q})$, which is a diffeomorphism by Proposition 5.5. The Legendre transform is then given by

$$\Phi : \left\{ \begin{array}{ccc} \mathcal{A}_0 \times \Pi_{\mathcal{A}_0} \mathcal{F} & \longrightarrow & \mathcal{A}_0 \times \mathcal{F}^*, \\ (q, \dot{q}) & \longmapsto & (q, [\phi_h \circ \Pi_{\mathcal{A}}](\dot{q})), \end{array} \right.$$

and therefore, $\mathcal{L}$ is converted into the Hamiltonian system associated with

$$\mathcal{H} : \left\{ \begin{array}{ccc} \mathcal{A}_0 \times \mathcal{F}^* & \longrightarrow & \mathbb{R}, \\ (q,p) & \longmapsto & \langle p, [\phi_h \circ \Pi_{\mathcal{A}}]^{-1}(p) \rangle - \mathcal{L}(q, [\phi_h \circ \Pi_{\mathcal{A}}]^{-1}(p)). \end{array} \right. \tag{5.9}$$

Let us now introduce the concept of a completely integrable Hamiltonian system. The Poisson bracket of two real-valued functions f_1, f_2 on $\mathbb{R}^{2r}$ is given by $\{f_1, f_2\} = \sum_{i=1}^{r} \frac{\partial f_1}{\partial p_i}\frac{\partial f_2}{\partial q_i} - \frac{\partial f_1}{\partial q_i}\frac{\partial f_2}{\partial p_i}$. Notice that, from the definitions, we have $\{f_1, f_2\} = X_{f_1}(f_2)$ and $X_{\{f_1,f_2\}} = [X_{f_1}, X_{f_2}]$, where $[\cdot,\cdot]$ is the standard *bracket product* of vector fields [35]. Now, the system (5.7) is called *completely integrable* if there exist r functions $f_1, \dots, f_r$ with $f_1 = \mathcal{H}$, satisfying

$$\begin{cases} \{f_i, f_j\} = 0 \quad \forall i, j = 1, \dots, r, \\ df_1(x), \dots, df_r(x) \text{ are linearly independent at any } x \in \mathbb{R}^{2r}. \end{cases}$$

As a motivation for completely integrable systems, we will just point out the following: the functions f_i are called *integrals of motions* because $X_{\mathcal{H}}(f_i) = \{h, f_i\} = 0$, which means that any trajectory of $X_{\mathcal{H}}$ lies on the level sets of each f_i (the same holds for all X_{f_j}). Also, the trajectory passing through (q_0, p_0) lies in the set $\bigcap_{i=1,\dots,r} f_i^{-1}(\{f_i(q_o, p_0)\})$. Additionally, $[X_{f_i}, X_{f_j}] = 0$ implies that we can find, at least locally, coordinates $(x_1, \dots, x_r)$ on this set such that $X_{\mathcal{H}} = \frac{\partial}{\partial x_1}, X_{f_2} = \frac{\partial}{\partial x_2}, \dots, X_{f_r} = \frac{\partial}{\partial x_r}$; that is, in these coordinates, the trajectories of X_{f_i} are straight lines.

THEOREM 5.13. *Suppose* $\Pi_{\mathcal{A}_0} c \neq 0$. *The Lagrangian system on* $\mathcal{A}_0 \times \Pi_{\mathcal{A}_0}\mathcal{F}$ *associated with* (5.6), (5.8) *gives rise, by the Legendre transform, to a completely integrable Hamiltonian system on* $\mathcal{A}_0 \times \mathcal{F}^*$ *with Hamiltonian given by* (5.9).

Proof. There remains only to prove the complete integrability of the system. To this end, we adapt the proof of [6, Theorem II.12.2] to our abstract framework. Take the integrals of motion to be $f_1 = \mathcal{H}$, $f_i(q,p) = \langle v_i, p\rangle$, $i = 2, \dots, r$, where $r = n - m$ and $\{\Pi_{\mathcal{A}_0} c, v_2, \dots, v_r\}$ is chosen to be an orthonormal basis of $\mathcal{A}_0$. For any $i, j \in \{2, \dots, r\}$, $\{f_i, f_j\}$ is zero since f_i and f_j depend only on p. Let $\phi_{h,l}^{-1}(q,p)$ (resp., $(\Pi_{\mathcal{A}_0} c)_l$) stand for the lth component of $\phi_h^{-1}(q,p)$ (resp., the lth component of $\Pi_{\mathcal{A}_0} c$), and take some $k \in \{1, \dots, r\}$. Since

$$\begin{aligned} \frac{\partial \mathcal{H}}{\partial q_k}(q,p) &= \frac{\partial(\sum_{l=1}^{r} p_l \phi_{h,l}^{-1})}{\partial q_k}(q,p) - \frac{\partial(\mathcal{L}\circ\Phi^{-1})}{\partial q_k}(q,p) \\ &= \sum_{l=1}^{r} p_l \frac{\partial \phi_{h,l}^{-1}}{\partial q_k}(p,q) - \frac{\partial \mathcal{L}}{\partial q_k}(q, \phi_h^{-1}(q,p)) - \sum_{l=1}^{r} \frac{\partial \mathcal{L}}{\partial \dot{q}_l}(q, \phi_h^{-1}(q,p)) \frac{\partial \phi_{h,l}}{\partial q_k}(q,p) \\ &= -(\Pi_{\mathcal{A}_0} c)_k, \end{aligned}$$

we deduce that for all $i \in \{2, \dots, r\}$, $\{\mathcal{H}, f_i\} = \sum_{k=1}^{r} -\frac{\partial f_i}{\partial p_k}\frac{\partial \mathcal{H}}{\partial q_k} = \langle \Pi_{\mathcal{A}_0} c, v_i\rangle = 0$. The second condition for complete integrability is satisfied too, as the $r \times 2r$ matrix

$$\left(\left[\frac{\partial f_i}{\partial q_1}, \dots, \frac{\partial f_i}{\partial q_r}, \frac{\partial f_i}{\partial p_1}, \dots, \frac{\partial f_i}{\partial p_r}\right]\right)_{i=1,\dots,r} = \begin{pmatrix} \Pi_{\mathcal{A}_0} c^T & \star \\ & v_2^T \\ 0 & \dots \\ & v_r^T \end{pmatrix}$$

is full rank. ∎

REFERENCES

[1] E. AKIN, *The geometry of population genetics*, Lecture Notes in Biomath. 31, Springer-Verlag, Berlin, 1979.

[2] H. ATTOUCH, *Viscosity solutions of minimization problems*, SIAM J. Optim., 6 (1996), pp. 769–806.

[3] H. Attouch and M. Teboulle, *A regularized Lotka–Volterra dynamical system as a continuous proximal-like method in optimization*, J. Optim. Theory Appl., to appear.
[4] A. Auslender, R. Cominetti, and M. Haddou, *Asymptotic analysis for penalty and barrier methods in convex and linear programming*, Math. Oper. Res., 22 (1997), pp. 43–62.
[5] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Bregman monotone optimization algorithms*, SIAM J. Control Optim., 42 (2003), pp. 596–636.
[6] D. A. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming* I. *Affine and projective scaling trajectories*. Trans. Amer. Math. Soc., 314 (1989), pp. 499–526.
[7] D. A. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming* II. *Legendre transform coordinates and central trajectories*, Trans. Amer. Math. Soc., 314 (1989), pp. 527–581.
[8] J. Bolte and M. Teboulle, *Barrier operators and associated gradient-like dynamical systems for constrained minimization problems*, SIAM J. Control Optim., 42 (2003), pp. 1266–1292.
[9] L. M. Bregman, *The relaxation method for finding the common point of convex sets and its application to the solution of problems in convex programming*, Zh. Vychisl. Mat. Mat. Fiz., 7 (1967), pp. 620–631 (in Russian); English translation in Comput. Math. Math. Phys., 7 (1967), pp. 200–217.
[10] R. W. Brockett, *Dynamical systems that sort lists and solve linear programming problems*, in Proceedings of IEEE Conference Decision and Control, Austin, Texas, 1988, IEEE Press, Piscataway, NJ, pp. 779–803.
[11] R. W. Brockett, *Dynamical systems that sort lists, diagonalize matrices and solve linear programming problems*, Linear Algebra Appl., 146 (1991), pp. 79–91.
[12] R. E. Bruck, *Asymptotic convergence of non linear contraction semi-groups in Hilbert space*, J. Funct. Anal., 18 (1974), pp. 15–26.
[13] Y. Censor and A. Lent, *An iterative row action method for interval convex programming*, J. Optim. Theory Appl., 34 (1981), pp. 321–353.
[14] Y. Censor and S. A. Zenios, *Proximal minimization algorithm with D-functions*, J. Optim. Theory Appl., 73 (1992), pp. 451–464.
[15] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.
[16] R. Cominetti, *Nonlinear average and convergence of penalty trajectories in convex programming*, in Ill-Posed Variational Problems and Regularization Techniques (Trier, 1998), Lecture Notes in Econom. and Math. Systems 477, Springer, Berlin, 1999, pp. 65–78.
[17] R. Cominetti and J. San Martín, *Asymptotic analysis of the exponential penalty trajectory in linear programming*, Math. Programming, 67 (1994), pp. 169–187.
[18] M. P. do Carmo, *Riemannian Geometry (Mathematics, Theory, and Applications)*, Birkhäuser Boston, Cambridge, MA, 1992.
[19] J. J. Duistermaat, *On Hessian Riemannian structures*, Asian J. Math., 5 (2001), pp. 79–91.
[20] L. E. Faybusovich, *Dynamical systems which solve optimization problems with linear constraints*, IMA J. Math. Control Inform., 8 (1991), pp. 135–149.
[21] L. E. Faybusovich, *Hamiltonian structure of dynamical systems which solve linear programming problems*, Phys. D, 53 (1991), pp. 217–232.
[22] L. E. Faybusovich, *Interior point methods and entropy*, in Proceedings of the IEEE Conference on Decision and Control, Tucson, Arizona, IEEE Press, Piscataway, NJ, 1992, pp. 1626–1631.
[23] A. V. Fiacco, *Perturbed variations of penalty function methods. Example: Projective SUMT*, Ann. Oper. Res., 27 (1990), pp. 371–380.
[24] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.
[25] S. Herzel, M. C. Recchini, and F. Zirilli, *A quadratically convergent method for linear programming*, Linear Algebra Appl., 151 (1991), pp. 255–290.
[26] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II, Springer-Verlag, Berlin, 1996.
[27] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.
[28] A. N. Iusem and R. D. C. Monteiro, *On dual convergence of the generalized proximal point method with Bregman distances*, Math. Oper. Res., 25 (2000), pp. 606–624.
[29] A. N. Iusem, B. F. Svaiter, and J. X. da Cruz Neto, *Central paths, generalized proximal point methods, and Cauchy trajectories in Riemannian manifolds*, SIAM J. Control Optim., 37 (1999), pp. 566–588.
[30] N. Karmarkar, *A new polynomial time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[31] N. Karmarkar, *Riemannian geometry underlying interior point methods for linear programming*, in Mathematical Developments Arising from Linear Programming, Contemp. Math. 114, J. C. Lagarias and M. J. Todd, eds., AMS, Providence, RI, 1990, pp. 51–76.
[32] N. Kenmochi and I. Pawlow, *A class of doubly nonlinear elliptic-parabolic equations with time dependent constraints*, Nonlinear Anal., 10 (1986), pp. 1181–1202.
[33] K. C. Kiwiel, *Free-steering relaxation methods for problems with strictly convex costs*, Math. Oper. Res., 22 (1997), pp. 326–349.
[34] K. C. Kiwiel, *Proximal minimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), pp. 1142–1168.
[35] S. Lang, *Differential and Riemannian Manifolds*, Springer-Verlag, New York, 1995.
[36] G. P. McCormick, *The continuous Projective SUMT method for convex programming*, Math. Oper. Res., 14 (1989), pp. 203–223.
[37] Y. Nesterov and M. J. Todd, *On the Riemannian geometry defined by self-concordant barriers and interior-point methods*, Found. Comput. Math., 2 (2002), pp. 333–361.
[38] J. Palis and W. De Melo, *Geometric Theory of Dynamical Systems*, Springer, New York, 1982.
[39] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[40] R. T. Rockafellar and J-B. R. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1998.
[41] S. T. Smith, *Optimization techniques on Riemannian manifolds*, in Hamiltonian and Gradient Flows, Algorithms and Control, Fields Inst. Commun. 3, AMS, Providence, RI, 1994, pp. 113–136.
[42] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–690.
[43] C. Udrişte, *Convex Functions and Optimization Methods on Riemannian Manifolds*, Math. Appl., 297, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 502–531

AN APPLICATION OF STOCHASTIC CONTROL THEORY TO FINANCIAL ECONOMICS*

WENDELL H. FLEMING† AND TAO PANG‡

Abstract. We consider a portfolio optimization problem which is formulated as a stochastic control problem. Risky asset prices obey a logarithmic Brownian motion, and interest rates vary according to an ergodic Markov diffusion process. The goal is to choose optimal investment and consumption policies to maximize the infinite horizon expected discounted hyperbolic absolute risk aversion (HARA) utility of consumption. A dynamic programming principle is used to derive the dynamic programming equation (DPE). The subsolution–supersolution method is used to obtain existence of solutions of the DPE. The solutions are then used to derive the optimal investment and consumption policies.

Key words. portfolio optimization, dynamic programming equations, subsolutions, supersolutions

AMS subject classifications. 93E20, 60H30

DOI. 10.1137/S0363012902419060

1. Introduction. In the classical Merton portfolio optimization problem, an investor dynamically allocates wealth between a risky and a riskless asset and chooses a consumption rate, with the goal of maximizing total expected discounted utility of consumption. For a hyperbolic absolute risk aversion (HARA) utility function the Merton problem has a simple explicit solution. See, for example, Fleming and Soner [16, Example 5.2]. In the Merton model, the interest rate r of the riskless asset is a constant and the risky asset price fluctuates randomly according to a logarithmic Brownian motion. However, in our real world, even for money in the bank, the interest rate may fluctuate from time to time. Therefore, in the present paper we assume that the "riskless" interest rate r_t is an ergodic Markov diffusion process on the real line $-\infty < r < \infty$. A typical example is the Vasicek model, in which r_t is of Ornstein–Uhlenbeck type. In addition, the change of interest rate could be correlated with price fluctuation of the risky asset. A recent example is that the U.S. Federal Reserve has lowered the interest rate several times since 2000, due to the poor performance of the U.S. stock market. We also take this into account in this paper; see section 2. For the case where the change of interest rate is independent of the price change of the risky asset, similar problems were considered in [27, 28].

Another motivation for our work comes from models for optimal investment, production, and consumption, of a kind considered by Fleming and Stein [18]. This interpretation of our model will be explained at the end of section 2. See also Fleming and Pang [12].

We use the dynamic programming method. The stochastic control problem we consider has state variables x_t, r_t, where x_t is wealth. The controls are the fraction u_t of wealth in the risky asset and $c_t = \frac{C_t}{x_t}$, where C_t is the consumption rate. The state

*Received by the editors December 3, 2002; accepted for publication (in revised form) December 5, 2003; published electronically July 23, 2004. This research was partially supported by NSF grant DMS-9970852 through Brown University.

http://www.siam.org/journals/sicon/43-2/41906.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (whf@cfm.brown.edu).

‡Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (tpang@math.ncsu.edu).

dynamics are the SDEs (2.1)–(2.4). For a HARA utility, the value function $V(x,r)$ is a homogeneous function of x: $V(x,r) = \frac{1}{\gamma}x^\gamma W(r)$, where γ is the HARA parameter. For $\gamma > 0$, a source of technical difficulty is that $W(r)$ increases rapidly to infinity as $|r| \to \infty$. In fact, $Z(r) = \log W(r)$ should grow quadratically as $|r| \to \infty$. The dynamic programming equation (DPE) (2.14) for $V(x,r)$ is equivalent to a nonlinear ODE (2.22) for $Z(r)$. We call (2.22) the reduced DPE.

We use a method of subsolution and supersolution to show that the reduced DPE (2.22) has a solution $\tilde{Z}(r)$ with appropriate behavior as $|r| \to \infty$. The subsolution–supersolution method is developed in section 3. It is applied in section 4 with $\gamma > 0$ to find a classical solution $\tilde{Z}(r)$ to (2.22) which is bounded below and which grows at most quadratically as $|r| \to \infty$. A verification result (Theorem 4.13) then shows that $\tilde{Z}(r) = Z(r)$ and that the corresponding control policies $u^*(r), c^*(r)$ in formulas (4.60) are optimal. These results require that $0 < \gamma < \bar{\gamma}$ for suitable $\bar{\gamma} \leq 1$. In section 5 we consider $\gamma < 0$. In this case $\tilde{W}(r) = \exp(\tilde{Z}(r))$ decays to 0 as $|r| \to \infty$ like $|r|^{2(\gamma-1)}$. The verification result is Theorem 5.7 in this case.

The results in this paper are adapted from the second author's Ph.D. thesis [27]. In Chapter 2 of [27], a related optimal investment problem on a finite time horizon $0 \leq t \leq T$ was also considered. The goal is then to choose an investment control u_t to maximize the expected HARA utility of final wealth $E[\gamma^{-1}x_T^\gamma]$. This model is of a type previously considered by Bielecki and Pliska [2], Zariphopoulou [32], and Fleming and Sheu [14]. The analysis for that finite horizon stochastic control problem is considerably simpler than for the optimal investment-consumption model considered in the present paper.

Fleming and Hernandez-Hernandez [10] considered an investment-consumption model in which the interest rate is constant but the volatility of the risky asset price is stochastic. The approach in [10] has some features in common with the present paper. However, the methods and technical issues to be resolved in the two papers are different.

Our methods should apply to a wider class of stochastic control problems in which the DPE reduces to an ODE of the form $-LZ = h(r,Z)$ as in (4.3). The function $h(r,Z)$ in (4.2) is the sum of a term $\gamma Q(r) - \beta$ and a decreasing function of Z. The function $Q(r)$ grows quadratically as $|r| \to \infty$. This feature significantly complicated the analyses in sections 4 and 5, in the cases $\gamma > 0$ and $\gamma < 0$.

2. The DPE. We use a logarithmic Brownian motion to describe the price P_t of the risky asset:

$$\frac{dP_t}{P_t} = bdt + \sigma_1 dw_{1,t},$$

where b, σ_1 are positive constants and $w_{1,t}$ is a standard one-dimensional Brownian motion. Let x_t be the wealth at time t. The investment control u_t at time t is the fraction of wealth invested in the risky asset. So $(1-u_t)$ is the fraction of the wealth invested in the riskless asset. Denote by C_t the consumption rate at time t. For technical reasons, we take $c_t \equiv \frac{C_t}{x_t}$ as a control instead of C_t. Suppose the initial wealth is $x > 0$. Then the SDE for the process x_t is

$$dx_t = x_t[r_t + (b - r_t)u_t - c_t]dt + \sigma_1 u_t x_t dw_{1,t}, \tag{2.1}$$

$$x_0 = x, \tag{2.2}$$

where r_t is the interest rate of the riskless asset at time t. Instead of a constant interest rate in the classical Merton's model, we consider a randomly fluctuating interest rate

model,

$$dr_t = f(r_t)dt + \sigma_2 d\tilde{w}_t, \tag{2.3}$$

$$r_0 = r, \tag{2.4}$$

where σ_2 is a constant and $\tilde{w}_t$ is a standard one-dimensional Brownian motion. In some cases, the fluctuation of the interest rate is correlated with the price change of the risky asset. To describe this, we let $w_t = (w_{1,t}, w_{2,t})'$ be a standard two-dimensional Brownian motion. Define $\tilde{w}_t$ such that

$$d\tilde{w}_t = \rho dw_{1,t} + \sqrt{1-\rho^2} dw_{2,t}, \tag{2.5}$$

where $\rho \in [-1, 1]$ is a constant. Since $w_{1,t}$ and $w_{2,t}$ are independent, we have

$$E[dw_{1,t} \cdot d\tilde{w}_t] = \rho dt. \tag{2.6}$$

So ρ is the correlation coefficient.

In this paper, we will consider the generalized Vasicek model,

$$f(r) \in \mathbf{C}^2(\mathbf{R}), \tag{2.7}$$

$$|f_{rr}(r)| \le K(1 + |r|^\alpha), \tag{2.8}$$

$$-c_2 \le f_r(r) \le -c_1, \tag{2.9}$$

where K, α, c_1, and c_2 are positive constants.

We consider a HARA utility function $U(\cdot)$:

$$U(C) = \frac{1}{\gamma} C^\gamma, \quad -\infty < \gamma < 1, \quad \gamma \neq 0. \tag{2.10}$$

Our goal is then to maximize the objective function

$$J(x, r, u_\cdot, c_\cdot) \equiv E_{x,r} \int_0^\infty e^{-\beta t} U(c_t x_t) dt, \tag{2.11}$$

where $(u_\cdot, c_\cdot)$ belong to a class Π of admissible controls. Then our value function is

$$V(x, r) = \sup_{u_\cdot, c_\cdot} E_{x,r} \int_0^\infty e^{-\beta t} U(c_t x_t) dt. \tag{2.12}$$

We require that the control $(u_t, c_t; t \ge 0)$ is an $\mathbf{R}^2$-valued process. In addition, we require that it is $\mathcal{F}_t$-progressively measurable for some $(w_{1,t}, \tilde{w}_t)$-adapted increasing family of σ-algebras $(\mathcal{F}_t, t \ge 0)$. See Fleming and Soner [16, Chapter 4] for details. In certain cases, (u_t, c_t) may be obtained from locally Lipschitz continuous control policies $(\hat{u}, \hat{c})$,

$$u_t = \hat{u}(t, x_t, r_t), \quad c_t = \hat{c}(t, x_t, r_t),$$

where x_t is obtained by substituting these policies in (2.1).

We also assume that $c_t \ge 0$, and there is no constraint for the value of u_t. In other words, we take the u-value space $\mathbf{U} = (-\infty, \infty)$ in this paper. The negative value of u_t corresponds to disinvestment such as short selling.

In addition, we require that

$$P\left(\int_0^T u_t^2 dt < \infty\right) = 1 \quad \forall T > 0. \tag{2.13}$$

Given this, we can use Ito's differential rule to verify that

$$x_t = x \exp\left\{\int_0^t \left[r_s + (b - r_s)u_s - c_s - \frac{1}{2}\sigma_1^2 u_s^2\right] ds + \int_0^t \sigma_1 u_s dw_{1,s}\right\}$$

is a solution of (2.1)–(2.2). We can see that $x_t > 0$ as long as $x > 0$.

Remark 2.1. The admissible control space Π will be specified later in Definition 4.11 ($\gamma > 0$ case) and Definition 5.6 ($\gamma < 0$ case). For fixed $\beta > 0$, there exists a constant $\bar{\gamma} \leq 1$ such that $0 < \gamma < \bar{\gamma}$ will ensure that $V(x, r) < \infty$. For a constant interest r, a condition about β and γ is given in [16, p. 176].

Remark 2.2. The log utility case, which corresponds to a HARA utility with $\gamma = 0$, is studied in Pang [27, section 1.4]. It is much easier to deal with.

By the definition of $V(x, r)$, using the dynamic programming principle, we can obtain that the corresponding DPE is

$$\begin{aligned}\beta V = &\sup_u \left[(b - r)uxV_x + \frac{1}{2}\sigma_1^2 u^2 x^2 V_{xx} + \rho\sigma_1\sigma_2 uxV_{xr}\right] + rxV_x \\ &+ f(r)V_r + \frac{1}{2}\sigma_2^2 V_{rr} + \sup_{c \geq 0}\left[-cxV_x + \frac{1}{\gamma}(cx)^\gamma\right].\end{aligned} \tag{2.14}$$

For details, refer to [16, section 4.5].

Since we consider a HARA utility function that is homogeneous in x with an order of γ, it is not hard to get the following lemma.

LEMMA 2.3. *$V(x, r)$ is homogeneous in x with an order of γ.*

Proof. According to (2.1)–(2.2), for any $k > 0$, we have

$$\begin{aligned} dkx_t &= kx_t[r_t + (b - r_t)u_t - c_t]dt + \sigma_1 u_t kx_t dw_{1,t}, \\ kx_0 &= kx. \end{aligned}$$

Therefore,

$$J(kx, r, u_\cdot, c_\cdot) = E_{x,r}\int_0^\infty e^{-\beta t}\frac{1}{\gamma}(c_t kx_t)^\gamma dt = k^\gamma J(x, r, u_\cdot, c_\cdot).$$

Thus we have

$$V(x, r) = \sup_{u_\cdot, c_\cdot} J(x, r, u_\cdot, c_\cdot) = \sup_{u_\cdot, c_\cdot} x^\gamma J(1, r, u_\cdot, c_\cdot) = x^\gamma V(1, r).$$

That is, $V(x, r)$ is homogeneous in x. ∎

From Lemma 2.3, we can assume that

$$V(x, r) = \frac{1}{\gamma}x^\gamma W(r). \tag{2.15}$$

Then, the differential equation for $W(r)$ can be written as

$$\begin{aligned}\frac{\beta}{\gamma}W = &\sup_u \left[(b - r)uW + \frac{1}{2}(\gamma - 1)\sigma_1^2 u^2 W + \rho\sigma_1\sigma_2 uW_r\right] + rW \\ &+ \frac{1}{\gamma}f(r)W_r + \frac{1}{2\gamma}\sigma_2^2 W_{rr} + \sup_{c \geq 0}\left[-cW + \frac{1}{\gamma}c^\gamma\right].\end{aligned}$$

By the definition of $V(x,r)$, it is not hard to see that the suitable $W(r)$ should be positive.

Actually, if $W(r) > 0$ and smooth enough, we can define

$$u^*(r) \equiv \frac{(b-r)W(r) + \rho\sigma_1\sigma_2 W_r(r)}{(1-\gamma)\sigma_1^2 W(r)}, \tag{2.16}$$

$$c^*(r) \equiv W(r)^{\frac{1}{\gamma-1}}. \tag{2.17}$$

Then we have

$$u^*(r) \in \arg\max_u \left[(b-r)uW + \frac{1}{2}(\gamma-1)\sigma_1^2 u^2 W + \rho\sigma_1\sigma_2 u W_r\right],$$
$$c^*(r) \in \arg\max_c \left[-cW + \frac{1}{\gamma}c^\gamma\right].$$

Actually, (u^*, c^*) will be verified to be the optimal control policy later in section 4 and section 5 for $\gamma > 0$ and $\gamma < 0$, respectively.

Now we can rewrite the differential equation of $W(r)$ as

$$\begin{aligned}\frac{1}{2}\sigma_2^2 W_{rr} + \left[\frac{\gamma\rho\sigma_2(b-r)}{\sigma_1(1-\gamma)} + f(r)\right]W_r + \frac{\gamma\rho^2\sigma_2^2 W_r^2}{2(1-\gamma)W} \\ + [\gamma Q(r) - \beta]W + (1-\gamma)W^{\frac{\gamma}{\gamma-1}} = 0,\end{aligned} \tag{2.18}$$

where

$$Q(r) = \frac{(b-r)^2}{2(1-\gamma)\sigma_1^2} + r. \tag{2.19}$$

We can see that $Q(r)$ is quadratic with respect to r. Let

$$Z(r) \equiv \log W(r). \tag{2.20}$$

Then the ODE for $Z(r)$ is

$$\begin{aligned}\frac{\sigma_2^2}{2}Z_{rr} + \frac{\sigma_2^2}{2}\left[1 + \frac{\gamma\rho^2}{1-\gamma}\right]Z_r^2 + \left[\frac{\gamma\rho\sigma_2(b-r)}{\sigma_1(1-\gamma)} + f(r)\right]Z_r \\ + \gamma Q(r) - \beta + (1-\gamma)e^{\frac{Z}{\gamma-1}} = 0.\end{aligned}$$

Define

$$\begin{aligned}H(r,z,p) \equiv -\frac{\sigma_2^2}{2}\left[1 + \frac{\gamma\rho^2}{1-\gamma}\right]p^2 - \left[\frac{\gamma\rho\sigma_2(b-r)}{\sigma_1(1-\gamma)} + f(r)\right]p \\ - \gamma Q(r) + \beta - (1-\gamma)e^{\frac{z}{\gamma-1}};\end{aligned} \tag{2.21}$$

then the equation for $Z(r)$ can be rewritten as

$$\frac{\sigma_2^2}{2}Z_{rr} = H(r, Z, Z_r). \tag{2.22}$$

We call (2.22) the reduced DPE. Our goal is to find a suitable solution $\tilde{V}(x,r)$ of the DPE (2.14) and verify that $\tilde{V}(x,r)$ is equal to the value function defined by (2.12). To obtain $\tilde{V}(x,r)$, it is sufficient to find a suitable solution $Z(r)$ of (2.22).

Then, $\tilde{V}(x,r) = \frac{1}{\gamma}x^\gamma e^{Z(r)}$ will be the desired solution of (2.14). Although (2.22) is a nonlinear equation, we can get some existence results by using a subsolution–supersolution method.

Investment, production, and consumption model. In addition to Merton-type, small-investor portfolio optimization problems with randomly fluctuating interest rates, another motivation for our work comes from considering models of the following kind. An economic unit has productive capital and also liabilities in the form of debt. Let K_t denote the worth of capital at time t and L_t the debt. K_t changes through investment, at rate I_t. Debt changes through interest payments, investment, consumption C_t, and income from production Y_t:

$$dK_t = I_t dt, \tag{2.23}$$

$$dL_t = (r_t L_t + I_t + C_t - Y_t)dt. \tag{2.24}$$

It is assumed that productivity of capital fluctuates randomly about a mean rate b. This is expressed by writing (formally)

$$Y_t dt = K_t(bdt + \sigma_1 dw_{1,t}), \tag{2.25}$$

where $w_{1,t}$ is a Brownian motion as above. The constraints imposed are $K_t \geq 0, C_t \geq 0, x_t > 0$, where $x_t = K_t - L_t$ is the net worth of the economic unit. By subtracting (2.24) from (2.23) we find that x_t satisfies the SDE (2.1) with

$$u_t = x_t^{-1}K_t, \quad c_t = x_t^{-1}C_t. \tag{2.26}$$

If no bounds are imposed on the investment rate I_t, then u_t can be taken as the investment control and c_t the consumption control. The constraint $K_t \geq 0$ is equivalent to the "no short selling" constraint $u_t \geq 0$. We will ignore this constraint in the sections to follow. To include this constraint, it requires rather easy modifications. For example, in (2.14), the first sup would be taken over $u \geq 0$ rather than over all u.

In [18], a similar international finance and debt model was considered. In that interpretation the economic unit is a nation. Y_t represents the national gross domestic product and L_t is the foreign debt. However, instead of a "mean reverting" model (2.3) for the interest rate r_t, it is assumed in [18] that (formally)

$$r_t dt = rdt + \sigma_2 dw_{2,t},$$

where $w_{2,t}$ is a Brownian motion. As in the Merton problem, there is an explicit solution in the model considered in [18]. However, if the interest rate r_t satisfies the SDE (2.3), then the optimal investment and consumption policies $u^*(r), c^*(r)$ depend on the solution $W(r)$ to a reduced DPE as in (2.16) and (2.17). This differential equation, or the equivalent differential equation for $Z(r) = \log W(r)$, can be solved numerically.

3. Method of subsolution and supersolution. In this section, we will give an existence result for some type of ODEs, which includes (2.22). The method of subsolution and supersolution will be used. This idea is partially from [29], [3], and [31].

Consider a second order differential equation

$$Z_{rr} = \bar{H}(r, Z, Z_r). \tag{3.1}$$

First let us define subsolutions and supersolutions of (3.1).

DEFINITION 3.1. *A function $\hat{Z}$ is said to be a subsolution of* (3.1) *on the whole real line if*

$$\hat{Z}_{rr} \geq \bar{H}(r, \hat{Z}, \hat{Z}_r).$$

$\bar{Z}$ is a supersolution if

$$\bar{Z}_{rr} \leq \bar{H}(r, \bar{Z}, \bar{Z}_r).$$

In addition, $(\hat{Z}, \bar{Z})$ is said to be an ordered pair of subsolution/supersolution of (3.1) *if they also satisfy*

$$\hat{Z}(r) \leq \bar{Z}(r) \quad \forall r \in \mathbf{R}.$$

We also want to define supersolutions and subsolutions of the corresponding boundary value problem on a finite interval $[r_1, r_2]$,

$$\begin{cases} Z_{rr} = \bar{H}(r, Z, Z_r), \\ Z(r_1) = Z_1, \quad Z(r_2) = Z_2. \end{cases} \tag{3.2}$$

DEFINITION 3.2. *A function $\hat{Z}$ is said to be a subsolution of* (3.2) *if*

$$\hat{Z}_{rr} \geq \bar{H}(r, \hat{Z}, \hat{Z}_r), \quad \hat{Z}(r_1) \leq Z_1, \quad \hat{Z}(r_2) \leq Z_2.$$

$\bar{Z}$ is a supersolution of (3.2) *if*

$$\bar{Z}_{rr} \leq \bar{H}(r, \bar{Z}, \bar{Z}_r), \quad \bar{Z}(r_1) \geq Z_1, \quad \bar{Z}(r_2) \geq Z_2.$$

In addition, $\hat{Z}$ and $\bar{Z}$ are said to be an ordered pair of subsolution/supersolution if they also satisfy

$$\hat{Z}(r) \leq \bar{Z}(r) \quad \forall r \in [r_1, r_2].$$

First we will show that a similar existence result holds for (3.2). Then we will extend the result to the whole real line and get an existence result of (3.1). Equation (2.22) will be a special case. The following lemma is needed.

LEMMA 3.3. *Let $F(r, z, p)$ be continuous and bounded on $J \times \mathbf{R}^2$, where $J = [r_1, r_2]$. Then the boundary value problem*

$$\begin{cases} Z_{rr} = F(r, Z, Z_r), \\ Z(r_1) = Z_1, \quad Z(r_2) = Z_2 \end{cases}$$

has at least one solution.

Proof. This is a direct result of Walter [31, p. 262, Existence Theorem XX]. □

Some a priori estimates are needed to get the existence results for the boundary value problem (3.2).

LEMMA 3.4. *Suppose $Z(r)$ is a classical $\mathbf{C}^2$ solution of* (3.2) *on $J = [r_1, r_2]$, and it satisfies*

$$\hat{Z}(r) \leq Z(r) \leq \bar{Z}(r) \quad \text{on } J,$$

where $\hat{Z}(r)$ and $\bar{Z}(r)$ are subsolution and supersolution of (3.2)*, respectively. Define*

$$M \equiv \max\left\{ \sup_J |\bar{Z}(r)|, \sup_J |\hat{Z}(r)| \right\}. \tag{3.3}$$

Suppose that

$$|\bar{H}(r,z,p)| \le C_1(p^2 + C_2) \tag{3.4}$$

for $r \in J$ *and* $|z| \le 3M$, *where* M *is given by* (3.3) *and* $C_1 > 0, C_2 \ge 0$ *are two constants. Then there exists a constant* Λ, *which depends only on* M, C_1, *and* C_2, *such that*

$$|Z_r| \le \Lambda \quad \text{on } J.$$

Proof. Take $\bar{\mu} \equiv \max\left\{2C_1, \sqrt{C_2}\right\}$. Then, by the above definition, we can get that if $|p| \ge \bar{\mu}$, we have

$$|\bar{H}(r,z,p)| \le C_1(p^2 + C_2) \le C_1(p^2 + p^2) \le \bar{\mu}p^2. \tag{3.5}$$

Take constants k, δ such that

$$k \ge \bar{\mu}^2 e^{2\bar{\mu}M}, \quad k\delta = e^{2\bar{\mu}M} - 1.$$

Fix an $r_0 \in [r_1, r_2]$. For $r \in [r_0, r_0 + \delta]$, define

$$w(r) \equiv \frac{1}{\bar{\mu}} \log[1 + k(r - r_0)] + Z(r_0).$$

Then we can verify that

$$\begin{aligned} w(r_0) = Z(r_0), \quad & w(r_0 + \delta) = 2M + Z(r_0) \ge M \ge Z(r_0 + \delta), \\ & |w(r)| \le 2M + |Z(r_0)| \le 3M, \quad |w_r(r)| \ge \bar{\mu}. \end{aligned}$$

Given this, noting (3.4), we can show that

$$w_{rr} - \bar{H}(r, w, w_r) \le -\bar{\mu}w_r^2 + |\bar{H}(r, w, w_r)| \le -\bar{\mu}w_r^2 + \bar{\mu}w_r^2 = 0.$$

Now, by virtue of Gilbarg and Trudinger [20, Theorem 10.1, p. 263], we can get

$$w(r) \ge Z(r) \quad \forall r \in [r_0, r_0 + \delta].$$

Similarly, for $\breve{w}(r) \equiv -w(r)$, using the same method, we can get

$$-w(r) = \breve{w}(r) \le Z(r) \quad \forall r \in [r_0, r_0 + \delta].$$

Therefore, for any $r \in (r_0, r_0 + \delta)$, we have

$$\frac{|Z(r) - Z(r_0)|}{|r - r_0|} \le \frac{|w(r) - w(r_0)|}{|r - r_0|}.$$

Let $r \to r_0^+$; we can get

$$|Z_r(r_0)| \le \left|\frac{k}{\bar{\mu}}\right| \equiv \Lambda.$$

Since $r_0 \in J$ is arbitrary, we are done. □

LEMMA 3.5. *Suppose* $\bar{H}(r,z,p)$ *is strictly increasing with respect to* z, *and it satisfies* (3.4). *If* $\hat{Z}$ *and* $\bar{Z}$ *are an ordered pair of subsolution/supersolution of* (3.2)

on $J = [r_1, r_2]$, then the boundary value problem (3.2) *has at least one solution on J such that*

$$\hat{Z}(r) \le Z(r) \le \bar{Z}(r) \quad \forall r \in J.$$

Proof. Define

$$\Omega \equiv \{(r, z, p) : r \in J,\ z \in [\hat{Z}, \bar{Z}],\ |p| < \Lambda_0\}, \tag{3.6}$$

where $\Lambda_0 \equiv \max\{\Lambda, \max_J \hat{Z}_r, \max_J \bar{Z}_r\}$ and Λ is a constant as in Lemma 3.4.

Since $\bar{H}(r, z, p)$ is strictly increasing with respect to z, and it satisfies (3.4), it is not hard to extend $\bar{H}$ to the domain $J \times \mathbf{R}^2$, such that it is a continuous, bounded function and it is strictly increasing with respect to z. Denote the extension to be $\tilde{H}$. In addition, we can suppose that $\tilde{H}$ satisfies (3.4). For example, we can take

$$\tilde{H}_1(r, z, p) = \begin{cases} \bar{H}(r, z, p) & \text{if } r \in J,\ \hat{Z} \le z \le \bar{Z}, \\ \bar{H}(r, \hat{Z}, p) + e^z - e^{\hat{Z}} & \text{if } r \in J,\ z < \hat{Z}, \\ \bar{H}(r, \bar{Z}, p) + e^{-\bar{Z}} - e^{-z} & \text{if } r \in J,\ z \ge \bar{Z}, \end{cases}$$

and

$$\tilde{H}(r, z, p) = \begin{cases} \tilde{H}_1(r, z, p) & \text{if } |p| \le \Lambda_0, \\ \tilde{H}_1(r, z, -\Lambda_0) & \text{if } p < -\Lambda_0, \\ \tilde{H}_1(r, z, \Lambda_0) & \text{if } p > \Lambda_0. \end{cases}$$

It is not hard to verify that $\tilde{H}(r, z, p)$ is a bounded continuous function on $J \times \mathbf{R}^2$. In addition, $\tilde{H}(r, z, p)$ is strictly increasing with respect to z, and it satisfies (3.4).

Take constants Z_1, Z_2 such that

$$\hat{Z}(r_i) \le Z_i \le \bar{Z}(r_i), \quad i = 1, 2.$$

Now according to Lemma 3.3, we know that the boundary value problem

$$\begin{cases} Z_{rr} = \tilde{H}(r, Z, Z_r), \\ Z(r_1) = Z_1, \quad Z(r_2) = Z_2 \end{cases}$$

has a solution, say, $Z(r)$. Now we need to show that $\hat{Z} \le Z \le \bar{Z}$ and $|Z_r| \le \Lambda_0$. Assume that $Z \le \bar{Z}$ does not always hold on J. Then $\bar{Z} - Z$ is negative in an open set I_0 and is nonnegative at its endpoints. Suppose $\bar{Z} - Z$ reaches its minimum at $r_0 \in I_0$; then we have

$$\bar{Z}_r(r_0) = Z_r(r_0), \quad \bar{Z}(r_0) < Z(r_0).$$

Noting that $\tilde{H}$ is strictly increasing with respect to z, we can get

$$(\bar{Z}_{rr} - Z_{rr})(r_0) \le \tilde{H}(r_0, \bar{Z}(r_0), \bar{Z}_r(r_0)) - \tilde{H}(r_0, Z(r_0), Z_r(r_0)) < 0.$$

So $(\bar{Z} - Z)$ cannot reach its minimum in I_0. This is a contradiction. Therefore, we must have $Z \le \bar{Z}$ on J. A similar argument gives $\hat{Z} \le Z$. Further, since $\tilde{H}$ satisfies (3.4), following the same procedure in the proof of Lemma 3.4, we can show that $|Z_r| \le \Lambda \le \Lambda_0$ on J. Therefore, we can get that $\tilde{H}(r, Z, Z_r) = \bar{H}(r, Z, Z_r)$. Therefore, Z is a solution of (3.2). ∎

The following uniqueness result is needed later.

LEMMA 3.6 (uniqueness). *Suppose $\bar{H}(r,z,p)$ is strictly increasing with respect to z, and it satisfies* (3.4). *If two $\mathbf{C}^2$ functions $Z(r)$ and $\tilde{Z}(r)$ are solutions of* (3.1) *on $J=[r_1,r_2]$, such that*

$$\hat{Z}(r) \le Z(r), \quad \tilde{Z}(r) \le \bar{Z}(r), \tag{3.7}$$

then

$$Z(r) \equiv \tilde{Z}(r) \quad \text{on } J. \tag{3.8}$$

Proof. Let $\psi(r) \equiv Z(r) - \tilde{Z}(r)$. Then we have that $\psi(r_1) = \psi(r_2) = 0$. Assume that ψ reaches it minimum at $r_0 \in (r_1, r_2)$, such that $\psi(r_0) < 0$, that is, $Z(r_0) < \tilde{Z}(r_0)$, and $Z_r(r_0) = \tilde{Z}_r(r_0)$. Then, by virtue of (3.1) and the definition of ψ, noting that $\bar{H}(r,z,p)$ is strictly increasing with z, we can get

$$\psi_{rr}(r_0) < 0.$$

This contradicts the assumption that ψ reaches its minimum at $r_0 \in (r_1, r_2)$. Therefore, we must have $Z(r) \ge \tilde{Z}(r)$ on J. The same argument for $\psi = \tilde{Z} - Z$ will lead to $Z(r) \le \tilde{Z}(r)$ on J. □

Let $(\hat{Z}(r), \bar{Z}(r))$ be an ordered pair of subsolution/supersolution of (3.1) on the whole real line, that is, $\forall r \in \mathbf{R}$,

$$\hat{Z}_{rr} \ge \bar{H}(r, \hat{Z}, \hat{Z}_r), \tag{3.9}$$

$$\bar{Z}_{rr} \le \bar{H}(r, \bar{Z}, \bar{Z}_r), \tag{3.10}$$

$$\hat{Z}(r) \le \bar{Z}(r). \tag{3.11}$$

According to the definitions, it is immediate that $\hat{Z}$ and $\bar{Z}$ are an ordered pair of subsolution/supersolution of the following problem on any $I_m \equiv [-m, m]$:

$$\begin{cases} Z_{rr} = \bar{H}(r, Z, Z_r), \\ Z(-m) = \bar{Z}(-m), \quad Z(m) = \bar{Z}(m). \end{cases} \tag{3.12}$$

Now by virtue of Lemma 3.5, the above problem has at least one solution $\tilde{Z}_m^0(r)$, such that

$$\hat{Z}(r) \le \tilde{Z}_m^0(r) \le \bar{Z}(r) \quad \text{on } I_m.$$

Define its extension on $\mathbf{R}$ by

$$\tilde{Z}_m(r) = \begin{cases} \tilde{Z}_m^0(r) & \text{if } r \in I_m, \\ \bar{Z}(r) & \text{otherwise.} \end{cases}$$

Then $\tilde{Z}_m$ is continuous. Further, we have the following lemma.

LEMMA 3.7. *For any m, we have*

$$\hat{Z}(r) \le \tilde{Z}_{m+1}(r) \le \tilde{Z}_m(r) \le \bar{Z}(r). \tag{3.13}$$

Proof. By definition, for any m, we must have

$$\hat{Z} \le \tilde{Z}_m \le \bar{Z}.$$

So we only need to show that

$$\tilde{Z}_{m+1}(r) \le \tilde{Z}_m(r) \quad \forall r. \tag{3.14}$$

By the definitions of $\{Z_m, m = 1, 2, 3, \ldots\}$, it is sufficient to show that the above inequality holds on I_m. Actually, it is not hard to verify that $\tilde{Z}_{m+1}$ is a subsolution of (3.12)–(3.13) on I_m. Then by virtue of Lemma 3.5, there exists a solution $\tilde{Z}^*(r)$ of (3.12)–(3.13), such that

$$\tilde{Z}_{m+1}(r) \le \tilde{Z}^*(r) \le \bar{Z}(r) \quad \forall r \in I_m.$$

Noting the result of Lemma 3.6, we must have

$$\tilde{Z}^*(r) \equiv \tilde{Z}_m(r) \quad \forall r \in I_m,$$

which implies that (3.14) holds on I_m. This completes our proof. □

Finally, we have the following existence result.

THEOREM 3.8. *Suppose $\bar{H}(r, z, p)$ is strictly increasing with respect to z, and it satisfies* (3.4). *Let $(\hat{Z}, \bar{Z})$ be an ordered pair of subsolution/supersolution of* (3.1) *on* **R**. *Then* (3.1) *has a solution $Z(r)$ such that*

$$\hat{Z}(r) \le Z(r) \le \bar{Z}(r). \tag{3.15}$$

Proof. Consider the sequence $\{\tilde{Z}_m\}$ as in Lemma 3.7. It is easy to show that $\tilde{Z}_m$ converges in pointwise sense to a function Z as $m \to \infty$.

Since any bounded interval J is contained in I_m for some m, a $\mathbf{C}^2$ function Z is a solution of (3.1) if it satisfies (3.1) in I_m for any m. Let m be fixed, and let $k > m$ be arbitrary. Then for $r \in I_m$, $\tilde{Z}_k(r)$ satisfies

$$\frac{\partial^2 \tilde{Z}_k}{\partial r^2} = \bar{H}\left(r, \tilde{Z}_k, \frac{\partial \tilde{Z}_k}{\partial r}\right), \quad \tilde{Z}_k(-m) \le \bar{Z}(-m), \quad \tilde{Z}_k(m) \le \bar{Z}(m).$$

Since $\hat{Z} \le \tilde{Z}_k \le \bar{Z}$ $\forall r \in I_m$, we know that $\{\tilde{Z}_k\}$ is uniformly bounded on I_m. In addition, noting Lemma 3.4, we can get that $\{\frac{\partial \tilde{Z}_k}{\partial r}\}$ is uniformly bounded on I_m. Finally, by virtue of (3.1), (3.4), and (3.15), it is not hard to show that $\{\frac{\partial^2 \tilde{Z}_k}{\partial r^2}\}$ and $\{[\frac{\partial^2 \tilde{Z}_k}{\partial r^2}]_{\alpha; I_m}\}$ are uniformly bounded on I_m.

Given the above results, using the Arzela–Ascoli theorem, we can show that $\{\tilde{Z}_m\}$ contains a subsequence that converges in $\mathbf{C}^2(I_m)$ to a function $\tilde{Z} \in \mathbf{C}^{2,\alpha}(I_m)$. Since $\{\tilde{Z}_k\}$ converges to Z in pointwise sense, $\tilde{Z}$ must coincide with Z. Moreover, the whole sequence $\{\tilde{Z}_k\}$ converges in $\mathbf{C}^2(I_m)$ to Z as $k \to \infty$. Let $k \to \infty$, and we can get that Z is a solution of (3.1) on I_m. By the arbitrariness of I_m, Z is a solution of (3.1) on **R**. □

Now we need only find an ordered pair of subsolution/supersolution to get the existence of the classical solution $\tilde{Z}(r) = Z(r)$. Then we can obtain the classical solution $\tilde{V}(x, r) = \frac{1}{\gamma} x^\gamma e^{\tilde{Z}(r)}$. This will be done for the $\gamma > 0$ case in section 4 and for the $\gamma < 0$ case in section 5. The solution will be verified to be the value function in both cases. These verification results imply that the solution $Z(r)$ to (3.1) satisfying the bounds (3.16) is unique.

It is not hard to show that the function $H(r, z, p)$ defined by (2.21) is strictly increasing with respect to z, and it satisfies (3.4). Therefore, we have the following lemma.

LEMMA 3.9. *Let* $(\hat{Z}, \bar{Z})$ *be an ordered pair of subsolution/supersolution of* (2.22) *on* $\mathbf{R}$. *Then* (2.22) *has a solution* $Z(r)$ *such that*

$$\hat{Z}(r) \leq Z(r) \leq \bar{Z}(r). \tag{3.16}$$

4. $\gamma > 0$ case. In this section, we will find an ordered pair of subsolution/supersolution when $\gamma > 0$ under some conditions, which will be specified in Lemma 4.1 and Lemma 4.2. Then we can get the existence of the solution of the reduced DPE (2.22) by using Lemma 3.9. Further, we need to verify that this solution is actually our value function. This result is given in Theorem 4.4. The admissible control space is defined by Definition 4.11.

Define

$$LZ = \frac{\sigma_2^2}{2} Z_{rr} + \frac{\sigma_2^2}{2}\left[1 + \frac{\gamma\rho^2}{1-\gamma}\right] Z_r^2 + \left[\frac{\gamma\rho\sigma_2(b-r)}{\sigma_1(1-\gamma)} + f(r)\right] Z_r, \tag{4.1}$$

$$h(r, Z) = [\gamma Q(r) - \beta] + (1-\gamma) e^{\frac{Z}{\gamma-1}}. \tag{4.2}$$

Then (2.22) for Z can be written as

$$-LZ = h(r, Z). \tag{4.3}$$

It is easy to verify that Z is a subsolution (supersolution) of (2.22) if and only if

$$-LZ \leq (\geq) h(r, Z).$$

LEMMA 4.1. *Suppose*

$$\beta > \gamma b - \frac{\sigma_1^2}{2}\gamma(1-\gamma). \tag{4.4}$$

Define K_1 *as*

$$K_1 \equiv \log \tilde{K}_1, \tag{4.5}$$

where $\tilde{K}_1$ *is a positive constant defined by*

$$\tilde{K}_1^{\frac{1}{\gamma-1}} = \frac{1}{1-\gamma}\left[\beta - b\gamma + \frac{\sigma_1^2}{2}\gamma(1-\gamma)\right]. \tag{4.6}$$

Then, any constant $K_2 \leq K_1$ *is a subsolution of* (2.22).

Proof. Since K_2 is a constant, we have

$$-LK_2 = 0.$$

On the other hand, since $Q(r)$ is quadratic, by the definition of K_1, it is not hard to verify that

$$h(r, K_2) > 0$$

for any constant $K_2 \leq K_1$. Thus, we have

$$-LK_2 < h(r, K_2).$$

Therefore, K_2 is a subsolution of (2.22). □

The constant K_1 has the following interpretation. The constant investment control $u_t = 1\ \forall t$ (no wealth in the riskless asset) is suboptimal. The solution to the optimal consumption problem with this special choice for u_t has value function $\gamma^{-1}K_1 x^\gamma$. Condition (4.4) is equivalent to $K_1 > 0$.

A formal asymptotic analysis suggests (but does not prove) that $Z(r)$ in (2.20) grows quadratically as $|r| \to \infty$. With this in mind, we next seek a quadratic supersolution $\bar{Z}(r)$ of the form (4.13), where the constants a_1 and a_2 are to be suitably chosen. The bounds (4.8) on the risk sensitivity parameter γ and the lower bound (4.14) on the discount factor β give sufficient conditions that such a supersolution $\bar{Z}(r)$ exists. Later in the section, further restrictions on a_1, a_2, and β will be imposed to ensure that the solution $\tilde{V}(x,r)$ to the DPE obtained by the subsolution–supersolution method is indeed the value function $V(x,r)$. See Theorem 4.13.

LEMMA 4.2. *Define*

$$\gamma_1 \equiv \frac{\sigma_1^2 c_1^2}{\sigma_1^2 c_1^2 + \sigma_2^2 - 2c_1\rho\sigma_1\sigma_2}. \tag{4.7}$$

Assume that

$$0 < \gamma < \min\{1, \gamma_1\}. \tag{4.8}$$

In addition, define

$$\mu_1 \equiv -2\sigma_2^2 \left[1 + \frac{\gamma\rho^2}{1-\gamma}\right], \tag{4.9}$$

$$\mu_2 \equiv 2c_1 + \frac{2\gamma\rho\sigma_2}{\sigma_1(1-\gamma)}, \tag{4.10}$$

$$\mu_3 \equiv -\frac{\gamma}{2\sigma_1^2(1-\gamma)}. \tag{4.11}$$

Let a^+, a^- be the real roots of $\mu_1 a^2 + \mu_2 a + \mu_3 = 0$. Then we have

$$0 < a^- < a^+. \tag{4.12}$$

Moreover, for any $a_1 \in I_1 \equiv (a^-, a^+)$, there exist constants $a_2 > K_1$ and $C_1(a_1)$, where K_1 is given by (4.5) *and $C_1(\cdot)$ are given by* (4.20)*, such that*

$$\bar{Z}(r) \equiv a_1 r^2 + a_2 \tag{4.13}$$

is a supersolution of (2.22)*, provided that*

$$\beta > -C_1(a_1). \tag{4.14}$$

Proof. Since $|\rho| \le 1$, by (4.7) we can get $\gamma_1 > 0$. Moreover, under condition (4.8), it is not hard to verify that (4.12) holds.

On the other hand, for $\bar{Z}(r)$ defined by (4.13), it is easy to verify that

$$\bar{Z}_r = 2a_1 r, \quad \bar{Z}_{rr} = 2a_1.$$

Then we have

$$-L\bar{Z} = -2a_1^2\sigma_2^2\left[1+\frac{\gamma\rho^2}{1-\gamma}\right]r^2 - 2a_1 f(r)r - \frac{2a_1\gamma\rho\sigma_2}{\sigma_1(1-\gamma)}(b-r)r - a_1\sigma_2^2.$$

By virtue of (2.9), there exists a $\xi \in [0,r]$ such that

$$\begin{aligned} -2a_1 r f(r) &= -2a_1 r[f(0) + f_r(\xi)r] \\ &= -2a_1 f_r(\xi)r^2 - 2a_1 f(0)r \\ &\geq 2c_1 a_1 r^2 - 2a_1 f(0)r. \end{aligned}$$

Therefore, we have

$$\begin{aligned} -L\bar{Z} \geq &\left[2a_1\left(c_1 + \frac{\gamma\rho\sigma_2}{\sigma_1(1-\gamma)}\right) - 2a_1^2\sigma_2^2\left(1+\frac{\gamma\rho^2}{1-\gamma}\right)\right]r^2 \\ &-2a_1\left[f(0) + \frac{\gamma\rho\sigma_2 b}{\sigma_1(1-\gamma)}\right]r - a_1\sigma_2^2. \end{aligned} \tag{4.15}$$

To ensure that $\bar{Z}(r)$ is a supersolution of (2.22), we need only show that

$$-L\bar{Z} \geq h(r,\bar{Z}). \tag{4.16}$$

Define

$$\lambda_1(a_1) \equiv \mu_1 a_1^2 + \mu_2 a_1 + \mu_3, \tag{4.17}$$

$$\lambda_2(a_1) \equiv -\left[2f(0) + \frac{2\gamma\rho\sigma_2}{\sigma_1(1-\gamma)}\right]a_1 + \frac{b\gamma}{\sigma_1^2(1-\gamma)} - \gamma, \tag{4.18}$$

$$\lambda_3(a_1) \equiv -a_1\sigma_2^2 - \frac{\gamma b^2}{2\sigma_1^2(1-\gamma)}, \tag{4.19}$$

$$C_1(a_1) \equiv \frac{4\lambda_1(a_1)\lambda_3(a_1) - \lambda_2^2(a_1)}{4\lambda_1(a_1)}, \tag{4.20}$$

where μ_1, μ_2, μ_3 are given by (4.9)–(4.11). Then, by virtue of (4.15), to show (4.16), it is sufficient to show that

$$\lambda_1(a_1)r^2 + \lambda_2(a_1)r + \lambda_3(a_1) \geq -\beta + (1-\gamma)e^{\frac{a_1 r^2 + a_2}{\gamma-1}}. \tag{4.21}$$

A basic calculation implies that $\lambda_1(a_1) > 0$, provided that $a_1 \in I_1$. Then it is not hard to verify that the left-hand side of (4.21) is bounded below by $C_1(a_1)$. From the definition, we know that $C_1(a_1)$ depends only on $a_1, c_1, b, \rho, \sigma_1, \sigma_2, \gamma$, and $f(0)$. Since $0 < \gamma < \min\{\gamma_1, 1\}$ and $a_1 > 0$, we have

$$e^{\frac{a_1 r^2}{\gamma-1}} \leq 1.$$

Thus, if (4.14) holds, we can take $a_2 > K_1$ large enough such that

$$(1-\gamma)e^{\frac{a_2}{\gamma-1}}e^{\frac{a_1 r^2}{\gamma-1}} \leq (1-\gamma)e^{\frac{a_2}{\gamma-1}} \leq \beta - C_1(a_1),$$

which implies (4.21). □

Remark 4.3. From (4.7), we can get that $\gamma_1 \leq 1$ if and only if $\sigma_2 \geq 2c_1\rho\sigma_1$.

We have the following existence results for (2.22).

THEOREM 4.4. *Suppose* (4.4), (4.8), *and* (4.14) *hold. Then* (2.22) *possesses a classical solution* $\tilde{Z}(r)$ *such that*

$$K_1 \le \tilde{Z}(r) \le \bar{Z}(r), \tag{4.22}$$

where K_1 *and* $\bar{Z}$ *are given by* (4.5) *and* (4.13), *respectively. Define*

$$\tilde{V}(x,r) \equiv \frac{1}{\gamma} x^\gamma e^{\tilde{Z}(r)}. \tag{4.23}$$

Then $\tilde{V}(x,r)$ *is a classical solution of* (2.14).

Proof. It is not hard to verify that $\big(K_1, \bar{Z}(r)\big)$ is an ordered pair of subsolution/supersolution. Then by Lemma 3.9, there exists a classical solution $\tilde{Z}(r)$ of (2.22) such that (4.22) holds. By virtue of (4.23), it is not hard to verify that $\tilde{V}(x,r)$ is a classical solution of (2.14). □

Now we need to verify that $\tilde{V}(x,r)$ is equal to our value function. This will be done in Theorem 4.13. We will also specify the admissible control space in Definition 4.11. Before we go to the verification theorem, we need some lemmas. In those lemmas, we always suppose that $(r_t, t \ge 0)$ is a solution of (2.3)–(2.4).

LEMMA 4.5. *Suppose* $v(r) \in \mathbf{C}^2(\mathbf{R})$ *is bounded. In addition, suppose* v_r *and* v_{rr} *are all bounded. Then* $\phi(r,T) \equiv E_r e^{\int_0^T v(r_t)dt}$ *is in* $\mathbf{C}^{2,1}(\mathbf{R}, [0,\infty))$ *and it is a classical solution of*

$$\begin{cases} \phi_T = \frac{1}{2}\sigma_2^2 \phi_{rr} + f(r)\phi_r + v(r)\phi, \\ \phi(r,0) = 1. \end{cases} \tag{4.24}$$

The proof is rather standard; see Pang [27, Lemma 1.12] for details.

LEMMA 4.6. *Suppose* $\check{v}(r) \in \mathbf{C}^2(\mathbf{R})$. *In addition, suppose* $\check{v}, \check{v}_r, \check{v}_{rr}$ *are all bounded. Then* $\eta(r,T) \equiv E_r e^{\check{v}(r_T)}$ *is in* $\mathbf{C}^{2,1}(\mathbf{R}, [0,\infty))$ *and it is a classical solution of*

$$\begin{cases} \eta_T = \frac{1}{2}\sigma_2^2 \eta_{rr} + f(r)\eta_r, \\ \eta(r,0) = e^{\check{v}(r)}. \end{cases} \tag{4.25}$$

This is a direct corollary of Theorem 5.6.1 of Friedman [19].

LEMMA 4.7. *Let*

$$\check{Q}(r) \equiv \nu_1 r^2 + \nu_2 r + \nu_3, \tag{4.26}$$

where ν_1, ν_2, *and* ν_3 *are constants and* ν_1 *satisfies*

$$\nu_1 < \frac{c_1^2}{2\sigma_2^2}. \tag{4.27}$$

Define $\check{a}^-, \check{a}^+$ *as the real roots of the equation* $2\sigma_2^2 \check{a}^2 - 2c_1 \check{a} + \nu_1 = 0$. *Then we have*

$$0 < \check{a}^- < \check{a}^+. \tag{4.28}$$

Moreover, suppose $(r_t, t \ge 0)$ *is a solution of* (2.3)–(2.4) *and suppose that*

$$\check{a}^- \le \check{a}_1 \le \check{a}^+. \tag{4.29}$$

Then we have

$$e^{-\beta t}Ee^{\int_0^t \check{Q}(r_s)ds} \le \Lambda \quad \forall t \in [0,T], \tag{4.30}$$

where Λ *is a constant, provided that*

$$\beta > -C_2(\check{a}_1), \tag{4.31}$$

where $C_2(\check{a}_1)$ *is given by* (4.36).

Proof. Define a sequence of functions $\{\check{Q}_M(r), M = 1,2,3,\ldots\}$ such that

$$\check{Q}_M \in \mathbf{C}^\infty; \quad 0 \le \check{Q}_M \le M; \quad \left|\frac{\partial \check{Q}_M(r)}{\partial r}\right| \le \check{M}; \quad \left|\frac{\partial^2 \check{Q}_M(r)}{\partial r^2}\right| \le \check{M};$$
$$\check{Q}_{M_1}(r) \le \check{Q}_{M_2}(r) \le \check{Q}(r), \ M_1 < M_2; \quad \lim_{M\to\infty} \check{Q}_M(r) = \check{Q}(r),$$

where $\check{M}, M_1, M_2$ are constants. Define

$$\psi(r,t) \equiv e^{-\beta t}Ee^{\int_0^t \check{Q}_M(r_s)ds}.$$

Then according to Lemma 4.5, $\psi \in \mathbf{C}^{2,1}(\mathbf{R},[0,\infty))$ and it is a solution of the problem

$$\begin{cases} \frac{\partial \psi}{\partial t} = \frac{\sigma_2^2}{2}\frac{\partial^2 \psi}{\partial r^2} + f(r)\frac{\partial \psi}{\partial r} + [\check{Q}_M(r) - \beta]\psi, \\ \psi(r,0) = 1. \end{cases} \tag{4.32}$$

It is easy to verify that, under condition (4.27), $\check{a}^-$ and $\check{a}^+$ are real, positive numbers. So we can take an $\check{a}_1$ such that (4.29) holds.

Define

$$\check{\lambda}_1(\check{a}_1) = -2\sigma_2^2\check{a}_1^2 + 2c_1\check{a}_1 - \nu_1, \tag{4.33}$$
$$\check{\lambda}_2(\check{a}_1) = -2\check{a}_1 - \nu_2, \tag{4.34}$$
$$\check{\lambda}_3(\check{a}_1) = -\sigma_2^2\check{a}_1 - \nu_3, \tag{4.35}$$
$$C_2(\check{a}_1) = \frac{4\check{\lambda}_1(\check{a}_1)\check{\lambda}_3(\check{a}_1) - \check{\lambda}_2(\check{a}_1)}{4\check{\lambda}_1(\check{a}_1)}. \tag{4.36}$$

Following the same procedure in the proof of Lemma 4.2, it is not hard to verify that, under conditions (4.29) and (4.31), $\bar{\psi}(r) \equiv e^{\check{a}_1 r^2}$ satisfies

$$\begin{cases} \frac{\partial \bar{\psi}}{\partial T} \ge \frac{\sigma_2^2}{2}\frac{\partial^2 \bar{\psi}}{\partial r^2} + f(r)\frac{\partial \bar{\psi}}{\partial r} + [\check{Q}_M(r) - \beta]\bar{\psi}, \\ \bar{\psi}(r,0) \ge 1. \end{cases}$$

Define $\xi(r,T) \equiv \psi(r,T) - \bar{\psi}(r)$. Then it satisfies

$$\begin{cases} \frac{\partial \xi}{\partial T} \le \frac{\sigma_2^2}{2}\frac{\partial^2 \xi}{\partial r^2} + f(r)\frac{\partial \xi}{\partial r} + [\check{Q}_M(r) - \beta]\xi, \\ \xi(r,0) \le 0. \end{cases}$$

Since $\check{Q}_M$ is bounded, there exists a constant $B > 0$ such that $\check{Q}_M(r) - \beta < B$. Define

$$\tilde{\xi}(r,T) \equiv e^{-BT}\xi(r,T), \quad \tilde{Q}_M(r) \equiv \check{Q}_M(r) - \beta - B.$$

Then $\tilde{Q}_M(r) < 0$ and $\tilde{\xi}$ satisfies

$$\begin{cases} \frac{\partial \tilde{\xi}}{\partial T} \le \frac{\sigma_2^2}{2}\frac{\partial^2 \tilde{\xi}}{\partial r^2} + f(r)\frac{\partial \tilde{\xi}}{\partial r} + \tilde{Q}_M(r)\tilde{\xi}, \\ \tilde{\xi}(r,0) \le 0. \end{cases}$$

Since $\check{Q}_M(r)$ is bounded, by definitions of $\psi, \bar{\psi}, \xi$, and $\tilde{\xi}$, we can get

$$\lim_{|r|\to\infty} \tilde{\xi}(r,T) = -\infty.$$

If $\tilde{\xi}(r,T)$ reaches its maximum on $\mathbf{R}\times[0,T_1]$ at a point (r_0,T_0), such that $\tilde{\xi}(r_0,T_0) > 0$, then we must have $T_0 > 0$ and

$$\tilde{\xi}_r(r_0,T_0) = 0, \quad \tilde{\xi}_{rr}(r_0,T_0) \le 0, \quad \tilde{\xi}_T(r_0,T_0) \ge 0.$$

This contradicts

$$\frac{\partial \tilde{\xi}}{\partial T} \le \frac{\sigma_2^2}{2}\frac{\partial^2 \tilde{\xi}}{\partial r^2} + f(r)\frac{\partial \tilde{\xi}}{\partial r} + \tilde{Q}_M(r)\tilde{\xi}.$$

Therefore, we must have

$$\tilde{\xi}(r,T) \le 0 \quad \forall r, T.$$

By definitions of $\tilde{\xi}$ and ξ, we can get

$$\psi(r,T) \le \bar{\psi}(r) \quad \forall r, T.$$

Define $\Lambda \equiv \bar{\psi}(r)$. Then Λ is a constant that does not depend on $M, \check{M}$, or T. Thus, by the monotone convergence theorem, we can get (4.30). □

LEMMA 4.8. *Define*

$$\gamma_2 \equiv \frac{\sigma_1^2 c_1^2}{2\sigma_1^2 c_1^2 + 4\sigma_2^2}, \tag{4.37}$$

and suppose that

$$0 < \gamma < \gamma_2. \tag{4.38}$$

Define

$$I_2 \equiv \left(\frac{c_1}{2\sigma_2^2} - \frac{1}{2\sigma_2^2}\sqrt{c_1^2 - \frac{4\sigma_2^2\gamma}{\sigma_1^2(1-2\gamma)}},\ \frac{c_1}{2\sigma_2^2} + \frac{1}{2\sigma_2^2}\sqrt{c_1^2 - \frac{4\sigma_2^2\gamma}{\sigma_1^2(1-2\gamma)}} \right), \tag{4.39}$$

and assume that

$$\check{a}_2 \in I_2. \tag{4.40}$$

Define

$$\nu_1 = \frac{2\gamma}{\sigma_1^2(1-2\gamma)}, \quad \nu_2 = -\frac{4b\gamma}{\sigma_1^2(1-2\gamma)} + 4\gamma, \quad \nu_3 = \frac{2b^2\gamma}{\sigma_1^2(1-2\gamma)}. \tag{4.41}$$

Then for any

$$\beta > -\frac{1}{2}C_2(\check{a}_2), \tag{4.42}$$

where $C_2(\check{a}_2)$ *is given by* (4.36) *with* ν_1, ν_2, ν_3 *defined above, there is a constant* Λ*, which is independent of* T*, such that*

$$e^{-2\beta T} E_r e^{\int_0^T 4\gamma Q_1(r_t)dt} \leq \Lambda, \tag{4.43}$$

where

$$Q_1(r) \equiv \frac{(b-r)^2}{2\sigma_1^2(1-2\gamma)} + r. \tag{4.44}$$

Proof. This is a direct corollary of Lemma 4.7. □

LEMMA 4.9. *Define*

$$I_3 \equiv \left(0, \frac{c_1}{K\sigma_2^2}\right), \tag{4.45}$$

where $K > 8$ *is a constant. Assume that*

$$a_3 \in I_3. \tag{4.46}$$

Define

$$C_3(a_3) \equiv \frac{Kf(0)^2 a_3}{2K\sigma_2^2 a_3 - 2c_1} - K\sigma_2^2 a_3. \tag{4.47}$$

Then for any

$$\beta > -\frac{1}{2}C_3(a_3), \tag{4.48}$$

there is a constant Λ*, which is independent of* T*, such that*

$$e^{-2\beta T} E_r e^{Ka_3 r_T^2} \leq \Lambda.$$

The proof is almost the same as the proof of Lemma 4.7, and so we omit it here. Refer to Pang [27, Lemma 1.15] for details.

LEMMA 4.10. *Suppose* (4.8) *holds. Define*

$$k_1 \equiv \frac{K\rho\sigma_2}{\sigma_1} - 2c_1\rho^2, \quad k_2 \equiv \frac{K^2(2c_1\rho\sigma_1\sigma_2 - \sigma_2^2)}{\sigma_1^2}, \quad \bar{\nu}_1 \equiv \frac{(2-K)c_1}{k_1},$$

$$\gamma_3 \equiv \frac{\bar{\nu}_1}{1+\bar{\nu}_1}, \tag{4.49}$$

$$\bar{\nu}_2 \equiv \frac{-2(K-2)c_1k_1 + k_2}{2k_1} + \frac{\sqrt{(2(K-2)c_1k_1 - k_2)^2 + 4(4K-4)c_1^2k_1^2}}{2k_1^2},$$

$$\gamma_4 \equiv \frac{\bar{\nu}_2}{1+\bar{\nu}_2}, \tag{4.50}$$

where K *is the constant in Lemma* 4.9*. Then if*

$$k_1 < 0, \quad 0 < \gamma < \gamma_3, \tag{4.51}$$

or

$$k_1 > 0, \quad 0 < \gamma < \gamma_4, \tag{4.52}$$

we have

$$I_1 \cap I_3 \neq \emptyset. \tag{4.53}$$

Proof. It can be verified by virtue of some basic calculations. For details, see the proof of Lemma 1.16 in Pang [27]. □

DEFINITION 4.11. *The admissible control space* Π *is*

$$\Pi \equiv \left\{ (u_t, c_t) : P\left(\int_0^T u_t^2 dt < \infty \right) = 1 \quad \forall T > 0,\ c_t \geq 0 \right\}. \tag{4.54}$$

We have the following lemma.

LEMMA 4.12. *For* $(u_t, c_t) \in \Pi$, *define*

$$Y_t \equiv e^{2\gamma\sigma_1 \int_0^t u_s dw_{1,s} - 2\gamma^2\sigma_1^2 \int_0^t u_s^2 ds}, \tag{4.55}$$

and define τ_R *to be the exit time of* (x_t, r_t) *from the ball* $\{x^2 + r^2 \leq R^2\}$. *Then we have*

$$EY_{T\wedge\tau_R} \leq 1 \quad \forall T > 0. \tag{4.56}$$

Proof. Since $(u_t, c_t) \in \Pi$, we can get that $P(Y_t < \infty) = 1$. Then, by virtue of Ito's rule, we can get

$$Y_{T\wedge\tau_R} = 1 + 2\gamma\sigma_1 \int_0^{T\wedge\tau_R} u_s Y_s dw_{1,s}.$$

Denote $\tau_n \equiv \inf\{t \leq T : \int_0^t u_s^2 Y_s^2 ds \geq n^2\}$. Then it is easy to verify that $Y_{T\wedge\tau_R\wedge\tau_n}$ is a martingale for any $n > 0$, and it satisfies $EY_{T\wedge\tau_R\wedge\tau_n} = 1$. Since Y_t is nonnegative, by virtue of Fatou's lemma, we can get (4.56). □

THEOREM 4.13. *Suppose that* (4.4), (4.8), (4.38) *hold and either* (4.51) *or* (4.52) *holds. In addition, assume*

$$a_1 \in I_1 \cap I_3, \quad a_2 \in I_2, \tag{4.57}$$

and

$$\beta > \max\left\{ -C_1(a_1), -\frac{1}{2}C_2(a_2), -\frac{1}{2}C_3(a_1) \right\}, \tag{4.58}$$

where $C_1(\cdot), C_2(\cdot)$, *and* $C_3(\cdot)$ *are given by* (4.20), (4.36), *and* (4.47), *respectively.*

Define $V(x, r)$ *as in* (2.12) *and define* $\tilde{V}(x, r), \tilde{Z}(r)$ *as in Theorem* 4.4. *Then we have*

$$\tilde{V}(x, r) \equiv V(x, r). \tag{4.59}$$

In addition, $J(x, r, u_\cdot, c_\cdot)$ *reaches its maximum at*

$$u^*(r) = \frac{(b - r)}{\sigma_1^2(1 - \gamma)} + \frac{\rho\sigma_2 \tilde{Z}_r(r)}{\sigma_1(1 - \gamma)}, \quad c^*(r) = e^{\frac{\tilde{Z}(r)}{\gamma - 1}}. \tag{4.60}$$

Proof. For any admissible control $(u_t, c_t) \in \Pi$, denote $\mathcal{G}^{u_t,c_t}$ as the generator of the process (x_t, r_t) under control (u_t, c_t). Then, by virtue of Ito's rule, we can get

$$\begin{aligned} d\left[e^{-\beta t}\tilde{V}(x_t, r_t)\right] &= e^{-\beta t}\left[d\tilde{V}(x_t, r_t) - \beta\tilde{V}(x_t, r_t)dt\right] \\ &= e^{-\beta t}\left[\mathcal{G}^{u_t,c_t}\tilde{V}(x_t, r_t) - \beta\tilde{V}(x_t, r_t)\right]dt + dm_{1,t} + dm_{2,t}, \end{aligned}$$

where $m_{1,t}$ and $m_{2,t}$ are local martingales under P.

Integrate it on $[0, T]$. Since $\tilde{V}$ is a classical solution of (2.14), we have

$$e^{-\beta T}\tilde{V}(x_T, r_T) - \tilde{V}(x, r) \leq -\int_0^T e^{-\beta t}\frac{1}{\gamma}(c_t x_t)^\gamma dt + m_{1,T} + m_{2,T}.$$

Let τ_R define the exit time of (x_t, r_t) from the ball $\{x^2 + r^2 < R^2\}$. Then, for every finite T, we have

$$\tilde{V}(x, r) \geq E\int_0^{T\wedge\tau_R} e^{-\beta t}\frac{1}{\gamma}(c_t x_t)^\gamma dt + E\left[e^{-\beta T\wedge\tau_R}\tilde{V}(x_{T\wedge\tau_R}, r_{T\wedge\tau_R})\right]. \tag{4.61}$$

Noting $\tilde{V} > 0$, by virtue of Fatou's lemma as $R \to \infty$, we can take $\liminf$ to get

$$\tilde{V}(x, r) \geq E\int_0^T e^{-\beta t}\frac{1}{\gamma}(c_t x_t)^\gamma dt.$$

Now, let $T \to \infty$; then we have

$$\tilde{V}(x, r) \geq E\int_0^\infty e^{-\beta t}\frac{1}{\gamma}(c_t x_t)^\gamma dt,$$

which holds for any admissible control $(u_t, c_t) \in \Pi$. By the definition of $V(x, r)$, we must have, for any r,

$$\tilde{V}(x, r) \geq V(x, r). \tag{4.62}$$

On the other hand, for control (u^*, c^*) defined by (4.60), it is not hard to verify that $(u_t^*, c_t^*) \in \Pi$. Then, instead of (4.61), we can get

$$\tilde{V}(x, r) = E\int_0^{T\wedge\tau_R} e^{-\beta t}\frac{1}{\gamma}(c_t^* x_t)^\gamma dt + E\left[e^{-\beta T\wedge\tau_R}\tilde{V}(x_{T\wedge\tau_R}, r_{T\wedge\tau_R})\right]. \tag{4.63}$$

Using the monotone convergence theorem, for any fixed $T > 0$, we can show that

$$\lim_{R\to\infty} E\int_0^{T\wedge\tau_R} e^{-\beta t}\frac{1}{\gamma}(c_t x_t)^\gamma dt = E\int_0^T e^{-\beta t}\frac{1}{\gamma}(c_t x_t)^\gamma dt. \tag{4.64}$$

Define $\tilde{W}(r) \equiv e^{\tilde{Z}(r)}$. Then by virtue of (4.23), we can rewrite $\tilde{V}(x, r)$ as

$$\tilde{V}(x, r) \equiv \frac{1}{\gamma}x^\gamma\tilde{W}(r). \tag{4.65}$$

Define

$$\tilde{l}(r, u) = r + (b - r)u - \frac{1}{2}\sigma_1^2 u^2.$$

Then for any admissible control $(u_t, c_t) \in \Pi$, using Ito's formula, we can get

$$x_t = x \exp\left\{ \int_0^t [\tilde{l}(r_s, u_s) - c_s]ds + \int_0^t \sigma_1 u_s dw_{1,s} \right\}.$$

Given the above equality, by virtue of the Cauchy–Schwarz inequality, we can get

$$\begin{aligned}
&E\left[e^{-\beta T\wedge\tau_R}\tilde{V}(x_{T\wedge\tau_R}, r_{T\wedge\tau_R})\right] \\
&\quad = \frac{1}{\gamma}E\left[e^{-\beta T\wedge\tau_R}x^{\gamma}_{T\wedge\tau_R}\tilde{W}(r_{T\wedge\tau_R})\right] \\
&\quad \le \frac{1}{\gamma}x^{\gamma}\left(Ee^{\int_0^{T\wedge\tau_R} 2[\gamma l(r_s,u_s)-\gamma c_s-\beta]ds}\tilde{W}^2(r_{T\wedge\tau_R})\right)^{\frac{1}{2}} \\
&\qquad \cdot\left(Ee^{\int_0^{T\wedge\tau_R}[2\gamma\sigma_1 u_t dw_{1,t}-2\gamma^2\sigma_1^2 u_t^2 dt]}\right)^{\frac{1}{2}},
\end{aligned}$$

where

$$l(r,u) \equiv r + (b-r)u + \left(\gamma - \frac{1}{2}\right)\sigma_1^2 u^2. \tag{4.66}$$

Using the result of Lemma 4.12, we have

$$Ee^{\int_0^{T\wedge\tau_R}[2\gamma\sigma_1 u_t dw_{1,t}-2\gamma^2\sigma_1^2 u_t^2 dt]} \le 1. \tag{4.67}$$

Thus, we can get

$$\begin{aligned}
\tilde{V}(x,r) &= E\int_0^{T\wedge\tau_R} e^{-\beta t}\frac{1}{\gamma}(c_t x_t)^{\gamma}dt \\
&\quad + \frac{1}{\gamma}x^{\gamma}\left(Ee^{\int_0^{T\wedge\tau_R} 2[\gamma l(r_s,u_s)-\gamma c_s-\beta]ds}\tilde{W}^2(r_{T\wedge\tau_R})\right)^{\frac{1}{2}}.
\end{aligned} \tag{4.68}$$

From (4.37), we can get that $\gamma_2 < \frac{1}{2}$. Since $0 < \gamma < \gamma_2 < \frac{1}{2}$, by virtue of (4.44) and (4.66), we can get

$$l(r,u) \le Q_1(r).$$

Therefore, for $0 < \gamma < \gamma_2$, $\gamma Q_1(r) - \beta$ is lower bounded. Choose B such that $\gamma Q_1(r) - \beta - B \ge 0$. Noting that $c^* > 0, \tilde{W}(r) \le e^{a_1 r^2 + a_2}$ and using the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\left[e^{\int_0^{T\wedge\tau_R} 2[\gamma l(r_s,u_s^*)-\gamma c_s^*-\beta]ds}\tilde{W}^2(r_{T\wedge\tau_R})\right] \\
&\le \left[e^{\int_0^{T\wedge\tau_R} 2[\gamma Q_1(r_s)-\beta]ds}\tilde{W}^2(r_{T\wedge\tau_R})\right] \\
&= e^{2B(T\wedge\tau_R)}\left[e^{\int_0^{T\wedge\tau_R} 2[\gamma Q_1(r_s)-\beta-B]ds}\sup_{0\le t\le T}\tilde{W}^2(r_t)\right] \\
&\le e^{2B(T\wedge\tau_R)}\left[e^{\int_0^{T} 2[\gamma Q_1(r_s)-\beta-B]ds}\sup_{0\le t\le T}\tilde{W}^2(r_t)\right]
\end{aligned}$$

$$
\begin{aligned}
&= e^{2B(T\wedge\tau_R - T)}\left[e^{\int_0^T 2[\gamma Q_1(r_s)-\beta]ds}\sup_{0\le t\le T}\tilde{W}^2(r_t)\right]\\
&\le e^{2|B|T}\left[e^{\int_0^T 2[\gamma Q_1(r_s)-\beta]ds}\sup_{0\le t\le T}\tilde{W}^2(r_t)\right]\\
&\le \frac{1}{2}e^{2(|B|-\beta)T}\left[e^{\int_0^T 4\gamma Q_1(r_s)ds} + \sup_{0\le t\le T}\tilde{W}^4(r_t)\right]\\
&\le \frac{1}{2}e^{2(|B|-\beta)T}\left[e^{\int_0^T 4\gamma Q_1(r_s)ds} + A^2\sup_{0\le t\le T}e^{4a_1 r_t^2}\right]\\
&\equiv \eta_T,
\end{aligned}
$$

where $A \equiv e^{2a_2}$. Next, we are going to show that

$$E_r\eta_T < \infty, \tag{4.69}$$

which ensures that we can use Fatou's lemma in (4.68) to get rid of the stopping time.

First, by virtue of Lemma 4.8, we have

$$E_r e^{\int_0^T 4\gamma Q_1(r_s)ds} < \infty \tag{4.70}$$

provided that $0 < \gamma < \gamma_2$. Define $\psi(r_t) \equiv e^{4a_1 r_t^2}$. Then we also need to show that

$$E_r\left[\sup_{0\le t\le T}\psi(r_t)\right] < \infty. \tag{4.71}$$

By virtue of Ito's lemma, we have

$$
\begin{aligned}
d\psi(r_t) &= 8a_1 r_t\psi(r_t)[f(r_t)dt + \sigma_2 dw_{2,t}] + \frac{\sigma_2^2}{2}[8a_1\psi(r_t) + 64a_1^2 r_t^2\psi(r_t)]dt\\
&\le [8a_1(4a_1\sigma_2^2 - c_1)r_t^2 + 8a_1 f(0)r_t + 4a_1\sigma_2^2]\psi(r_t)dt + dm_t,
\end{aligned}
$$

where

$$m_t \equiv 8a_1\sigma_2\int_0^t r_s\psi(r_s)dw_{2,s}. \tag{4.72}$$

By the definition of a_1, we know that $4a_1\sigma_2^2 - c_1 < 0$. Therefore, $8a_1(4a_1\sigma_2^2 - c_1)r_t^2 + 8a_1 f(0)r_t + 4a_1\sigma_2^2$ is upper bounded. Suppose $N > 0$ is an upper bound; then we have

$$\psi(r_t) \le \psi(r) + \int_0^t N\psi(r_s)ds + m_t. \tag{4.73}$$

By the definition of m_t, we have

$$E_r m_t^2 = 64a_1^2\sigma_2^2 E\int_0^t r_s^2\psi^2(r_s)ds.$$

For any $s \in [0,t]$, by virtue of Lemma 4.9, we have

$$E_r\left[r_s^2\psi^2(r_s)\right] = E_r\left[r_s^2 e^{8a_1 r_s^2}\right] \le \Lambda_1[E_r e^{(8+\epsilon)a_1 r_s^2}] < \infty, \tag{4.74}$$

where $\Lambda_1 > 0$ is a constant and ϵ can be any positive number. Therefore, we must have $E_r m_t^2 < \infty$. So m_t is a martingale. Using fundamental martingale inequalities, we can show that

$$E_r\left[\sup_{0\le t\le T} m_t^2\right] \le 4E_r m_T^2 \le \Lambda_2 < \infty, \tag{4.75}$$

where Λ_1 is a positive constant. Given this and (4.73), using the Chebyshev inequality, we can show that

$$E_r\left[\sup_{0\le t\le T}\psi(r_t)\right] \le \psi(r) + \Lambda_2 + \int_0^T NE_r\left[\sup_{0\le s\le t}\psi(r_s)\right]dt. \tag{4.76}$$

Then, by virtue of Gronwall's inequality, it is easy to get (4.71). Combined with (4.70), this implies (4.69). Now, when we let $R\to\infty$ and take lim sup in (4.68), we can use the monotone convergence theorem and Fatou's lemma to get

$$\begin{aligned}\tilde V(x,r) \le{}& E\int_0^T e^{-\beta t}\frac{1}{\gamma}(c_t^* x_t)^\gamma dt \\ &+ \frac{1}{\gamma}x^\gamma\left(E\left[e^{\int_0^T 2[\gamma Q_1(r_s)-\gamma c_s^*-\beta]ds}\tilde W^2(r_T)\right]\right)^{\frac12}.\end{aligned} \tag{4.77}$$

Denote $\delta = \beta - \max\{-C_1(a_1), C_2(\tilde a), C_3(a_1)\}$. Then, from (4.58), we can get that $\delta > 0$. Take $\check\beta = \beta - \frac{\delta}{2}$. Then, by virtue of $c^* \ge 0$, using the Cauchy–Schwarz inequality, we can get

$$\begin{aligned}&E_r\left[e^{\int_0^T 2[\gamma Q_1(r_s)-\gamma c_s^*-\beta]ds}\tilde W^2(r_T)\right]\\ &< e^{-2\beta T}\left[E_r e^{\int_0^T 4\gamma Q_1(r_s)ds}\right]^{\frac12}\cdot\left[E_r\tilde W^4(r_T)\right]^{\frac12}\\ &\le e^{-\delta T}e^{4a_2}\left[e^{-2\check\beta T}E_r e^{\int_0^T 4\gamma Q_1(r_s)ds}\right]^{\frac12}\left[e^{-2\check\beta T}E_r e^{4a_1 r_T^2}\right]^{\frac12}.\end{aligned}$$

Given the above inequality, by virtue of Lemmas 4.8 and 4.9, we can show that, for any fixed $T > 0$,

$$\lim_{T\to\infty} E_r\left[e^{\int_0^T 2[\gamma Q(r_s)-\gamma c_s^*-\beta]ds}\tilde W^2(r_T)\right] = 0.$$

Now in (4.77), let $T\to\infty$; then we have

$$\tilde V(x,r) \le E\int_0^\infty e^{-\beta t}\frac{1}{\gamma}(c_t^* x_t)^\gamma dt \le V(x,r). \tag{4.78}$$

Combined with (4.62), this implies

$$\tilde V(x,r) = E\int_0^\infty e^{-\beta t}\frac{1}{\gamma}(c_t^* x_t)^\gamma dt = V(x,r).$$

Thus, (u^*, c^*) is optimal and $\tilde V(x,r)\equiv V(x,r)$. □

5. $\gamma < 0$ case. In this section, we will investigate the $\gamma < 0$ case. The existence results will be given in Theorem 5.4 and the verification results will be given in Theorem 5.7. The admissible control space will be specified in Definition 5.6.

Using the same notations as in the last section, we can write the equation of $Z(r)$ as

$$-LZ = h(r, Z). \tag{5.1}$$

For the $\gamma < 0$ case, we have the following results. A formal asymptotic analysis suggests that, for $\gamma < 0$, $Z(r)$ in (2.20) behaves like $2(\gamma - 1)\log r$ as $|r| \to \infty$. This leads to the choice of $\hat{Z}(r)$ in (5.3) and the choice of $\bar{Z}(r)$ in (5.7). Condition (5.6) is sufficient for the existence of a supersolution of the form (5.7).

LEMMA 5.1. *Suppose $\gamma < 0$. Define*

$$a_1 \equiv \frac{-2\gamma}{3\sigma_1^2(1-\gamma)^2}, \qquad a_2 \equiv b - \sigma_1^2(1-\gamma). \tag{5.2}$$

Then there exists a constant $\bar{a}_3 > 0$ such that for any $a_3 \geq \bar{a}_3$

$$\hat{Z}(r) \equiv \log\left[(a_1(r-a_2)^2 + a_3)^{\gamma-1}\right] \tag{5.3}$$

is a subsolution of (5.1).

It can be verified by virtue of direct calculations. See Pang [27, Lemma 1.18] for details.

LEMMA 5.2. *Suppose $\gamma < 0$. Define*

$$b_1 \equiv \frac{-\gamma}{2\sigma_1^2(1-\gamma)^2}, \qquad b_2 \equiv b - \sigma_1^2(1-\gamma), \tag{5.4}$$

$$b_3 \equiv b_1 \frac{2\sigma_2^2[\frac{3}{2} - \gamma + \gamma\rho^2] - 2\rho\sigma_1^3\sigma_2\gamma(1-\gamma) + |f(b_2)|}{2c_2 + |f(b_2)|}. \tag{5.5}$$

If

$$\begin{aligned}\beta \geq b\gamma + (1-\gamma)\left[2c_2|f(b_2)| - \frac{\sigma_1^2\gamma}{2}\right] \\ - \frac{2\gamma\sigma_2^2[\frac{3}{2} - \gamma + \gamma\rho^2] - 2\rho\sigma_1^3\sigma_2\gamma^2(1-\gamma) + \gamma|f(b_2)|}{2\sigma_1^2(1-\gamma)[2c_2 + |f(b_2)|]},\end{aligned} \tag{5.6}$$

then

$$\bar{Z}(r) \equiv \log\left[(b_1(r-b_2)^2 + b_2)^{\gamma-1}\right] \tag{5.7}$$

is a supersolution of (5.1).

The proof involves a lot of calculations. The techniques used in the proof are very similar to those used in the proof of Lemma 4.2. See Pang [27, Lemma 1.19] for details.

Remark 5.3. From (5.2), (5.3), (5.6), and (5.7) we can see that

$$a_1 > b_1, \quad a_2 = b_2. \tag{5.8}$$

In addition, we can take a_3 large enough such that

$$a_3 > b_3. \tag{5.9}$$

Then

$$\bar{Z}(r) > \hat{Z}(r) \quad \forall r. \tag{5.10}$$

Given the above results, we can get the following theorem.

THEOREM 5.4. *Suppose* $\gamma < 0$ *and* (5.6) *holds. Then* (5.1) *possesses a classical solution* $\tilde{Z}(r)$ *such that*

$$\hat{Z}(r) \le \tilde{Z}(r) \le \bar{Z}(r), \tag{5.11}$$

where $\hat{Z}(r)$ *and* $\bar{Z}(r)$ *are given by* (5.3) *and* (5.7)*, respectively. Define*

$$\tilde{V}(x,r) \equiv \frac{1}{\gamma} x^\gamma e^{\tilde{Z}(r)}. \tag{5.12}$$

Then $\tilde{V}(x,r)$ *is a classical solution of* (2.14)*, and it satisfies*

$$\frac{1}{\gamma} x^\gamma e^{\hat{Z}(r)} \le \tilde{V}(x,r) \le \frac{1}{\gamma} x^\gamma e^{\bar{Z}(r)}. \tag{5.13}$$

The proof is almost the same as the proof of Theorem 4.4, and we omit it here.

LEMMA 5.5. *Suppose* $\gamma < 0$*. Let* $\tilde{Z}(r)$ *be a solution of* (5.1) *which satisfies* (5.11)*. Then we have*

$$\lim_{|r|\to\infty} \tilde{Z}_r(r) = 0. \tag{5.14}$$

Proof. By the definitions of $\bar{Z}$ and $\hat{Z}$, we can get

$$(\gamma-1)\log(a_1(r-a_2)^2 + a_3) \le \tilde{Z}(r) \le (\gamma-1)\log(b_1(r-b_2)^2 + b_3). \tag{5.15}$$

The above inequality implies

$$\liminf_{|r|\to\infty} |\tilde{Z}_r(r)| = 0. \tag{5.16}$$

Otherwise, $\tilde{Z}(r)$ will have at least a linear growth as $|r| \to \infty$, which contradicts (5.15).

If (5.14) does not hold, $\tilde{Z}_r(r)$ must have a sequence of either positive local maxima or negative local minima at points $\{r_m, m = 1, 2, 3, \ldots\}$, which tend to either $+\infty$ or $-\infty$ with the following property: there exists $\delta > 0$, such that

$$|\tilde{Z}_r(r_m)| \ge \delta. \tag{5.17}$$

Suppose that $\tilde{Z}_r(r)$ has a positive local maximum at r_m. Since $\tilde{Z}(r) \in \mathbf{C}^2(\mathbf{R})$, by virtue of (5.1), we have that $\tilde{Z}(r) \in \mathbf{C}^3(\mathbf{R})$. Therefore, $\tilde{Z}_{rr}(r_m) = 0$, $\tilde{Z}_{rrr}(r_m) \le 0$. Define

$$\check{\sigma}_2^2 \equiv \sigma_2^2 \left(1 + \frac{\gamma\rho}{1-\gamma}\right), \quad \check{f}(r) \equiv f(r) + \frac{\gamma\rho\sigma_2(b-r)}{\sigma_1(1-\gamma)}, \quad \check{c}_1 \equiv c_1 + \frac{\gamma\rho\sigma_2}{\sigma_1(1-\gamma)}.$$

Noting (2.9), we can get that

$$\check{f}_r(r) \le -\check{c}_1. \tag{5.18}$$

By virtue of (5.1), we can get

$$0 = \frac{\sigma_2^2}{2}\tilde{Z}_{rrr} + \check{\sigma}_2^2\tilde{Z}_r\tilde{Z}_{rr} + \check{f}\tilde{Z}_{rr} + \check{f}_r\tilde{Z}_r + \gamma Q_r - e^{-\frac{\tilde{Z}}{1-\gamma}}\tilde{Z}_r,$$

where $Q(r)$ is defined by (2.19) and Q_r stands for its derivative. Since $\tilde{Z}_{rr}(r_m) = 0$, $\tilde{Z}_{rrr}(r_m) \le 0$, noting that $\tilde{Z} \le \bar{Z}$, we have

$$0 \le (\check{f}_r(r_m) - b_1(r_m - b_2)^2 - b_3)\tilde{Z}_r(r_m) + \gamma\left(1 + \frac{r_m - b}{\sigma_2^2(1-\gamma)}\right). \tag{5.19}$$

If $|r_m|$ is big enough, by virtue of (5.18), we will have

$$-\check{f}_r(r_m) + b_1(r_m - b_2)^2 + b_3 \ge \check{c}_1 + b_1(r_m - b_2)^2 + b_3 > 0.$$

Therefore, by virtue of (5.19), we can get

$$\tilde{Z}_r(r_m) \le \frac{\left|\gamma\left(1 + \frac{r_m - b}{\sigma_2^2(1-\gamma)}\right)\right|}{\check{c}_1 + b_1(r_m - b_2)^2 + b_3}.$$

But the right-hand side of the above inequality goes to 0 as $|r_m|$ goes to $+\infty$, so we must have

$$\lim_{|r_m|\to\infty} \tilde{Z}_r(r_m) = 0. \tag{5.20}$$

This contradicts our assumption (5.17). Similarly, if $\tilde{Z}_r$ has a negative local minimum at $\{r_m, m = 1, 2, 3, \ldots\}$, and we can also get (5.20). Therefore, (5.17) holds. □

Define the admissible control space Π as follows.

DEFINITION 5.6 (admissible control space). *A control* $(u_t, c_t) \in \mathbf{R}^2$ *is in the admissible control space* Π*, if the following hold:*

$$0 \le c_t \le A_1(r_t - A_2)^2 + A_3 \quad \forall t \ge 0, \tag{5.21}$$

$$E\int_0^T u_t^2 dt < \infty \quad \forall T \ge 0, \tag{5.22}$$

$$E\int_0^T e^{-2\beta t}u_t^2 x_t^{2\gamma} dt < \infty \quad \forall T \ge 0, \tag{5.23}$$

where $A_1 > 0, A_3 > 0$*, and* A_2 *are some constants.*

We have the following verification theorem.

THEOREM 5.7. *Suppose* $\gamma < 0$ *and* (5.6) *holds. In addition, assume that*

$$\frac{4\gamma(1+3\gamma)}{\sigma_1^2(1-\gamma)^2} \le \frac{c_1^2}{2\sigma_2^2}. \tag{5.24}$$

Define $V(x, r)$ *as in* (2.12) *and define* $\tilde{V}(x, r)$ *and* $\tilde{Z}(r)$ *as in Theorem* 5.4. *Then we have*

$$\tilde{V}(x, r) \equiv V(x, r). \tag{5.25}$$

In addition, $J(x, r, u_\cdot, c_\cdot)$ *reaches its maximum at*

$$u^*(r) = \frac{(b-r)}{\sigma_1^2(1-\gamma)} + \frac{\rho\sigma_2\tilde{Z}_r(r)}{\sigma_1(1-\gamma)}, \quad c^*(r) = e^{\frac{\tilde{Z}(r)}{\gamma-1}}. \tag{5.26}$$

Proof. For any admissible control $(u_t, c_t) \in \Pi$, denote $\mathcal{G}^{u_t,c_t}$ as the generator of the process (x_t, r_t) under control (u_t, c_t). Then, by Ito's rule, we can get

$$d\left[e^{-\beta t}\tilde{V}(x_t, r_t)\right] = e^{-\beta t}\left[\mathcal{G}^{u_t,c_t}\tilde{V}(x_t, r_t) - \beta\tilde{V}(x_t, r_t)\right]dt + dm_{1,t} + dm_{2,t}, \tag{5.27}$$

where

$$m_{1,t} \equiv \int_0^t e^{-\beta s}\sigma_1 u_s x_s^\gamma \tilde{W}(r_s)dw_{1,s}, \quad m_{2,t} \equiv \frac{1}{\gamma}\int_0^t e^{-\beta s}\sigma_2 x_s \gamma \tilde{W}_r(r_s)d\tilde{w}_s,$$

and $\tilde{W}(r)$ is defined by $\tilde{W}(r) \equiv e^{\tilde{Z}(r)}$. It is not hard to verify that $\tilde{W}(r)$ is a classical solution of (2.18).

From the definition of $\tilde{W}(r)$, we know that $e^{\hat{Z}(r)} \le \tilde{W}(r) \le e^{\bar{Z}(r)}$. Thus, by virtue of the definitions of $\hat{Z}(r)$ and $\bar{Z}(r)$, we can get that $\tilde{W}(r)$ is bounded. In addition, from Lemma 5.5, we know that $\tilde{Z}_r(r)$ is bounded. Since $\tilde{W}_r(r) = \tilde{W}(r)Z_r(r)$, $\tilde{W}_r(r)$ is also bounded. Therefore, it is not hard to show that $m_{1,t}, m_{2,t}$ are both martingales.

Now integrate (5.27) on $[0, T]$. Since $\tilde{W}(r)$ is a classical solution of (2.18), it is not hard to verify that $\tilde{V}(x, r)$ is a classical solution of (2.14). Then we have

$$e^{-\beta T}\tilde{V}(x_T, r_T) - \tilde{V}(x, r) \le -\int_0^T e^{-\beta t}\frac{1}{\gamma}(c_t x_t)^\gamma dt + m_{1,T} + m_{2,T}.$$

Take expectation for both sides, and we can get

$$\begin{aligned}\tilde{V}(x, r) &\ge E\int_0^T e^{-\beta t}\frac{1}{\gamma}(c_t x_t)^\gamma dt + E\left[e^{-\beta T}\tilde{V}(x_T, r_T)\right] \\ &= E\int_0^T e^{-\beta t}\frac{1}{\gamma}(c_t x_t)^\gamma dt + \frac{1}{\gamma}E\left[e^{-\beta T}x_T^\gamma \tilde{W}(r_T)\right].\end{aligned} \tag{5.28}$$

If $J(x, r, u_\cdot, c_\cdot) = -\infty$, then we must have

$$\tilde{V}(x, r) \ge J(x, r, u_\cdot, c_\cdot). \tag{5.29}$$

Otherwise, if $J(x, r, u_\cdot, c_\cdot) > -\infty$, i.e.,

$$\int_0^\infty E\left[e^{-\beta t}c_t^\gamma x_t^\gamma\right]dt < \infty, \tag{5.30}$$

we must have

$$\liminf_{T\to\infty} E\left[e^{-\beta T}c_T^\gamma x_T^\gamma\right] = 0. \tag{5.31}$$

In addition, it not hard to find a constant Λ such that

$$\Lambda[b_1(r - b_2)^2 + b_3] \ge \left[A_1(r - A_2)^2 + A_3\right].$$

Therefore, since $\gamma < 0$, by virtue of (5.21), we can get

$$\begin{aligned}c_t^\gamma &\ge \left[A_1(r_t - A_2)^2 + A_3\right]^\gamma \\ &\ge \Lambda^\gamma\left[b_1(r_t - b_2)^2 + b_3\right]^\gamma \\ &\ge b_3\Lambda^\gamma b_3^{-1}\left[b_1(r_t - b_2)^2 + b_3\right]^\gamma \\ &\ge b_3\Lambda^\gamma\left[b_1(r_t - b_2)^2 + b_3\right]^{\gamma-1} \\ &\ge b_3\Lambda^\gamma\tilde{W}(r_t).\end{aligned}$$

Combined with (5.31), this implies

$$\liminf_{T\to\infty} E\left[e^{-\beta T}x_T^{\gamma}\tilde{W}(r_T)\right]=0. \tag{5.32}$$

Then, let $T\to\infty$ in (5.28) and take lim inf, and we can get

$$\tilde{V}(x,r)\geq J(x,r,u_\cdot,c_\cdot). \tag{5.33}$$

On the other hand, for u_t^*, c_t^* defined by (5.26), since $\tilde{Z}_r(r)$ is bounded, it is not hard to verify that

$$0\leq c_t^*\leq a_1(r_t-a_2)^2+a_3 \quad \forall t\geq 0, \tag{5.34}$$

$$E\int_0^T (u_t^*)^2 dt<\infty \quad \forall T\geq 0. \tag{5.35}$$

So (5.21), (5.22) hold if we take $A_1\geq a_1, A_2=a_2$, and $A_3\geq a_3$. Thus, to ensure that $(u^*,c^*)\in\Pi$, we need to show that (5.23) holds for u_t^*, c_t^*. By Ito's rule, define $\tilde{l}(r,u)\equiv r+(b-r)u-\frac{1}{2}\sigma_1^2u^2$. Then using Ito's rule, we can get

$$x_t=x\exp\left\{\int_0^t[\tilde{l}(r_s,u_s^*)-c_s^*]ds+\int_0^t\sigma_1u_s^*dw_{1,s}\right\}.$$

It is not hard to verify that $e^{\int_0^t[4\gamma\sigma_1u_s^*dw_{1,s}-8\gamma^2\sigma_1^2(u_s^*)^2ds]}$ is a positive supermartingale which satisfies

$$Ee^{\int_0^t[4\gamma\sigma_1u_s^*dw_{1,s}-8\gamma^2\sigma_1^2(u_s^*)^2ds]}\leq 1.$$

Given the above equality and by virtue of the Cauchy–Schwarz inequality, we can get

$$\begin{aligned}
&E\left[e^{-2\beta t}(u_t^*)^2x_t^{2\gamma}\right]\\
&\leq x^{2\gamma}e^{-2\beta t}\left(E\left[(u_t^*)^4e^{\int_0^t 4\gamma[l(r_s,u_s^*)-c_s^*]ds}\right]\right)^{\frac{1}{2}}\cdot\left(Ee^{\int_0^t[4\gamma\sigma_1u_s^*dw_{1,s}-8\gamma^2\sigma_1^2(u_s^*)^2ds]}\right)^{\frac{1}{2}}\\
&\leq x^{2\gamma}e^{-2\beta t}\left(E\left[(u_t^*)^4e^{\int_0^t 4\gamma[l(r_s,u_s^*)-c_s^*]ds}\right]\right)^{\frac{1}{2}}\\
&\leq x^{2\gamma}e^{-2\beta t}\left(E[(u_t^*)^8]\right)^{\frac{1}{4}}\cdot\left(Ee^{\int_0^t 8\gamma[l(r_s,u_s^*)-c_s^*]ds}\right)^{\frac{1}{4}},
\end{aligned}$$

where

$$l(r,u)\equiv r+(b-r)u+\left(2\gamma-\frac{1}{2}\right)\sigma_1^2u^2. \tag{5.36}$$

Since $\tilde{Z}_r$ is bounded, using the Cauchy–Schwarz inequality, we can get

$$l(r_s,u_s^*)\geq\Lambda_1+r_s+\frac{(b-r_s)^2}{2\sigma_1^2(1-\gamma)^2}\left(1+\frac{5}{2}\gamma\right). \tag{5.37}$$

In addition, by virtue of (5.34), (5.2), we can get

$$8\gamma[l(r_s,u_s^*)-c_s^*]\leq 8\gamma\left[\Lambda_1+\Lambda_2r_s+\frac{(1+3\gamma)r_s^2}{2\sigma_1^2(1-\gamma)^2}\right]. \tag{5.38}$$

By Lemma 4.7, we can get that if (5.24) holds, then

$$Ee^{\int_0^t 8\gamma[l(r_s,u_s^*)-c_s^*]ds} \le \Lambda(T) < \infty \quad \forall t \in [0,T]. \tag{5.39}$$

In addition, since $\tilde{Z}_r(r)$ is bounded, we can get

$$E\left[(u_t^*)^8\right] \le \Lambda(T) < \infty \quad \forall t \in [0,T]. \tag{5.40}$$

Therefore, we now have

$$E\left[e^{-2\beta t}(u_t^*)^2 x_t^{2\gamma}\right] \le \Lambda(T) \quad \forall t \in [0,T], \tag{5.41}$$

which implies (5.23). Thus, we have shown that $(u_t^*, c_t^*) \in \Pi$. Given this, instead of (5.28), we can now get

$$\tilde{V}(x,r) = E\int_0^T e^{-\beta t}\frac{1}{\gamma}(c_t^* x_t)^\gamma dt + E\left[e^{-\beta T}\tilde{V}(x_T, r_T)\right]. \tag{5.42}$$

Since $\tilde{V}(x_T, r_T) \le 0$, we can get

$$\tilde{V}(x,r) \le E\int_0^T e^{-\beta t}\frac{1}{\gamma}(c_t^* x_t)^\gamma dt.$$

Let T go to $+\infty$; then we have

$$\tilde{V}(x,r) \le E\int_0^\infty e^{-\beta t}\frac{1}{\gamma}(c_t^* x_t)^\gamma dt,$$

i.e.,

$$\tilde{V}(x,r) \le J(x,r,u_\cdot^*, c_\cdot^*). \tag{5.43}$$

This completes the proof. □

REFERENCES

[1] A. Bensoussan and J. Frehse, *Regularity Results for Nonlinear Elliptic Systems and Applications*, Springer, New York, 2002.

[2] T. R. Bielecki and S. R. Pliska, *Risk sensitive dynamic asset management*, Appl. Math. Optim., 37 (1997), pp. 337–360.

[3] P. B. Bailey, L. F. Shampine, and P. E. Waltman, *Nonlinear Two Point Boundary Value Problems*, Academic Press, New York, 1968.

[4] D. Bainov and P. Simeonov, *Integral Inequalities and Applications*, Kluwer, Norwell, MA, 1992.

[5] V. S. Borkar, *Optimal Control of Diffusion Processes*, Longman Scientific and Technical, Essex, UK, 1989.

[6] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. AMS, 27 (1992), pp. 1–67.

[7] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[8] B. G. Fitzpatrick and W. H. Fleming, *Numerical methods for an optimal investment-consumption model*, Math. Oper. Res., 16 (1991), pp. 823–841.

[9] W. H. Fleming, *Optimal investment models and risk-sensitive stochastic control*, in Mathematical Finance, IMA Vol. Math. Appl. 65, Springer, New York, 1995, pp. 75–88.

[10] W. H. Fleming and D. Hernandez-Hernandez, *An optimal consumption model with stochastic volatility*, Finance Stoch., 7 (2003), pp. 245–262.

[11] W. H. Fleming and W. M. McEneaney, *Risk-sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.
[12] W. H. Fleming and T. Pang, *A Stochastic Control Model of Investment, Production and Consumption*, preprint, 2003.
[13] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer, New York, 1975.
[14] W. H. Fleming and S. J. Sheu, *Optimal long term growth rate of expected utility of wealth*, Ann. Appl. Probab., 9 (1999), pp. 871–903.
[15] W. H. Fleming and S. J. Sheu, *Risk-sensitive control and optimal investment model*, Math. Finance, 10 (2000), pp. 197–213.
[16] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer, New York, 1992.
[17] W. H. Fleming and J. L. Stein, *Stochastic inter-temporal optimization in discrete time*, in Economic Theory, Dynamics and Markets: Essays in Honor of Ryuzo Sato, Research Monographs in Japan–U.S. Business & Economics 5, T. Negishi, R. Ramachandran, and K. Mino, eds., Kluwer, Norwell, MA, 2001.
[18] W. H. Fleming and J. L. Stein, *Stochastic optimal control in international finance and debt*, J. Banking and Finance, 28 (2004), pp. 979–996.
[19] A. Friedman, *Stochastic Differential Equations and Applications*, Vol. 1, Academic Press, New York, 1975.
[20] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, New York, 1983.
[21] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed., Springer, New York, 2001.
[22] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer, New York, 1991.
[23] P.-L. Lions, *Optimal control of diffusion processes and Hamiltion–Jacobi–Bellman equations, part* II: *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.
[24] R. S. Lipster and A. N. Shiryayev, *Statistics of Random Processes* I, Springer, New York, 1977.
[25] R. C. Merton, *Continuous Time Finance*, rev. ed., Blackwell, Cambridge, MA, 1992.
[26] B. Øksendal, *Stochastic Differential Equations*, 5th ed., Springer, New York, 1998.
[27] T. Pang, *Stochastic Control Theory and Its Applications to Financial Economics*, Ph.D. thesis, Brown University, Providence, RI, 2002.
[28] T. Pang, *Portfolio optimization models on infinite time horizon*, J. Optim. Theory Appl., to appear.
[29] C. V. Pao, *Nonlinear Parabolic and Elliptic Equations*, Plenum Press, New York, 1992.
[30] A. S. Üstünel and M. Zakai, *Transformation of Measure of Weiner Space*, Springer, New York, 2000.
[31] W. Walter, *Ordinary Differential Equations*, Springer, New York, 1998.
[32] T. Zariphopoulou, *A solution approach to valuation with unhedgeable risks*, Finance Stoch., 5 (2001), pp. 61–82.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 532–548

ON THE BELLMAN EQUATION FOR THE MINIMUM TIME PROBLEM IN INFINITE DIMENSIONS*

PIERMARCO CANNARSA† AND OVIDIU CÂRJĂ‡

Abstract. We consider the time optimal control problem for a semilinear parabolic control system, where the target is the closed ball with center 0 and radius $R \geq 0$ in a Hilbert space X. In particular, we allow the origin of X to be the target. Using an appropriate Kružkov-type transformation, we give an existence and uniqueness result for the associated Hamilton–Jacobi–Bellman equation, even when the reachable set is not the whole space.

Key words. Hamilton–Jacobi–Bellman equation, viscosity solutions, optimality principle, time optimal control

AMS subject classifications. 49L20, 49K20, 93C10

DOI. 10.1137/S0363012902419011

1. Introduction.

This paper is concerned with the time optimal control problem for the system

$$y'(t) = Ay(t) + f(y(t)) + u(t), \quad t > 0, \quad y(0) = x, \tag{1}$$

where the operator A generates a C_0-semigroup on a Hilbert space X, f is a Lipschitz continuous function, and $u(\cdot)$ is a control taking values in B_r, the closed ball of center 0 and radius $r > 0$ in X. Here, the problem we consider is to reach a given *target set*, starting from the initial point x in minimum time $T(x)$, by trajectories of the state equation (1). We restrict our attention to the case where the target set is B_R, the closed ball of center 0 and radius $R \geq 0$ in X. Since the literature on this problem is huge, we refer the reader to some of the most recent monographs on this subject, such as [2], [15], and [13].

The goal of this paper is to study the Hamilton–Jacobi–Bellman equation associated with our time optimal control problem. Such equations have been analyzed by many authors using various approaches and—at times—different frameworks within the same approach, as in [8], [9], [10], [11], [12], [14], [16], [17] and [5], where variants of the viscosity solution method are proposed. The aforementioned results, however, cannot be applied to the problem we are interested in due to the presence of boundary conditions.

The viscosity solution technique was adapted to the specific case of Hamilton–Jacobi–Bellman equations related to the minimum time problem in [3], for a nonlinear case, as well as in [4], for a linear problem, obtaining existence and uniqueness results. We point out that, in both [3] and [4], the reachable set $\mathcal{R} = \{x \in X : T(x) < \infty\}$ is the whole space X, and the authors apply the Kružkov transformation $W(x) = 1 - e^{-T(x)}$ in order to arrive at a new Hamiltonian, more suitable for getting uniqueness under appropriate boundary conditions.

*Received by the editors November 29, 2002; accepted for publication (in revised form) December 9, 2003; published electronically July 23, 2004.

http://www.siam.org/journals/sicon/43-2/41901.html

†Dipartimento di Matematica, Università di Roma "Tor Vergata," 00133 Roma, Italy (cannarsa@mat.uniroma2.it).

‡Department of Mathematics, University of Iaşi, Iaşi, 700506, Romania (ocarja@uaic.ro). The work of this author was partially supported by MECT-CNCSIS. Part of this work was completed while the author was visiting the University of Rome "Tor Vergata," funded by Istituto Nazionale di Alta Matematica. The author wishes to thank these institutions for their warm hospitality.

The main purpose of this paper is to characterize W as the unique solution of the associated Hamilton–Jacobi–Bellman equation, in case the reachable set is not necessarily the whole space. A major difficulty to overcome is related to the behavior of W at the boundary of the reachable set. More precisely, our technique for uniqueness requires W to be uniformly continuous, as well as sequentially weakly continuous, on $\partial\mathcal{R}$.

The key idea of our approach is to modify the Kružkov transformation as follows:

$$W(x) = \begin{cases} 1 - e^{-kT(x)} & \text{for } x \in \mathcal{R}, \\ 1 & \text{for } x \notin \mathcal{R}, \end{cases}$$

where $k > 0$ is a constant to be chosen so that W be differentiable on $X \setminus \mathcal{R}$. This is indeed possible since, as we will show, $T(x)$ has a logarithmic behavior near $\partial\mathcal{R}$; see Proposition 3.3.

Some technical tools of our method are built on the results of [1], where regularity properties of the minimum time function are obtained for system (1) with a target of the form B_R, $R > 0$. In order to recover the smoothness of W on $\partial\mathcal{R}$ that is needed to handle the Hamilton–Jacobi–Bellman equation, we had to produce a finer version of the Lipschitz regularity result of [1]. It is worth noting that such a regularity result holds for a point target ($R = 0$) as well.

The outline of this paper is the following. In section 2 we recall known results on semiconcave functions and generalized gradients. Section 3 is focussed on the regularity properties of the minimum time function, while section 4 contains our existence and uniqueness result for the Hamilton–Jacobi–Bellman equation.

2. Preliminaries. Throughout this paper we denote by X a Hilbert space with scalar product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$. We denote by $\|\cdot\|$ the standard norm of a linear operator between Banach spaces. For any $R \geq 0$ and $x \in X$, $B_R(x)$ stands for the *closed* ball of radius R centered at x, that is,

$$B_R(x) = \{y \in X \,:\, |y - x| \leq R\}.$$

We abbreviate $B_R = B_R(0)$.

Let us recall the definition of semiconcave functions, a basic notion for our approach. Indeed, if the target is a closed ball B_R with $R > 0$, then both the minimum time function and its Kružkov transformation are semiconcave on the reachable set.

DEFINITION 2.1. *Given $\alpha \in (0, 1]$ and Ω an open subset of X, a function $g : \Omega \to \mathbb{R}$ is said to be* semiconcave *with exponent α (or with modulus Cr^α) if, for any point $x_0 \in \Omega$, there exist $\rho > 0$ and $C \geq 0$ such that*

$$g(x) + g(y) - 2g\left(\frac{x+y}{2}\right) \leq C|x - y|^{1+\alpha}$$

for all $x, y \in B_\rho(x_0)$.

Semiconcave functions share many properties with concave functions; see, e.g., [1]. We now recall two such properties that will be used later on in the paper.

Recall first the notion of *superdifferential* of g in $x \in \Omega$, $D^+g(x)$. Namely,

$$D^+g(x) = \left\{ p \in X \;:\; \limsup_{y \to x} \frac{g(y) - g(x) - \langle p, y - x \rangle}{|y - x|} \leq 0 \right\}.$$

Similarly, the *subdifferential* of g in $x \in \Omega$ is defined as

$$D^-g(x) = \left\{ p \in X \;:\; \liminf_{y \to x} \frac{g(y) - g(x) - \langle p, y - x \rangle}{|y - x|} \geq 0 \right\}.$$

PROPOSITION 2.2. *For a semiconcave function g with exponent α, the superdifferential $D^+g(x)$ is nonempty for any $x \in \Omega$. Moreover, for any $x_0 \in \Omega$ there exist $\gamma > 0$ and $C > 0$ such that*

$$g(y) - g(x) - \langle p, y - x\rangle \le C|y - x|^{1+\alpha} \quad \forall x, y \in B_\gamma(x_0),\ p \in D^+g(x). \tag{2}$$

Another property of the minimum time function that plays a prominent role in our approach is a stronger variant of the local Lipschitz continuity. It is a Lipschitz property, labelled below as (P_θ), with respect to a weaker norm on X, $x \mapsto |(-A)^{-\theta}x|$, where $\theta \in (0, 1]$, $(-A)^{-\theta}$ is the inverse of the fractional power of $-A$ and A is the generator of an analytic semigroup of type $-\omega$ with $\omega > 0$. Recall that the semigroup $S(t)$ is of type $-\omega$ if it satisfies the property

$$\|S(t)\| \le e^{-\omega t} \quad \forall t \ge 0. \tag{3}$$

PROPOSITION 2.3. *Let $\theta \in (0, 1]$ and let $g : \Omega \to \mathbb{R}$. Suppose that*

(P_θ) *for any $x_0 \in \Omega$ there exist $\delta, K > 0$ such that*

$$|g(x) - g(z)| \le K|(-A)^{-\theta}(x - z)| \qquad \forall x, z \in B_\delta(x_0).$$

Then

$$D^+g(x),\ D^-g(x) \subset D((-A)^\theta) \qquad \forall x \in \Omega$$

and

$$|(-A)^\theta p| \le K \qquad \forall x \in B_\delta(x_0),\ \forall p \in D^+g(x),\ \forall p \in D^-g(x).$$

Proof. A proof for the case of $\theta = 1$ is given in [4]. To assist the reader, we outline here the proof of the general case. Given $y \in B_1$ and $\lambda > 0$ small enough, we have

$$-\langle p, y\rangle \le \frac{g(x + \lambda y) - g(x) - \langle p, \lambda y\rangle}{\lambda} + \frac{|g(x + \lambda y) - g(x)|}{\lambda}$$

for all x in a neighborhood of x_0. By letting $\lambda \to 0$, we obtain

$$-\langle p, y\rangle \le K|(-A)^{-\theta}y| \qquad \forall p \in D^+g(x).$$

Since this holds for any $y \in B_1$, we get that $p \in D((-A)^\theta)$ and $|(-A)^\theta p| \le K$. For $p \in D^-g(x)$ we start with the inequality

$$-\langle p, y\rangle \ge \frac{g(x + \lambda y) - g(x) - \langle p, \lambda y\rangle}{\lambda} - \frac{|g(x + \lambda y) - g(x)|}{\lambda}$$

and proceed as above. □

Property (P_θ) and the semiconcavity of g yield a useful result for the graph of D^+g.

PROPOSITION 2.4. *Suppose that a function $g : \Omega \to \mathbb{R}$ is semiconcave with some exponent $\alpha > 0$ and satisfies property (P_θ) for some $\theta \in (0, 1]$. Let $\{x_n\}$, $\{p_n\}$ be given sequences in X such that $p_n \in D^+g(x_n)$ and $x_n \to x$ strongly. Then, at least on a subsequence, $(-A)^\theta p_n \to q$ weakly for some $q \in X$, $p_n \to (-A)^{-\theta}q$ weakly, and $(-A)^{-\theta}q \in D^+g(x)$.*

Proof. By Proposition 2.3 we get that $\{(-A)^\theta p_n\}$ is a bounded sequence in X. Hence, it admits a subsequence (still labelled $\{(-A)^\theta p_n\}$) that weakly converges to some $q \in X$. Since operator $(-A)^{-\theta}$ is bounded, we have $p_n \to (-A)^{-\theta} q$ weakly. Next, by (2) we obtain

$$g(y) - g(x_n) - \langle p_n, y - x_n \rangle \le C|y - x_n|^{1+\alpha}$$

for any y in a neighborhood of x and for any n sufficiently large. Letting $n \to \infty$ and then $y \to x$, we conclude that $(-A)^{-\theta} q \in D^+ g(x)$, as claimed. □

We end this section recalling some facts about *proximal gradients*, a notion that plays a basic role in the study of the Bellman equation in case the target is the origin of X. We refer the reader to [7, p. 27] for details. For our purposes here, it is suitable to take the characterization given in [7, Theorem 2.5] as the definition of the *proximal subdifferential* of g at x, namely,

$$\partial_P g(x) = \{\zeta \in H : \exists \sigma, \rho > 0;\ g(y) - g(x) + \sigma|y - x|^2 \ge \langle \zeta, y - x \rangle \ \ \forall y \in B_\rho(x)\}.$$

Clearly $\partial_P g(x) \subseteq D^- g(x)$. Further, the following proposition, proved in [7, pp. 37–38], will be useful in what follows.

PROPOSITION 2.5. *Assume that $g : X \to \mathbb{R}$ is lower semicontinuous.*

(a) *If g has a local minimum at x, then $0 \in \partial_P g(x)$.*

(b) *If $h : X \to \mathbb{R}$ is of class C^2 in a neighborhood of x, then*

$$\zeta \in \partial_P(g + h)(x) \ \Rightarrow\ \zeta - Dh(x) \in \partial_P g(x),$$

where $Dh(x)$ is the Fréchet derivative of h in x.

The *proximal superdifferential* of g at x is defined as

$$\partial^P g(x) = \{\zeta \in H : \exists \sigma, \rho > 0;\ g(y) - g(x) - \sigma|y - x|^2 \le \langle \zeta, y - x \rangle \ \ \forall y \in B_\rho(x)\}.$$

Since $\partial^P g(x) = -\partial_P(-g)(x)$, a similar result as in Proposition 2.5 holds with upper instead of lower and maximum instead of minimum.

3. The minimum time function. First, let us list the basic assumptions we shall refer to in what follows.

(H1) X is a Hilbert space with scalar product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$, and $A : D(A) \subset X \to X$ is the infinitesimal generator of an analytic semigroup $S(t)$, $t \ge 0$, on X. Further, A is self-adjoint and satisfies

$$\langle Ax, x \rangle \le -\omega |x|^2 \quad \forall x \in D(A),$$

for some constant $\omega > 0$.

(H2) $f : X \to X$ is a Lipschitz continuous function satisfying, for some constant $L > 0$,

$$|f(x) - f(y)| \le L|x - y| \quad \forall x, y \in X,$$

$$f(0) = 0.$$

(H3) Two constants $r > 0$ and $R \ge 0$ are given such that $(L - \omega)R < r$.

(H4) The function f above is Gâteaux differentiable on X and there exist $\theta_1 \in (0, 1/2)$, $\alpha \in (0, 1]$, and C' such that

$$|f(x + y) - f(x) - \delta f(x) y| \le C' |(-A)^{\theta_1} y|^{1+\alpha}$$

for all $x, y \in D((-A)^{\theta_1})$. Moreover, the map $x \mapsto \delta f(x)$ is strongly continuous on X; that is, for any sequence $\{x_n\}$ in X,

$$\lim_{n\to\infty} x_n = x_\infty \quad \Rightarrow \quad \lim_{n\to\infty} \delta f(x_n)x = \delta f(x_\infty)x \qquad \forall x \in X.$$

(H5) $S(\cdot)$ is a compact semigroup; i.e., for any $t > 0$, $S(t)$ is a compact operator.

Let $\mathcal{U}$ be the set of all *control strategies* (or, briefly, *controls*), i.e., measurable functions $u : [0, \infty) \to B_r$. For $x \in X$ and $u \in \mathcal{U}$, consider the mild solution of (1) which satisfies the initial condition $y(0) = x$, i.e., a function $y \in C([0, \infty); X)$ satisfying

$$y(t) = S(t)x + \int_0^t S(t-s)[f(y(s)) + u(s)]ds. \tag{4}$$

We call such a solution the *trajectory* of system (1) starting from x with control u and denote it by $y(\cdot, x, u)$. For any $x \in X$, and any control u we define

$$\tau(x, u) = \min\{t \geq 0 : \ y(t, x, u) \in B_R\} \in [0, \infty],$$

called the transition time from x to B_R. Define now the *reachable set* $\mathcal{R}$ as the set of all points x such that $\tau(x, u) < \infty$ for some u. A control u at which $\tau(x, \cdot)$ attains its minimum is called *optimal* for x, and the corresponding solution $y(\cdot, x, u)$ of (1) is called the *optimal trajectory*. Finally, define the *minimum time function* as

$$T : \mathcal{R} \to [0, \infty), \quad T(x) = \inf_{u\in\mathcal{U}} \tau(x, u).$$

It is well known that assumption (H1) implies that semigroup $S(\cdot)$ satisfies (3). Moreover, the fractional powers of $-A$, denoted by $(-A)^\theta$, are well defined for any $\theta \in [0, 1]$ and satisfy

$$|(-A)^\theta S(t)x| \leq \frac{M_\theta}{t^\theta}|x| \qquad \forall x \in X, \ t > 0,$$

$$|x| \leq M'_\theta |(-A)^\theta x| \qquad \forall x \in D((-A)^\theta)$$

for some constants $M_\theta, M'_\theta > 0$. These facts are used to prove that the minimum time function satisfies property (P_θ).

Assumption (H4) is used to prove that, in case the target is B_R with $R > 0$, the minimum time function is *semiconcave* with exponent α; see [1, Theorem 4.3].

In the next propositions we gather some basic properties of the trajectories of (1), the reachable set, and the associated minimum time function.

PROPOSITION 3.1. *Assume* (H1), (H2), *and* (H3). *Then the following properties hold:*

(i) *In case $L - \omega > 0$, for every $x \in X$ satisfying $R < |x| < r/(L - \omega)$ a control strategy u^* exists such that the corresponding trajectory of* (1) *reaches the target B_R with $R \geq 0$ and satisfies*

$$|y(t, x, u^*)| \leq e^{(L-\omega)t}\left(|x| - \frac{r}{L-\omega}\right) + \frac{r}{L-\omega} \tag{5}$$

for any

$$0 \leq t \leq \tau(x, u^*) \leq \frac{1}{L-\omega} \log \frac{\frac{r}{L-\omega} - R}{\frac{r}{L-\omega} - |x|}.$$

(ii) *In case* $L - \omega \le 0$, *for any* $x \in X$ *a control strategy* u^* *exists such that the corresponding trajectory of* (1) *reaches* B_R *with* $R \ge 0$ *and satisfies*

$$|y(t, x, u^*)| \le |x| - rt \tag{6}$$

for any $0 \le t \le \tau(x, u^*) \le (|x| - R)r^{-1}$.

Remark 3.1. The above proposition is a considerable extension of [1, Lemma 3.6 (i)], where the authors proved (5) for $t \le \omega^{-1} \log(1 + r^{-1}\omega|x|)$. Indeed, under the last restriction one does not get controllability to the target B_R with $R > 0$ for all points satisfying $|x| < r/(L - \omega)$. Nor does one get null controllability at all. Finally, the results provided by Proposition 3.1 hold in any Banach space and for any C_0-semigroup satisfying $\|S(t)\| \le e^{-\omega t}$ for some $\omega \in \mathbb{R}$.

Proof of Proposition 3.1. (i) First, let us consider the case $R > 0$. Observe that the function

$$F_R(y) = \begin{cases} -\frac{r}{R} y & \text{if } |y| \le R, \\ -r\frac{y}{|y|} & \text{if } |y| \ge R \end{cases}$$

is Lipschitz continuous in X. Then, for every $x \in D(A)$, the equation

$$y'(t) = Ay(t) + f(y(t)) + F_R(y(t)), \quad t > 0, \tag{7}$$

has a unique classical solution on each interval $[0, T]$ satisfying $y(0) = x$ (see, e.g., [6, p. 60]). Multiplying (7) by $y(t)$ and taking into account that

$$\langle y'(t), y(t) \rangle = \frac{1}{2} \frac{d}{dt} |y(t)|^2 \quad \forall t > 0,$$

we easily get

$$|y(t)| \frac{d}{dt} |y(t)| \le (L - \omega)|y(t)|^2 - r\,|y(t)|$$

for $t \in [0, \tilde{t}]$, where $\tilde{t}$ is the minimum of t for which $y(t) \in B_R$. Then,

$$|y(t)| \le |x| - rt + (L - \omega) \int_0^t |y(t)|\, ds \quad \forall t \in [0, \tilde{t}],$$

which, if $L - \omega \le 0$, immediately gives (6). In case $L - \omega > 0$, (5) follows from the above inequality and Gronwall's lemma. Notice that the existence of time $\tilde{t}$ is a consequence of (5) and (6).

Therefore, (5) and (6) have been proven for every $x \in D(A)$. By a density argument, they also hold for any $x \in X$. Thus, the desired control is

$$u^*(t) = \frac{-ry(t)}{|y(t)|}. \tag{8}$$

In the case of $R = 0$, we observe that, if $R_1 < R_2$, then the corresponding times $\tilde{t}_1$ and $\tilde{t}_2$ satisfy $\tilde{t}_1 > \tilde{t}_2$, while the corresponding solutions of (7), $y_1(\cdot)$ and $y_2(\cdot)$, satisfy

$$y_1(t) = y_2(t) \quad \forall t \in [0, \tilde{t}_2].$$

In other words, the trajectory of (7) does not depend on $R > 0$ until it reaches B_R. Taking a sequence R_n decreasing to 0, we get a sequence $\tilde{t}_n$ increasing to some $\bar{t}$ that

satisfies the inequalities requested in (i) or (ii). The proof ends by considering the control given by (8) on $[0, \bar{t}]$. □

Remark 3.2. The above proof shows that for any $x \in B_R$ there exists a control strategy u^* keeping the corresponding trajectory $y(t, x, u^*)$ inside B_R for every $t \geq 0$. Indeed, for $R = 0$ it suffices to take $u^* \equiv 0$, while, for $R > 0$, one can take the feedback control u^* given in (8) until y reaches the origin, and then 0.

In the next proposition $d(x)$ denotes the distance of x from the target, that is,

$$d(x) = \max\{|x| - R, 0\}, \qquad x \in X\,.$$

PROPOSITION 3.2. *Assume* (H1), (H2), *and* (H3).
(i) *In case* $L - \omega > 0$, *for any* $\rho \in (0, r/(L-\omega) - R)$ *we have*

$$T(x) \leq \frac{d(x)}{r - (R+\rho)(L-\omega)}$$

for all points x *satisfying* $d(x) \leq \rho$.
(ii) *In case* $L - \omega \leq 0$ *we have*

$$T(x) \leq \frac{d(x)}{r}$$

for every $x \in X$.

The proof goes as in [1, Proposition 3.8] and is based on Proposition 3.1 above.

Remark 3.3. The Lipschitz continuity of T around the target, which plays a basic role in our analysis, is essentially a consequence of the fact that we have full control for the state equation (1). The same property can be obtained, using the same idea, in case the state equation is

$$y'(t) = Ay(t) + f(y(t)) + Du(t), \tag{9}$$

where D is a bounded linear operator from a Banach space U *onto* X. In fact, assuming that controls take values in B_r, the closed ball of center 0 and radius r in U, let us note that, by the open mapping theorem, there exists a ball B_p in X, contained in the image of B_r under D. Then, reasoning as above and using Filippov's selection theorem, we obtain the Lipschitz property around the target for the minimum time function for (9) with target B_R, $R \geq 0$.

Let us now show that the surjectivity of D is a necessary condition in order for T to be Lipschitz around the point target 0. Indeed, assume that there exist $\rho > 0$ and $\gamma > 0$ such that $T(x) \leq \gamma|x|$ for every $x \in B_\rho$. This implies that, if $|x| \leq t/\gamma$, then x can be transferred to 0 in time t by some control strategy $\bar{u}$. Hence,

$$S(t)x = -\int_0^t S(t-s)[f(y(s, x, \bar{u})) + D\bar{u}(s)]ds.$$

Taking the scalar product of the both sides with a fixed element x^* in X (or in the dual of X, should X be a reflexive Banach space), we get

$$\langle x, S^*(t)x^*\rangle \leq L\int_0^t |S^*(t-s)x^*||y(s, x, u)|ds + r\int_0^t |D^*S^*(s)x^*|ds. \tag{10}$$

Notice that, arguing by Gronwall's inequality as in [1, Lemma 3.6], we have

$$|y(t, x, \bar{u})| \leq e^{(L-\omega)t}\left(|x| + \frac{r\|D\|}{L-\omega}\right) - \frac{r\|D\|}{L-\omega} \qquad \forall t \geq 0.$$

Using this inequality in (10), we obtain

$$\langle x, S^*(t)x^*\rangle \le \frac{Lt}{\gamma}|x^*|\int_0^t e^{\omega(s-t)+(L-\omega)t}ds + L|x^*|\frac{r\|D\|}{L-\omega}\int_0^t e^{\omega(s-t)}[e^{(L-\omega)t}-1]ds + r\int_0^t |D^*S^*(s)x^*|ds.$$

Finally, take the supremum for $x \in B_{t/\gamma}$, divide by t, and let $t \downarrow 0$ to obtain

$$C|x^*| \le |D^*x^*| \quad \forall x^* \in X$$

for some constant $C > 0$. This implies the surjectivity of D, as claimed.

It is well known that, in finite dimensional control theory, the Lipschitz continuity of T around the target is equivalent to the so-called Petrov condition. It is easy to see that, when $\dim X < \infty$, the Petrov condition for (9) reduces to the surjectivity of D.

The next result shows that the Lipschitz property around the target yields the local Lipschitz continuity of T on the whole reachable set, as well as a precise estimate of the behavior of T near the boundary, $\partial\mathcal{R}$, of the reachable set. To this end, the following well-known dynamic programming principle is a main tool: for any $x \in \mathcal{R}$ and for any control u,

$$T(x) \le t + T(y(t,x,u)) \quad \forall t \in [0, \tau(x,u)]. \tag{11}$$

Furthermore, equality holds in (11) if and only if u is optimal for x.

PROPOSITION 3.3. *Assume* (H1), (H2), *and* (H3).

(i) *Suppose* $L-\omega > 0$, *let* $\rho \in (0, r/(L-\omega) - R)$, *and define*

$$\gamma = \frac{1}{r-(R+\rho)(L-\omega)}.$$

If $x \in \mathcal{R}$ *and* z *is such that*

$$|z-x| \le \rho\, e^{-(L-\omega)T(x)},$$

then $z \in \mathcal{R}$ *and*

$$T(z) \le T(x) + \gamma\, e^{(L-\omega)T(x)}|x-z|.$$

Moreover, the reachable set $\mathcal{R}$ *is open, and the minimum time function is locally Lipschitz continuous on* $\mathcal{R}$.

(ii) *If* $L-\omega \le 0$, *then the minimum time function is globally Lipschitz continuous on* X. *More precisely, for any* $x, z \in X$ *we have*

$$|T(z) - T(x)| \le \frac{1}{r}|x-z|.$$

(iii) *For every* $x \in \mathcal{R}$ *and* $z \notin \mathcal{R}$ *we have*

$$T(x) \ge -\frac{1}{L-\omega}\log\frac{|x-z|}{\rho}.$$

Consequently,

$$\lim_{x\to z} T(x) = \infty \quad \forall z \in \partial\mathcal{R}.$$

Proof. (i) Recall first that, for any $x, z \in X$ and any control strategy u, we have

$$|y(t,x,u) - y(t,z,u)| \le e^{(L-\omega)t}|x-z| \quad \forall t \ge 0. \tag{12}$$

This follows easily by the Gronwall inequality. Without loss of generality, we can assume the existence of an optimal control u for x. In the general case, the conclusion follows by an approximation argument. Since $y(T(x),x,u) \in B_R$, by using (12) we get

$$d(y(T(x),z,u)) \le e^{(L-\omega)T(x)}|x-z| \le \rho,$$

so $y(T(x),z,u) \in \mathcal{R}$, which, in turn, implies that $z \in \mathcal{R}$ and

$$T(y(T(x),z,u)) \le \gamma e^{(L-\omega)T(x)}|x-z|.$$

Here, we have used Proposition 3.2. Now, apply the dynamic programming principle to obtain the first part of the claim in (i). To conclude the proof of (i), for each $x_0 \in \mathcal{R}$, take

$$\delta = \frac{\rho}{2} e^{-(L-\omega)(T(x_0)+\gamma\rho)}$$

and observe that, from the first part, we have

$$|T(x_1) - T(x_2)| \le \gamma e^{(L-\omega)(T(x_0)+\gamma\rho)}|x_1 - x_2|$$

for every $x_1, x_2 \in B_\delta(x_0)$.

Similar arguments can be used to prove (ii). For the proof of (iii), let us recall, first, that we are assuming $L - \omega > 0$. Suppose $(L-\omega)T(x) < -\log|x-z|/\rho$. Then we would have $|x-z| < \rho\, e^{-(L-\omega)T(x)}$. This, along with (i), implies the contradiction $z \in \mathcal{R}$. □

In the next proposition we show that, if $S(\cdot)$ is compact, then the minimum time function is sequentially weakly continuous. For this, we will use a well-known compactness result.

LEMMA 3.4. *Let $0 < a < t$, let $\{x_n\}$ be a sequence, weakly convergent to x, and $\{u_n\}$ be a sequence weakly convergent to u in $L^2(0,t;X)$. Then, $y(\cdot,x_n,u_n) \to y(\cdot,x,u)$ in $C([a,t];X)$.*

Proof. The proof is standard and relies on the compactness of the operator $\varphi \mapsto P(\varphi)$ from $L^p(0,t;X)$, $p > 1$, to $C([0,t];X)$, where

$$P(\varphi)(s) = \int_0^s S(s-\tau)\varphi(\tau)\, d\tau$$

(see [15, p. 104]). □

PROPOSITION 3.5. *Assume* (H1), (H2), (H3), *and* (H5).

(i) *If $x \in \mathcal{R}$ and $x_n \to x$ weakly, then, for n large enough, we have $x_n \in \mathcal{R}$ and $T(x_n) \to T(x)$.*

(ii) *For any $x \notin \mathcal{R}$ and any sequence $\{x_n\}$, contained in $\mathcal{R}$ and weakly convergent to x,*

$$\lim_{n\to\infty} T(x_n) = \infty.$$

Proof. In this proof we shall assume $L-\omega>0$, the reasoning being simpler for $L-\omega\le 0$.

In order to prove the first statement of (i), let $\bar t>0$ be such that $y(\bar t,x,0)\in\mathcal{R}$ (this is possible because $\mathcal{R}$ is open). Then, by Lemma 3.4, $y(\bar t,x_n,0)\in\mathcal{R}$ for n sufficiently large. This clearly implies that $x_n\in\mathcal{R}$.

Now, take a subsequence $\{x_{n_k}\}$ such that

$$T(x_{n_k})\to T^*=\limsup_{n\to\infty}T(x_n),$$

and fix $\varepsilon>0$. Again by Lemma 3.4, we have that $y(\varepsilon,x_{n_k},0)\to y(\varepsilon,x,0)$. Thus, $T(y(\varepsilon,x_{n_k},0))$ converges to $T(y(\varepsilon,x,0))$ as $k\to\infty$. By the dynamic programming principle (11), we get

$$T(x_{n_k})\le\varepsilon+T(y(\varepsilon,x_{n_k},0))\,.$$

Therefore, $T^*\le\varepsilon+T(y(\varepsilon,x,0))$. Let $\varepsilon\to 0$ to obtain $\limsup_{n\to\infty}T(x_n)\le T(x)$. We have thus proven that $T(\cdot)$ is sequentially weakly upper semicontinuous at x.

Next, let us show that $T(\cdot)$ is sequentially weakly lower semicontinuous at x. To this end, take a subsequence $\{x_{n_k}\}$ such that

$$T(x_{n_k})\to T^\#=\liminf_{n\to\infty}T(x_n).$$

First, suppose $T^\#>0$ and, having fixed $0<\varepsilon<T^\#$, let u_{n_k} be the time optimal control for x_{n_k}. By Lemma 3.4, we may assume with no loss of generality (by extracting a subsequence if necessary) that there exists some control strategy u such that $y(\varepsilon,x_{n_k},u_{n_k})\to y(\varepsilon,x,u)$ strongly. Taking into account the dynamic programming principle, we have

$$T(x_{n_k})=\varepsilon+T(y(\varepsilon,x_{n_k},u_{n_k})),$$

which implies that $T(y(\varepsilon,x_{n_k},u_{n_k}))\le M$ for some $M>0$. Let $\rho\in(0,r/(L-\omega)-R)$, and let k_0 be such that

$$|y(\varepsilon,x,u)-y(\varepsilon,x_{n_k},u_{n_k})|\le\rho\,e^{-(L-\omega)M}$$

for every $k>k_0$. By Proposition 3.3 we obtain that $y(\varepsilon,x,u)\in\mathcal{R}$ and

$$T(y(\varepsilon,x,u))\le T(y(\varepsilon,x_{n_k},u_{n_k}))+\gamma\,e^{(L-\omega)M}|y(\varepsilon,x_{n_k},u_{n_k})-y(\varepsilon,x,u)|$$

for every $k>k_0$. Therefore,

$$T(x_{n_k})=\varepsilon+T(y(\varepsilon,x_{n_k},u_{n_k}))\ge\varepsilon+T(y(\varepsilon,x,u))-\gamma\,e^{(L-\omega)M}|y(\varepsilon,x_{n_k},u_{n_k})-y(\varepsilon,x,u)|.$$

Let $k\to\infty$ and then $\varepsilon\to 0$ to obtain $\liminf_{n\to\infty}T(x_n)\ge T(x)$. Now, let us suppose $T^\#=0$ and prove that $T(x)=0$. To this end, fix $\varepsilon>0$ and note that $T(x_{n_k})<\varepsilon$ for k sufficiently large. Then, for x_{n_k} define a control strategy, v_k, on $[0,\varepsilon]$ as follows: on $[0,T(x_{n_k})]$, v_k equals the time optimal control steering x_{n_k} to B_R, and on $[T(x_{n_k}),\varepsilon]$, v_k is any control strategy that forces the trajectory to remain in B_R. This is possible in view of Remark 3.2. Taking into account that (at least on a subsequence) $y(\varepsilon,x_{n_k},v_k)\to y(\varepsilon,x,u)$ for some control strategy u, we deduce that $y(\varepsilon,x,u)\in B_R$, and hence $T(x)\le\varepsilon$. Since this fact holds for any $\varepsilon>0$, we get $T(x)=0$, as claimed.

To prove (ii), consider $x \notin \mathcal{R}$, take a sequence $\{x_n\}$ in $\mathcal{R}$ weakly convergent to x, and assume, by contradiction, that

$$\liminf_{n\to\infty} T(x_n) < \infty.$$

Arguing as in the above proof of the lower semicontinuity of T, we get that $y(\varepsilon, x, u) \in \mathcal{R}$. This clearly implies that $x \in \mathcal{R}$, a contradiction. Thus, $\liminf_{n\to\infty} T(x_n) = \infty$. □

We end this section by recalling some results of [1] concerning the semiconcavity of the minimum time function and property (P_θ).

PROPOSITION 3.6. *Assume* (H1), (H2), *and* (H3). *Then the minimum time function satisfies property* (P_θ) *for any* $\theta \in [0, 1)$. *If, in addition,* (H4) *holds and the target is* B_R *with* $R > 0$, *then the minimum time function is semiconcave with exponent* α.

We note that, in [1], only the case of $R > 0$ is considered. However, taking into account Proposition 3.3, one can use the same arguments to recover property (P_θ) in the case of $R = 0$.

4. The Bellman equation. Let us introduce a variant of the Kružkov transformation that we shall use in this section, namely the function $W : X \to \mathbb{R}$ defined by

$$W(x) = \begin{cases} 1 - e^{-2(L-\omega)T(x)} & \text{if } x \in \mathcal{R}, \\ 1 & \text{if } x \notin \mathcal{R}. \end{cases} \tag{13}$$

Since the value function is known to be semiconcave only in the case of $R > 0$, we shall give two results for the Hamilton–Jacobi–Bellman equation. The former assumes semiconcavity and therefore applies only if $R > 0$; the latter applies to the general case $R \geq 0$.

THEOREM 4.1. *Assume* (H1)–(H5), *and consider the target* B_R *with* $R > 0$. *Then the Kružkov transformation* W *defined by* (13) *is the unique function from* X *to* $\mathbb{R}$ *which satisfies the following properties:*

(a) W *is sequentially weakly continuous.*

(b) $\emptyset \neq D^+W(x) \subset D((-A)^\theta))$ *for any* $x \in X \setminus B_R$ *and for any* $\theta \in (0, 1)$.

(c) *There exist* $\rho > 0$ *and* $M > 0$ *such that if* $|z - x| \leq \rho\sqrt{1 - W(x)}$, *then*

$$W(z) - W(x) \leq M\sqrt{1 - W(x)}|z - x|.$$

(d) $W(x) = 0$ *for any* $x \in B_R$, *and* $W(x) \in [0, 1]$ *for any* $x \in X$.

(e) W *satisfies the Hamilton–Jacobi–Bellman equation in the following sense: for any* $x \in X \setminus B_R$ *with* $x \in \cap_{\theta\in(0,1)} D((-A)^\theta)$ *and for any* $\gamma \in (0, 1)$,

(i) *there exists* $p \in D^+W(x)$ *such that*

$$r\,|p| + \langle(-A)^{1-\gamma}p, (-A)^\gamma x\rangle - \langle p, f(x)\rangle + 2(L - \omega)W(x) \geq 2(L - \omega);$$

(ii) *for any* $p \in D^+W(x)$ *we have*

$$r\,|p| + \langle(-A)^{1-\gamma}p, (-A)^\gamma x\rangle - \langle p, f(x)\rangle + 2(L - \omega)W(x) \leq 2(L - \omega).$$

Proof. Let us show, first, that the function W defined by (13) satisfies properties (a)–(e). To begin with, we note that property (a) follows from Proposition 3.5, while

(d) is obvious. In order to prove (c), we use the elementary inequality $1 - e^t \le -t$ to obtain

$$W(z) - W(x) \le 2(L-\omega)e^{-2(L-\omega)T(x)}(T(z) - T(x)) .$$

This, in view of Proposition 3.3, gives

$$W(z) - W(x) \le 2(L-\omega)\gamma\, e^{-(L-\omega)T(x)}|z-x|,$$

and hence the conclusion. Let us now prove (b). For any $x \in \mathcal{R}$, we have $\emptyset \neq D^+W(x)$ because, when restricted to $\mathcal{R}$, W is semiconcave together with T on $\mathcal{R}$ (see Proposition 3.6). Furthermore, $D^+W(x) \subset D((-A)^\theta)$ for any $\theta \in (0,1)$ because of Proposition 2.3. To complete the proof of (b), we observe that

$$0 \le W(x) - W(z) \le \frac{1}{\rho^2}|z-x|^2 \qquad \forall x \notin \mathcal{R},\ z \in \mathcal{R}. \tag{14}$$

This follows by Proposition 3.3(iii) and easily implies that

$$D^+W(x) = \{0\} \qquad \forall x \notin \mathcal{R}. \tag{15}$$

Next, let us prove (e). Let $x \in X \setminus B_R$ be such that $x \in \cap_{\theta\in(0,1)} D((-A)^\theta)$, and take $\gamma \in (0,1)$. Let us observe first that, owing to (15), both (i) and (ii) are verified in case $x \notin \mathcal{R}$. Now, take $x \in \mathcal{R}$. To prove (i), let $u(\cdot)$ be an optimal control strategy for x and apply the Lebourg's mean value theorem to deduce that, for any t small enough, there exists a point x_t in the segment joining $y(t) = y(t,x,u)$ to x, and an element $p_t \in D^+W(x_t)$ such that

$$W(x) - W(y(t)) = \langle p_t, x - y(t)\rangle.$$

By the dynamic programming principle and the definition of W we obtain

$$\frac{e^{kt}-1}{t}(1 - W(x)) = \left\langle p_t, \frac{x-y(t)}{t}\right\rangle .$$

Hereafter, we set $k = 2(L-\omega)$ for simplicity. By Propositions 2.4 and 3.6, there exists a sequence $\{t_n\}$ such that $t_n \to 0$ and $(-A)^{1-\gamma}p_n \to (-A)^{1-\gamma}p$ weakly, for some $p \in D^+W(x)$. In particular, $p_n \to p$ weakly. (We have written, for simplicity, p_n instead of p_{t_n}.) Reasoning as in [1, p. 943], we first get

$$\begin{aligned}\lim_{n\to\infty}\left\langle p_n, \frac{S(t)x - x}{t}\right\rangle &= -\lim_{n\to\infty}\frac{1}{t_n}\int_0^{t_n}\langle (-A)^{1-\gamma}p_n, (-A)^\gamma S(s)x\rangle\, ds \\ &= -\langle (-A)^{1-\gamma}p, (-A)^\gamma x\rangle\end{aligned}$$

and then

$$\liminf_{n\to\infty}\left\langle p_n, \frac{x - y(t_n)}{t_n}\right\rangle \le \langle (-A)^{1-\gamma}p, (-A)^\gamma x\rangle - \langle p, f(x)\rangle + r|p|.$$

This ends the proof of (i).

To prove (ii), take $p \in D^+W(x)$, $v \in B_r$ and denote by $y(\cdot)$ the trajectory of (1) associated with the constant control $u(t) = v$. We have

$$W(y(t)) - W(x) = (1 - W(x))(1 - e^{-k(T(y(t))-T(x))}) \ge (1 - W(x))(1 - e^{kt});$$

thus

$$\frac{W(y(t))-W(x)}{t} \geq (1-W(x))\frac{1-e^{kt}}{t}.$$

In order to evaluate the left-hand side of the above inequality we use (2), taking into account that W is semiconcave on $\mathcal{R}$. We have

$$\frac{W(y(t))-W(x)}{t} \leq \left\langle p, \frac{y(t)-x}{t} \right\rangle + C\frac{1}{t}|y(t)-x|^{1+\alpha} \tag{16}$$

for t sufficiently small. Reasoning as above, we get

$$\lim_{t\downarrow 0} \left\langle p, \frac{y(t)-x}{t} \right\rangle \leq -\langle (-A)^{1-\gamma}p, (-A)^{\gamma}x\rangle + \langle p, f(x)\rangle + \langle p, u\rangle.$$

Now,

$$\lim_{t\downarrow 0} \frac{1}{t}|y(t)-x|^{1+\alpha} = 0$$

since $x \in D((-A)^{\theta})$ with $\theta > 1/(1+\alpha)$. Thus, the last three inequalities imply

$$kW(x) + \langle (-A)^{1-\gamma}p, (-A)^{\gamma}x\rangle - \langle p, f(x)\rangle + \langle -p, u\rangle \leq k$$

for each $u \in B_r$, yielding the conclusion. We have thus proven that the Kružkov transformation (13) satisfies all the properties (a)–(e).

To prove uniqueness, we will use the viscosity approach and borrow an idea from [4]. Let us show that two functions W_1 and W_2 satisfying (a)–(e) must coincide. Since we can exchange the role of W_1 and W_2, it is enough to prove that $W_1 \leq W_2$. For any $0 < \varepsilon < 1$, consider the function

$$\Phi_{\varepsilon}(x,y) = W_1(x) - W_2(y) - \frac{1}{2\varepsilon}|x-y|^2 - \frac{\varepsilon}{2}(|x|^2+|y|^2) \qquad \forall x, y \in X.$$

The function Φ_{ε} is weakly upper semicontinuous and tends to $-\infty$ as $|(x,y)| \to \infty$. Therefore, it attains its maximum at some point $(x_{\varepsilon}, y_{\varepsilon})$. With no loss of generality, we can restrict the analysis to the case when $\max \Phi_{\varepsilon}$ is positive and ε is small enough, since otherwise we get the conclusion immediately. We proceed by a sequence of steps, beginning with an estimate on $|x_{\varepsilon} - y_{\varepsilon}|$.

Step 1. Let us prove that

$$\frac{1}{\varepsilon}|x_{\varepsilon}-y_{\varepsilon}|^2 \leq C\sqrt{\varepsilon}. \tag{17}$$

(Hereafter, we denote by C any positive constant independent of ε.) An easy computation, based on the inequality $2\Phi_{\varepsilon}(x_{\varepsilon}, y_{\varepsilon}) \geq \Phi_{\varepsilon}(x_{\varepsilon}, x_{\varepsilon}) + \Phi_{\varepsilon}(y_{\varepsilon}, y_{\varepsilon})$, shows that

$$\frac{1}{\varepsilon}|x_{\varepsilon}-y_{\varepsilon}|^2 \leq W_1(x_{\varepsilon}) - W_1(y_{\varepsilon}) + W_2(x_{\varepsilon}) - W_2(y_{\varepsilon}), \tag{18}$$

whence

$$\frac{1}{\varepsilon}|x_{\varepsilon}-y_{\varepsilon}|^2 \leq 2.$$

In order to prove (17), suppose $W_1(x_{\varepsilon}) - W_1(y_{\varepsilon}) \leq W_2(x_{\varepsilon}) - W_2(y_{\varepsilon})$. In case $W_2(x_{\varepsilon}) - W_2(y_{\varepsilon}) \leq W_1(x_{\varepsilon}) - W_1(y_{\varepsilon})$, we work with W_1 instead of W_2 as follows. We distinguish two cases, as follows:

Case 1: $\sqrt{2\varepsilon} \le \rho\sqrt{1 - W_2(y_\varepsilon)}$. In this case we have $|x_\varepsilon - y_\varepsilon| \le \rho\sqrt{1 - W_2(y_\varepsilon)}$, which, by (c), gives

$$W_2(x_\varepsilon) - W_2(y_\varepsilon) \le M\,|x_\varepsilon - y_\varepsilon| \le M\sqrt{2\varepsilon}.$$

Hence, by (18), we obtain

$$\frac{1}{\varepsilon}|x_\varepsilon - y_\varepsilon|^2 \le 2M\sqrt{2\varepsilon}.$$

Case 2: $\sqrt{2\varepsilon} \ge \rho\sqrt{1 - W_2(y_\varepsilon)}$. Since $W_2(x_\varepsilon) - W_2(y_\varepsilon) \le 1 - W_2(y_\varepsilon)$, we easily obtain

$$\frac{1}{\varepsilon}|x_\varepsilon - y_\varepsilon|^2 \le \frac{4\varepsilon}{\rho^2},$$

and (17) follows.

Step 2. Suppose $y_\varepsilon \in B_R$. Then, for any $x \in X$,

$$W_1(x) - W_2(x) \le \Phi_\varepsilon(x_\varepsilon, y_\varepsilon) + \varepsilon\,|x|^2 \le W_1(x_\varepsilon) + \varepsilon\,|x|^2\,.$$

By Step 1, we have that $|x_\varepsilon - y_\varepsilon| \le \rho$ for ε small, so, by (c), we get

$$W_1(x_\varepsilon) - W_1(y_\varepsilon) \le M|x_\varepsilon - y_\varepsilon| \le\ M\,\sqrt{2\varepsilon}.$$

Since $W_1(y_\varepsilon) = 0$, we finally obtain

$$W_1(x) - W_2(x) \le\ M\,\sqrt{2\varepsilon} + \varepsilon|x|^2.$$

Step 3. Suppose now that $y_\varepsilon \notin B_R$, and observe that $x_\varepsilon \notin B_R$ since $\max \Phi_\varepsilon > 0$. Define

$$\phi(x) = W_2(y_\varepsilon) + \frac{1}{2\varepsilon}|x - y_\varepsilon|^2 + \frac{\varepsilon}{2}(|x|^2 + |y_\varepsilon|^2).$$

Then $W_1 - \phi$ attains its maximum at x_ε. This implies that

$$D\phi(x_\varepsilon) \in D^+W_1(x_\varepsilon). \tag{19}$$

Analogously, for

$$\psi(y) = W_2(x_\varepsilon) - \frac{1}{2\varepsilon}|x_\varepsilon - y|^2 - \frac{\varepsilon}{2}(|x_\varepsilon|^2 + |y|^2)$$

we get

$$D\psi(y_\varepsilon) \in D^-W_2(y_\varepsilon). \tag{20}$$

Since, by (b), $D^+W_2(y_\varepsilon) \ne \emptyset$, we deduce that

$$D^+W_2(y_\varepsilon) = D^-W_2(y_\varepsilon) = D\psi(y_\varepsilon).$$

Furthermore, since $D\phi(x_\varepsilon) = \varepsilon^{-1}(x_\varepsilon - y_\varepsilon) + \varepsilon\,x_\varepsilon$ and $D\psi(y_\varepsilon) = \varepsilon^{-1}(x_\varepsilon - y_\varepsilon) - \varepsilon\,y_\varepsilon$, and since, again by (b), $D\phi(x_\varepsilon)$, $D\psi(y_\varepsilon) \subset D((-A)^\theta)$ for any $\theta \in (0,1)$, we obtain x_ε, $y_\varepsilon \in D((-A)^\theta)$ for any $\theta \in (0,1)$. Now, apply (e) with some $\gamma \in (0,1)$ to obtain

$$r\,|D\psi(y_\varepsilon)| + \langle(-A)^{1-\gamma}D\psi(y_\varepsilon), (-A)^\gamma y_\varepsilon\rangle - \langle D\psi(y_\varepsilon), f(y_\varepsilon)\rangle + kW_2(y_\varepsilon) \ge k$$

and

$$r\,|D\phi(x_\varepsilon)| + \langle (-A)^{1-\gamma} D\phi(x_\varepsilon), (-A)^{\gamma} x_\varepsilon\rangle - \langle D\phi(x_\varepsilon), f(x_\varepsilon)\rangle + kW_1(x_\varepsilon) \le k.$$

Taking the difference of the last two estimates and recalling assumption (H2), we conclude that

$$W_1(x_\varepsilon) - W_2(y_\varepsilon) \le \frac{r}{k}|D\phi(x_\varepsilon) - D\psi(y_\varepsilon)| + C\varepsilon(|x_\varepsilon|^2 + |y_\varepsilon|^2) + \frac{L}{k\varepsilon}|x_\varepsilon - y_\varepsilon|^2$$

$$\le \varepsilon C(|x_\varepsilon| + |y_\varepsilon| + |x_\varepsilon|^2 + |y_\varepsilon|^2) + C\sqrt{\varepsilon}.$$

Step 4. We make estimates on $|x_\varepsilon|, |y_\varepsilon|$. Given $\delta > 0$, let $x_\delta \in X$ be such that

$$W_1(x_\delta) - W_2(x_\delta) > \sup_{x\in X}(W_1(x) - W_2(x)) - \delta.$$

Then

$$W_1(x_\varepsilon) - W_2(y_\varepsilon) - \frac{1}{2\varepsilon}|x_\varepsilon - y_\varepsilon|^2 \le W_1(x_\varepsilon) - W_2(x_\varepsilon) + W_2(x_\varepsilon) - W_2(y_\varepsilon)$$

$$\le W_1(x_\delta) - W_2(x_\delta) + \delta + W_2(x_\varepsilon) - W_2(y_\varepsilon).$$

Therefore,

$$\begin{aligned}\frac{\varepsilon}{2}(|x_\varepsilon|^2 + |y_\varepsilon|^2) &\le -\Phi_\varepsilon(x_\varepsilon, y_\varepsilon) + W_1(x_\delta) - W_2(x_\delta) + \delta + W_2(x_\varepsilon) - W_2(y_\varepsilon)\\ &\le -\Phi_\varepsilon(x_\delta, x_\delta) + W_1(x_\delta) - W_2(x_\delta) + \delta + W_2(x_\varepsilon) - W_2(y_\varepsilon)\\ &= \varepsilon|x_\delta|^2 + \delta + W_2(x_\varepsilon) - W_2(y_\varepsilon).\end{aligned}$$

If $W_2(y_\varepsilon) = 1$, then $W_2(x_\varepsilon) - W_2(y_\varepsilon) \le 0$, so that

$$\frac{\varepsilon}{2}(|x_\varepsilon|^2 + |y_\varepsilon|^2) \le \varepsilon|x_\delta|^2 + \delta.$$

If, on the contrary, $W_2(y_\varepsilon) < 1$, then we proceed as in Step 1 by considering two cases. In case $\sqrt{2\varepsilon} \le \rho\sqrt{1 - W_2(y_\varepsilon)}$, we get

$$\frac{\varepsilon}{2}(|x_\varepsilon|^2 + |y_\varepsilon|^2) \le \varepsilon|x_\delta|^2 + \delta + M\sqrt{2\varepsilon},$$

whereas, in case $\sqrt{2\varepsilon} \ge \rho\sqrt{1 - W_2(y_\varepsilon)}$, we obtain

$$\frac{\varepsilon}{2}(|x_\varepsilon|^2 + |y_\varepsilon|^2) \le \varepsilon|x_\delta|^2 + C\varepsilon.$$

Since $\delta > 0$ is arbitrary, in both cases we have

$$\lim_{\varepsilon\to 0} \varepsilon(|x_\varepsilon|^2 + |y_\varepsilon|^2) = 0.$$

Step 5. Conclusion: for any $x \in X$ we have

$$W_1(x) - W_2(x) \le \Phi_\varepsilon(x_\varepsilon, y_\varepsilon) + \varepsilon\,|x|^2 \le W_1(x_\varepsilon) - W_2(y_\varepsilon) + \varepsilon|x|^2$$

$$\le C\varepsilon(|x_\varepsilon| + |y_\varepsilon| + |x_\varepsilon|^2 + |y_\varepsilon|^2) + C\sqrt{\varepsilon} + \varepsilon|x|^2.$$

Letting $\varepsilon \to 0$, we obtain $W_1(x) - W_2(x) \le 0$ for any $x \in X$, which proves the theorem. □

Remark 4.1. Theorem 4.1 remains true if (e) is weakened requiring that (i) and (ii) hold for $\gamma \in (1/(1+\alpha), 1)$, and also if (i) is replaced by

(i′) for any $p \in D^- W(x)$ we have

$$r\,|p| + \langle (-A)^{1-\gamma} p, (-A)^{\gamma} x\rangle - \langle p, f(x)\rangle + 2(L-\omega)W(x) \geq 2(L-\omega).$$

This remark prepares our second result on the Hamilton–Jacobi–Bellman equation, which can be applied even when the target is the origin of X.

THEOREM 4.2. *Assume* (H1), (H2), (H3), *and* (H5). *Suppose that the target is* B_R *with* $R \geq 0$. *Then* W *defined by* (13) *is the unique function from* X *to* $\mathbb{R}$ *which satisfies the following properties:*

(a) W *is sequentially weakly continuous.*

(b) $\partial_P W(x), \partial^P W(x) \subset D((-A)^{\theta}))$ *for any* $x \in X \setminus B_R$ *and for any* $\theta \in [0,1)$.

(c) *There exist* $\rho > 0$ *and* $M > 0$ *such that if* $|z-x| \leq \rho\sqrt{1-W(x)}$, *then*

$$W(z) - W(x) \leq M\sqrt{1-W(x)}|z-x|.$$

(d) $W(x) = 0$ *for any* $x \in B_R$, *and* $W(x) \in [0,1]$ *for any* $x \in X$.

(e) W *satisfies the Hamilton–Jacobi–Bellman equation in the following sense: for any* $x \in X \setminus B_R$ *with* $x \in \cap_{\theta\in(0,1)} D((-A)^{\theta})$ *and for any* $\gamma \in (0,1)$,

(i) *for any* $p \in \partial_P W(x)$ *we have*

$$r\,|p| + \langle (-A)^{1-\gamma} p, (-A)^{\gamma} x\rangle - \langle p, f(x)\rangle + 2(L-\omega)W(x) \geq 2(L-\omega);$$

(ii) *for any* $p \in \partial^P W(x)$ *we have*

$$r\,|p| + \langle (-A)^{1-\gamma} p, (-A)^{\gamma} x\rangle - \langle p, f(x)\rangle + 2(L-\omega)W(x) \leq 2(L-\omega).$$

Proof. To prove that W satisfies properties (a), (c), and (d), we argue as in Theorem 4.1. For $x \in \mathcal{R}$, (b) is satisfied because of Proposition 2.3, since $\partial_P g(x) \subseteq D^- g(x)$ and $\partial^P g(x) \subseteq D^+ g(x)$. In case $x \notin \mathcal{R}$, it follows by (14) that

$$\partial_P W(x) = \partial^P W(x) = \{0\}, \tag{21}$$

and hence (b) is satisfied. Let us prove (e). In virtue of (21), W satisfies both (i) and (ii) in case $x \notin \mathcal{R}$. To prove (ii) in case $x \in \mathcal{R}$, we argue as in Theorem 4.1. The only difference is in the use of inequality (16): in Theorem 4.1 it is a consequence of semiconcavity, while here we have the inequality (16) with $\alpha = 1$ as a consequence of the definition of $\partial^P W(x)$. A similar argument goes for the proof of (i). For the uniqueness part, the only modification in the proof of Theorem 4.1 is to set

$$D\phi(x_\varepsilon) \in \partial^P W_1(x_\varepsilon)$$

instead of (19), and

$$D\psi(y_\varepsilon) \in \partial_P W_2(y_\varepsilon)$$

instead of (20). This is possible because of Proposition 2.5. □

Remark 4.2. Theorem 4.2 holds true if in (e) we set that (i) and (ii) hold for any $\gamma \in (1/2, 1)$ and for any $x \in D((-A)^{\gamma})$.

Acknowledgement. The authors are grateful to the referees for valuable comments that improved the quality of this paper.

REFERENCES

[1] P. Albano, P. Cannarsa, and C. Sinestrari, *Regularity results for the minimum time function of a class of semilinear evolution equations of parabolic type*, SIAM J. Control Optim., 38 (2000), pp. 916–946.

[2] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser Boston, Cambridge, MA, 1997.

[3] V. Barbu, *The dynamic programming equation for the time-optimal control problem in infinite dimensions*, SIAM J. Control Optim., 29 (1991), pp. 445–456.

[4] P. Cannarsa and C. Sinestrari, *An infinite dimensional time optimal control problem*, Contemp. Math., 209 (1997), pp. 29–41.

[5] P. Cannarsa and M. E. Tessitore, *Infinite-dimensional Hamilton–Jacobi equations and Dirichlet boundary control problems of parabolic type*, SIAM J. Control Optim., 34 (1996), pp. 1831–1847.

[6] T. Cazenave and A. Haraux, *An Introduction to Semilinear Evolution Equations*, Clarendon Press, Oxford, UK, 1998.

[7] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer, New York, 1998.

[8] M. G. Crandall and P. L. Lions, *Hamilton–Jacobi equations in infinite dimensions. Part* I: *Uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379–396.

[9] M. G. Crandall and P. L. Lions, *Hamilton–Jacobi equations in infinite dimensions. Part* II: *Existence of viscosity solutions*, J. Funct. Anal., 65 (1986), pp. 368–405.

[10] M. G. Crandall and P. L. Lions, *Hamilton–Jacobi equations in infinite dimensions. Part* III., J. Funct. Anal., 68 (1986), pp. 214–247.

[11] M. G. Crandall and P. L. Lions, *Hamilton–Jacobi equations in infinite dimensions. Part* IV: *Hamiltonians with unbounded linear terms*, J. Funct. Anal., 90 (1990), pp. 237–283.

[12] M. G. Crandall and P. L. Lions, *Hamilton–Jacobi equations in infinite dimensions. Part* V: *Unbounded linear terms and B-continuous solutions*, J. Funct. Anal., 97 (1991), pp. 417–465.

[13] H. O. Fattorini, *Infinite Dimensional Optimization and Control Theory*, Cambridge University Press, New York, 1996.

[14] H. Ishii, *Viscosity solutions for a class of Hamilton–Jacobi equations in Hilbert spaces*, J. Funct. Anal., 105 (1992), pp. 301–341.

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

[16] D. Tataru, *Viscosity solutions of Hamilton–Jacobi equations with unbounded linear terms*, J. Math. Anal. Appl., 163 (1992), pp. 345–392.

[17] D. Tataru, *Viscosity solutions for the dynamic programming equation*, Appl. Math. Optim., 25 (1992), pp. 109–126.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 549–569

GLOBAL STEADY-STATE CONTROLLABILITY OF ONE-DIMENSIONAL SEMILINEAR HEAT EQUATIONS*

JEAN-MICHEL CORON† AND EMMANUEL TRÉLAT†

Abstract. We investigate the problem of exact boundary controllability of semilinear one-dimensional heat equations. We prove that it is possible to move from any steady-state to any other by means of a boundary control, provided that both are in the same connected component of the set of steady-states. The proof is based on an effective feedback stabilization procedure, which is implemented.

Key words. heat equation, controllability, pole shifting, Lyapunov functional

AMS subject classifications. 93B05, 93C20, 35B37

DOI. 10.1137/S036301290342471X

1. Introduction.

1.1. Statement of the main result. Let $L > 0$ be fixed and $f : \mathbb{R} \to \mathbb{R}$ be a function of class C^2. Let us consider the boundary control system

$$\begin{cases} \dfrac{\partial y}{\partial t} = \dfrac{\partial^2 y}{\partial x^2} + f(y), \\ y(t,0) = 0,\ y(t,L) = u(t), \end{cases} \tag{1.1}$$

where the state is $y(t,.) : [0,L] \to \mathbb{R}$ and the control is $u(t) \in \mathbb{R}$.

Concerning the global controllability problem, one of the main results [5] asserts that if f is globally Lipschitzian, then this control system is approximately globally controllable (see also [11] for exact controllability). When f is superlinear, the situation is still widely open, in particular because of possible blowing up. Indeed, it is well known that if $yf(y) > 0$ as $y \neq 0$, then blow-up phenomena may occur for the Cauchy problem

$$\begin{cases} \dfrac{\partial y}{\partial t} = \dfrac{\partial^2 y}{\partial x^2} + f(y), \\ y(t,0) = 0,\ y(t,L) = 0, \\ y(0,x) = y_0(x). \end{cases} \tag{1.2}$$

For example, if $f(y) = y^3$, then for numerous initial data there exists $T > 0$ such that the unique solution to the previous Cauchy problem is well defined on $[0,T) \times [0,L]$ and satisfies

$$\lim_{t \to T} \|y(t,.)\|_{L^\infty(0,L)} = +\infty$$

(see, for instance, [1, 8, 2, 12, 14, 15, 18] and references therein).

*Received by the editors March 23, 2003; accepted for publication (in revised form) December 8, 2003; published electronically July 23, 2004.
http://www.siam.org/journals/sicon/43-2/42471.html
†Département de Mathématiques, Université de Paris-Sud, Bâtiment 425, 91405 Orsay, France (Jean-Michel.Coron@math.u-psud.fr, Emmanuel.Trelat@math.u-psud.fr).

One may ask if, acting on the boundary of $[0, L]$, one could avoid the blow-up phenomenon. Actually the answer to this question is negative in general (see [7]; see also [6] for a weaker nonlinearity): for some nonlinear functions f satisfying

$$|f(y)| \sim |y| \log^p(1 + |y|) \quad \text{as } |y| \to +\infty,$$

with $p > 2$, and for any time $T > 0$, there exist initial data which lead to blow-up before time T, whatever the control function u is. Notice, however, that if

$$|f(y)| = \mathrm{o}\left(|y| \log^{3/2}(1 + |y|)\right) \quad \text{as } |y| \to +\infty,$$

then the blow-up (which could occur in the absence of control) can be avoided by means of boundary control (see [7]).

Nevertheless, in the first case where the blow-up phenomenon cannot be compensated by means of boundary control, the situation is not completely desperate. In fact, as we shall see in this paper, we can move from any given steady-state to any other belonging to the same connected component of the set of steady-states. More precisely, let us define the notion of steady-state.

DEFINITION 1.1. *A function $y \in C^2([0, L])$ is a steady-state of the control system* (1.1) *if*

$$\frac{d^2y}{dx^2} + f(y) = 0, \ y(0) = 0.$$

We denote by $\mathcal{S}$ the set of steady-states endowed with the C^2 topology.

Let us also introduce the Banach space

$$Y_T = \left\{ y(t,x), (t,x) \in (0,T) \times (0,L) \ / \ y \in L^2(0,T, W^{2,2}(0,L)) \right. \\ \left. \text{and } \frac{\partial y}{\partial t} \in L^2((0,T) \times (0,L)) \right\} \tag{1.3}$$

endowed with the norm

$$\|y\|_{Y_T} = \|y\|_{L^2(0,T,W^{2,2}(0,L))} + \left\| \frac{\partial y}{\partial t} \right\|_{L^2((0,T)\times(0,L))}.$$

Notice that Y_T is continuously imbedded in $L^\infty((0,T) \times (0,L))$.

The main result of the paper is the following.

THEOREM 1.2. *Let y_0 and y_1 be two steady-states belonging to a same connected component of $\mathcal{S}$. There exist a time $T > 0$ and a control function $u \in L^2(0,T)$ such that the solution $y(t,x)$ in Y_T of*

$$\begin{cases} \dfrac{\partial y}{\partial t} = \dfrac{\partial^2 y}{\partial x^2} + f(y), \\ y(t,0) = 0, \ y(t,L) = u(t), \\ y(0,x) = y_0(x) \end{cases} \tag{1.4}$$

satisfies $y(T,.) = y_1(.)$.

Remark 1.3. In fact, we prove the following result: for all neighborhood V of y_1 in H^1-topology, there exists a positive real number ε_0 such that for all $\varepsilon \in (0, \varepsilon_0)$

there exists a control function $u \in H^1(0, 1/\varepsilon)$ such that the solution $y(t, x)$ in Y_T of the Cauchy–Dirichlet problem (1.4) satisfies $y(1/\varepsilon, .) \in V$.

In the proof of this result, which represents the main part of the paper, we give an explicit construction of the control u in a *feedback-type form* and also of a *Lyapunov functional*. We stress that the procedure is effective and consists actually of solving a stabilization problem in finite dimension. Indeed, in order to construct u we need to compute only a finite number of quantities related to a Hilbertian expansion of the solution. The procedure has been implemented numerically, and simulations are presented in the last section of the paper.

Remark 1.4. For any $T > 0$ and $u \in L^2(0, T)$, there is at most one solution of (1.4) in the Banach space Y_T.

Remark 1.5. This is a (partial) global exact controllability result. The time needed in our proof is large, but on the other hand there are indeed cases where the time T of controllability cannot be taken arbitrarily small. For instance, in the case where $f(y) = -y^3$, any solution of (1.4) starting from 0 satisfies the inequality

$$\int_0^L (L - x)^4 y(T, x)^2 dx \leq 8LT,$$

and hence, if $y_0 = 0$, a minimal time is needed to reach a given $y_1 \neq 0$. This result is due to Bamberger [10] (see also [9, Lemma 2.1]).

Remark 1.6. In section 3 we prove that if y_0 and y_1 belong to distinct connected components of $\mathcal{S}$, then it is actually impossible to move either from y_0 to y_1 or from y_1 to y_0, whatever the time and the control are. In the same section we also investigate the connectedness of the set $\mathcal{S}$ of steady-states.

Remark 1.7. The result of Theorem 1.2 may be achieved directly by using repeatedly a local exact controllability theorem (see [9, Theorem 4.4] or [11, Theorem 3.3]). Here we present a new controllability strategy based on a feedback stabilization procedure, which is more effective. It is clear also that this approach may be applied to other problems without requiring controllability of the linearized system around an equilibrium (see [3]).

1.2. The idea of the proof. The method we shall use to prove Theorem 1.2 stems from classical Lyapunov stability theory together with quasi-static deformation theory. For the sake of simplicity we explain it in finite dimension. Let us consider in $\mathbb{R}^n$ a general control system of the form

$$\dot{y}(t) = g(y(t), u(t)), \tag{1.5}$$

where $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is of class C^1, $u(t) \in \mathcal{U}$, and $\mathcal{U}$ denotes the set of measurable essentially bounded admissible controls. Let $y_0, y_1 \in \mathbb{R}^n$ be two equilibrium points of system (1.5); that is,

$$g(y_i, u_i) = 0, \ i = 0, 1,$$

for some $u_0, u_1 \in \mathbb{R}^m$. We assume that (y_0, u_0) and (y_1, u_1) belong to the same connected component of the zero set of g in $\mathbb{R}^n \times \mathbb{R}^m$. Our aim is to steer the system from y_0 to y_1 in some (large) time $T > 0$. The method splits into four steps.

First step. Construct a C^1-path $(\bar{y}(\tau), \bar{u}(\tau))$, with $\tau \in [0, 1]$, connecting (y_0, u_0) to (y_1, u_1) and such that

$$\forall \tau \in [0, 1], \quad g(\bar{y}(\tau), \bar{u}(\tau)) = 0.$$

Of course, this path is not in general a solution of system (1.5), but if $\varepsilon > 0$ is small enough, then the C^1-path $(y^\varepsilon, u^\varepsilon)$

$$\begin{array}{rcl} [0, 1/\varepsilon] & \to & \mathbb{R}^n \times \mathbb{R}^m, \\ t & \mapsto & (y^\varepsilon(t), u^\varepsilon(t)) = (\bar{y}(\varepsilon t), \bar{u}(\varepsilon t)) \end{array}$$

is "almost" a solution of system (1.5). Indeed,

$$\|\dot{y}^\varepsilon - g(y^\varepsilon, u^\varepsilon)\| = \mathrm{O}(\varepsilon).$$

Second step. This quasi-static trajectory is not in general stable, and thus has to be stabilized. To this aim, introduce the following change of variable:

$$\begin{aligned} z(t) &= y(t) - y^\varepsilon(t), \\ v(t) &= u(t) - u^\varepsilon(t), \end{aligned}$$

where $t \in [0, 1/\varepsilon]$. In the new variables z, v, the control system writes, at least if $\|z(t)\| + \|v(t)\|$ is small enough,

$$\dot{z}(t) = A(\varepsilon t)z(t) + B(\varepsilon t)v(t) + \mathrm{O}(\|z(t)\|^2 + \|v(t)\|^2 + \varepsilon),$$

where $t \in [0, 1/\varepsilon]$, and where

$$A(\tau) = \frac{\partial g}{\partial y}(\bar{y}(\tau), \bar{u}(\tau)) \quad \text{and} \quad B(\tau) = \frac{\partial g}{\partial u}(\bar{y}(\tau), \bar{u}(\tau)),$$

with $\tau = \varepsilon t \in [0, 1]$. Therefore we have to stabilize near the origin a *slowly-varying in time* linear control system; we refer to [13] for this classical theory.

Third step. Under mild controllability assumptions, namely,

$$\forall \tau \in [0, 1], \quad \operatorname{rank}\big(B(\tau), A(\tau)B(\tau), \dots, A(\tau)^{n-1}B(\tau)\big) = n$$

(Kalman condition), it is actually possible to stabilize the system by *pole shifting* and to construct a quadratic Lyapunov function. Notice that this does not work in general if the system is not slowly-varying. So if ε is small enough, then using this Lyapunov function we infer that $y(1/\varepsilon)$ belongs to some prescribed neighborhood of the target y_1. At this stage, a stabilization result is achieved.

Fourth step. If the system (1.5) is *locally controllable* near the point y_1, we conclude that it is possible to steer the system in finite time from the point $y(1/\varepsilon)$ to the desired target y_1. Usually such a local controllability result is achieved by using an implicit function argument, after proving that the linearized system is controllable.

Remark 1.8. The use of quasi-static deformation for the controllability of a nonlinear partial differential control system has already been used in [3]. But note that in [3] the quasi-static trajectory $(y^\varepsilon, u^\varepsilon)$ was stable, so it was not necessary to perform steps 2 and 3.

2. Proof of the main results. In order to prove Theorem 1.2 we shall follow exactly the steps described previously.

2.1. Construction of a path of steady-states. The following lemma is obvious.

LEMMA 2.1. *Let $\phi_0, \phi_1 \in \mathcal{S}$. Then ϕ_0 and ϕ_1 belong to the same connected component of $\mathcal{S}$ if and only if, for any real number α between $\phi_0'(0)$ and $\phi_1'(0)$, the maximal solution of*

$$\frac{d^2y}{dx^2} + f(y) = 0, \ y(0) = 0, y'(0) = \alpha,$$

denoted by $y^\alpha(.)$, is defined on $[0, L]$.

Now let y_0 and y_1 be in the same connected component of $\mathcal{S}$. Let us construct in $\mathcal{S}$ a C^1-path $(\bar{y}(\tau, .), \bar{u}(\tau))$, $0 \le \tau \le 1$, joining y_0 to y_1. For each $i = 0, 1$, set

$$\alpha_i = y_i'(0).$$

Then with our previous notation, $y_i(.) = y^{\alpha_i}(.), i = 0, 1$. Now set

$$\bar{y}(\tau, x) = y^{(1-\tau)\alpha_0 + \tau\alpha_1}(x) \text{ and } \bar{u}(\tau) = \bar{y}(\tau, L),$$

where $\tau \in [0, 1]$ and $x \in [0, L]$. By construction we have

$$\bar{y}(0, .) = y_0(.),\ \bar{y}(1, .) = y_1(.) \text{ and } \bar{u}(0) = \bar{u}(1) = 0,$$

and thus $(\bar{y}(\tau, .), \bar{u}(\tau))$ is a C^1-path in $\mathcal{S}$ connecting y_0 to y_1.

2.2. Reduction of the problem. Let $\varepsilon > 0$. We set, for any $t \in [0, 1/\varepsilon]$ and any $x \in [0, L]$,

$$\begin{aligned} z(t, x) &= y(t, x) - \bar{y}(\varepsilon t, x), \\ v(t) &= u(t) - \bar{u}(\varepsilon t). \end{aligned} \tag{2.1}$$

Then from the definition of $(\bar{y}, \bar{u})$ we infer that z satisfies the initial boundary problem

$$\begin{cases} z_t = z_{xx} + f'(\bar{y})z + z^2 \displaystyle\int_0^1 (1-s) f''(\bar{y} + sz) ds - \varepsilon \bar{y}_\tau, \\ z(t, 0) = 0,\ z(t, L) = v(t), \\ z(0, x) = 0. \end{cases} \tag{2.2}$$

Now, in order to deal with a Dirichlet-type problem instead, we set

$$w(t, x) = z(t, x) - \frac{x}{L} v(t), \tag{2.3}$$

and we suppose that the control v is derivable. This leads to the equation

$$\begin{cases} w_t = w_{xx} + f'(\bar{y})w + \dfrac{x}{L} f'(\bar{y})v - \dfrac{x}{L} v' + r(\varepsilon, t, x), \\ w(t, 0) = w(t, L) = 0, \\ w(0, x) = -\dfrac{x}{L} v(0), \end{cases} \tag{2.4}$$

where

$$r(\varepsilon, t, x) = -\varepsilon \bar{y}_\tau + \left(w + \frac{x}{L} v\right)^2 \int_0^1 (1-s) f'' \left(\bar{y} + s\left(w + \frac{x}{L} v\right)\right) ds, \tag{2.5}$$

and the next step is to prove that there exist ε small enough and a pair (v, w) solution of (2.4) such that $w(1/\varepsilon, .)$ belongs to some arbitrary neighborhood of 0 in H_0^1-topology. To achieve this we shall construct an appropriate control function and a Lyapunov functional which stabilizes system (2.4) to 0.

In fact, as we shall see, the control will be chosen in $H^1(0, 1/\varepsilon)$ and such that $v(0) = 0$.

2.3. Construction of a Lyapunov functional. This is the most technical part of the work. In order to motivate what follows, let us first notice that if the residual term r and the control v were equal to zero, then (2.4) would reduce to

$$w_t = w_{xx} + f'(\bar{y})w,$$
$$w(t,0) = w(t,L) = 0.$$

This suggests that we introduce the *one-parameter family of linear operators*

$$A(\tau) = \Delta + f'(\bar{y}(\tau,.))Id,\ \tau \in [0,1], \tag{2.6}$$

defined on $H^2(0,L) \cap H_0^1(0,L)$. Let $(e_j(\tau,.))_{j\geq 1}$ be a Hilbertian basis of $L^2(0,L)$ of eigenfunctions of $A(\tau)$, such that for each $j \geq 1$ and each $\tau \in [0,1]$,

$$e_j(\tau,.) \in H_0^1(0,L) \cap C^2([0,L]),$$

and let $(\lambda_j(\tau))_{j\geq 1}$ denote the corresponding eigenvalues. A standard application of the minimax principle (see, for instance, [16]) shows that these eigenfunctions and eigenvalues are C^1 functions of τ. Moreover, for each $\tau \in [0,1]$,

$$-\infty < \cdots < \lambda_n(\tau) < \cdots < \lambda_1(\tau) \quad \text{and} \quad \lambda_n(\tau) \underset{n\to+\infty}{\longrightarrow} -\infty.$$

From the continuity of the eigenvalues on $[0,1]$, we can define n as the maximal number of eigenvalues taking at least a nonnegative value as $\tau \in [0,1]$; i.e., there exists $\eta > 0$ such that

$$\forall t \in [0,1/\varepsilon], \quad \forall k > n, \quad \lambda_k(\varepsilon t) < -\eta < 0. \tag{2.7}$$

Remark 2.2. Note that the integer n can be arbitrarily large. For example, if $f(y) = y^3$ and if $y_1'(0) \to +\infty$, then $n \to +\infty$.

We also set, for any $\tau \in [0,1]$ and $x \in [0,L]$,

$$a(\tau,x) = \frac{x}{L} f'(\bar{y}(\tau,x)) \ \text{ and } \ b(x) = -\frac{x}{L}.$$

In this notation, system (2.4) leads to

$$w_t(t,.) = A(\varepsilon t)w(t,.) + a(\varepsilon t,.)v(t) + b(.)v'(t) + r(\varepsilon,t,.). \tag{2.8}$$

Any solution $w(t,.) \in H^2(0,L) \cap H_0^1(0,L)$ of (2.8) can be expanded as a series in the eigenfunctions $e_j(\varepsilon t,.)$, convergent in $H_0^1(0,L)$:

$$w(t,.) = \sum_{j=1}^{\infty} w_j(t) e_j(\varepsilon t,.).$$

In fact, the w_j's depend on ε and should be called, for example, w_j^ε. For simplicity we omit the index ε, and we shall also omit the index ε for other functions.

In what follows we are going to move, by means of an appropriate *feedback control*, the n first eigenvalues of the operator A, without moving the others, in order to make all eigenvalues negative. This pole shifting process is the first part of the stabilization procedure (see [17, p. 711]).

For any $\tau \in [0,1]$, let $\pi_1(\tau)$ denote the orthogonal projection onto the subspace of $L^2(0,L)$ spanned by $e_1(\tau,.),\dots,e_n(\tau,.)$, and let

$$w^1(t) = \pi_1(\varepsilon t)w(t,.) = \sum_{j=1}^{n} w_j(t)e_j(\varepsilon t,.). \tag{2.9}$$

It is clear that for any τ, the operators $\pi_1(\tau)$ and $A(\tau)$ commute, and moreover, for any $y \in L^2(0,L)$, we have

$$\pi_1'(\tau)y = \sum_{j=1}^{n} \langle y, e_j(\tau,.)\rangle_{L^2(0,L)} \frac{\partial e_j}{\partial \tau}(\tau,.) + \sum_{j=1}^{n} \left\langle y, \frac{\partial e_j}{\partial \tau}(\tau,.)\right\rangle_{L^2(0,L)} e_j(\tau,.).$$

Hence, derivating (2.9) with respect to t, we get

$$\sum_{j=1}^{n} w_j'(t)e_j(\varepsilon t,.) = \pi_1(\varepsilon t)w_t(t,.) + \varepsilon \sum_{j=1}^{n} \left\langle w(t,.), \frac{\partial e_j}{\partial \tau}(\varepsilon t,.)\right\rangle_{L^2(0,L)} e_j(\varepsilon t,.).$$

On the other hand

$$A(\varepsilon t)w^1(t) = \sum_{j=1}^{n} \lambda_j(\varepsilon t)w_j(t)e_j(\varepsilon t,.),$$

and thus (2.8) yields

$$\begin{aligned}\sum_{j=1}^{n} w_j'(t)e_j(\varepsilon t,.) = &\sum_{j=1}^{n} \lambda_j(\varepsilon t)w_j(t)e_j(\varepsilon t,.) + \pi_1(\varepsilon t)a(\varepsilon t,.)v(t)\\ &+ \pi_1(\varepsilon t)b(.)v'(t) + r^1(\varepsilon,t,.),\end{aligned} \tag{2.10}$$

where

$$r^1(\varepsilon,t,.) = \pi_1(\varepsilon t)r(\varepsilon,t,.) + \varepsilon \sum_{j=1}^{n} \left\langle w, \frac{\partial e_j}{\partial \tau}(\varepsilon t,.)\right\rangle_{L^2(0,L)} e_j(\varepsilon t,.). \tag{2.11}$$

Let us set an upper bound to the residual term r^1. First, it is not difficult to check that there exists a constant C such that, if $|v(t)|$ and $\|w(t,.)\|_{L^\infty(0,L)}$ are less than 1, then the inequality

$$\|r(\varepsilon,t,.)\|_{L^\infty(0,L)} \le C(\varepsilon + v(t)^2 + \|w(t,.)\|^2_{L^\infty(0,L)})$$

holds, where r is defined by (2.5). Therefore we get easily

$$\|r^1(\varepsilon,t,.)\|_{L^\infty(0,L)} \le C_1(\varepsilon + v(t)^2 + \|w(t,.)\|^2_{L^\infty(0,L)}).$$

Moreover, since $H^1(0,L)$ is continuously imbedded in $C^0([0,L])$, we can assert that there exists a constant C_2 such that, if $|v(t)|$ and $\|w(t,.)\|_{L^\infty(0,L)}$ are less than 1, then

$$\|r^1(\varepsilon,t,.)\|_{L^\infty(0,L)} \le C_2(\varepsilon + v(t)^2 + \|w(t,.)\|^2_{H_0^1(0,L)}). \tag{2.12}$$

Now projecting (2.10) on each $e_i, i = 1,\dots,n$, one comes to

$$w_i'(t) = \lambda_i(\varepsilon t)w_i(t) + a_i(\varepsilon t)v(t) + b_i(\varepsilon t)v'(t) + r_i^1(\varepsilon,t), \quad i = 1,\dots,n, \tag{2.13}$$

where

$$
\begin{aligned}
r_i^1(\varepsilon,t) &= \langle r^1(\varepsilon,t,.), e_i(\varepsilon t,.)\rangle_{L^2(0,L)},\\
a_i(\varepsilon t) &= \langle a(\varepsilon t,.), e_i(\varepsilon t,.)\rangle_{L^2(0,L)} = \frac{1}{L}\int_0^L x f'(\bar y(\varepsilon t,x)) e_i(\varepsilon t,x)dx,\\
b_i(\varepsilon t) &= \langle b(.), e_i(\varepsilon t,.)\rangle_{L^2(0,L)} = -\frac{1}{L}\int_0^L x e_i(\varepsilon t,x)dx.
\end{aligned}
\tag{2.14}
$$

The n equations (2.13) form a differential system controlled by v, v'. Set

$$
\alpha(t) = v'(t), \tag{2.15}
$$

and consider now $v(t)$ as a state and $\alpha(t)$ as a control. Then the former finite dimensional system may be rewritten as

$$
\begin{cases}
v' = \alpha,\\
w_1' = \lambda_1 w_1 + a_1 v + b_1\alpha + r_1^1,\\
\quad\vdots\\
w_n' = \lambda_n w_n + a_n v + b_n\alpha + r_n^1.
\end{cases}
\tag{2.16}
$$

If we introduce the matrix notation

$$
X_1(t) = \begin{pmatrix} v(t)\\ w_1(t)\\ \vdots\\ w_n(t)\end{pmatrix},\ R_1(\varepsilon,t) = \begin{pmatrix} 0\\ r_1^1(\varepsilon,t)\\ \vdots\\ r_n^1(\varepsilon,t)\end{pmatrix},
$$

$$
A_1(\tau) = \begin{pmatrix} 0 & 0 & \cdots & 0\\ a_1(\tau) & \lambda_1(\tau) & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ a_n(\tau) & 0 & \cdots & \lambda_n(\tau)\end{pmatrix},\ B_1(\tau) = \begin{pmatrix} 1\\ b_1(\tau)\\ \vdots\\ b_n(\tau)\end{pmatrix},
$$

then equations (2.16) yield the *finite dimensional linear control system*

$$
X_1'(t) = A_1(\varepsilon t)X_1(t) + B_1(\varepsilon t)\alpha(t) + R_1(\varepsilon,t). \tag{2.17}
$$

Let us now prove the following lemma.

LEMMA 2.3. *For each $\tau \in [0,1]$, the pair $(A_1(\tau), B_1(\tau))$ satisfies the Kalman condition, i.e.,*

$$
\operatorname{rank}\big(B_1(\tau), A_1(\tau)B_1(\tau), \dots, A_1(\tau)^{n-1}B_1(\tau)\big) = n. \tag{2.18}
$$

Proof. Let $\tau \in [0,1]$ be fixed. We compute directly

$$
\det\big(B_1, A_1B_1, \dots, A_1^{n-1}B_1\big) = \prod_{j=1}^n (a_j + \lambda_j b_j)\ \mathrm{VdM}(\lambda_1,\dots,\lambda_n), \tag{2.19}
$$

where $\mathrm{VdM}(\lambda_1, \ldots, \lambda_n)$ is a Van der Monde determinant and thus is never equal to zero, since the $\lambda_i(\tau), i = 1, \ldots, n$, are distinct for any $\tau \in [0,1]$. On the other hand, using the fact that each $e_j(\tau, .)$ is an eigenfunction of $A(\tau)$ and belongs to $H_0^1(0,L)$, we compute

$$\begin{aligned} a_j(\tau) + \lambda_j(\tau) b_j(\tau) &= \frac{1}{L} \int_0^L x \left(f'(\bar{y}(\tau,x)) e_j(\tau,x) - \lambda_j(\tau) e_j(\tau,x) \right) dx \\ &= -\frac{1}{L} \int_0^L x \frac{\partial^2 e_j}{\partial x^2}(\tau,x) dx \\ &= -\frac{\partial e_j}{\partial x}(\tau, L). \end{aligned}$$

But this quantity is never equal to zero since $e_j(\tau, L) = 0$ and $e_j(\tau, .)$ is a nontrivial solution of a linear second-order scalar differential equation. Therefore the determinant (2.19) is never equal to zero and we are done. □

It is a standard fact that the Kalman condition (2.18) implies a *pole shifting* result, and we get the following corollary (see [13]).

COROLLARY 2.4. *For each $\tau \in [0,1]$, there exist scalars $k_0(\tau), \ldots, k_n(\tau)$ such that, if we denote*

$$K_1(\tau) = (k_0(\tau), \ldots, k_n(\tau)),$$

then the matrix $A_1(\tau) + B_1(\tau) K_1(\tau)$ admits -1 as an eigenvalue with order $n+1$.

Moreover, there exists a C^1 application $\tau \mapsto P(\tau)$ on $[0,1]$, where $P(\tau)$ is an $(n+1) \times (n+1)$ symmetric positive definite matrix, such that the identity

$$P(\tau) \left(A_1(\tau) + B_1(\tau) K_1(\tau) \right) + {}^t\left(A_1(\tau) + B_1(\tau) K_1(\tau) \right) P(\tau) = -I \tag{2.20}$$

holds for any $\tau \in [0,1]$.

We are now able to construct a control Lyapunov functional in order to stabilize system (2.8). Leave $c > 0$ to be chosen later. For any $t \in [0, 1/\varepsilon]$, $v \in \mathbb{R}$, and $w \in H^2(0,L) \cap H_0^1(0,L)$, we set

$$V(t,v,w) = c\, {}^t X_1(t) P(\varepsilon t) X_1(t) - \frac{1}{2} \langle w, A(\varepsilon t) w \rangle_{L^2(0,L)}, \tag{2.21}$$

where $X_1(t)$ denotes the matrix vector in $\mathbb{R}^{n+1}$,

$$X_1(t) = \begin{pmatrix} v \\ w_1(t) \\ \vdots \\ w_n(t) \end{pmatrix},$$

and

$$w_i(t) = \langle w, e_i(\varepsilon t, .) \rangle_{L^2(0,L)}.$$

In particular, we have

$$V(t,v,w) = c\, {}^t X_1(t) P(\varepsilon t) X_1(t) - \frac{1}{2} \sum_{j=1}^{\infty} \lambda_j(\varepsilon t) w_j(t)^2. \tag{2.22}$$

In what follows we will repeatedly use the equivalence of norms in finite dimension. The following notation will thus be useful.

NOTATION. *Let Λ be a set and let $\Delta = \{(\varepsilon, t) \ / \ 0 < \varepsilon \leq 1,\ 0 \leq t \leq 1/\varepsilon\}$. Let F_1, F_2 be two real functions defined on $\Delta \times \Lambda$. The notation $F_1 \lesssim F_2$ means that $F_2 \geq 0$ and that there exists a positive constant C such that*

$$F_1(\varepsilon, t, \lambda) \leq C F_2(\varepsilon, t, \lambda) \quad \forall (\varepsilon, t) \in \Delta, \quad \forall \lambda \in \Lambda.$$

We say that $F_1 \sim F_2$ if both $F_1 \lesssim F_2$ and $F_2 \lesssim F_1$. Moreover, if F_3 is a real function defined on $\Delta \times \Lambda$ and if $\theta \in [0, +\infty)$, the notation $F_1 \lesssim F_2$ for $F_3 \leq \theta$ means that $F_2 \geq 0$ and that there exists a positive constant C such that

$$\forall (\varepsilon, t) \in \Delta, \quad \forall \lambda \in \Lambda, \quad (F_3(\varepsilon, t, \lambda) \leq \theta) \Rightarrow (F_1(\varepsilon, t, \lambda) \leq C F_2(\varepsilon, t, \lambda)) \,.$$

For simplicity, when the set Λ is clear from the context it will not be given explicitly.

Let $\| \ \|_2$ denote the euclidean norm in $\mathbb{R}^{n+1}$. Since $P(\tau)$ is symmetric positive definite, we can write (with $\Lambda = \mathbb{R} \times (H^2(0, L) \cap H_0^1(0, L))$)

$$^t X_1(t) P(\varepsilon t) X_1(t) \sim \|X_1(t)\|_2^2 = v^2 + \sum_{j=1}^{n} w_j(t)^2.$$

From (2.7) we know that, except the n first ones, the eigenvalues of A are all negative, less than $-\eta < 0$. By continuity the n first eigenvalues are bounded as $\tau \in [0, 1]$ and thus we can assert that, if c is large enough in the definition of V, then

$$V(t, v, w) \sim \|X_1(t)\|_2^2 - \sum_{j=n+1}^{\infty} \lambda_j(\varepsilon t) w_j(t)^2, \tag{2.23}$$

where $t \in [0, 1/\varepsilon]$. In particular, $V(t, ., .)$ is positive definite. Let us further prove the following lemma.

LEMMA 2.5. *The equivalence*

$$V(t, v, w) \sim v^2 + \|w\|_{H_0^1(0,L)}^2 \tag{2.24}$$

holds with $\Lambda = \{(v, w) \ / \ v \in \mathbb{R}, w \in H^2(0, L) \cap H_0^1(0, L)\}$, where $\|w\|_{H_0^1(0,L)} = \|w_x\|_{L^2(0,L)}$. Moreover,

$$V(t, v, w) \lesssim \|X_1(t)\|_2^2 + \|Aw\|_{L^2(0,L)}^2. \tag{2.25}$$

Proof. Any $w \in H^2(0, L) \cap H_0^1(0, L)$ can be expanded as a series in the eigenfunctions of $A(\varepsilon t)$, convergent in $H_0^1(0, L)$,

$$w(.) = \sum_i w_i(t) e_i(\varepsilon t, .).$$

Hence

$$\|w\|_{H_0^1(0,L)}^2 = \sum_{i,j} w_i(t) w_j(t) \int_0^L e_{ix}(\varepsilon t, x) e_{jx}(\varepsilon t, x) dx.$$

Integrating by parts and using the definition of e_j, we compute

$$\int_0^L e_{ix}(\varepsilon t,x)e_{jx}(\varepsilon t,x)dx=\int_0^L f'(\bar y(\varepsilon t,x))e_i(\varepsilon t,x)e_j(\varepsilon t,x)dx-\lambda_j\delta_{ij},$$

and thus

$$\|w\|^2_{H_0^1(0,L)}=\int_0^L f'(\bar y(\varepsilon t,x))w(x)^2dx-\sum_{j=1}^{\infty}\lambda_j(\varepsilon t)w_j(t)^2.$$

Therefore, since $f'(\bar y)$ is bounded on $[0,1/\varepsilon]\times[0,L]$ uniformly in $\varepsilon\in(0,1]$,

$$\|w\|^2_{H_0^1(0,L)}\ \lesssim\ \|w\|^2_{L^2(0,L)}-\sum_{j=n+1}^{\infty}\lambda_j(\varepsilon t)w_j(t)^2\ \lesssim\ V(t,v,w).$$

Conversely, we have

$$\begin{aligned}-\sum_{j=n+1}^{\infty}\lambda_j(\varepsilon t)w_j(t)^2&=\|w\|^2_{H_0^1(0,L)}-\int_0^L f'(\bar y(\varepsilon t,x))w(x)^2dx+\sum_{j=1}^{n}\lambda_j(\varepsilon t)w_j(t)^2\\&\lesssim\|w\|^2_{H_0^1(0,L)}+\|w\|^2_{L^2(0,L)}\\&\lesssim\|w\|^2_{H_0^1(0,L)},\end{aligned}$$

and using (2.23) we conclude easily that (2.24) holds.

On the other hand, notice that

$$\|w\|^2_{H_0^1(0,L)}\lesssim\sum_{j=n+1}^{\infty}|\lambda_j|w_j^2+\sum_{j=1}^{n}w_j^2.$$

Therefore, using (2.7),

$$\|w\|^2_{H_0^1(0,L)}\lesssim\sum_{j=1}^{n}w_j^2+\sum_{j=1}^{\infty}\lambda_j^2w_j^2=\sum_{j=1}^{n}w_j^2+\|Aw\|^2_{L^2(0,L)},$$

and hence the estimate (2.25) follows. □

Let now $(v(t),w(t,.))$ denote a solution of (2.8) in which we choose the control in the feedback form suggested from Corollary 2.4, namely,

$$\alpha(t)=K_1(\varepsilon t)X_1(t),$$

such that $v(0)=0$ and $w(0,.)=0$, i.e., $(v(t),w(t,.))$ satisfies

$$w_t=Aw+av+bK_1(\varepsilon t)X_1(t)+r,\quad v'(t)=K_1(\varepsilon t)X_1(t), \tag{2.26}$$

$$v(0)=0,\quad w(0,x)=0. \tag{2.27}$$

We set

$$V_1(t)=V(t,v(t),w(t,.))=c\,{}^tX_1(t)P(\varepsilon t)X_1(t)-\frac{1}{2}\langle w(t,.),A(\varepsilon t)w(t,.)\rangle_{L^2(0,L)}.$$

Let us compute $V_1'(t)$ and state a differential inequality satisfied by V_1. We have

$$\begin{aligned} V_1'(t) &= c\left({}^tX_1'(t)P(\varepsilon t)X_1(t) + {}^tX_1(t)P(\varepsilon t)X_1'(t)\right) + \varepsilon c\,{}^tX_1(t)P'(\varepsilon t)X_1(t) \\ &\quad -\frac{1}{2}\langle w_t(t,.), A(\varepsilon t)w(t,.)\rangle_{L^2(0,L)} - \frac{1}{2}\langle w(t,.), A(\varepsilon t)w_t(t,.)\rangle_{L^2(0,L)} \\ &\quad -\frac{1}{2}\varepsilon\langle w(t,.), A'(\varepsilon t)w(t,.)\rangle_{L^2(0,L)}. \end{aligned} \tag{2.28}$$

Note the following facts:

- From (2.17) and (2.20), we infer

$$ {}^tX_1'PX_1 + {}^tX_1PX_1' = -\|X_1\|_2^2 + {}^tR_1PX_1 + {}^tX_1PR_1. $$

- The operator A is self-adjoint in L^2, hence

$$ \langle w, Aw_t\rangle_{L^2(0,L)} = \langle Aw, w_t\rangle_{L^2(0,L)}. $$

- Equation (2.26) leads to

$$\begin{aligned} \langle Aw, w_t\rangle_{L^2(0,L)} &= \langle Aw, Aw + av + bK_1X_1 + r\rangle_{L^2(0,L)} \\ &= \|Aw\|_{L^2(0,L)} + \langle Aw, a\rangle_{L^2(0,L)}v \\ &\quad + \langle Aw, b\rangle_{L^2(0,L)}K_1X_1 + \langle Aw, r\rangle_{L^2(0,L)}. \end{aligned}$$

- From the definition of $A(\tau)$, we have

$$ A'(\tau) = f''(\bar y(\tau,.))\bar y_\tau(\tau,.)Id, $$

and thus

$$ \langle w(t,.), A'(\varepsilon t)w(t,.)\rangle_{L^2(0,L)} = \langle w(t,.), f''(\bar y(\varepsilon t,.))\bar y_\tau(\varepsilon t,.)w(t,.)\rangle_{L^2(0,L)}. $$

Therefore, turning back to (2.28),

$$\begin{aligned} V_1' =& -c\,\|X_1\|_2^2 - \|Aw\|_{L^2(0,L)}^2 - \langle Aw, a\rangle_{L^2(0,L)}v - \langle Aw, b\rangle_{L^2(0,L)}K_1X_1 \\ &- \langle Aw, r\rangle_{L^2(0,L)} + \varepsilon c\,{}^tX_1P'X_1 + c\left({}^tR_1PX_1 + {}^tX_1PR_1\right) \\ &- \frac{1}{2}\varepsilon\langle w, f''(\bar y)\bar y_\tau w\rangle_{L^2(0,L)}. \end{aligned} \tag{2.29}$$

Let us set an upper bound for the terms of the second line of (2.29), as follows:

- From Corollary 2.4, the application $\tau \mapsto P'(\tau)$ is bounded on $[0,1]$, hence

$$ \left|\varepsilon c{}^tX_1P'X_1\right| \lesssim \varepsilon\|X_1\|_2^2 \lesssim \varepsilon V_1. $$

- Inequality (2.12) yields, for $|v(t)| + \|w(t,.)\|_{L^\infty(0,L)} \le 1$,

$$ \|R_1(\varepsilon,t)\|_{L^\infty(0,L)} \lesssim \varepsilon + v(t)^2 + \|w(t,.)\|_{H_0^1(0,L)}^2, $$

and thus, still for $|v(t)| + \|w(t,.)\|_{L^\infty(0,L)} \le 1$,

$$\begin{aligned} {}^tR_1PX_1 + {}^tX_1PR_1 &\lesssim \|X_1\|_2\left(\varepsilon + v^2 + \|w\|_{H_0^1(0,L)}^2\right) \\ &\lesssim \sqrt{V_1}\,(\varepsilon + V_1) = \varepsilon\sqrt{V_1} + V_1^{3/2}. \end{aligned}$$

- Since f is of class C^2, we can assert

$$\varepsilon\left|\frac{1}{2}\langle w, f''(\bar{y})\bar{y}_\tau w\rangle_{L^2(0,L)}\right| \lesssim \varepsilon\|w\|^2_{L^2(0,L)} \lesssim \varepsilon\|w\|^2_{H_0^1(0,L)} \lesssim \varepsilon V_1.$$

- The term $\langle Aw, r\rangle_{L^2(0,L)}$ is the most difficult to handle. Using (2.5), write

$$\begin{aligned}\langle Aw, r\rangle_{L^2(0,L)} &= \left\langle Aw, -\varepsilon\bar{y}_\tau + \Big(w+\frac{x}{L}v\Big)^2\int_0^1(1-s)f''\Big(\bar{y}+s(w+\frac{x}{L}v)\Big)ds\right\rangle_{L^2(0,L)}\\ &= \left\langle Aw, \Big(w+\frac{x}{L}v\Big)^2\int_0^1(1-s)f''\Big(\bar{y}+s(w+\frac{x}{L}v)\Big)ds\right\rangle_{L^2(0,L)}\\ &\quad -\varepsilon\langle Aw, \bar{y}_\tau\rangle_{L^2(0,L)}.\end{aligned}$$

We clearly have

$$|\varepsilon\langle Aw, \bar{y}_\tau\rangle_{L^2(0,L)}| \lesssim \varepsilon\|Aw\|_{L^2(0,L)}.$$

Let us now deal with the integral term. First of all, using the continuous imbedding of H_0^1 in C^0, we estimate

$$\begin{aligned}\left\|\Big(w(t,x)+\frac{x}{L}v(t)\Big)^2\right\|_{L^\infty(0,L)} &\lesssim \|w(t,.)\|^2_{L^\infty(0,L)} + v(t)^2\\ &\lesssim \|w(t,.)\|^2_{H_0^1(0,L)} + \|X_1(t)\|_2^2\\ &\lesssim V_1(t)\end{aligned}$$

since $V_1 \sim \|w\|^2_{H_0^1(0,L)} + \|X_1\|_2^2$. For $|v(t)| + \|w(t,.)\|_{L^\infty(0,L)} \le 1$, one has

$$\left\langle Aw, \Big(w+\frac{x}{L}v\Big)^2\int_0^1(1-s)f''\left(\bar{y}+s\left(w+\frac{x}{L}v\right)\right)ds\right\rangle_{L^2(0,L)} \lesssim \|Aw\|_{L^2(0,L)}V_1,$$

and we arrive at the estimate

$$\big|\langle Aw, r\rangle_{L^2(0,L)}\big| \lesssim \varepsilon\|Aw\|_{L^2(0,L)} + \|Aw\|_{L^2(0,L)}V_1$$

for $|v(t)| + \|w(t,.)\|_{L^\infty(0,L)} \le 1$.

Let us also estimate the terms of the principal part of (2.29). We clearly have

$$\big|\langle Aw, a\rangle_{L^2(0,L)}v\big| \le \frac{1}{4}\|Aw\|^2_{L^2(0,L)} + \|a\|^2_{L^2(0,L)}\|X_1\|_2^2$$

and

$$\big|\langle Aw, b\rangle_{L^2(0,L)}K_1X_1\big| \le \frac{1}{4}\|Aw\|^2_{L^2(0,L)} + M\|X_1\|_2^2,$$

where

$$M = \|b\|_{L^2(0,L)}\max\left\{\sum_{i=0}^n k_i(\tau)^2/\tau \in [0,1]\right\}.$$

Hence, concerning the principal part of (2.29), we first get

$$\begin{aligned} &- c\|X_1\|_2^2 - \|Aw\|^2_{L^2(0,L)} - \langle Aw, a\rangle_{L^2(0,L)} v - \langle Aw, b\rangle_{L^2(0,L)} K_1 X_1 \\ &\quad \leq \ -c_1\|X_1\|_2^2 - \frac{1}{2}\|Aw\|^2_{L^2(0,L)}, \end{aligned}$$

where $c_1 = c - \|a\|^2_{L^2(0,L)} - M$. We choose c so that $c_1 > 0$.

The previous estimates and (2.29) now yield, for $|v(t)| + \|w(t,.)\|_{L^\infty(0,L)} \leq 1$,

(2.30)
$$V_1' + \|X_1\|_2^2 + \|Aw\|^2_{L^2(0,L)} \ \lesssim \ \varepsilon\sqrt{V_1} + V_1^{3/2} + \varepsilon\|Aw\|_{L^2(0,L)} + V_1\|Aw\|_{L^2(0,L)}.$$

Note that, for every $\theta \in (0, +\infty)$,

$$\begin{aligned} \varepsilon\sqrt{V_1} &\leq \frac{\theta}{2}V_1 + \frac{1}{2\theta}\epsilon^2, \\ \varepsilon\|Aw\|_{L^2(0,L)} &\leq \frac{\theta}{2}\|Aw\|^2_{L^2(0,L)} + \frac{1}{2\theta}\varepsilon^2, \\ V_1\|Aw\|_{L^2(0,L)} &\leq \frac{\theta}{2}\|Aw\|^2_{L^2(0,L)} + \frac{1}{2\theta}V_1^2. \end{aligned}$$

Hence, taking $\theta > 0$ small enough, we get, using (2.25) and (2.30), the existence of $\sigma > 0$ and $\rho \in (0, \sigma]$ such that, for every $\varepsilon \in (0, 1]$ and for every $t \in [0, 1/\varepsilon]$ such that $V_1(t) \leq \rho$,

$$V_1'(t) \leq \sigma\varepsilon^2.$$

Hence, since $V_1(0) = 0$, we get, if $\varepsilon \in (0, \rho/\sigma]$, that

$$V_1(t) \leq \sigma\varepsilon \quad \forall t \in [0, 1/\varepsilon]$$

and, in particular,

$$V_1\left(\frac{1}{\varepsilon}\right) \leq \sigma\varepsilon.$$

Coming back to definitions (2.1) and (2.3), we have proved

$$\left\| y\left(\frac{1}{\varepsilon}, .\right) - y_1(.)\right\|_{H^1(0,L)} \leq \gamma\varepsilon, \tag{2.31}$$

where $y_1(.) = \bar{y}(1, .)$ is the final target and γ is a positive constant which does not depend on $\varepsilon \in (0, \rho/\sigma]$. This concludes the third step, and thus the proof of the stabilization part of Theorem 1.2 (see Remark 1.3).

2.4. End of the proof. The last step consists of solving a local exact controllability result: from the previous section, $y(\frac{1}{\varepsilon}, .)$ belongs to an arbitrarily small neighborhood of $y_1(.)$ in H^1-topology if ε is small enough, and our aim is now to construct a trajectory $q(t, x)$ solution of the control system steering $y(\frac{1}{\varepsilon}, .)$ to $y_1(.)$ in some time $T > 0$ (for instance, $T = 1$), i.e.,

$$\begin{cases} q_t = q_{xx} + f(q), \\ q(t, 0) = 0, \ q(t, L) = u(t), \\ q(0, x) = y\left(\frac{1}{\varepsilon}, x\right), \ q(T, x) = y_1(x). \end{cases}$$

Existence of such a solution q is given by [11, Theorem 3.3]. Actually, in [11] the function f is assumed to be globally Lipschitzian, but the local result we need here follows readily from the proofs and the estimates contained in this paper. Indeed, let $T > 0$ and let $\tilde{f}$ be a globally Lipschitzian mapping such that

$$\tilde{f}(s) = f(s) \quad \forall s \in [-\|y_1\|_{L^\infty} - 1, \|y_1\|_{L^\infty} + 1]. \tag{2.32}$$

From the proof of [11, Theorem 3.3], we get the existence of $\mu > 0$ such that there exists $z \in Y_T$ satisfying

$$\begin{cases} z_t = z_{xx} + \tilde{f}(z + y_1) - \tilde{f}(y_1), \\ z(t,0) = 0, \\ z(0,x) = y\left(\dfrac{1}{\varepsilon}, x\right) - y_1(x), \ z(T,x) = 0 \end{cases}$$

and the estimate

$$\|z\|_{Y_T} \leq \mu \left\| y\left(\frac{1}{\varepsilon}, \cdot\right) - y_1(\cdot) \right\|_{H^1(0,L)}, \tag{2.33}$$

which leads, with $q = z + \tilde{y}_1$, to

$$\begin{cases} q_t = q_{xx} + \tilde{f}(q), \\ q(t,0) = 0, \\ q(0,x) = y\left(\dfrac{1}{\varepsilon}, x\right), \ q(T,x) = y_1(x) \end{cases}$$

and

$$\|q - \tilde{y}_1\|_{Y_T} \leq \mu \left\| y\left(\frac{1}{\varepsilon}, \cdot\right) - y_1(\cdot) \right\|_{H^1(0,L)}, \tag{2.34}$$

where $\tilde{y}_1(t,x) := y_1(x)$. From (2.33) and (2.34), we get

$$\|q - \tilde{y}_1\|_{L^\infty((0,T)\times(0,L))} \leq 1 \tag{2.35}$$

for $\|y(1/\varepsilon, \cdot) - y_1(\cdot)\|_{H^1(0,L)}$ small enough. From (2.32) and (2.35), we infer that $\tilde{f}(q) = f(q)$, which ends the proof.

3. Controllability versus connectedness. Let us first give some sufficient conditions ensuring the connectedness of $\mathcal{S}$.

PROPOSITION 3.1. *In each of the following cases, the set of steady-states $\mathcal{S}$ is connected:*

- *The function F, defined as*

$$F(y) = \int_0^y f(s)ds,$$

satisfies the asymptotic condition

$$F(y) \underset{|y|\to+\infty}{\longrightarrow} +\infty.$$

- *For any $\alpha > 0$, the indefinite integral*

$$\int \frac{dy}{\sqrt{\alpha - F(y)}}$$

diverges in $-\infty$ and in $+\infty$ (if it makes sense).
- *The function f is odd, i.e., for any $y \in \mathbb{R}$,*

$$f(-y) = -f(y).$$

Remark 3.2. Notice that, contrary to the two first cases of the proposition, in the third case blow-up phenomena may occur. Nevertheless, the set of steady-states is connected.

On the other hand, we have the following result.

PROPOSITION 3.3. *If y_0 and y_1 belong to distinct connected components of $\mathcal{S}$, then it is not possible to move either from y_0 to y_1, or from y_1 to y_0, whatever the control $u \in L^2(0,T)$ and the time T are.*

Remark 3.4. If y_0 and y_1 are both periodic, then they are in the same connected component.

In order to prove these two propositions, let us first note some general facts about the maximal solutions of the scalar differential equation

$$y''(x) + f(y(x)) = 0,\ y(0) = 0. \tag{3.1}$$

LEMMA 3.5.

- *Any solution of* (3.1) *satisfies on its maximal interval of definition the conservation law*

$$y'(x)^2 + 2F(y(x)) = y'(0)^2. \tag{3.2}$$

- *Any solution of* (3.1) *such that y' vanishes at least at two distinct points is actually periodic.*
- *The phase portrait in the plane (y, y') of the associated differential system*

$$y' = z,\ z' = -f(y),$$

is symmetric with respect to the y axis, and moreover, all singular points of the system are located on this axis.

The proof of these facts is obvious. Now the key lemma for proving Propositions 3.1 and 3.3 is the following.

LEMMA 3.6. *Let y_0 and y_1 be two steady-states, extended on their maximal interval of definition as solutions of* (3.1)*, belonging to distinct connected components of $\mathcal{S}$, such that $y_0'(0) < y_1'(0)$. Then there exists an $l \in (0, L]$ and $\bar{y} \in C^2([0,l))$ solution of* (3.1) *such that either*

$$\bar{y}(x) \underset{x \to l}{\longrightarrow} +\infty \tag{3.3}$$

or

$$\bar{y}(x) \underset{x \to l}{\longrightarrow} -\infty. \tag{3.4}$$

In the first case we have, moreover (see Figure 3.1*), the following:*

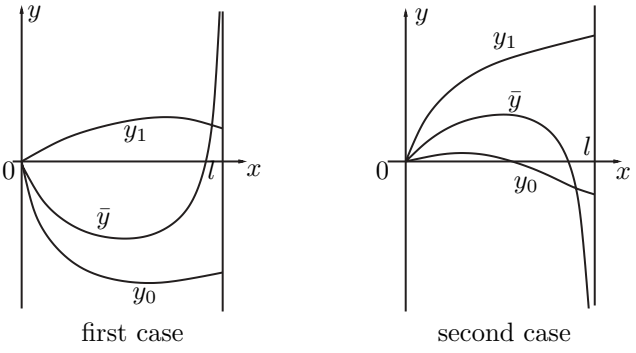


FIG. 3.1. *Existence of an explosive solution.*

1. $y_0(x) < \bar{y}(x)$ *for any* $x \in [0, l)$.
2. $y_0'(0) < \bar{y}'(0) < y_1'(0)$, $|y_1'(0)| < |\bar{y}'(0)|$, *and* $\bar{y}'(0) < 0$.
3. $\#\{x \in [0, l) \ / \ \bar{y}(x) = y_1(x)\} = 1$.
4. y_0 *is not periodic, and* $y_1(x)$ *does not tend to* $-\infty$ *as* x *tends to* b, *where* (a, b) *denotes the maximal interval of definition of* y_1.

In the second case we have the symmetric situation (see Figure 3.1*), as follows:*

1. $y_1(x) > \bar{y}(x)$ *for any* $x \in [0, l)$.
2. $y_0'(0) < \bar{y}'(0) < y_1'(0)$, $|y_0'(0)| < |\bar{y}'(0)|$, *and* $\bar{y}'(0) > 0$.
3. $\#\{x \in [0, l) \ / \ \bar{y}(x) = y_0(x)\} = 1$.
4. y_1 *is not periodic, and* $y_0(x)$ *does not tend to* $+\infty$ *as* x *tends to* b, *where* (a, b) *denotes the maximal interval of the definition of* y_0.

Proof of Lemma 3.6. Clearly one of the two cases (3.3) and (3.4) occurs, with, moreover, $y_0'(0) < \bar{y}'(0) < y_1'(0)$. Assume we are in the first case, and let us first prove the second of the four properties claimed in that case. To proceed, we have to distinguish between the following three possibilities:

- y_1 is monotonic on $[0, l]$. In this case the conservation law (3.2) immediately implies that $|y_1'(0)| < |\bar{y}'(0)|$.
- y_1 is not monotonic on $[0, l]$ and is not periodic on its maximal interval (a, b). That is, y_1' vanishes exactly once on (a, b). The only nonobvious case occurs when $y_1'(0) > 0$. But then, since the phase portrait is symmetric with respect to the y axis, either $y_1(x)$ tends to $-\infty$ as x tends to a and b or $y_1(x)$ tends to a finite limit which corresponds to a singular point on the phase portrait (see Figure 3.2). In both cases it is clear on the phase portrait that $\bar{y}(x)$ may tend to $+\infty$ as x tends to l only if $|y_1'(0)| < |\bar{y}'(0)|$.
- y_1 is periodic (i.e., y_1' vanishes at least two times). Again in this case the phase portrait immediately implies the desired inequality.

Before proving that $\bar{y}'(0) < 0$, let us prove the third point. The only nonobvious case occurs when y_1 is periodic and $\bar{y}$ is not monotonic on $[0, l)$. Notice that $\bar{y}'$ vanishes only once (if not, $\bar{y}$ would be periodic), and thus it decreases on an interval $[0, x_0]$ and increases on $[x_0, l)$. Now, on the one hand, the function y_1 cannot intersect $\bar{y}$ on the interval $[0, x_0]$, for this would contradict the conservation law (3.2). On the other hand, if y_1 would intersect $\bar{y}$ more than once on the interval $[x_0, l)$, then there would be at least three intersections, and again this leads to a contradiction with (3.2). This proves the third point.

Now the inequality $\bar{y}(0) < 0$ is an obvious consequence of (3.2).

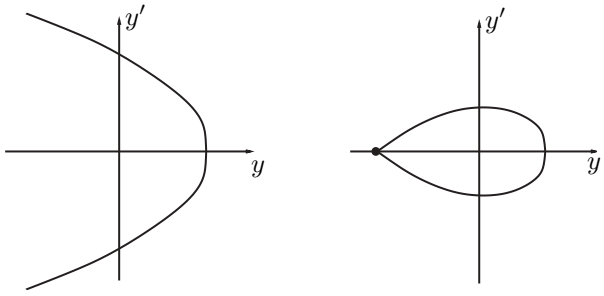


FIG. 3.2. *Behavior of* $(y_1(x), y_1'(x))$ *in the phase plane.*

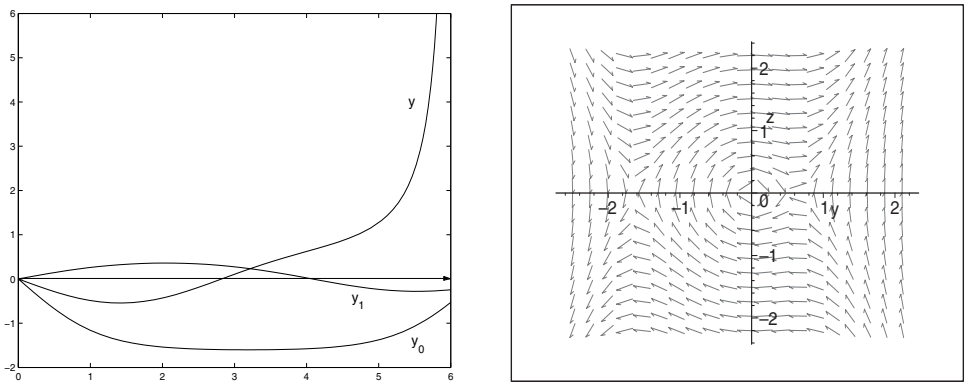


FIG. 3.3. *An example where* y_0 *and* y_1 *are not in the same connected component of* $\mathcal{S}$.

Let us now prove that $y_0 < \bar{y}$ on $[0, l)$. The same reasoning as above shows that $\bar{y}$ intersects y_0 at most once (notice that y_0' cannot vanish more than once). But such an intersection would contradict the fact that $\bar{y}(x)$ tends to $+\infty$ as x tends to l.

Finally, the last point of the lemma is proved by observing the phase portrait. □

Proof of Proposition 3.3. Proposition 3.3 follows from Lemma 3.6. Indeed, let us assume, for example, that we are in the first case of the lemma. Then for any $T > 0$ and $u \in L^2([0, T])$, the solution y of the control system (1.4) satisfies, as long as defined, the inequality

$$y(t, x) \leq \bar{y}(x)$$

(see [4] for this application of the classical maximum principle to similar control problems). In particular, $y(T, .) \neq y_1(.)$. □

Finally let us prove Proposition 3.1. The only difficult case is to prove that if f is odd, then the set $\mathcal{S}$ is connected. In this case the conservation law (3.2) implies that the phase portrait is symmetric with respect to the y axis and the y' axis. As a consequence, any solution of (3.1) such that y' vanishes at least once is necessarily periodic.

Now from Lemma 3.6 we know that if y_0 and y_1 are not in the same connected component, then there exists an explosive solution $\bar{y}$ of (3.1) such that $\bar{y}'$ vanishes at least once. Hence, $\bar{y}$ must be periodic, and we get a contradiction.

Example 3.7. An example where the situation of Proposition 3.3 and Lemma 3.6 occurs is given by

$$f(y) = y - y^2 - y^3.$$

The graph of y_0, y_1, an explosive y, and the phase portrait are drawn on Figure 3.3.

4. Numerical simulations. In this section we present numerical simulations using Matlab for the nonlinear function $f(y) = y^3$. Let $L = 1$; the set $\mathcal{S}$ of steady-states consists of all solutions of class C^2 on $[0,1]$ such that

$$y''(x) + y(x)^3 = 0,\ y(0) = 0. \tag{4.1}$$

It follows from Proposition 3.1 that this set is connected. Let y_0 be identically zero, and let y_1 denote the solution of (4.1) vanishing at 0, 1/2, and 1, and having no other zero on $[0,1]$ (see Figure 4.1).

For all $\tau \in [0,1]$, we define the function $\bar{y}(\tau, .)$ on $[0,1]$ as the solution of (4.1) such that

$$\frac{\partial \bar{y}}{\partial x}(\tau, 0) = \tau y_1'(0)$$

and we set $\bar{u}(\tau) = \bar{y}(\tau, 1)$. We then introduce on $H^2(0,1) \cap H_0^1(0,1)$ the one-parameter family of linear operators

$$A(\tau) = \Delta + 3\bar{y}(\tau, .)^2 Id,\ \tau \in [0,1].$$

For $\tau = 0$, we have $A(0) = \Delta$, and the eigenvalues and eigenvectors write

$$\lambda_i(0) = -i^2\pi^2,\ e_i(0,x) = \sqrt{2}\sin k\pi x.$$

Then, solving by homotopy as $\tau \in [0,1]$ boundary value problems, we compute numerically, using a standard finite difference code implemented in Matlab, the first eigenvalues and associated eigenvectors. In the present example, numerical experiments show that only the two first eigenvalues may take positive values as $\tau \in [0,1]$. In other words, with the notations of section 2.3, one has $n = 2$. Then we achieve a

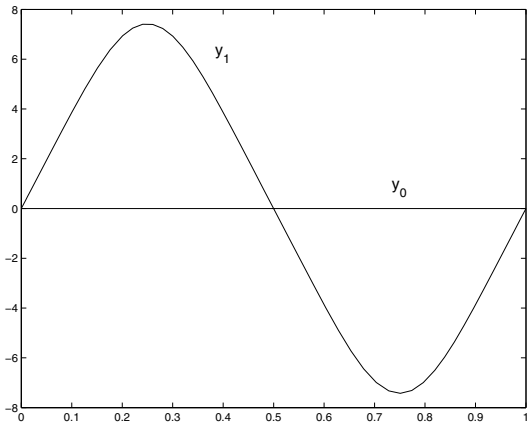


FIG. 4.1. *Definition of the steady-states* y_0 *and* y_1.

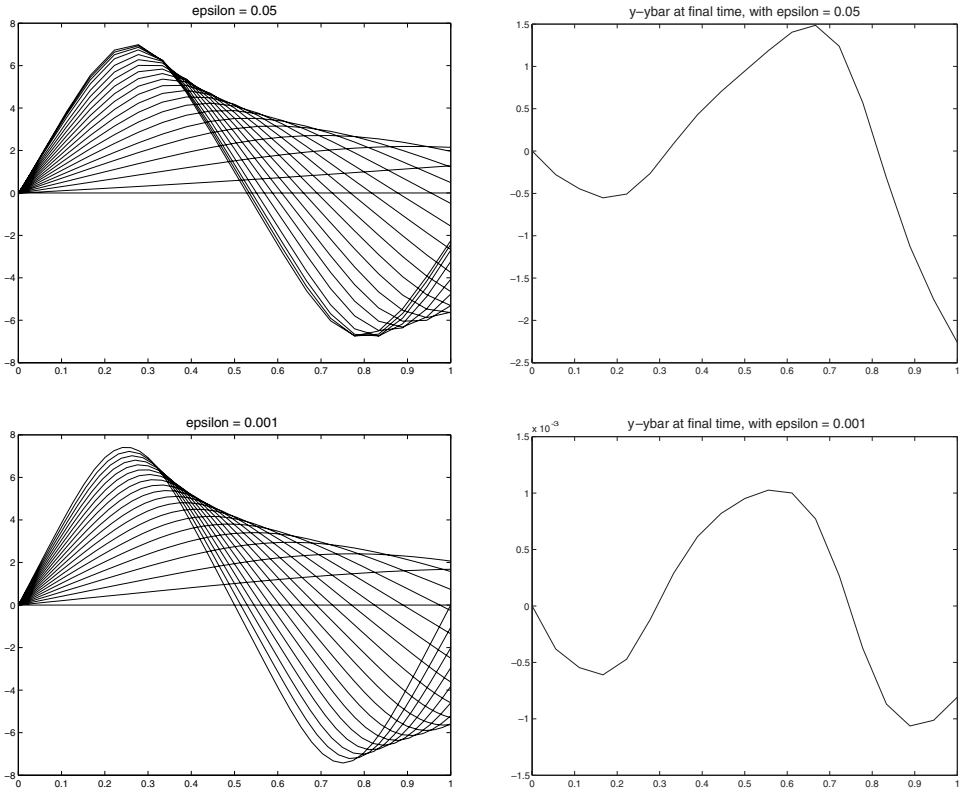


FIG. 4.2. *Simulations results for $y(t,.)$, where $t \in [0, 1/\varepsilon]$.*

pole placement on the finite dimensional system (2.16) by applying an LQR algorithm (see [13]). Notice that the finite dimensional system corresponding to these two first modes is very unstable: numerically, one has $\lambda_1(1) \simeq 89.743$ and $\lambda_2(1) \simeq 82.518$.

Results are drawn on Figure 4.2 for $\varepsilon = 0.05$ and $\varepsilon = 0.001$. Notice that if ε is too large, then the solution blows up.

REFERENCES

[1] J. BEBERNES AND D. EBERLY, *Mathematical Problems from Combustion Theory*, Appl. Math. Sci. 83, Springer-Verlag, New York, 1989.

[2] T. CAZENAVE AND A. HARAUX, *Introduction aux problèmes d'évolution semi-linéaires*, Math. Appl. 1, Ellipses, Paris, 1990.

[3] J. M. CORON, *Local controllability of a 1-D tank containing a fluid modeled by the shallow water equations*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 513–554.

[4] J. I. DIAZ, *Obstruction and some approximate controllability results for the Burgers equation and related problems*, in Control of Partial Differential Equations and Applications, Lecture Notes in Pure and Appl. Math. 174, Dekker, New York, 1996, pp. 63–76.

[5] C. FABRE, J. P. PUEL, AND E. ZUAZUA, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[6] E. FERNÁNDEZ-CARA, *Null controllability of the semilinear heat equation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 87–103.

[7] E. FERNÁNDEZ-CARA AND E. ZUAZUA, *Null and approximate controllability for weakly blowing up semilinear heat equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 583–616.

[8] H. Fujita, *On the blowing up of solutions to the Cauchy problem for* $u_t = \Delta u + u^{1+\alpha}$, J. Fac. Sci. Univ. Tokyo Sect. IA Math, 13 (1966), pp. 109–124.

[9] A. Fursikov and O. Yu. Imanuvilov, *On controllability of certain systems simulating a fluid flow*, in Flow Control, IMA Vol. Math. Appl. 68, M. D. Gunzburger, ed., Springer-Verlag, New York, 1995, pp. 149–184.

[10] J. Henry, *Etude de la contrôlabilité de certaines équations paraboliques non linéaires*, Thèse, Paris, 1977.

[11] O. Yu. Imanuvilov, *Controllability of parabolic equations*, Mat. Sb. 186 (1995), pp. 879–900.

[12] S. Kaplan, *On the growth of solutions of quasilinear parabolic equations*, Comm. Pure Appl. Math., 16 (1963), pp. 305–330.

[13] H. K. Khalil, *Nonlinear Systems*, Macmillan, New York, 1992.

[14] H. A. Levine, *Some nonexistence and instability theorems for solutions of formally parabolic equations of the form* $Pu_t = -Au + F(u)$, Arch. Ration. Mech. Anal., 51 (1973), pp. 371–386.

[15] F. Merle and H. Zaag, *Stability of the blow-up profile for equations of the type* $u_t = \Delta u + |u|^{p-1}u$, Duke Math. J., 86 (1997), pp. 143–195.

[16] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, *Vol.* 4, *Analysis of Operators,* Academic Press, New York, London, 1978.

[17] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[18] H. Zaag, *A remark on the energy blow-up behavior for nonlinear heat equations*, Duke Math. J., 103 (2000), pp. 545–555.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 570–589

QUASI-VARIATIONAL INEQUALITIES WITH DIRICHLET BOUNDARY CONDITION RELATED TO EXIT TIME PROBLEMS FOR IMPULSE CONTROL*

A. L. AMADORI†

Abstract. We study degenerate-elliptic quasi-variational inequalities with Dirichlet boundary condition, which are related to the value function of the exit time problem for stochastic impulse control by means of the dynamic programming principle. The boundary condition in the viscosity solutions sense does not identify a unique solution, because in this nonlocal problem the boundary layer gives rise to a loss of information also at the interior points. The eventual discontinuities of solutions at the boundary of the domain play an essential role and cannot be removed. Therefore we superimpose a selection criterion which, enforcing the information coming from the boundary datum, picks up the value function among all possible viscosity solutions. As a result, we attain the continuity of the value function up to the boundary. In addition, we produce a monotone iterative scheme approximating the value function.

Key words. impulse stochastic control, quasi-variational inequality, boundary condition

AMS subject classifications. Primary, 49L25; Secondary, 49N25, 35J40, 35J85

DOI. 10.1137/S0363012902409738

1. Introduction.

This paper concerns a stochastic impulse control problem in an open bounded domain $\Omega \subset \mathbb{R}^{\mathrm{N}}$. The state of the controlled system is given by the solution $(X_t^x)_t$ of the following equation:

$$\begin{cases} dX_t^x = b(X_t^x, \alpha_t)\, dt + \sigma(X_t^x, \alpha_t)\, dW_t, & t \in]\theta_i, \theta_{i+1}[, \\ X_{\theta_i}^x = X_{\theta_i^-}^x + \xi_i, \\ X_0^x = x \in \Omega, \end{cases} \tag{1.1}$$

where $(W_t)_t$ is a D-dimensional Brownian motion (with, possibly, $\mathrm{D} \le \mathrm{N}$), $(\alpha_t)_t$ is any adapted process with values in a compact set $\mathcal{A} \subset \mathbb{R}^{\mathrm{P}}$, $(\theta_i)_i$ is an increasing sequence of stopping times (both with respect to the filtration generated by the Brownian motion W_t), $(\xi_i)_i$ is a sequence in $\mathbb{R}^{\mathrm{N}}_+$, and b, σ are continuous functions on $\overline{\Omega} \times \mathcal{A}$ taking values respectively in $\mathbb{R}^{\mathrm{N}}$ and in the space of $\mathrm{N} \times \mathrm{D}$ real matrices.

The control is the triplet $A = ((\alpha_t)_t, (\theta_i, \xi_i)_i)$, where the item $\theta_i = \infty$ for some $i \in \mathbb{N}$ is allowed. For any fixed x and A, the cost function is

$$\mathcal{J}(x, A) := \mathbb{E}\left[\int_0^{\tau_x} f(X_t^x, \alpha_t) e^{-\lambda t} dt + \sum_{i \le \tau_x} (c(\xi_i) + k) e^{-\lambda \theta_i} + 1_{\{\tau_x < \infty\}} \varphi(X_{\tau_x}^x) e^{-\lambda \tau_x} \right],$$

where f and c are real continuous functions, λ and k are positive numbers, and τ_x stands for the first exit time from $\overline{\Omega}$ of the trajectory $(X_t^x)_t$. We are interested in value function $\mathbf{U}$ of type

$$\mathbf{U}(x) := \inf_A \mathcal{J}(x, A).$$

*Received by the editors June 18, 2002; accepted for publication (in revised form) December 27, 2003; published electronically July 23, 2004. This work was partially supported by the TMR Network "Viscosity Solutions and Their Applications."

http://www.siam.org/journals/sicon/43-2/40973.html

†Istituto per le Applicazioni del Calcolo "Mauro Picone," C.N.R., Viale del Policlinico 137, 00161 Roma, Italy (a.amadori@iac.cnr.it).

It is easily seen that the minimization problem is equivalent to

$$\mathbf{U}(x) = \inf_{A} \mathbb{E}\left[\int_0^{\tau_x} f(X_t^x, \alpha_t) e^{-\lambda t} dt + \sum_{i\in\mathbb{N}} (c(\xi_i) + k) e^{-\lambda\theta_i} + 1_{\{\tau_x<\infty\}} \varphi(X_{\tau_x}^x) e^{-\lambda\tau_x}\right]. \tag{1.2}$$

Stochastic impulse control problems arise, for instance, in the modeling of portfolio management with transaction costs and have been studied, among others, by [17]: it is assumed that $\operatorname{rank}\sigma = \mathrm{N}$ so that the market is complete. The investigation of incomplete markets has to be addressed to control problems with $\operatorname{rank}\sigma < \mathrm{N}$, which is the topic of this present work.

The dynamic programming principle relates the optimization problem (1.1)–(1.2) to a Hamilton–Jacobi–Bellman equation, which now stands in the quasi-variational inequality

$$\max\{H(x, u, Du, D^2u), u - Mu\} = 0 \qquad \text{on } \Omega, \tag{1.3}$$

coupled with the Dirichlet boundary condition

$$u = \varphi \qquad \text{on } \partial\Omega. \tag{1.4}$$

Here, H is the (possibly degenerate) elliptic operator

$$H(x, u, p, X) := \sup_{\alpha\in\mathcal{A}} \left\{ -\frac{1}{2}\operatorname{tr}\left[a(x,\alpha)X\right] - b(x,\alpha)\cdot p + \lambda u - f(x,\alpha) \right\}$$

for any $x \in \overline{\Omega}$, $u \in \mathbb{R}$, $p \in \mathbb{R}^{\mathrm{N}}$, and $X \in \mathcal{S}^{\mathrm{N}}$ (the space of $\mathrm{N}\times\mathrm{N}$ symmetric matrices), where we have used the abbreviated notation $a(x,\alpha) = \sigma(x,\alpha)\sigma^T(x,\alpha)$ for any $x \in \overline{\Omega}$ and $\alpha \in \mathcal{A}$. Besides, M is a nonlocal operator of the form

$$Mu(x) := \inf\left\{u(x+\xi) + c(\xi) + k \,:\, \xi \in \mathbb{R}^{\mathrm{N}}_+, x+\xi \in \overline{\Omega}\right\}.$$

We refer to [6] for a detailed analysis of the relations between impulse control problems (1.1)–(1.2) and quasi-variational inequalities (1.3)–(1.4). If the matrix a is positively defined, the value function has been characterized as the unique strong solution to the quasi-variational inequality with Dirichlet boundary condition in [18]. In the general case when a may degenerate, it is necessary to turn to the theory of viscosity solutions. If one manages to establish that the value function is continuous on Ω and continuously extendable to $\overline{\Omega}$, one can characterize it as the unique continuous (viscosity) solution by combining the dynamic programming principle and uniqueness arguments. Such a program has been successfully performed in [1], dealing with deterministic impulse control in all $\mathbb{R}^{\mathrm{N}}$. However, in the case of exit time problems, it is well known that the value function is not continuous, in general. Concerning deterministic control, this topic has been discussed in [4, 7, 19, 20]. With respect to standard stochastic control, this problem has been solved quite generally by achieving a strong comparison result, i.e., a comparison among discontinuous sub/supersolutions satisfying the boundary condition in the relaxed sense, in [5, 15]. Both these results may handle cases when $\mathbf{U}$ is not equal to φ on the boundary points where σ degenerates along the normal direction, and they bring as a by-product that the value function is continuous; moreover, they rely heavily on the fact that the continuous extension of $\mathbf{U}$ to $\overline{\Omega}$ still solves the Dirichlet problem with relaxed boundary condition. Concerning stochastic impulse control, comparison results can be found in

[13, 14], both postulating the behavior of sub/supersolutions at the boundary. In [13] it is assumed that the sub/supersolutions are "strong," namely that they verify the boundary condition in the following sense:

$$\max\{u-Mu, u-\varphi\} \le 0 \quad \text{on } \partial\Omega, \tag{1.5}$$

$$\max\{u-Mu, u-\varphi\} \ge 0 \quad \text{on } \partial\Omega, \tag{1.6}$$

respectively. Next, by following the line of [15], [14] improves the previous comparison result, after assuming the nontangential semicontinuity of viscosity sub/supersolutions. These two comparison results allow characterization of the value function $\mathbf{U}$ as the "unique" viscosity solution to (1.3), in a restricted class of solutions which are known a priori to satisfy some technical regularity assumption at the boundary. Furthermore they apply only to the case when the value function is continuous in $\overline{\Omega}$.

Here, following the line of [5], we prefer to impose some "nondegeneracy"-type conditions on the boundary $\partial\Omega$, which can be read as "partial controllability" assumptions. As a first result we obtain comparison among any viscosity sub/supersolutions in the cases studied before in [13, 14]. Moreover we may deal with the general case, in which the value function is not continuous and its extension to $\overline{\Omega}$ does not solve (1.3). To be clear, we denote by $\Gamma_{in} \subset \partial\Omega$ the region of the boundary which can be reached only through a jump and not by using a standard continuous controlled trajectory (a precise definition is given later). Example 3.4 shows that, if the optimal trajectory jumps to Γ_{in}, the continuous extension of $\mathbf{U}$ to $\overline{\Omega}$ does not solve (1.3)–(1.4); nor is this the case if the boundary condition is understood in the relaxed sense

$$\min\left\{\max\left\{H\left(x,u,Du,D^2u\right), u-Mu\right\}, u-\varphi\right\} \le 0 \quad \text{on } \partial\Omega, \tag{1.7}$$

$$\max\left\{H\left(x,u,Du,D^2u\right), u-Mu, u-\varphi\right\} \ge 0 \qquad \text{on } \partial\Omega. \tag{1.8}$$

This matter breaks the analogy with previously mentioned control problems: the main difference is that not only are discontinuities at the boundary a by-product of the degeneracy of a, but they can be generated by the impulse control. In the case of standard stochastic control, the points of Γ_{in} behave as interior points, and imposing that the equation is satisfied at those points is sufficient to balance the loss of boundary data. On the contrary, in the case of impulse control, the points of Γ_{in} can be reached by jumping: as a result, the value function exactly takes the boundary value φ, but it is discontinuous. Such discontinuity is an integral part of the control problem and cannot be removed. The boundary condition (1.7)–(1.8) does not capture the information coming from the region Γ_{in}, so that the viscosity solution may even be discontinuous inside Ω and, in general, is not unique. These difficulties are overcome by giving a suitable notion of a boundary condition, which takes into account the jumps of the value function due to impulse control. The information about the boundary cost has to be strengthened by imposing that

$$\begin{aligned} u(x) &\le \tilde{M}\varphi(x) && \text{on } \Omega, \\ \limsup_{\Omega \ni y \to x} u(y) &\le \tilde{M}\varphi(x) && \text{on } \partial\Omega, \end{aligned} \tag{1.9}$$

where

$$\tilde{M}\varphi(x) := \inf\left\{\varphi(x+\xi) + c(\xi) + k \,:\, \xi \in \mathbb{R}^{\mathrm{N}}_{+}, x+\xi \in \Gamma_{in}\right\}. \tag{1.10}$$

This selection criterion picks up the value function among all the viscosity solutions and appears to be the natural way to read the boundary condition as more than a technical assumption. It is worth mentioning that the value function is neither the more regular solution nor the limit of an iterative procedure: indeed the standard iterative scheme possibly converges to the "wrong" solution.

This paper is organized as follows. In section 2, we introduce the assumptions that shall be in force in what follows and present some preliminary results concerning the behavior of any sub/supersolutions near the boundary. Section 3 concerns the case when $\Gamma_{in} = \emptyset$. We establish a comparison result which yields the continuity of the value function on $\overline{\Omega}$ and we show by means of a counterexample that it is not true anymore if $\Gamma_{in} \neq \emptyset$. Section 4 concerns the case when $\Gamma_{in} \neq \emptyset$: as shown in Example 4.1, comparison does not hold in general. Thus, we introduce a selection criterion and we obtain a comparison result between selected subsolutions and any supersolutions. As a by-product, we establish the continuity of the value function in $\overline{\Omega} \setminus \Gamma_{in}$ and we describe the eventual discontinuity in Γ_{in}. Finally, in section 5 we produce a monotone iterative scheme which approximates the value function.

2. Assumptions and preliminary results. We make the following standard assumptions:

H.1. The functions σ, b, and f are continuous on $\overline{\Omega} \times \mathcal{A}$. For any $\alpha \in \mathcal{A}$, $\sigma(\cdot, \alpha)$ and $b(\cdot, \alpha)$ are Lipschitz continuous on $\overline{\Omega}$; in addition,

$$\sup_{\alpha \in \mathcal{A}} \|\sigma_{ij}(\cdot, \alpha)\|_{\mathcal{C}^{0,1}(\overline{\Omega})}, \qquad \sup_{\alpha \in \mathcal{A}} \|b_i(\cdot, \alpha)\|_{\mathcal{C}^{0,1}(\overline{\Omega})} < \infty$$

as $1 \leq i \leq \mathrm{N}$, $1 \leq j \leq \mathrm{D}$.

H.2. $\lambda > 0$.

H.3. The function c is continuous, nonnegative, and subadditive, namely, $c(\xi_1 + \xi_2) \leq c(\xi_1) + c(\xi_2)$. In addition $c(0) = 0$ and $k > 0$.

H.4. $\varphi \in \mathcal{C}\big(\partial\Omega\big)$.

H.5. Ω is a bounded domain of $\mathbb{R}^{\mathrm{N}}$ with a $W^{3,\infty}$-boundary $\partial\Omega$.

Remark 2.1. The assumption H.2 can be easily relaxed by assuming that λ is a continuous function of x, α on $\overline{\Omega} \times \mathcal{A}$, with $\lambda > 0$.

The regularity assumption H.5 is not enough to handle the nonlocal term Mu. In order to avoid pathological shapes for Ω, we use the same hypothesis as in [13, 14], namely, the following:

H.6. There exists a mapping $P : \overline{\Omega} \times \mathbb{R}^{\mathrm{N}}_+ \to \mathbb{R}^{\mathrm{N}}_+$ satisfying

$$\begin{aligned} x + P(x, \xi) &\in \overline{\Omega} && \text{for all } (x, \xi) \in \overline{\Omega} \times \mathbb{R}^{\mathrm{N}}_+, \\ P(x, \xi) &= \xi && \text{if } x + \xi \in \overline{\Omega}, \\ P(\cdot, \xi) &\in \mathcal{C}\big(\overline{\Omega}\big) && \text{for any } \xi \in \mathbb{R}^{\mathrm{N}}_+. \end{aligned}$$

Assumption H.6 is not trivial. As shown in [14, Proposition 3.2], it guarantees that M preserves semicontinuity, and so it justifies the following definition in the framework of viscosity solutions to discontinuous equations started by [12].

Definition 2.1. *A function u which is locally bounded and upper semicontinuous on $\overline{\Omega}$ is a* viscosity subsolution *to* (1.3)–(1.4) *if it satisfies in the viscosity sense the inequalities*

$$\max\big\{H\big(x, u, Du, D^2u\big), u - Mu\big\} \leq 0 \qquad \text{on } \Omega \tag{2.1}$$

and (1.7) *on* $\partial\Omega$.

A function u which is locally bounded and lower semicontinuous on $\overline{\Omega}$ is a viscosity supersolution *to* (1.3)–(1.4) *if it satisfies in the viscosity sense the inequalities*

$$\max\{H(x,u,Du,D^2u),u-Mu\}\geq 0 \qquad \text{on } \Omega \tag{2.2}$$

and (1.8) *on* $\partial\Omega$.

Any locally bounded function u is a viscosity solution *to* (1.3)–(1.4) *if its upper and lower semicontinuous envelopes are, respectively, a subsolution and a supersolution.*

Remark 2.2. It is easily seen that any subsolution to (1.3)–(1.4) is a subsolution to the differential equation

$$H\left(x,u,Du,D^2u\right)=0 \qquad \text{on } \Omega, \tag{1.3.o}$$

which takes the boundary value (1.4) in the relaxed sense

$$\min\{H\left(x,u,Du,D^2u\right),u-\varphi\}\leq 0 \qquad \text{on } \partial\Omega.$$

Afterward, we use the same nondegeneracy-type assumptions as in [5]. To this end, we denote by d the distance to the boundary, which is a $W^{3,\infty}$ function in a neighborhood of $\partial\Omega$, by assumption H.5. For all $x\in\partial\Omega$, we set

$$\mathcal{A}_{\rm in}(x)=\left\{\alpha\in\mathcal{A}\,:\,\sigma^T(x,\alpha)Dd(x)=0,\frac{1}{2}\mathrm{tr}\left[a(x,\alpha)D^2d(x)\right]+b(x,\alpha)\cdot Dd(x)\geq 0\right\},$$

and we denote by $\Gamma_{\rm in}$ the set of all $x\in\partial\Omega$ such that $\mathcal{A}_{\rm in}(x)=\mathcal{A}$, by $\Gamma_{\rm out}$ the set of all $x\in\partial\Omega$ such that $\mathcal{A}_{\rm in}(x)=\emptyset$, and by $\Gamma=\partial\Omega\setminus(\Gamma_{\rm in}\cup\Gamma_{\rm out})$. We assume the following:

H.7. $\Gamma_{\rm in}$ and $\Gamma_{\rm out}$ are the union of connected components of $\partial\Omega$.

H.8. For every $x\in\Gamma$, one of the following assumptions holds:

H.8.i. There exist $\eta=\eta(x)>0$ and $\mathcal{V}=\mathcal{V}(x)$ an $\overline{\Omega}$-neighborhood of x and three subsets $\mathcal{A}_1(x),\mathcal{A}_2(x),\mathcal{A}_3(x)$ of $\mathcal{A}$ such that $\mathcal{A}=\mathcal{A}_1(x)\cup\mathcal{A}_2(x)\cup\mathcal{A}_3(x)$ and

(a) for any $\alpha\in\mathcal{A}_1(x)$, $\left|\sigma^T(x,\alpha)Dd(x)\right|\geq\eta$;

(b) for any $\alpha\in\mathcal{A}_2(x)\cup\mathcal{A}_3(x)$, the function $\phi_\alpha(y)=\sigma^T(y,\alpha)Dd(y)$ is of class $W^{2,\infty}$ on $\mathcal{V}$ and identically zero on $\mathcal{V}\cup\Gamma$; moreover, $\sup\{\|\phi_\alpha\|_{W^{2,\infty}(\mathcal{V})}:\alpha\in\mathcal{A}_2(x)\cup\mathcal{A}_3(x)\}<\infty$;

(c) for any $\alpha\in\mathcal{A}_2(x)$, $\frac{1}{2}\mathrm{tr}\left[a(x,\alpha)D^2d(x)\right]+b(x,\alpha)\cdot Dd(x)\leq-\eta$;

(d) for any $\alpha\in\mathcal{A}_3(x)$ and $y\in\mathcal{V}$, there exists $\alpha'=\alpha'(y)\in\mathcal{A}_3(x)$ such that $a(y,\alpha')=a(y,\alpha)$ and $\frac{1}{2}\mathrm{tr}\left[a(y,\alpha')D^2d(x)\right]+b(y,\alpha')\cdot Dd(x)\geq\eta$.

H.8.ii. There exists $\mathcal{V}=\mathcal{V}(x)$ an $\overline{\Omega}$-neighborhood of x and, for any $y\in\mathcal{V}$, $\alpha(y)\in\mathcal{A}$ depending continuously on y such that the function $\phi(y)=\sigma^T(y,\alpha'(y))$ is of class $W^{1,\infty}$ on $\mathcal{V}$ and identically zero on $\mathcal{V}\cup\Gamma$; moreover, $\sigma^T(x,\alpha(x))=0$ and $b(x,\alpha(x))\cdot Dd(x)>0$.

When dealing with impulse control problems, $\Gamma_{\rm in}$ plays a particular role, being the set of the points that can be reached only after a jump. In order to prevent tangential trajectories from causing some disturbing effects, we shall assume the following:

H.9. For every $x\in\Gamma_{\rm in}$, there exist $\eta=\eta(x)>0$ and $\mathcal{V}=\mathcal{V}(x)$ a $\partial\Omega$-neighborhood of x such that for all $y\in\mathcal{V}$ and $\alpha\in\mathcal{A}$

$$\frac{1}{2}\mathrm{tr}\left[a(y,\alpha)D^2d(y)\right]+b(y,\alpha)\cdot Dd(y)\geq\eta.$$

We now investigate the behavior of discontinuous sub/supersolutions near at the boundary, with the object of "cleaning" them by replacing their actual values on $\partial\Omega$ by their limits from inside Ω. Such a cleaning procedure avoids artificial discontinuities coming from the underdetermination of the boundary values due to the relaxation of the boundary condition. Furthermore, the solutions to quasi-variational inequalities may also present some discontinuities at the boundary which are not artificial, and these cannot be removed, unless some balancing condition is added. In view of this fact, subsolutions and supersolutions have to be treated in a completely different way.

2.1. Boundary behavior of subsolutions. With respect to subsolutions, the cleaning at the boundary does not change the nonlocal operator M.

LEMMA 2.3. *Let u be an upper semicontinuous function on $\overline{\Omega}$ and set*

$$\tilde{u}(x) = \begin{cases} \limsup_{\Omega \ni y \to x} u(y) & \text{if } x \in \partial\Omega, \\ \\ u(x) & \text{if } x \in \Omega. \end{cases} \tag{2.3}$$

Under assumptions H.3 *and* H.6*, then $M\tilde{u} = Mu$ on Ω.*

Proof. Since $\tilde{u} \le u$ on $\overline{\Omega}$ by construction, $M\tilde{u} \le Mu$; hence, it suffices to show that

$$\inf\{\tilde{u}(x+\xi)+c(\xi) : \xi \in \mathbb{R}^{\mathrm{N}}_+, x+\xi \in \partial\Omega\} \ge \inf\{u(x+\xi)+c(\xi) : \xi \in \mathbb{R}^{\mathrm{N}}_+, x+\xi \in \Omega\}$$

for all $x \in \Omega$. We now claim that the desired inequality follows by the following geometrical property of $\partial\Omega$: for any $x \in \Omega$ and $\xi \in \mathbb{R}^{\mathrm{N}}_+$ with $x + \xi \in \partial\Omega$, there is a sequence $r_n \to 0$ such that

$$\Omega \cap B(x+\xi, r_n) \cap \left(x + \mathbb{R}^{\mathrm{N}}_+\right) \ne \emptyset, \tag{2.4}$$

where $B(z, r)$ stands for the open ball in $\mathbb{R}^{\mathrm{N}}$ centered at z with radius r.

Indeed, let ξ_j be a sequence in $\mathbb{R}^{\mathrm{N}}_+$ such that $x + \xi_j \in \partial\Omega$ and $\tilde{u}(x+\xi_j)+c(\xi_j) \to \inf\{\tilde{u}(x+\xi)+c(\xi) : \xi \in \mathbb{R}^{\mathrm{N}}_+, x + \xi \in \partial\Omega\}$ as $j \to \infty$. If property (2.4) holds true, for any j there is a sequence $\xi_{j,n} \in \mathbb{R}^{\mathrm{N}}_+$ such that $x + \xi_{j,n} \in \Omega$ and $\xi_{j,n} \to \xi_j$ as $n \to \infty$; in particular for all j, n it holds that

$$u(x+\xi_{j,n})+c(\xi_{j,n}) \ge \inf\{u(x+\xi)+c(\xi) : \xi \in \mathbb{R}^{\mathrm{N}}_+, x + \xi \in \Omega\}.$$

Eventually the thesis follows by extracting the limit with respect to both n and j. Thus we need to check (2.4). The proof is trivial if $\xi \in \mathcal{I}nt(\mathbb{R}^{\mathrm{N}}_+)$. Otherwise, we assume for simplicity that the first component of ξ is zero. Since Ω is open, there is a sequence $x_n \in \Omega$ that converges to x and such that the first component of $x_n - x$ is strictly positive for all n. We set $\xi_n = x_n - x + P(x_n, \xi)$ and $r_n = 2|\xi_n - \xi|$. Notice also that the first component of ξ_n is strictly positive, so that $r_n > 0$. Besides, by force of assumption H.6, $r_n \to 0$ and $x+\xi_n \in \overline{\Omega} \cap B(x+\xi, r_n) \cap (x+\mathcal{I}nt(\mathbb{R}^{\mathrm{N}}_+))$. Finally, because $B(x+\xi, r_n) \cap (x + \mathcal{I}nt(\mathbb{R}^{\mathrm{N}}_+))$ is open, we get that (2.4) holds true. □

A relevant consequence of Lemma 2.3 is the following result, which will play a central role in the proofs of Theorems 3.1 and 4.6.

PROPOSITION 2.4. *Let assumptions* H.1–H.7 *hold and u be a subsolution to* (1.3)–(1.4)*. Then the function $\tilde{u}$ defined by* (2.3) *is a subsolution to* (1.3)–(1.4)*.*

Proof. By construction, $\tilde{u}$ is upper semicontinuous on $\overline{\Omega}$. At all $x \in \Omega$, the inequality $H(x, \tilde{u}, D\tilde{u}, D^2\tilde{u}) \le 0$ holds in the viscosity sense because $\tilde{u} \equiv u$ near x. Besides, since $\tilde{u}(x) = u(x) \le Mu(x)$, Lemma 2.3 implies that $\tilde{u}(x) \le M\tilde{u}(x)$.

If $x \in \partial\Omega$, two cases may happen: the first one is $u(x) \le \varphi(x)$. In that case, $\tilde{u}(x) \le u(x) \le \varphi(x)$, and then the proof is completed.

The second one is $u(x) > \varphi(x)$. Now, remembering Remark 2.2, [5, Proposition 4.1] yields that $x \in \Gamma_{\rm in}$. Next, by assumption H.7 there exists $r > 0$ such that $\partial\Omega \cap B_r(x) \subset \Gamma_{in}$. It suffices to check that $\max\{H\big(x, \tilde{u}, D\psi, D^2\psi\big),\ \tilde{u} - M\tilde{u}\} \le 0$ for all smooth test functions ψ such that x is a strict global maximum of $\tilde{u} - \psi$ on $\overline{\Omega} \cap B_r(x)$. To this aim, we introduce the function $y \mapsto \tilde{u}(y) - \psi(y) - \varepsilon/d(y)$. By classical arguments, for all $\varepsilon > 0$ this function achieves its maximum at $x_\varepsilon \in \Omega \cap \overline{B_r}(x)$, and one can prove that $x_\varepsilon \to x$, $u(x_\varepsilon) \to \tilde{u}(x)$, $\varepsilon/d(x_\varepsilon) \to 0$. In particular, x_ε is a local maximum point for $u - \psi - \varepsilon/d$ on $\Omega \cap B_r(x)$; therefore $\max\{H(x_\varepsilon, u, D\psi, D^2\psi), u - Mu\} \le 0$. So the same arguments of [5, Lemma 4.1] yield $H(x, \tilde{u}, D\psi, D^2\psi) \le 0$. Moreover, recalling Lemma 2.3, we get $\tilde{u}(x) = \lim_{\varepsilon\to 0} u(x_\varepsilon) \le \lim_{\varepsilon\to 0} M\tilde{u}(x_\varepsilon) \le M\tilde{u}(x)$ by the upper semicontinuity of $M\tilde{u}$ (see [14, Proposition 3.2]). □

Remark 2.5. It is easily seen that $\tilde{u}(x) \le M\tilde{u}(x)$ at all $x \in \overline{\Omega}$.

It is clear that the property stated by Lemma 2.3 does not hold for supersolutions.

EXAMPLE 2.6. *Take* $\Omega = (0,1)$, $\sigma \equiv 0$, $b \equiv -1$, c *any increasing function of class* $\mathcal{C}^1$, $f(x) > -c'(1-x) + \lambda\,[c(1-x) + k]$, *and*

$$\varphi(x) = \begin{cases} \mathrm{C}_o & \text{if } x = 0, \\ 0 & \text{if } x = 1, \end{cases}$$

where C_o *is any constant greater than or equal to* $c(1) + k$. *Then the function*

$$u(x) = \begin{cases} c(1-x) + k & \text{if } x \in [0,1), \\ 0 & \text{if } x = 1 \end{cases}$$

is a viscosity solution, which is discontinuous at the point $x = 1$. *Since* $Mu(x) = u(1) + c(1-x) + k$ *at all* $x \in \Omega$, *the discontinuity at the boundary point* $x = 1$ *plays an essential role and cannot be removed. Actually* $\tilde{u}(x) = c(1-x) + k$ *on* $[0,1]$ *is not a supersolution because it does not satisfy* (2.2) *at any* $x \in (0,1)$.

2.2. Boundary behavior of supersolutions. We begin by noticing that, under assumption H.9, a discontinuity at $\Gamma_{\rm in}$ may occur only if the optimal trajectory stands in jumping.

LEMMA 2.7. *We suppose that* H.1–H.7 *and* H.9 *hold. If* v *is a supersolution and* $x_o \in \Gamma_{\rm in}$ *is such that* $v(x_o) < \liminf_{\Omega \ni y \to x_o} v(y)$, *then* $v(x_o) \ge \min\{Mv(x_o), \varphi(x_o)\}$.

Proof. We assume by contradiction that $v(x_o) < \min\{Mv(x_o), \varphi(x_o)\}$. By semicontinuity arguments, we may suppose that $v(x_\varepsilon) < \min\{Mv(x_\varepsilon), \varphi(x_\varepsilon)\}$ for all sequences x_ε such that $x_\varepsilon \to x_o$ and $v(x_\varepsilon) \to v(x_o)$. Notice that we may assume without loss of generality that $x_\varepsilon \in \Gamma_{\rm in}$ for any ε small enough. Next, we take x_ε as a minimum point for the function

$$x \mapsto v(x) + \frac{|x - x_o|^2}{\varepsilon^2} - \frac{1}{\varepsilon^3} d(x).$$

Because v satisfies (1.8), then

$$H\left(x_\varepsilon, v(x_\varepsilon), -\frac{2}{\varepsilon^2}(x_\varepsilon - x_o) + \frac{1}{\varepsilon^3} Dd(x_\varepsilon), -\frac{2}{\varepsilon^2} I + \frac{1}{\varepsilon^3} D^2 d(x_\varepsilon)\right) \ge 0.$$

After computations we get

$$\sup_{\alpha\in\mathcal{A}}\left\{\operatorname{tr} a(x_\varepsilon,\alpha)+2b(x_\varepsilon,\alpha)\cdot(x_\varepsilon-x_o)+\varepsilon^2\left[\lambda v(x_\varepsilon)-f(x_\varepsilon,\alpha)\right]\right\}$$

$$\geq\frac{1}{\varepsilon}\inf_{\alpha\in\mathcal{A}}\left\{\frac{1}{2}\operatorname{tr}\left[a(x_\varepsilon,\alpha)D^2d(x_\varepsilon)\right]+b(x_\varepsilon,\alpha)\cdot Dd(x_\varepsilon)\right\}.$$

Passing to the limit as ε goes to zero gives a contradiction, because the left-hand side is bounded by virtue of H.1, while the right-hand side blows up because of H.9. □

In order to state a property of supersolutions that corresponds to Lemma 2.3 for subsolutions, we need to introduce the obstacle $\tilde{M}\varphi(x)$ defined by (1.10).

LEMMA 2.8. *Let v be a supersolution to* (1.3)–(1.4) *and set*

$$\tilde{v}(x)=\begin{cases}\liminf\limits_{\Omega\ni y\to x}v(y) & \text{if } x\in\Gamma_{\text{in}},\\ \\ v(x) & \text{if } x\in\overline{\Omega}\setminus\Gamma_{\text{in}}.\end{cases}\tag{2.5}$$

Under assumptions H.1–H.7, H.9*, then* $\min\{M\tilde{v},\tilde{M}\varphi\}\leq Mv$ *on* $\overline{\Omega}$.

Proof. We assume by contradiction that, at a certain $x\in\overline{\Omega}$, the desired inequality does not hold; equivalently, there exists $\xi_o\in\mathbb{R}^{\mathrm{N}}_+$, with $x+\xi_o\in\Gamma_{\text{in}}$, such that

$$Mv(x)=v(x+\xi_o)+c(\xi_o)+k,\tag{2.6}$$

$$v(x+\xi_o)<\liminf_{\Omega\ni y\to x+\xi_o}v(y),\tag{2.7}$$

$$v(x+\xi_o)+c(\xi_o)+k<\tilde{M}\varphi(x).\tag{2.8}$$

Since v is a supersolution, (2.7) and (2.8) imply that $v(x+\xi_o)\geq Mv(x+\xi_o)$, via Lemma 2.7. However, this is impossible because, in force of the subadditivity of c and of (2.6), we have that $Mv(x+\xi_o)\geq Mv(x)-c(\xi_o)>v(x+\xi_o)$. □

Eventually, we can remove the discontinuities at Γ_{in} of any supersolution to (1.3)–(1.4) and obtain a supersolution to the modified quasi-variational inequality

$$\max\left\{H\left(x,u,Du,D^2u\right),u-Mu,u-\tilde{M}\varphi\right\}=0\qquad\text{on }\Omega,\tag{2.9}$$

which satisfies the boundary condition (1.4) in the relaxed sense.

PROPOSITION 2.9. *Let assumptions* H.1–H.7, H.9 *hold and let v be a supersolution to* (1.3)–(1.4)*. Then the function $\tilde{v}$ defined by* (2.5) *is a supersolution to* (2.9)–(1.4).

Proof. By construction $\tilde{v}$ is lower semicontinuous on $\overline{\Omega}$ and $\tilde{v}\geq v$. Let $x\in\overline{\Omega}$ with $\tilde{v}(x)<\varphi(x)$ if $x\in\partial\Omega$ (otherwise there is nothing to prove): if $v(x)\geq Mv(x)$, then $\max\{\tilde{v}(x)-M\tilde{v}(x),\tilde{v}(x)-\tilde{M}\varphi(x)\}\geq 0$ by Lemma 2.8 and we are done. Otherwise, $H\left(x,v,Dv,D^2v\right)\geq 0$, and we need to check that $H(x,\tilde{v},D\tilde{v},D^2\tilde{v})\geq 0$ in the viscosity sense. $H(x,\tilde{v},D\tilde{v},D^2\tilde{v})\geq 0$ in the viscosity sense follows by the same argument of [5, Lemma 4.1] by force of the semicontinuity of Mv established by [14, Proposition 3.2]. □

We emphasize that, if u is any subsolution to (1.3)–(1.4), it is not true that $\tilde{u}$ is a subsolution to (2.9).

EXAMPLE 2.10. *Take Ω, σ, b, and c as in Example* 2.6*, $f(x)=-c'(1-x)+\lambda\left[c(1-x)+2k\right]$, and*

$$\varphi(x)=\begin{cases}c(1)+2k & \text{if } x=0,\\ 0 & \text{if } x=1.\end{cases}$$

In this case $u(x) = c(1-x) + 2k$ *is a viscosity solution, continuous on* $[0,1]$*, but it is not a subsolution to* (2.9) *because* $u(x) > \tilde{M}\varphi(x)$ *at all* $x \in [0,1]$*.*

3. The case when $\Gamma_{\rm in}$ is empty. This section concerns the case when, anytime that the diffusion degenerates along the normal direction to the boundary, there exists at least one continuous trajectory of (1.1) exiting from Ω with probability 1. So, any point of $\partial\Omega$ can be reached by means of a standard control, and the behavior at the boundary of the value function of the impulse control problem is similar to the standard one.

In view of attaining the continuity of the value function up to the boundary, we establish a strong comparison result for the related quasi-variational inequality.

THEOREM 3.1 (comparison principle). *Assume that* H.1–H.8 *hold and that* $\Gamma_{\rm in} = \emptyset$*. If* u *and* v *are a bounded upper semicontinuous subsolution and lower semicontinuous supersolution to* (1.3)–(1.4)*, then* $u \le v$ *in* Ω*. Moreover* $\tilde{u} \le v$ *in* $\overline{\Omega}$*, where* $\tilde{u}$ *is the subsolution defined in* (2.3)*.*

A relevant consequence of this comparison result is the characterization of the value function of the impulse control problem.

THEOREM 3.2. *Assume that* H.1–H.8 *hold and that* $\Gamma_{\rm in} = \emptyset$*. Then the value function* $\mathbf{U}$ *of the impulse control problem* (1.1)–(1.2) *is characterized in* Ω *as the unique viscosity solution to* (1.3)–(1.4)*. Moreover, it is continuous on* Ω *and can be continuously extended into* $\overline{\Omega}$*, by setting* $\mathbf{U}(x) = \limsup_{\Omega \ni y \to x} \mathbf{U}(y)$ *on* $\partial\Omega$*.*

Before entering into the details of the proofs, we want to emphasize that the assumption $\Gamma_{\rm in} = \emptyset$ cannot be dropped. To this end we sum up some properties of viscosity solutions that come out immediately by Theorem 3.1.

COROLLARY 3.3. *Assume that* H.1–H.8 *hold, that* $\Gamma_{\rm in} = \emptyset$*, and that* u *is a viscosity solution in the sense of Definition* 2.1*. Then* u *is continuous on* Ω *and*

$$u(x) \ge \limsup_{\Omega \ni y \to x} u(y) \qquad \text{on } \partial\Omega.$$

Moreover the function $\tilde{u}$ *defined as in* (2.3) *is the unique viscosity solution of* (1.3)–(1.4) *continuous on* $\overline{\Omega}$*.*

Next we show that the conclusions of Corollary 3.3 (and so the ones of Theorems 3.1 and 3.2) do not hold if there is some point of the boundary that can be reached only by jumping.

EXAMPLE 3.4. *Take* Ω*,* σ*,* b*,* c*,* f*, and* φ *as in Example* 2.6*. Now* $\Gamma_{\rm out} = \{0\}$*,* $\Gamma_{\rm in} = \{1\}$*, and for all* $\mathrm{C} \in [c(1) + k, \mathrm{C}_o]$*, the function*

$$u_{\mathrm{C}}(x) = \begin{cases} \mathrm{C} & \text{if } x = 0, \\ c(1-x) + k & \text{if } x \in (0,1), \\ 0 & \text{if } x = 1 \end{cases}$$

is a viscosity solution whose continuous extension to $\overline{\Omega}$ *is not a solution anymore.*

The phenomenon appearing on $\Gamma_{\rm out}$ *is the well-known boundary layer which characterizes viscosity solutions to the Dirichlet problem. In particular, choosing* $\mathrm{C} = c(1) + k$ *gives a solution which is continuous to* $\Gamma_{\rm out}$*. In contrast, the discontinuity coming up on* $\Gamma_{\rm in}$ *is intrinsic in the problem and cannot be removed, as observed in Example* 2.6*. When replacing the exact value of* u_{C} *with the limit from inside at* $\Gamma_{\rm in}$*, a bit of crucial information about the boundary datum is lost because* $M\tilde{u}_{\mathrm{C}}(x) > \tilde{M}\varphi(x)$ *near* $\Gamma_{\rm in}$*. Therefore, even though* $\tilde{u}_{\mathrm{C}}$ *satisfies* (1.8) *on* $\Gamma_{\rm in}$*, it fails to be a solution because* (2.2) *is no longer satisfied near* $\Gamma_{\rm in}$*.*

It is worth mentioning that the value function of the related impulse control problem is

$$\mathbf{U}(x) = \begin{cases} c(1-x)+k & \text{if } x \in [0,1), \\ 0 & \text{if } x = 1. \end{cases}$$

Thus, it is continuous in $\overline{\Omega} \setminus \Gamma_{\text{in}}$*, but its continuous extension to* $\overline{\Omega}$ *does not solve* (1.3).

Let us come to the proof of the strong comparison principle.

Proof of Theorem 3.1. The proof relies on establishing a comparison between the subsolution $\tilde{u}$ and the supersolution v. In order to deal with the nonlocal term M, we do not directly compare $\tilde{u}$ with v. Following an idea by [2], we define for any parameter $\mu \in (0,1)$

$$\hat{u} = (1-\mu)\,(\tilde{u} - \mathrm{B}), \qquad \hat{v} = v - \mathrm{B}, \qquad \hat{\varphi} = \varphi - \mathrm{B},$$

where $\mathrm{B} = \min\{0, \min\varphi, \min f/\lambda\}$. The functions $\hat{u}$ and $\hat{v}$ satisfy, respectively,

$$\max\{H(x,\hat{u},D\hat{u},D^2\hat{u}) + \lambda\mathrm{B}, \hat{u} - M\hat{u} + \mu k\} \le 0, \qquad x \in \Omega, \tag{3.1}$$

$$\min\{\max\{H(x,\hat{u},D\hat{u},D^2\hat{u}) + \lambda\mathrm{B}, \hat{u} - M\hat{u} + \mu k\}, \hat{u} - \hat{\varphi}\} \le 0, \quad x \in \partial\Omega, \tag{3.2}$$

$$\max\{H(x,\hat{v},D\hat{v},D^2\hat{v}) + \lambda\mathrm{B}, \hat{v} - M\hat{v}\} \ge 0, \qquad x \in \Omega, \tag{3.3}$$

$$\max\{H(x,\hat{v},D\hat{v},D^2\hat{v}) + \lambda\mathrm{B}, \hat{v} - M\hat{v}, \hat{v} - \hat{\varphi}\} \ge 0, \qquad x \in \partial\Omega. \tag{3.4}$$

The proof consists of showing that

$$\beta = \max_{\overline{\Omega}} (\hat{u} - \hat{v}) \le 0,$$

and it is achieved by making use of the standard technique of doubling variables and approaching the maximum point by means of two sequences x_n, y_n. The motivation for replacing $\tilde{u}$, v by $\hat{u}$, $\hat{v}$ is made clear by the following lemma.

LEMMA 3.5. *Let* x_n, y_n *be two sequences in* $\overline{\Omega}$ *such that* $x_n, y_n \to x_o \in \overline{\Omega}$ *and* $\hat{u}(x_n) - \hat{v}(y_n) \to \beta$*. Then* $\hat{v}(y_n) < M\hat{v}(y_n)$ *for all large* n*.*

Proof. We suppose by contradiction that there is a subsequence (that we still denote by n) such that $\hat{v}(y_n) \ge M\hat{v}(y_n)$, and we take $\xi_n \in \mathbb{R}^{\mathrm{N}}_+$ such that $y_n + \xi_n \in \overline{\Omega}$, $M\hat{v}(y_n) = \hat{v}(y_n + \xi_n) + c(\xi_n) + k$. Recalling Remark 2.5, we have

$$\hat{u}(x_n) - \hat{v}(y_n) \le \hat{u}(x_n + \eta_n) + c(\eta_n) + (1-\mu)k - \hat{v}(y_n + \xi_n) - c(\xi_n) - k$$

for all $\eta_n \in \mathbb{R}^{\mathrm{N}}_+$ such that $x_n + \eta_n \in \overline{\Omega}$. Besides, by classical compactness arguments we may assume without loss of generality that ξ_n converges to $\xi_o \in \mathbb{R}^{\mathrm{N}}_+$, with $x_o + \xi_o \in \overline{\Omega}$. Thus, choosing $\eta_n = P(x_n, \xi_o)$ and extracting the limit as $n \to 0$ gives the contradiction $\beta \le \hat{u}(x_o + \xi_o) - \hat{v}(x_o + \xi_o) - \mu k \le \beta - \mu k$. □

The inequality stated by Lemma 3.5 suggests that, when approaching a maximum point, $\hat{u}$ and $\hat{v}$ are close to being sub/supersolutions to the plain Dirichlet problem. Next, we suppose by contradiction that $\beta > 0$ and take a point $x_o \in \overline{\Omega}$ such that $\hat{u}(x_o) - \hat{v}(x_o) = \beta$. We reach a contradiction arguing separately from case to case.

Case $x_o \in \Omega$. We proceed as in the standard proof (see, for instance, [9]) by taking $(x_\varepsilon, y_\varepsilon)$ as a global maximum point for

$$\overline{\Omega}^2 \ni (x,y) \mapsto \hat{u}(x) - \hat{v}(y) - \frac{|x-y|^4}{\varepsilon^4} - |x - x_o|^4.$$

For all $\varepsilon > 0$, $H(x_\varepsilon, \hat{u}, D\hat{u}, D^2\hat{u}) + \lambda_B \leq 0$ by Remark 2.2, while $H(y_\varepsilon, \hat{v}, D\hat{v}, D^2\hat{v}) + \lambda_B \geq 0$ by Lemma 3.5. Eventually the thesis follows by classical arguments.

Case $x_o \in \Gamma_{\text{out}}$. Recalling Remark 2.2, [5, Proposition 4.1] yields that $\hat{u} \leq \hat{\varphi}$ on $\Gamma_{\text{out}} \cup \Gamma = \partial\Omega$. Hence, the assumption $\beta > 0$ implies that $\hat{v}(x_o) < \hat{\varphi}(x_o)$. Let us set

$$\psi_\varepsilon(y) = \frac{|y - x_o|^4}{\varepsilon^4} + \frac{1}{\varepsilon^2} d(y) - \frac{1}{2\varepsilon^3} d(y)^2.$$

It is easily seen that there exists a sequence y_ε of local minimum points for $\hat{v} + \psi_\varepsilon$ such that $y_\varepsilon \to x_o$ and $\hat{v}(y_\varepsilon) \to \hat{v}(x_o)$ as $\varepsilon \to 0$. In particular, we may suppose without loss of generality that $\hat{v}(y_\varepsilon) < \hat{\varphi}(y_\varepsilon)$ whenever $y_\varepsilon \in \partial\Omega$; in addition, Lemma 3.5 gives that $\hat{v}(y_\varepsilon) < M\hat{v}(y_\varepsilon)$ for small ε. Eventually (3.3)–(3.4) give that $H(x, \hat{v}, -D\psi_\varepsilon, -D^2\psi_\varepsilon) + \lambda_B \geq 0$ and, arguing as in the proof of [3, Proposition 1.1], one gets the contradiction $\mathcal{A}_{\text{in}}(x_o) \neq \emptyset$.

Case $x_o \in \Gamma$. As noticed while looking into the case $x_o \in \Gamma_{\text{out}}$, we already know that $\hat{u}(x_o) \leq \hat{\varphi}(x_o)$, and we have to check that the item $\hat{v}(x_o) < \hat{\varphi}(x_o)$ gives a contradiction. To begin, by replacing $\hat{v}$ with $\hat{v} + |\cdot - x_o|^4$, we may suppose that x_o is a strict maximum point. Next, we perturb $\hat{u}$ and $\hat{v}$ by means of a change of variable,

$$\hat{u}_\delta(x) = \omega_\delta^{-1}(\hat{u}(x)), \qquad \hat{v}_\delta(x) = \omega_\delta^{-1}\big(\hat{v}(x) + |x - x_o|^4\big),$$

where ω_δ is chosen close to identity, namely, $\omega_\delta'(s) = 1 - \delta e^{\omega_\delta(s)}$ and $\omega_\delta(0) = 0$. Moreover we set $\beta_\delta = \max\big\{\hat{u}_\delta(x) - \hat{v}_\delta(x) : x \in \overline{\Omega}\big\}$ and x_δ a maximum point. An easy computation gives that $x_\delta \to x_o$, $\beta_\delta \to \beta$, and $\hat{u}(x_\delta) - \hat{v}(x_\delta) \to \hat{u}(x_o) - \hat{v}(x_o)$, up to an extracted sequence. From now on, the proof proceeds as in [5, section 5.2], by force of Lemma 3.5. □

Remark 3.6. If $\partial\Omega = \Gamma_{\text{out}}$, Theorem 3.1 can be compared with [13, Theorem 3.1]: the main difference is that our proof does not need the sub/supersolutions to be "strong" in the sense of (1.5)–(1.6). Indeed, the comparison principle stated by Theorem 3.1 allows to enlarge the uniqueness class of [13, Theorem 4.1], including any viscosity solutions.

Next, if $\partial\Omega = \Gamma$, Theorem 3.1 can be compared with the result in [14]. Since we do not need to know a priori the behavior of the sub/supersolutions near the boundary, our nondegeneracy assumption H.8 is different from assumptions C.3 and C.4 of [14, Theorem 4.2]. On the other hand, following the proof of [14, Theorem 4.2], namely combining the iterative approximation scheme by Hanouzet and Joly [11] with the comparison principle stated by Theorem 3.1, gives the existence of a (unique) continuous viscosity solution.

Eventually we attain the characterization of the value function as the unique continuous viscosity solution to the quasi-variational inequality, as an easy consequence of Theorem 3.1.

Proof of Theorem 3.2. By virtue of the dynamic programming principle (see, for instance, [8, section III.1]), the value function satisfies

$$\mathbf{U}(x) = \inf_A \mathbb{E}\Bigg[\int_0^{T\wedge\tau_x} f(X_t^x, \alpha_t)\, e^{-\lambda t} dt + \sum_{\theta_i \leq T\wedge\tau_x} (c(\xi_i) + k)\, e^{-\lambda\theta_i} + 1_{\{T<\tau_x\}} \mathbf{U}\left(X_T^x\right) e^{-\lambda T} + 1_{\{T\geq\tau_x\}} \varphi\left(X_{\tau_x}^x\right) e^{-\lambda\tau_x} \Bigg]$$

for all $x \in \Omega$ and all stopping times T. Hence, standard arguments (see, for instance, [10, section V.3] and [4, section I]) give that $\mathbf{U}$ solves (1.3)–(1.4) in the viscosity

sense. Besides, Corollary 3.3 guarantees that any solution is continuous and that its extension, obtained by setting $\mathbf{U}(x) = \limsup_{\Omega \ni y \to x} \mathbf{U}(y)$ on $\partial\Omega$, is continuous in $\overline{\Omega}$ and still solves (1.3)–(1.4). Eventually, via Theorem 3.1, $\mathbf{U}$ is the unique continuous solution on Ω. □

4. The case when Γ_{in} is nonempty. Example 3.4 shows that, in general, if Γ_{in} is nonempty, we do not expect to have continuous solutions up to the boundary; in addition, we have to face the new difficulty that the continuous extension of the value function to $\overline{\Omega}$ does not satisfy (1.3). On the other hand, such situations are relevant for the applications to control theory, because a discontinuous value function at a boundary point reveals that the optimal trajectory stands in jumping to the boundary.

In this section, we first show that, in this case, it is possible to construct a number of viscosity solutions, which, in addition, are discontinuous inside Ω. This is due to the fact that the notion of relaxed boundary condition (1.7)–(1.8) does not capture all the information coming from (1.4) when the value function is discontinuous at Γ_{in}. Therefore, we enforce it by imposing the extra condition (1.9). Next, we characterize the value function as the unique viscosity solution fulfilling the strengthened notion of boundary condition, by means of a comparison result.

4.1. A suitable notion of boundary condition. With respect to the plain Dirichlet problem, the prescription of a boundary value is only a constraint at the boundary, which can be relaxed in the viscosity sense by substituting (1.4) with (1.7)–(1.8). When dealing with quasi-variational inequalities, the prescription of a boundary value is also a constraint inside Ω, because it is implicitly required that

$$u(x) \le \inf\left\{\varphi(x+\xi) + c(\xi) + k \,:\, \xi \in \mathbb{R}^{\mathrm{N}}_{+}, x+\xi \in \partial\Omega\right\} \qquad \text{in } \Omega. \tag{4.1}$$

If the diffusion is nondegenerate, the boundary condition (1.4) is fulfilled; therefore (4.1) is implied by requiring that $u \le Mu$ on Ω. However, if the diffusion degenerates at the boundary, a boundary layer arises and the constraint (4.1) can be violated. It is the case, for instance, if the optimal trajectory stands in jumping at a point $x \in \partial\Omega$: now $\mathbf{U}(x) = \varphi(x)$ and $\mathbf{U}(y) = M\mathbf{U}(y) = \mathbf{U}(x) + c(x-y) + k$ around x. The "standard" viscosity boundary condition only requires that the upper semicontinuous envelope of $\mathbf{U}$ satisfy the quasi-variational inequality at x and neglects that, when replacing $\mathbf{U}$ by its upper semicontinuous envelope, the obstacle $M\mathbf{U}$ is artificially increased all over Ω. Therefore this notion of boundary condition does not capture the constraint (4.1). As a consequence, it is not sufficient to pick up a unique solution, as shown by the following example.

EXAMPLE 4.1. *Take* $\Omega = (0,3)$, $\sigma \equiv 0$,

$$b(x) = \begin{cases} -(x-1)^2 & \text{if } x \in [0,1], \\ 0 & \text{if } x \in [1,2], \\ -(x-2)^2 & \text{if } x \in [2,3], \end{cases}$$

c *a strictly increasing function,* $f(x) \ge \lambda[c(3-x) + 2k]$, *and*

$$\varphi(x) = \begin{cases} c(3) + 2k & \text{if } x = 0, \\ 0 & \text{if } x = 3. \end{cases}$$

In that case $\Gamma_{\mathrm{out}} = \{0\}$, $\Gamma_{\mathrm{in}} = \{3\}$, *and jumping immediately to* Γ_{in} *is a better strategy than following any continuous trajectory. The value function is given by*

$$\mathbf{U}(x) = \begin{cases} c(3-x) + k & \text{if } x \in [0,3), \\ 0 & \text{if } x = 3. \end{cases}$$

Notice that the $\mathbf{U}$ *is continuous on* $\overline{\Omega} \setminus \Gamma_{\text{in}}$ *and lower semicontinuous on* $\overline{\Omega}$*, and that it fulfills the condition* $\mathbf{U} \leq \tilde{M}\varphi$ *on* $\overline{\Omega}$*. Moreover, its continuous extension on* $\overline{\Omega}$ *is not a solution anymore, by the same arguments of Example* 4.1.

On the other hand, for all $z \in (1,2)$ *the function*

$$u_z(x) = \begin{cases} c(3-x) + 2k & \text{if } x \in [0,z), \\ c(3-x) + k & \text{if } x \in [z,3), \\ 0 & \text{if } x = 3 \end{cases}$$

is a viscosity solution, which is discontinuous at the interior point z*. When taking the upper semicontinuous envelope of* u_z*, the information about the boundary datum on* Γ_{in} *is lost: removing the discontinuity at* Γ_{in} *artificially increases the nonlocal term* Mu_z^**, so that* $Mu_z^*(x) > \tilde{M}\varphi(x)$*.*

This example reveals that some selection criterion is needed for uniqueness. Replacing a subsolution with its upper semicontinuous envelope makes the term Mu increase and possibly exceed $M\phi$, so that some relevant information about the boundary datum is lost. Hence we need to give a more appropriate notion of the relaxed boundary condition.

Definition 4.1. *A function* u *which is locally bounded and upper semicontinuous on* $\overline{\Omega}$ *is a* selected viscosity subsolution *to* (1.3)–(1.4) *if it is a viscosity subsolution in the sense of Definition* 2.1 (*i.e., it fulfills* (2.2), (1.8)) *and, in addition, satisfies* (1.9)*.*

A locally bounded function u *is a* selected viscosity solution *to* (1.3)–(1.4) *if it is a viscosity solution in the sense of Definition* 2.1 *and, in addition, its upper semicontinuous envelope satisfies* (1.9)*.*

Remark 4.2. Let u be an upper semicontinuous function. Because in general $\limsup_{\Omega \ni y \to x} u(y) \leq u(x)$, the condition (1.9) is weaker than $u \leq \tilde{M}\varphi$ on $\overline{\Omega}$. On the other hand, since we do not ask a priori that the function $\overline{\Omega} \ni x \mapsto \tilde{M}\varphi(x)$ be upper semicontinuous, condition (1.9) is not implied by asking that $u(x) \leq \tilde{M}\varphi$ on Ω. Actually, the condition (1.9) is equivalent to asking that $\tilde{u} \leq \tilde{M}\varphi$ on $\overline{\Omega}$, where $\tilde{u}$ has been defined in (2.3).

We explicitly notice that any function is a selected viscosity solution if and only if its upper semicontinuous envelope is a selected viscosity subsolution and its lower semicontinuous envelope is a supersolution according to Definition 2.1. In view of Proposition 2.9, a selection criterion for supersolutions is not needed.

Up to modifying the values in the Γ_{in} part of the boundary, there is a one-to-one correspondence between the selected solutions to (1.3) and the solutions to (2.9) which do not jump downward at Γ_{in}.

Lemma 4.3. *If* u *is a selected solution to* (1.3)–(1.4)*, then*

$$w(x) = \begin{cases} u(x) & \text{if } x \in \overline{\Omega} \setminus \Gamma_{\text{in}}, \\ \liminf\limits_{\Omega \ni y \to x} u(y) & \text{if } x \in \Gamma_{\text{in}} \end{cases}$$

solves (2.9)–(1.4)*. Conversely, let* w *be a solution to* (2.9)–(1.4) *with the regularity property*

$$w_*(x) = \liminf_{\Omega \ni y \to x} w(y) \quad \text{for all } x \in \Gamma_{\text{in}}$$

(where w_* stands for the lower semicontinuous envelope of w) and set

$$u(x) = \begin{cases} w(x) & \text{if } x \in \overline{\Omega} \setminus \Gamma_{\text{in}}, \\ \min\left\{ \liminf_{\Omega \ni y \to x} w(y), \varphi(x) \right\} & \text{if } x \in \Gamma_{\text{in}}. \end{cases}$$

Then u is a selected solution to (1.3)–(1.4).

Proof. The first part of the statement is an immediate consequence of Proposition 2.9. With respect to the second part, we begin by noticing that w^* is a subsolution to (1.3)–(1.4). Moreover the same arguments of Proposition 2.4 yield that $u^*(x)$ is a subsolution to (1.3)–(1.4). Therefore, since u^* satisfies (1.9) by construction, it is a selected subsolution.

In order to check that the lower semicontinuous envelope of u is a supersolution to (1.3)–(1.4) in the sense of Definition 2.1, take $x \in \overline{\Omega}$ with $u_*(x) < \varphi(x)$ if $x \in \partial\Omega$ (otherwise there is nothing to prove). By construction, $u_*(x) = w_*(x)$ and $Mu_*(x) = \min\{Mw_*(x), \tilde{M}\varphi(x)\}$. If $w_*(x) \geq \min\{Mw_*(x), \tilde{M}\varphi(x)\}$, then $u_*(x) = w_*(x) \geq \min\{Mw_*(x), \tilde{M}\varphi(x)\} = Mu_*(x)$. Otherwise, $H(x, w_,* Dw_*, D^2 w_*) \geq 0$ and the same arguments of the proof of Proposition 2.9 yield that $H(x, u_*, Du_*, D^2 u_*) \geq 0$. □

As our goal relies on eliminating solutions which are not related to the impulse control problem, we have to verify that the value function is a selected viscosity solution.

PROPOSITION 4.4. *The value function of the impulse control problem* (1.1)–(1.2) *is a selected viscosity solution to* (1.3)–(1.4).

Proof. Proceeding as in the proof of Theorem 3.2, one gets that the value function $\mathbf{U}$ solves (1.3)–(1.4) according to Definition 2.1. On the other hand, at all $x \in \overline{\Omega}$ and for all $\xi \in \mathbb{R}^{\mathrm{N}}_+$ such that $x + \xi \in \overline{\Omega}$, one may chose the control $\theta_o = 0, \theta_1 = \infty, \xi_o = \xi$. Therefore (1.2) gives that $\mathbf{U}(x) \leq \varphi(x + \xi) + c(\xi) + k$, and taking the infimum over all ξ gives $\mathbf{U}(x) \leq \tilde{M}\varphi(x)$ at all $x \in \overline{\Omega}$. □

4.2. Comparison principle. We are now ready to characterize the value function as the unique selected viscosity solution.

THEOREM 4.5. *Under assumptions* H.1–H.9*, the value function* $\mathbf{U}$ *of the impulse control problem* (1.1)–(1.2) *is characterized in* Ω *as the unique selected viscosity solution to* (1.3)–(1.4)*. Moreover, it is continuous on* Ω *and can be continuously extended into* $\overline{\Omega} \setminus \Gamma_{\text{in}}$ *by setting*

$$\mathbf{U}(x) = \begin{cases} \limsup_{\Omega \ni y \to x} \mathbf{U}(y) & \text{if } x \in \Gamma_{\text{out}} \cup \Gamma, \\ \min\left\{ \liminf_{\Omega \ni y \to x} \mathbf{U}(y), \varphi(x) \right\} & \text{if } x \in \Gamma_{\text{in}}. \end{cases} \tag{4.2}$$

This characterization is achieved by means of the following comparison result.

THEOREM 4.6 (comparison principle). *Under assumptions* H.1–H.9*, if u is a bounded upper semicontinuous selected subsolution and v is a bounded lower semicontinuous supersolution to* (1.3)–(1.4)*, then $u \leq v$ in Ω. Moreover, $\tilde{u} \leq \tilde{v}$ in $\overline{\Omega}$, where $\tilde{u}$ and $\tilde{v}$ have been defined in* (2.3) and (2.5)*, respectively.*

Before entering into the details of the proof of the comparison principle, we show how it brings us to Theorem 4.5. To this end, we first list the properties of viscosity solutions that arise from Theorem 4.6.

COROLLARY 4.7. *Assume that* H.1–H.9 *hold and let* u *be a selected viscosity solution to* (1.3)–(1.4). *Then* u *is continuous on* Ω *and*

$$\begin{aligned} u(x) &\geq \limsup_{\Omega \ni y \to x} u(y) \qquad \text{on } \Gamma_{\text{out}} \cup \Gamma, \\ u(x) &\leq \liminf_{\Omega \ni y \to x} u(y) \qquad \text{on } \Gamma_{\text{in}}. \end{aligned}$$

Moreover, the function

$$\underline{u}(x) = \begin{cases} u(x) & \text{if } x \in \Omega, \\ \limsup_{\Omega \ni y \to x} u(y) & \text{if } x \in \Gamma_{\text{out}} \cup \Gamma, \\ \min\left\{ \liminf_{\Omega \ni y \to x} u(y), \varphi(x) \right\} & \text{if } x \in \Gamma_{\text{in}} \end{cases}$$

is continuous on $\overline{\Omega} \setminus \Gamma_{\text{in}}$, *still is a selected viscosity solution, and is the minimal bounded solution to* (1.3)–(1.4) *(according to Definition* 2.1*).*

Notice that, in view of Example 4.1, the continuous extension of a solution to $\overline{\Omega}$ generally is not a solution.

Proof. The first part of the statement is a trivial consequence of Theorem 4.6, and we prove only that $\underline{u}$ is the minimal solution. By taking advantage of Lemma 4.3, one may easily obtain that $\underline{u}$ is a selected solution by checking that

$$w(x) = \begin{cases} u(x) & \text{if } x \in \Omega, \\ \limsup_{\Omega \ni y \to x} u(y) & \text{if } x \in \Gamma_{\text{out}} \cup \Gamma, \\ \liminf_{\Omega \ni y \to x} u(y) & \text{if } x \in \Gamma_{\text{in}} \end{cases}$$

solves (2.9)–(1.4).

Next we prove that $\underline{u}$ is the minimal solution. First of all, we observe that any viscosity solution which lacks property (1.9) is bigger (at least at some point) than any selected viscosity solution. Therefore, it suffices to show that $\underline{u} \leq v$ on $\overline{\Omega}$ for any bounded selected viscosity solution v. However, the comparison principle stated by Theorem 4.6 implies that $\underline{u} = u = v$ on Ω and $v(x) \geq \underline{u}(x)$ for all $x \in \Gamma_{\text{out}} \cup \Gamma$. Hence, it remains to show that $\underline{u} \leq v$ on Γ_{in}.

To this end we suppose by contradiction that $v(x) < \underline{u}(x)$ at some $x \in \Gamma_{\text{in}}$; then Lemma 2.7 yields that $v(x) \geq \min\{Mv(x), \varphi(x)\}$ and, because $v(x) < \varphi(x)$, $v(x) \geq Mv(x)$, indeed. Set $\xi_o \in \mathbb{R}^{\text{N}}_+$ such that $Mv(x) = v(x+\xi_o) + c(\xi_o) + k$: in particular, $\xi_o \neq 0$ and $M\underline{u}(x) \geq \underline{u}(x) > v(x) \geq v(x+\xi_o) + c(\xi_o) + k$. Therefore $v(x+\xi_o) < \underline{u}(x+\xi_o)$ and $x+\xi_o \in \Gamma_{\text{in}}$. Iterating these arguments brings the existence of two sequences $\xi_n \in \mathbb{R}^{\text{N}}_+$, $\xi_n \neq 0$, and $y_n = x + \sum_{j=0}^{n+1} \xi_j \in \overline{\Omega}$ for which

$$v(x) \geq v(y_n) + \sum_{j=0}^{n+1} c(\xi_j) + (n+1)k.$$

However, this is absurd because it implies that

$$\inf_{\overline{\Omega}} v \le \lim_{n\to\infty} v(y_n) \le v(x) - \lim_{n\to\infty} \left[\sum_{j=0}^{n+1} c(\xi_j) + (n+1)k\right] = -\infty. \qquad \square$$

Corollary 4.7 immediately implies the result stated by Theorem 4.5. We conclude this section by proving the comparison result.

Proof of Theorem 4.6. We follow the same scheme of the proof of Theorem 3.1, by replacing $\tilde{u}$ and $\tilde{v}$ with

$$\hat{u} = (1-\mu)\,(\tilde{u} - \mathrm{B})\,, \qquad\qquad \hat{v} = \tilde{v} - \mathrm{B}.$$

We recall that $\hat{u}$ satisfies (3.1)–(3.2) and, in addition,

$$\hat{u} \le \tilde{M}\hat{\varphi} - \mu k \qquad \text{on } \overline{\Omega}. \tag{4.3}$$

Besides, $\hat{v}$ satisfies

$$\max\Big\{H(x,\hat{v},D\hat{v},D^2\hat{v}) + \lambda\mathrm{B}, \hat{v} - M\hat{v}, \hat{v} - \tilde{M}\hat{\varphi}\Big\} \ge 0, \qquad x \in \Omega, \tag{4.4}$$

$$\max\Big\{H(x,\hat{v},D\hat{v},D^2\hat{v}) + \lambda\mathrm{B}, \hat{v} - M\hat{v}, \hat{v} - \tilde{M}\hat{\varphi}, \hat{v} - \hat{\varphi}\Big\} \ge 0, \ x \in \partial\Omega. \tag{4.5}$$

The proof consists of checking that

$$\beta = \max_{\overline{\Omega}} (\hat{u} - \hat{v}) \le 0,$$

and it is achieved arguing by contradiction. The role of Lemma 3.5 is now played by the following result.

LEMMA 4.8. *Let x_n, y_n be two sequences in $\overline{\Omega}$ such that*

$$x_n, y_n \to x_o \in \overline{\Omega},$$

$$\hat{u}(x_n) - \hat{v}(y_n) \to \beta > 0.$$

Then $\hat{v}(y_n) < \min\{M\hat{v}(y_n), \tilde{M}\hat{\varphi}(y_n)\}$ for all large n.

Proof. We suppose by contradiction that there is a subsequence (that we still denote by n) such that $\hat{v}(y_n) \ge \min\{M\hat{v}(y_n), \tilde{M}\hat{\varphi}(y_n)\}$. Without loss of generality, we may assume that one of the following items is fulfilled:

(a) for all n, there exists $\xi_n \in \mathbb{R}^{\mathrm{N}}_+$ such that $y_n + \xi_n \in \overline{\Omega} \setminus \Gamma_{\mathrm{in}}$ and

$$\min\Big\{M\hat{v}(y_n), \tilde{M}\hat{\varphi}(y_n)\Big\} = \hat{v}(y_n + \xi_n) + c(\xi_n) + k;$$

(b) for all n, there exists $\xi_n \in \mathbb{R}^{\mathrm{N}}_+$ such that $y_n + \xi_n \in \Gamma_{\mathrm{in}}$ and

$$\min\Big\{M\hat{v}(y_n), \tilde{M}\hat{\varphi}(y_n)\Big\} = \hat{\varphi}(y_n + \xi_n) + c(\xi_n) + k.$$

In the first case, the thesis is obtained as in Lemma 3.5. Otherwise, because Γ_{in} is closed by hypothesis H.7, we may assume without loss of generality that ξ_n converges

to $\xi_o \in \mathbb{R}^N_+$ with $x_o+\xi_o \in \Gamma_{\rm in}$. Therefore, by making use of the property (4.3) at the point x_o, we obtain

$$\hat u(x_n) - \hat v(y_n) \le \hat u(x_n) - \hat u(x_o) + \hat\varphi(x_o+\xi_o) + c(\xi_o) + (1-\mu)k - \hat\varphi(y_n+\xi_n) - c(\xi_n) - k.$$

Eventually, by taking advantage of the semicontinuity of $\hat u$ and of the continuity of $\hat\varphi$, c, passing to the limit as $n \to \infty$ gives the contradiction $\beta \le -\mu k < 0$. □

Next, we take a point $x_o \in \overline\Omega$ such that $\hat u(x_o) - \hat v(x_o) = \beta > 0$, and we reach a contradiction arguing separately from case to case. In view of Lemma 4.8, the cases $x_o \in \Omega \cup \Gamma_{\rm out} \cup \Gamma$ may be dealt with as in the proof of Theorem 3.1. Therefore, it remains to study the case $x_o \in \Gamma_{\rm in}$.

Case $x_o \in \Gamma_{\rm in}$. Take $(x_\varepsilon, y_\varepsilon)$ as a global maximum point for

$$\overline\Omega^2 \ni (x,y) \mapsto \hat u(x) - \hat v(y) - \frac{|x-y|^4}{\varepsilon^4} - |y-x_o|^4 - \frac{\varepsilon}{d(y)},$$

and notice that $x_\varepsilon, y_\varepsilon \to x_o$ and $\hat u(x_\varepsilon) - \hat v(y_\varepsilon) \to \beta$ as ε goes to zero; in particular, $x_\varepsilon \in \Omega \cup \Gamma_{\rm in}$ and $y_\varepsilon \in \Omega$. Next, via Remark 2.2, [5, Proposition 4.2] yields that $H(x_\varepsilon, \hat u, D\hat u, D^2\hat u) + \lambda_{\rm B} \le 0$. In addition (4.4) and Lemma 4.8 imply that $H(y_\varepsilon, \hat v, D\hat v, D^2\hat v) + \lambda_{\rm B} \ge 0$. Therefore the thesis follows by standard arguments. □

5. Existence of solution: A constructive procedure. We conclude this paper by producing a selected viscosity solution by making use of the iterative approximation scheme introduced by Hanouzet and Joly in [11]. This procedure has been used before in [14, Theorem 4.2], where a solution to (1.3)–(1.4), continuous on $\overline\Omega$, has been produced as the uniform limit of the sequence:

- u_o is the solution to the plain Dirichlet problem (1.3.o)–(1.4),
- for all $n \ge 1$, u_n is the solution to the variational inequality

$$(1.3.n) \qquad \max\left\{H\left(x, u_n, Du_n, D^2u_n\right), u_n - Mu_{n-1}\right\} = 0 \qquad \text{on } \Omega,$$

satisfying the boundary condition (1.4).

If $\Gamma_{\rm in} \neq \emptyset$, in general the sequence u_n does not converge to the value function of the control problem.

EXAMPLE 5.1. *Take Ω, σ, b, φ as in Example* 4.1, *c of class $\mathcal{C}^1$ with $c' > 0$, and $f(x) = b(x)c'(3-x) + \lambda\,[c(3-x) + 2k]$. It is easy to check that $u_o(x) = c(3-x) + 2k$ is the unique solution in $\mathcal{C}(\overline\Omega)$ to the Dirichlet problem* (1.3.o)–(1.4). *In addition, $Mu_o = u_o + k$ on $\overline\Omega$, so that u_o also solves* (1.3.1)–(1.4) *and, by iteration, $u_n = u_o$ for all $n \in \mathbb{N}$. Actually u_o is a solution to* (1.3)–(1.4) *continuous on $\overline\Omega$, but $\tilde M\varphi = u_o(x) - k$. Hence u_o does not satisfy* (1.9) *and differs from the value function of the related control problem (whose explicit form has been given in Example* 4.1*).*

Notice that, if we do not superimpose the condition (1.9), *any information about the boundary cost on $\Gamma_{\rm in} = \{3\}$ is forgotten just in the zeroth step of the iterative scheme: indeed, u_o solves the Dirichlet problem with any boundary cost $\hat\varphi$ such that $\hat\varphi(3) \le 2k$.*

Therefore, in order to achieve convergence to the value function, it is necessary to add the information coming from the $\Gamma_{\rm in}$-part of the boundary into the iterative scheme. This allows us to exploit the stability of a modified version of the monotone iterative scheme by Hanouzet and Joly.

THEOREM 5.2. *Assume that* H.1–H.9 *hold and that $\tilde M\varphi \in \mathcal{C}(\overline\Omega)$. Then there exists a selected viscosity solution $u \in \mathcal{C}\left(\overline\Omega \setminus \Gamma_{\rm in}\right)$ to* (1.3)–(1.4). *Such a solution has*

the form

$$u(x) = \begin{cases} w(x), & x \in \overline{\Omega} \setminus \Gamma_{\text{in}}, \\ \min\{w(x), \varphi(x)\}, & x \in \Gamma_{\text{in}}, \end{cases}$$

where $w \in \mathcal{C}(\overline{\Omega})$ *is the uniform limit of the sequence:*

- u_o *is the solution to the plain Dirichlet problem* (1.3.o)–(1.4)*,*
- *for all* $n \geq 1$*,* u_n *is the solution to the variational inequality*

$$\max\Big\{H\big(x, u_n, Du_n, D^2u_n\big), u_n - Mu_{n-1}, u_n - \tilde{M}\varphi\Big\} = 0 \quad \text{on } \Omega, \tag{2.9.n}$$

satisfying the boundary condition (1.4)*.*

The proof of Theorem 5.2 requires the study of variational inequalities of type

$$\max\big\{H\big(x, v, Dv, D^2v\big), v - \psi_1, v - \psi_2\big\} = 0 \qquad \text{on } \Omega. \tag{5.1}$$

We begin by noticing that (5.1)–(1.4) may be interpreted as a Hamilton–Jacobi–Bellman equation after redefining the set of the controls as $\hat{\mathcal{A}} = \mathcal{A} \times \{0, 1, 2\}$ and the coefficients of the dynamics and the running cost as

$$\hat{\sigma}(x, \alpha, i) = \begin{cases} \sigma(x, \alpha) & \text{if } i = 0, \\ 0 & \text{if } i = 1, 2, \end{cases}$$

$$\hat{b}(x, \alpha, i) = \begin{cases} b(x, \alpha) & \text{if } i = 0, \\ 0 & \text{if } i = 1, 2, \end{cases}$$

$$\hat{f}(x, \alpha, i) = \begin{cases} f(x, \alpha) & \text{if } i = 0, \\ \psi_i(x) & \text{if } i = 1, 2. \end{cases}$$

In particular, the strong comparison principle [5, Theorem 2.1] applies, under assumptions H.1–H.5, H.7, H.8, whenever $\psi_i \in \mathcal{C}(\overline{\Omega})$, as $i = 1, 2$. Therefore, the proof of the next well-posedness result follows from a standard application of the technique of half-relaxed limits by Barles and Perthame [4].

PROPOSITION 5.3. *Let* $\psi_1, \psi_2 \in \mathcal{C}(\overline{\Omega})$*. Under assumptions* H.1–H.5*,* H.7*,* H.8*, there exists a unique continuous viscosity solution to the variational inequality* (5.1)–(1.4)*.*

Finally we establish the convergence of the modified Hanouzet and Joly scheme $(2.9.n)_n$.

Proof of Theorem 5.2. Because of Lemma 4.3, it suffices to show that w solves (2.9)–(1.4). We begin by noticing that, by virtue of the Perron method [9, Theorem 4.1] and of the strong comparison principle [5, Theorem 2.1], the Dirichlet problem (1.3.o)–(1.4) has a unique solution u_o in $\mathcal{C}(\overline{\Omega})$.

Next, [14, Proposition 3.2] yields that $Mu_o \in \mathcal{C}(\overline{\Omega})$, while $\tilde{M}\varphi \in \mathcal{C}(\overline{\Omega})$ by hypothesis. Hence, Proposition 5.3 guarantees the existence of a solution $u_1 \in \mathcal{C}(\overline{\Omega})$ to (2.9.1)–(1.4). In addition, u_1 is a subsolution to (1.3.o)–(1.4), so that [5, Theorem 2.1] gives that $u_1 \leq u_o$.

Again, [14, Proposition 3.2] yields that $Mu_1 \in \mathcal{C}(\overline{\Omega})$, and Proposition 5.3 guarantees the existence of a solution $u_2 \in \mathcal{C}(\overline{\Omega})$ to (1.3.2)–(1.4). Since $u_1 \leq u_o$, $Mu_1 \leq Mu_o$,

$$\max\Big\{H\big(x, u_2, Du_2, D^2u_2\big), u_2 - Mu_o, u_2 - \tilde{M}\varphi\Big\}$$

$$\leq \max\Big\{H\big(x, u_2, Du_2, D^2u_2\big), u_2 - Mu_1, u_2 - \tilde{M}\varphi\Big\} \leq 0.$$

Thus u_2 is a subsolution to (2.9.1)–(1.4), and [5, Theorem 2.1] gives that $u_2 \le u_1$. By iterating these arguments, we obtain a decreasing sequence $u_n \in \mathcal{C}(\overline{\Omega})$.

On the other hand, up to replacing φ by $\varphi - \mathrm{B}$ and f by $f + \lambda \mathrm{B}$, with $\mathrm{B} = \min\{0, \min \varphi, \min f/\lambda\}$, we may suppose without loss of generality that $\varphi, f \ge 0$, so that 0 is a subsolution to all problems (2.9.n)–(1.4), and [5, Theorem 2.1] guarantees that $u_n \ge 0$ on $\overline{\Omega}$ for all n.

Afterward, the same arguments of the proof of [14, Theorem 4.2] give that

$$u_{n+1} - u_{n+2} \le (1-\mu)^n \|u_o\|_{\mathcal{C}(\overline{\Omega})} \qquad \text{on } \overline{\Omega},$$

where $\mu < 1$ such that $\mu \|u_o\|_{\mathcal{C}(\overline{\Omega})} \le k$. Therefore the sequence u_n converges to a function $w \in \mathcal{C}(\overline{\Omega})$, uniformly on $\overline{\Omega}$. Eventually, the stability result [16, Proposition I.3] implies that w is a solution to (2.9)–(1.4). □

6. Conclusions. The counterexamples 4.1 and 5.1 show that reading the boundary condition (1.4) in the "standard" viscosity sense (1.7)–(1.8) is underdetermining when dealing with nonlocal problems such as (1.3).

In these cases, the loss of information caused by the boundary layer propagates into the interior of Ω. Thus the additional request $H(x, u, Du, D^2u) = 0$ at the boundary points where loss of boundary data may occur is not sufficient to offset such loss of information. Actually, it is needed to introduce the selection criterion (1.9) in order to spread the information coming from the boundary condition into all Ω.

Eventually, the triplet of conditions (1.7)–(1.8)–(1.9) seems to be the suitable way to relax the boundary condition, with respect to a nonlocal problem such as (1.3).

Acknowledgment. The author wishes to thank Prof. Guy Barles for having proposed to her the study of this problem and for his useful advice.

REFERENCES

[1] G. Barles, *Deterministic impulse control problems*, SIAM J. Control Optim., 23 (1985), pp. 419–432.

[2] G. Barles, *Quasivariational inequalities and first-order Hamilton–Jacobi equations*, Nonlinear Anal., 9 (1985), pp. 131–148.

[3] G. Barles and J. Burdeau, *The Dirichlet problem for semilinear second-order degenerate elliptic equations and applications to stochastic exit time control problems*, Comm. Partial Differential Equations, 20 (1995), pp. 129–178.

[4] G. Barles and B. Perthame, *Exit time problems in optimal control and vanishing viscosity method*, SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

[5] G. Barles and E. Rouy, *A strong comparison result for the Bellman equation arising in stochastic exit time control problems and its applications*, Comm. Partial Differential Equations, 23 (1998), pp. 1995–2033.

[6] A. Bensoussan and J.-L. Lions, *Controle impulsionnel et inequations quasi variationnelles*, Methodes Mathematiques de l'Informatique 11, Dunod, Paris, 1982.

[7] A.-P. Blanc, *Deterministic exit time control problems with discontinuous exit costs*, SIAM J. Control Optim., 35 (1997), pp. 399–434.

[8] V. S. Borkar, *Optimal Control of Diffusion Processes*, Pitman Res. Notes Math. Series 203, John Wiley & Sons, New York, 1989.

[9] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[10] W. H. Fleming and H. Soner, *Controlled Markov Processes and Viscosity Solutions*, Appl. Math. 25, Springer-Verlag, New York, 1993.

[11] B. Hanouzet and J.-L. Joly, *Convergence uniforme des itérés définissant la solution d'une inéquation quasi variationnelle abstraite*, C. R. Acad. Sci. Paris Sér. A, 286 (1978), pp. 735–738.

[12] H. Ishii, *Hamilton-Jacobi equations with discontinuous Hamiltonians on arbitrary open sets*, Bull. Fac. Sci. Eng. Chuo Univ., 28 (1985), pp. 33–77.
[13] K. Ishii, *Viscosity solutions of nonlinear second order elliptic PDEs associated with impulse control problems*, Funkcial. Ekvac., 36 (1993), pp. 123–141.
[14] K. Ishii, *Viscosity solutions of nonlinear second order elliptic PDEs associated with impulse control problems*. II, Funkcial. Ekvac., 38 (1995), pp. 297–328.
[15] M. A. Katsoulakis, *Viscosity solutions of second order fully nonlinear elliptic equations with state constraints*, Indiana Univ. Math. J., 43 (1994), pp. 493–519.
[16] P.-L. Lions, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations.* II: *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.
[17] B. Øksendal and A. Sulem, *Optimal consumption and portfolio with both fixed and proportional transaction costs: A combined stochastic control and impulse control model*, SIAM J. Control Optim., 40 (2002), pp. 1765–1790.
[18] B. Perthame, *Quasivariational inequalities and Hamilton–Jacobi–Bellman equations in a bounded region*, Comm. Partial Differential Equations, 9 (1984), pp. 561–595.
[19] P. Soravia, *Discontinuous viscosity solutions to Dirichlet problems for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 18 (1993), pp. 1493–1514.
[20] J. J. Ye, *Discontinuous solutions of the Hamilton–Jacobi equation for exit time problems*, SIAM J. Control Optim., 38 (2000), pp. 1067–1085.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 590–623

OBSERVATION AND CONTROL OF VIBRATIONS IN TREE-SHAPED NETWORKS OF STRINGS*

RENÉ DÁGER†

Abstract. In this paper we study the controllability problem for a system that models the vibrations of a controlled tree-shaped network of vibrating elastic strings. The control acts through one of the exterior nodes of the network. With the help of the d'Alembert representation formula for the solutions of the one-dimensional wave equation, we find certain linear relations between the traces of the solutions at the nodes of the network. These relations allow us to prove a weighted observability inequality with weights that may be explicitly computed in terms of the eigenvalues of the associated elliptic problem. We characterize the class of trees for which all those weights are different from zero, which leads to the spectral controllability of the system. Additionally, we consider the same one-node control problem for several networks that are controlled simultaneously.

Key words. controllability, observability, string network, wave equation

AMS subject classifications. 35L05, 35J05, 35L20

DOI. 10.1137/S0363012903421844

Introduction. In this paper we study the one-node controllability property of the vibrations of a planar network (i.e., several strings connected at their ends) of elastic homogeneous strings. That is, we analyze the possibility of driving to rest the motion of the network produced by an initial deformation of its strings, by means of a control applied through one of the nodes.

The network considered in this paper coincides at rest with a planar, finite, connected graph without closed paths, whose edges are straight segments. The deformations of the strings of the network are assumed to be transverse to the plane determined by the at-rest graph. The deformation of every string is expressed by means of a scalar function, defined on the edge of the graph that corresponds to the string and satisfying the one-dimensional (1-d) wave equation. At the interior nodes of the network, i.e., those where the strings are coupled, it is assumed that the displacements of all the strings coincide and that the sum of their tensions is equal to zero. These conditions express the continuity of the network and the balance of forces at the junction points. A control acts on one of the exterior nodes that regulates its displacement. The remaining exterior nodes are supposed to be clamped; that is, their displacements are equal to zero.

By now, there is an extensive literature devoted to the nodal controllability and stabilizability of networks (see, e.g., [7, 8] and the references therein). The main tool used to solve these problem has been the Hilbert uniqueness method, introduced by Lions [10], combined with multiplier and characteristic methods.

The question of the nodal controllability of networks of strings was first raised by Rolewicz in [12], where a network controlled at all its nodes was studied. Later on, Schmidt introduced in [13] the model of a network controlled at exterior (simple)

*Received by the editors January 22, 2003; accepted for publication (in revised form) November 18, 2003; published electronically July 23, 2004. This work has been partially supported by grants BFM2002-03345 of the MCYT (Spain) and the EU project "Homogenization and Multiple Scales."
http://www.siam.org/journals/sicon/43-2/42184.html
†Departamento de Matématica Aplicada, Universidad Complutense de Madrid, 28040, Madrid, Spain (rene_dager@mat.ucm.es).

nodes as considered in this paper, and proved the exact controllability of networks supported by tree-shaped graphs if all but one of the exterior nodes are controlled.

More general hyperbolic systems on graphs, including Timoshenko-beam models, have been considered by Lagnese, Leugering, and Schmidt in a series of papers which are collected in [8], where again exact controllability and stabilizability from exterior nodes were studied. These authors also noticed that the controllability property, even in an approximate sense, may fail for systems with rationally dependent physical characteristics supported on graphs containing cycles, or when two or more exterior nodes are left uncontrolled. This fact led to the question of characterizing those networks which are controllable under the action of a small number of controls. This problem was then studied by Leugering and Zuazua in [9] for a three-string star-shaped network controlled from a single exterior node. They found that the network is controllable if and only if the ratio of the lengths of the uncontrolled strings is an irrational number.

This paper extends the results in [9] for general tree-shaped networks controlled from the root. Our main results are related to the possibility of proving weighted observability inequalities for the solutions of the system modelling the network. Under certain conditions imposed over the lengths of the strings of the networks, which are verified in a generic sense, those observability results allow one to obtain information on the controllability properties of the tree-shaped networks from one exterior node. Thus, we provide a complete answer to the one-node controllability problem for tree-shaped networks.

1. Notation and statement of the problem.

1.1. Notation for the elements of the graph. In this section, we introduce precise notation for the elements of the rest configuration graph. This is needed to write the equations of the motion of the network in a way that takes into account the topological structure of the graph.

Let $\mathcal{A}$ be a planar, connected graph without closed paths. According to the usual terminology in graph theory, those graphs will be called *trees*. By the multiplicity of a vertex of $\mathcal{A}$ we mean the number of edges that branch out from that vertex. If the multiplicity is equal to one, the vertex is called exterior; otherwise, it is said to be interior. We assume that the graph $\mathcal{A}$ does not contain vertices of multiplicity two, since they are irrelevant for our model.

In what follows, we describe a procedure for indexing the edges and vertices of the graph. In Figure 1.1 an example is given of a tree with indices defined according to this rule. First, we choose an exterior vertex and denote it by $\mathcal{R}$. It is called the root of $\mathcal{A}$. The remaining edges and vertices will be denoted by $\mathbf{e}_{\bar\alpha}$ and $\mathcal{O}_{\bar\alpha}$, respectively, where $\bar\alpha = (\alpha_1, \dots, \alpha_k)$ is a multi-index (possibly empty) of variable length k defined by recurrence for every edge in the following way.

For the edge containing the root $\mathcal{R}$ we choose the empty index. Thus, that edge is denoted by $\mathbf{e}$ and its vertex different from $\mathcal{R}$ is denoted by $\mathcal{O}$.

Assume now that the interior vertex $\mathcal{O}_{\bar\alpha}$, contained in the edge $\mathbf{e}_{\bar\alpha}$, has multiplicity equal to $m_{\bar\alpha}+1$. This means that there are $m_{\bar\alpha}$ edges, different from $\mathbf{e}_{\bar\alpha}$, that branch out from $\mathcal{O}_{\bar\alpha}$. We denote these edges by $\mathbf{e}_{\bar\alpha\circ\beta}$, $\beta = 1, \dots, m_{\bar\alpha}$, and the other vertex of the edge $\mathbf{e}_{\bar\alpha\circ\beta}$ by $\mathcal{O}_{\bar\alpha\circ\beta}$. Here, $\bar\alpha\circ\beta$ represents the index $(\alpha_1, \dots, \alpha_k, \beta)$, obtained by adding a new component β to the index $\bar\alpha = (\alpha_1, \dots, \alpha_k)$. In general, if $\bar\alpha = (\alpha_1, \dots, \alpha_k)$ and $\bar\beta = (\beta_1, \dots, \beta_m)$, then $\bar\alpha\circ\bar\beta$ will denote the multi-index of length $k+m$ defined by $\bar\alpha\circ\bar\beta = (\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_m)$.

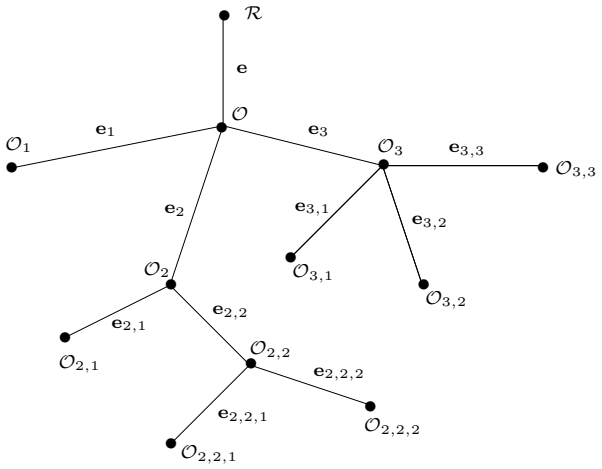


FIG. 1.1. *A tree with indices for its vertices and edges.*

Now let $\mathcal{M}$ be the set of the interior vertices of $\mathcal{A}$, and $\mathcal{S}$ the set of exterior vertices, except $\mathcal{R}$, and define

$$\mathcal{I}_{\mathcal{M}} = \{\bar{\alpha}; \quad \mathcal{O}_{\bar{\alpha}} \in \mathcal{M}\}, \qquad \mathcal{I}_{\mathcal{S}} = \{\bar{\alpha}; \quad \mathcal{O}_{\bar{\alpha}} \in \mathcal{S}\},$$

which are the sets of the indices of the interior and exterior vertices (except $\mathcal{R}$), respectively. Note that with these notations, we admit the empty multi-index, which corresponds to the vertex $\mathcal{O}$ and belongs to one of the sets $\mathcal{I}_{\mathcal{M}}$ or $\mathcal{I}_{\mathcal{S}}$. Finally, $\mathcal{I} = \mathcal{I}_{\mathcal{S}} \bigcup \mathcal{I}_{\mathcal{M}}$ is the set of the indices of all the vertices, except that of the root $\mathcal{R}$.

Further, for $\bar{\alpha} \in \mathcal{I}_{\mathcal{M}}$, the sets

$$\mathcal{A}_{\bar{\alpha}} = \left\{\mathbf{e}_{\bar{\alpha}\circ\bar{\beta}}; \quad \bar{\alpha}\circ\bar{\beta} \in \mathcal{I}\right\}$$

are called subtrees of $\mathcal{A}$. Note that $\mathcal{A}_{\bar{\alpha}}$ is formed by the edges having indices with a common initial part $\bar{\alpha}$. This means that $\mathcal{A}_{\bar{\alpha}}$ is also a tree branching out from the vertex of $\mathbf{e}_{\bar{\alpha}}$ different from $\mathcal{O}_{\bar{\alpha}}$. Then, if one chooses that vertex as the root $\mathcal{R}_{\bar{\alpha}}$ of $\mathcal{A}_{\bar{\alpha}}$ and denotes by $\mathbf{e}^{\bar{\alpha}}_{\bar{\beta}}$ the edge with index $\bar{\beta}$ in $\mathcal{A}_{\bar{\alpha}}$ according to the numbering rule defined above for trees, it holds that

$$\mathbf{e}_{\bar{\alpha}\circ\bar{\beta}} = \mathbf{e}^{\bar{\alpha}}_{\bar{\beta}}, \qquad \mathcal{O}_{\bar{\alpha}\circ\bar{\beta}} = \mathcal{O}^{\bar{\alpha}}_{\bar{\beta}}.$$

In order to prove properties of trees, we shall often proceed by induction with respect to the largest length of the indices $\bar{\alpha}$ used to number the edges according to the procedure described above. To do this we should prove that (1) the property is true for the simplest case of a one-edged tree (i.e., the corresponding network is formed by a single string) and also that (2) if the property is true for all the subtrees $\mathcal{A}_1, \ldots, \mathcal{A}_m$ branching out from $\mathcal{O}$, then it is also true for the whole tree $\mathcal{A}$. In what follows, such a process will be called simply induction.

Furthermore, the length of the edge $\mathbf{e}_{\bar{\alpha}}$ will be denoted by $\ell_{\bar{\alpha}}$. Then, $\mathbf{e}_{\bar{\alpha}}$ may be parameterized by its arc length by means of the functions $\pi_{\bar{\alpha}}$, defined in $[0, \ell_{\bar{\alpha}}]$ such that $\pi_{\bar{\alpha}}(\ell_{\bar{\alpha}}) = \mathcal{O}_{\bar{\alpha}}$ and $\pi_{\bar{\alpha}}(0)$ is the other vertex of this edge.

Finally, we denote by $L_{\mathcal{A}}$ and $L_{\bar{\alpha}}$, $\bar{\alpha} \in \mathcal{I}$, the sum of the lengths of all the edges of the tree $\mathcal{A}$ (i.e., the total length of $\mathcal{A}$) and of its subtrees $\mathcal{A}_{\bar{\alpha}}$, respectively.

1.2. The equations of the motion of the network. With the notation introduced above we can write the equations that describe the motion of the network.

Let $u^{\bar\alpha}(t,x):\ \mathbb{R}\times[0,\ell_{\bar\alpha}]\to\mathbb{R}$ be the transverse displacement of the string with index $\bar\alpha$ with respect to the rest position. These functions allow us to identify the network with its rest graph. In this sense, the vertices of $\mathcal{A}$ are called *nodes* and the edges, *strings*.

For every $\bar\alpha\in\mathcal{I}$, the function $u^{\bar\alpha}(t,x)$ is assumed to satisfy the 1-d wave equation

$$u^{\bar\alpha}_{tt}(t,x)=u^{\bar\alpha}_{xx}(t,x)\qquad\text{in }\mathbb{R}\times[0,\ell_{\bar\alpha}], \tag{1.1}$$

where the subscripts t, x denote the derivatives with respect to those variables.

The coupling conditions between the strings are given by

$$u^{\bar\alpha\circ\beta}(t,0)=u^{\bar\alpha}(t,\ell_{\bar\alpha})\qquad\text{in }\mathbb{R},\quad\beta=1,\dots,m_{\bar\alpha}, \tag{1.2}$$

$$\sum_{\beta=1}^{m_{\bar\alpha}}u^{\bar\alpha\circ\beta}_x(t,0)=u^{\bar\alpha}_x(t,\ell_{\bar\alpha})\qquad\text{in }\mathbb{R}, \tag{1.3}$$

for $\bar\alpha\in\mathcal{I}_{\mathcal{M}}$. The conditions (1.2), (1.3) express the network continuity and the balance of forces at the interior nodes, respectively. At the root $\mathcal{R}$ the deformation of the string $\mathbf{e}$ is determined by a "control" v:

$$u(t,0)=v(t)\quad\text{in }\mathbb{R}, \tag{1.4}$$

while the remaining exterior nodes are fixed:

$$u^{\bar\alpha}(t,\ell_{\bar\alpha})=0\quad\text{in }\mathbb{R}\quad\text{if }\bar\alpha\in\mathcal{I}_{\mathcal{S}}. \tag{1.5}$$

Besides, the initial displacement and velocity of the strings (i.e., the initial state of the system) are given:

$$u^{\bar\alpha}(0,x)=u^{\bar\alpha}_0(x),\quad u^{\bar\alpha}_t(0,x)=u^{\bar\alpha}_1(x)\quad\text{in }[0,\ell_{\bar\alpha}],\ \bar\alpha\in\mathcal{I}. \tag{1.6}$$

Now we provide a functional setting for problem (1.1)–(1.6). Assume the set of indices $\mathcal{I}$ has been ordered in some way. Then denote by $\bar u$ the vector function with components $u^{\bar\alpha},\ \bar\alpha\in\mathcal{I}$, and define the Hilbert spaces

$$V=\Bigg\{\bar u\in\prod_{\bar\alpha\in\mathcal{I}}H^1(0,\ell_{\bar\alpha}):u^{\bar\alpha\circ\beta}(0)=u^{\bar\alpha}(\ell_{\bar\alpha})\ \text{ for }\bar\alpha\in\mathcal{I}_{\mathcal{M}},\ \ \beta=1,\dots,m_{\bar\alpha};$$
$$u^{\bar\alpha}(\ell_{\bar\alpha})=0\ \text{ for }\bar\alpha\in\mathcal{I}_{\mathcal{S}}\text{ and }u(0)=0\Bigg\},$$

$$H=\prod_{\bar\alpha\in\mathcal{I}}L^2(0,\ell_{\bar\alpha}),$$

endowed with the natural product structures of those of H^1 and L^2, respectively.

When $v\in L^2(0,T)$ and $\bar u_0=(u^{\bar\alpha}_0)_{\bar\alpha\in\mathcal{I}}\in H,\ \bar u_1=(u^{\bar\alpha}_1)_{\bar\alpha\in\mathcal{I}}\in V'$ (the dual space of V), (1.1)–(1.6) has a unique weak solution, defined by transposition, that verifies

$$\bar u\in C([0,T]:H)\cap C^1([0,T]:V').$$

If the function v is equal to zero in the interval $[0,T]$ and $\bar{u}_0 = (u_0^{\bar{\alpha}})_{\bar{\alpha}\in\mathcal{I}} \in V$, $\bar{u}_1 = (u_1^{\bar{\alpha}})_{\bar{\alpha}\in\mathcal{I}} \in H$, then the system (1.1)–(1.6) has a unique solution that satisfies

$$\bar{u} \in C([0,T]:V) \cap C^1([0,T]:H)$$

(see [8, Chapter III] for details).

In this latter case, the solution $\bar{u}$ of (1.1)–(1.6) is expressed in terms of the initial data $\bar{u}_0$, $\bar{u}_1$ by the formula

$$\bar{u}(t) = \sum_{k\in\mathbf{Z}_+} \left(u_{0,k} \cos \lambda_k t + \frac{u_{1,k}}{\lambda_k} \sin \lambda_k t \right) \bar{\theta}_k. \tag{1.7}$$

In (1.7), $\lambda_k := \sqrt{\mu_k}$, where $\{\mu_k\}_{k\in\mathbf{Z}_+}$ is the increasing (and positive) sequence of the eigenvalues of the elliptic operator defined by the system (1.1)–(1.5) and $\{\bar{\theta}_k\}_{k\in\mathbf{Z}_+}$ is the corresponding sequence of eigenfunctions, chosen to be an orthonormal basis of H, while $\{u_{0,k}\}_{k\in\mathbf{Z}_+}$, $\{u_{1,k}\}_{k\in\mathbf{Z}_+}$ are the sequences of Fourier coefficients of $\bar{u}_0$ and $\bar{u}_1$ with respect to that basis.

1.3. Statement of the problem. Our main goal is to study the controllability problem for the system (1.1)–(1.6). That is, given $T > 0$ and $\bar{u}_0 \in H$, $\bar{u}_1 \in V'$, to choose the control function $v \in L^2(0,T)$ such that the solution of (1.1)–(1.6) with initial data $\bar{u}_0$, $\bar{u}_1$ verifies

$$\bar{u}(T,x) = \bar{u}_t(T,x) = \bar{0},$$

i.e., it reaches rest at time T. When this is possible, we say that the initial state $(\bar{u}_0, \bar{u}_1)$ is controllable in time T.

Let $W_T \subset H\times V'$ be the set of controllable initial states in time T. If $W_T = H\times V'$, the system is said to be *exactly controllable* in time T, and if W_T is dense in $H \times V'$, it is said to be *approximately controllable*.

A simple application of the Hilbert uniqueness method (HUM) of Lions (see [10]) allows us to show that if there exist positive numbers c_n, $n \in \mathbb{Z}_+$, such that

$$\int_0^T |u(t,x)|^2 dt \geq \sum_{n\in\mathbb{Z}_+} c_n^2 \left(\mu_n |u_n^0|^2 + |u_n^1|^2 \right) \tag{1.8}$$

for every solution $\bar{u}$ of (1.1)–(1.6) with $v = 0$ and initial data $\bar{u}_0 = \sum_{n\in\mathbb{Z}_+} u_n^0 \bar{\theta}_n \in V$, $\bar{u}_1 = \sum_{n\in\mathbb{Z}_+} u_n^1 \bar{\theta}_n \in H$, then the space

$$\mathcal{W} = \left\{ (\bar{u}_0, \bar{u}_1) \in V \times H' : \ \|(\bar{u}_0, \bar{u}_1)\|^2 = \sum_{n\in\mathbb{Z}_+} \frac{1}{c_n^2} \left(|u_n^0|^2 + \frac{1}{\mu_n} |u_n^1|^2 \right) < \infty \right\} \tag{1.9}$$

verifies $\mathcal{W} \subset W_T$; i.e., all the initial states from $\mathcal{W}$ are controllable in time T.

In particular, if (1.8) holds, W_T contains, for every $T' \geq T$, the space $Z\times Z$, where Z is the set of all finite linear combinations of the eigenfunctions $\{\bar{\theta}_k\}_{k\in\mathbf{Z}_+}$. In that case it is said that the system is *spectrally controllable in time* T. The space $Z \times Z$ being dense in $H \times V'$, the spectral controllability obviously implies the approximate controllability and is apparently stronger. We shall show, however, that these two properties coincide for the networks considered here. This will imply that a network is approximately controllable in some time T if and only if all the initial states (ϕ, ψ), where ϕ and ψ are eigenfunctions, are controllable in time T.

The main result of this paper is contained in Theorem 5.1 of section 5, where we prove an inequality like (1.8) for any $T \geq 2L_{\mathcal{A}}$ (i.e., when the control time is at least twice the sum of all the lengths of the strings of the network) and with coefficients c_n, which may be explicitly computed in terms of the eigenvalues μ_n. In general, a tree is said to be *nondegenerate* if an inequality like (1.8) is true with coefficients which are all different from zero. In the subsection 5.2 we give a characterization, in terms of the eigenfunctions of the associated elliptic problem, of nondegenerate trees.

Moreover, it turns out that, if some of the coefficients vanish (i.e., if the tree degenerates), then there exist eigenfunctions vanishing identically at the controlled string. It is very well known that this fact implies that the space of controllable initial data is not dense in $H \times V'$. Therefore, if some inequality (1.8) takes place in this case, then necessarily some of the coefficients vanish (otherwise the system would be approximately controllable). Thus, Theorem 5.1 will provide an inequality (1.8) with nonzero coefficients whenever such an inequality exists.

If we were able to establish uniform lower estimates of the form

$$|c_n| \geq C\lambda_n^{-\alpha},$$

it would imply that $\mathcal{W} \supset W^{\alpha} \times W^{\alpha-\frac{1}{2}}$, where W^{α} is the domain of the α-power of the elliptic problem defined by (1.1)–(1.5), and we would obtain an explicitly characterized space of controllable initial states. This has been done in [4] for star-shaped networks under suitable assumptions on the lengths of the strings. However, even in that simple case, the results are based on deep facts from number theory concerning the uniform approximation by rational numbers, and thus it is unlikely that one could obtain similar results in the case of general trees.

The results of this paper were essentially announced in [6]. We should also remark that the similar problem of simultaneous controllability of networks (see section 7 below) has been considered by several authors for the case of one-string networks in, e.g., [1], [2], and [3]. The proofs given in those papers are based on generalizations of the Ingham inequality for sums of complex exponentials. In [5] we used the method presented here, which considerably simplified the solution of the problem.

1.4. Additional notation. For technical purposes, we consider also solutions $\bar{u}$ of (1.1) such that $\bar{u}_{\bar{\alpha}} \in C^2(\mathbb{R} \times [0, \ell_{\bar{\alpha}}])$, which verify (1.2), (1.3), and (1.5), but not necessarily (1.4). That is, $\bar{u}$ is a smooth solution that satisfies the boundary conditions at all the nodes, except possibly at the root $\mathcal{R}$. We shall briefly refer to such solutions as *solutions of* (N).

To simplify the notation we introduce for a solution $\bar{u}$ of (N) the functions

$$G_{\bar{\alpha}}(t) := u_t^{\bar{\alpha}}(t,0), \qquad F_{\bar{\alpha}}(t) := u_x^{\bar{\alpha}}(t,0), \tag{1.10}$$

$$\widehat{G}_{\bar{\alpha}}(t) := u_t^{\bar{\alpha}}(t,\ell_{\bar{\alpha}}), \qquad \widehat{F}_{\bar{\alpha}}(t) := u_x^{\bar{\alpha}}(t,\ell_{\bar{\alpha}}), \tag{1.11}$$

for $\bar{\alpha} \in \mathcal{I}$. These functions are the velocity and the tension at the extremes of the edge $\mathbf{e}_{\bar{\alpha}}$.

According to the coupling conditions (1.2)–(1.3), it holds that

$$G_{\bar{\alpha}\circ\beta}(t) = \widehat{G}_{\bar{\alpha}}(t), \qquad \sum_{\beta=1}^{m_{\bar{\alpha}}} F_{\bar{\alpha}\circ\beta}(t) = \widehat{F}_{\bar{\alpha}}(t), \tag{1.12}$$

for all $t \in \mathbb{R}$, $\bar{\alpha} \in \mathcal{I}_{\mathcal{M}}$, and $\beta = 1, \ldots, m_{\bar{\alpha}}$.

If $\bar{w}(t)$ is a function on the tree $\mathcal{A}$, we define the energy of $\bar{w}$ in the edge $\mathbf{e}_{\bar{\alpha}}$ by

$$E^{\bar{\alpha}}_{\bar{w}}(t) := \int_0^{\ell_{\bar{\alpha}}} \left(|w^{\bar{\alpha}}_t(t,x)|^2 + |w^{\bar{\alpha}}_x(t,x)|^2\right) dx,$$

and by

$$\mathbf{E}_{\bar{w}}(t) := \sum_{\bar{\alpha}\in\mathcal{I}} E^{\bar{\alpha}}_{\bar{w}}(t)$$

the energy of $\bar{w}$ in the whole tree. We will also use the notation $\mathbf{E}^{\bar{\alpha}}_{\bar{\omega}}$ for the total energy of $\bar{\omega}$ in the subtree $\mathcal{A}_{\bar{\alpha}}$.

For the solutions of (N) a simple expression for the energy is obtained. It takes the following form.

PROPOSITION 1.1. *If $\bar{u}$ is a solution of* (N)*, then for any $t \in \mathbb{R}$*

$$\frac{d}{dt}\mathbf{E}_{\bar{u}}(t) = -G(t)F(t). \tag{1.13}$$

Proof. If we multiply (1.1) by $u^{\bar{\alpha}}_t$, it follows, after integration by parts in $[0, \ell_{\bar{\alpha}}]$, that

$$\frac{d}{dt}E^{\bar{\alpha}}(t) = \widehat{G}_{\bar{\alpha}}(t)\widehat{F}_{\bar{\alpha}}(t) - G_{\bar{\alpha}}(t)F_{\bar{\alpha}}(t), \qquad \bar{\alpha}\in\mathcal{I}_{\mathcal{M}},$$
$$\frac{d}{dt}E^{\bar{\alpha}}(t) = -G_{\bar{\alpha}}(t)F_{\bar{\alpha}}(t), \qquad \bar{\alpha}\in\mathcal{I}_{\mathcal{S}}.$$

Formula (1.13) is then obtained from the definition of $\mathbf{E}_{\bar{u}}$, taking into account the coupling conditions at the interior vertices. □

Remark 1. From the previous proposition it follows that, for the solutions of (1.1)–(1.6) with $v = 0$, the energy is conserved. In this case, as a consequence of the expansion (1.7), the energy of the solution $\bar{u}$ with initial data $\bar{u}_0 = \sum_{n\in\mathbb{Z}_+} u^0_n \bar{\theta}_n$, $\bar{u}_1 = \sum_{n\in\mathbb{Z}_+} u^1_n \bar{\theta}_n$ may be expressed by the formula

$$\mathbf{E}_{\bar{u}} = \frac{1}{2}\sum_{k\in\mathbf{Z}_+} \left(\lambda^2_k u^2_{0,k} + u^2_{1,k}\right). \tag{1.14}$$

2. The operators $\mathcal{P}$ and $\mathcal{Q}$. In this section we define two linear operators $\mathcal{P}$ and $\mathcal{Q}$ that allow us to express the relation

$$\mathcal{P}G + \mathcal{Q}F = 0 \tag{2.1}$$

between the traces of the velocity and the tension of the solutions of (N) at the root of the tree. These operators will play an essential role in the proof of the main results, and so we study them in detail. In particular, we need information on how they act on the traces $F_{\bar{\alpha}}$ and $G_{\bar{\alpha}}$ of the other components of the solution at the interior nodes.

First, $\mathcal{P}$ and $\mathcal{Q}$ are constructed for a string. Then, using a recursive argument, they are obtained for general trees.

2.1. The case of a single string. Assume that $u(t,x)$ satisfies the wave equation in $\mathbb{R}\times[0,\ell]$. Then $u(t,x)$ can be expressed by the d'Alembert formula

$$u(t,x) = \frac{1}{2}\left(u(t+x,0) + u(t-x,0)\right) + \frac{1}{2}\int_{t-x}^{t+x} u_x(\tau,0)d\tau, \tag{2.2}$$

from which, after differentiation, we get the equalities

$$\begin{aligned} u_x(t,x) &= \tfrac{1}{2}\left(u_t(t+x,0) - u_t(t-x,0)\right) + \tfrac{1}{2}\left(u_x(t+x,0) + u_x(t-x,0)\right), \\ u_t(t,x) &= \tfrac{1}{2}\left(u_t(t+x,0) + u_t(t-x,0)\right) + \tfrac{1}{2}\left(u_x(t+x,0) - u_x(t-x,0)\right). \end{aligned} \tag{2.3}$$

Now let the functions F, G, $\widehat{F}$, $\widehat{G}$ be defined as in (1.10), (1.11), that is,

$$G(t) = u_t(t,0), \quad F(t) = u_x(t,0), \qquad \widehat{G}(t) = u_t(t,\ell), \quad \widehat{F} = u_x(t,\ell).$$

With this notation, formulas (2.3) for $x = \ell$ can be written as

$$\widehat{F} = \ell^+ F + \ell^- G, \tag{2.4}$$

$$\widehat{G} = \ell^- F + \ell^+ G, \tag{2.5}$$

where ℓ^+, ℓ^- are the linear operators acting on a time-dependent function f according to

$$\ell^\pm f(t) = \frac{f(t+\ell) \pm f(t-\ell)}{2}. \tag{2.6}$$

If u is a solution of (N) for a one-string network, then u satisfies for $x = \ell$ the homogeneous Dirichlet boundary condition $\widehat{G} = 0$, and the equality (2.5) becomes

$$\ell^- F + \ell^+ G = 0,$$

which provides a relation of type (2.1) with $\mathcal{P} = \ell^+$, $\mathcal{Q} = \ell^-$.

The following proposition is a classical result on the observability of 1-d waves from the boundary. It is easily proved using the d'Alembert representation formula.

PROPOSITION 2.1. *If $u(t,x)$ satisfies the wave equation $u_{tt} = u_{xx}$ in $\mathbb{R} \times [0,\ell]$, then*

$$\mathbf{E}_u(t) \le \frac{1}{4}\int_{t-\ell}^{t+\ell} \left(|u_x(\tau,0)|^2 + |u_t(\tau,0)|^2\right) d\tau.$$

Proof. In view of (2.3), it holds that

$$\begin{aligned} \mathbf{E}_u(t) &= \frac{1}{8}\int_0^\ell \big\{|u_t(t+x,0) - u_t(t-x,0) + u_x(t+x,0) + u_x(t-x,0)|^2 \\ &\qquad + |u_t(t+x,0) + u_t(t-x,0) + u_x(t+x,0) - u_x(t-x,0)|^2\big\}\, dx \\ &\le \frac{1}{4}\int_0^\ell \left\{|u_t(t+x,0)|^2 + |u_t(t-x,0)|^2 + |u_x(t+x,0)|^2 + |u_x(t-x,0)|^2\right\} dx \\ &= \frac{1}{4}\int_{t-\ell}^{t} \left\{|u_t(\tau,0)|^2 + |u_x(\tau,0)|^2\right\} dx + \frac{1}{4}\int_{t}^{t+\ell} \left\{|u_t(\tau,0)|^2 + |u_x(\tau,0)|^2\right\} dx \\ &= \frac{1}{4}\int_{t-\ell}^{t+\ell} \left(|u_x(\tau,0)|^2 + |u_t(\tau,0)|^2\right) dt. \qquad \square \end{aligned}$$

We remark in the next proposition a simple fact that is widely used in what follows.

PROPOSITION 2.2. *Let $w(t,x)$ be a function defined in $\mathbb{R}\times[0,\ell]$. Then*

$$\mathbf{E}_{\ell^\pm w}(t)\le \ell^+\mathbf{E}_w(t).$$

(Here, the operators $\ell^\pm$ act on the variable t; i.e., $\ell^\pm w(t,x)=\frac{1}{2}(w(t+\ell,x)\pm w(t-\ell,x))$.)

Proof. For every $t\in\mathbb{R}$

$$\begin{aligned}\mathbf{E}_{\ell^\pm w}&=\frac{1}{8}\int_0^\ell \left\{|w_x(t+\ell,x)\pm w_x(t-\ell,x)|^2+|w_t(t+\ell,x)\pm w_t(t-\ell,x)|^2\right\}dx\\ &\le\ \frac{1}{4}\int_0^\ell \left\{|w_x(t+\ell,x)|^2+|w_x(t-\ell,x)|^2+|w_t(t+\ell,x)|^2+|w_t(t-\ell,x)|^2\right\}dx\\ &=\frac{1}{2}(\mathbf{E}_w(t+\ell)+\mathbf{E}_w(t-\ell))=\ell^+\mathbf{E}_w(t).\qquad\square\end{aligned}$$

2.2. Operators of type S. As stated above, we are interested not only in the existence of the operators $\mathcal{P}$ and $\mathcal{Q}$ satisfying (2.1), but also in their structure. That is why we consider a class of linear operators that are linear combinations of certain shift operators. This allows us to describe some properties of the operators $\mathcal{P}$ and $\mathcal{Q}$ that we shall use later.

For the real number h we denote by τ_h the shift operator defined by $\tau_h f(t):=f(t+h)$. As we shall be concerned only with algebraic properties of those operators, we may assume τ_h to act on the vector spaces of mappings $f=f(t):\mathbb{R}\to\mathbf{W}$, where $\mathbf{W}$ is a vector space.

Let $\Lambda=\{\ell_1,\dots,\ell_n\}$ be a set of positive numbers, not necessarily different. In what follows, whenever a set is denoted by Λ we tacitly assume that it may contain repeated elements. If $\tilde\Lambda=\{\tilde\ell_1,\dots,\tilde\ell_{n'}\}$ is another such set, we use the notation $\Lambda\sqcup\tilde\Lambda$ for the set $\{\ell_1,\dots,\ell_n,\tilde\ell_1,\dots,\tilde\ell_{n'}\}$, which once again may contain repeated elements. We set

$$S(\Lambda):=\operatorname{span}\{\tau_h : h\in\mathcal{H}_\Lambda\}$$

(the set of all linear combinations of shift operators τ_h with $h\in\mathcal{H}_\Lambda$), where

$$\mathcal{H}_\Lambda=\left\{h=\sum_{i=1}^n \varepsilon_i\ell_i,\ \varepsilon_i=\pm1\right\}.$$

Observe that the set $\mathcal{H}_\Lambda$ contains at most 2^n elements, so $S(\Lambda)$ is of finite dimension.

For an operator $\mathcal{B}\in S(\Lambda)$ we shall write $s(\mathcal{B}):=s(\Lambda):=\sum_{i=1}^n\ell_i$. We say that $\mathcal{B}$ *is of type* S if $\mathcal{B}\in S(\Lambda)$ for some set Λ.

The operators ℓ^+ and ℓ^-, defined in the previous section by (2.6), can be expressed as

$$\ell^\pm=\frac{\tau_\ell\pm\tau_{-\ell}}{2},$$

and thus they belong to $S(\{\ell\})$.

We gather in the following two propositions some elementary properties of the operators of type S, which will be used in the proof of our main results.

PROPOSITION 2.3. (i) $\mathcal{B} \in S(\Lambda)$ *if and only if it may be written as a linear combination of operators of the form* $\ell_1^{\pm}\ell_2^{\pm}\cdots\ell_n^{\pm}$, *where* $\pm$ *means that we choose one of the operators* ℓ_i^+ *or* ℓ_i^-.

(ii) *If* $\mathcal{B}_1 \in S(\Lambda_1)$ *and* $\mathcal{B}_2 \in S(\Lambda_2)$, *then* $\mathcal{B}_1\mathcal{B}_2 = \mathcal{B}_2\mathcal{B}_1 \in S(\Lambda_1 \sqcup \Lambda_2)$ *and* $s(\mathcal{B}_1\mathcal{B}_2) = s(\mathcal{B}_1) + s(\mathcal{B}_2)$.

PROPOSITION 2.4. *If* $\mathcal{B}$ *is an operator of type* S *with* $s(\mathcal{B}) = s$, *then there exist positive constants* C_1, C_2, *depending only on the coefficients of* $\mathcal{B}$, *such that*

$$\text{(i)} \qquad \int_a^b |\mathcal{B}f(t)|^2 dt \le C_1 \int_{a-s}^{b+s} |f(t)|^2 dt$$

for all the functions f *for which the integrals are defined.*[1]

(ii) *If the function* $w(t,x)$ *is defined in* $\mathbb{R} \times [0,\ell]$ *and there exists a constant* M *such that* $\mathbf{E}_w(t) \le M$ *for* $t \in [a,b]$, *then* $\mathbf{E}_{\mathcal{B}w}(t) \le C_2 M$ *for* $t \in [a+s, b-s]$.

Proof. (i) For $n = 1$, $\Lambda = \{\ell\}$ we have that $\mathcal{B} = c_1\ell^+ + c_2\ell^-$ and $s(\mathcal{B}) = \ell$. Then

$$\begin{aligned}
\int_a^b |\mathcal{B}f(t)|^2 dt &= \int_a^b \left|c_1\ell^+ f(t) + c_2\ell^- f(t)\right|^2 dt \\
&= \int_a^b \left|\frac{c_1+c_2}{2} f(t+\ell) + \frac{c_1-c_2}{2} f(t-\ell)\right|^2 dt \\
&\le \left(\frac{c_1+c_2}{2}\right)^2 \int_a^b |f(t+\ell)|^2 dt + \left(\frac{c_1-c_2}{2}\right)^2 \int_a^b |f(t-\ell)|^2 dt \\
&\le \left(\frac{c_1+c_2}{2}\right)^2 \int_{a+\ell}^{b+\ell} |f(t)|^2 dt + \left(\frac{c_1-c_2}{2}\right)^2 \int_{a-\ell}^{b-\ell} |f(t)|^2 dt \\
&\le (c_1^2 + c_2^2) \int_{a-\ell}^{b+\ell} |f(t)|^2 dt.
\end{aligned}$$

When $n \ge 2$, it suffices to iterate this inequality, taking into account Proposition 2.3(i). Note that C_1 may be chosen as the maximum of the squares of the coefficients of $\mathcal{B}$ in the representation of $\mathcal{B}$ given by Proposition 2.3(i), and thus C_1 depends only on $\mathcal{B}$.

(ii) This point is an immediate consequence of Proposition 2.2. □

The next proposition plays a crucial role in this paper.

PROPOSITION 2.5. *Let* $\Lambda = \{\ell_1, \dots, \ell_m\}$ *with* $\ell_1 \le \cdots \le \ell_m$, *and denote* $T_\Lambda = 2s(\Lambda) = 2\sum_{i=1}^m \ell_i$. *Assume that* $\mathcal{B} = \sum_{h\in\mathcal{H}_\Lambda} c_h\tau_h \in S(\Lambda)$ *and that the coefficient* $c_{\ell_1+\cdots+\ell_m}$ *is different from zero. Then, for any* $T > 0$ *there exists a constant* $C_T > 0$ *such that*

$$\int_0^T |u(t)|^2 dt \le C_T \int_0^{T_\Lambda} |u(t)|^2 dt$$

for any continuous function u *satisfying* $\mathcal{B}u \equiv 0$.

Proof. We shall prove that, for any natural number n and any function u satisfying $\mathcal{B}u \equiv 0$, it holds that

$$\int_0^{T_\Lambda + 2n\ell_1} |u(t)|^2 dt \le \gamma^n \int_0^{T_\Lambda} |u(t)|^2 dt, \tag{2.7}$$

[1] In other words, the operator $\mathcal{B}$ is bounded from $L^2[a-s, b+s]$ to $L^2[a,b]$.

where γ is a positive constant depending only on $\mathcal{B}$. Clearly, the assertion of the proposition immediately follows from inequality (2.7).

If $\mathcal{B}u \equiv 0$, i.e., $0 = \sum_{h\in\mathcal{H}_\Lambda} c_h\tau_h u(t) = \sum_{h\in\mathcal{H}_\Lambda} c_h u(t+h)$, then, replacing the variable t by $t-(\ell_1+\cdots+\ell_m)$ and taking into account that $c_{\ell_1+\cdots+\ell_m}\neq 0$, we get

$$u(t) = \sum_{h'\in\mathcal{H}^*_\Lambda} \delta_{h'} u(t-h'), \tag{2.8}$$

where $\mathcal{H}^*_\Lambda = \{h' = h-(\ell_1+\cdots+\ell_m) : h\in\mathcal{H}_\Lambda,\ h\neq(\ell_1+\cdots+\ell_m)\}$ and $\delta_{h'} = -\frac{c_{h'+(\ell_1+\cdots+\ell_m)}}{c_{\ell_1+\cdots+\ell_m}}$.

From (2.8) and the Cauchy–Schwarz inequality it follows that

$$|u(t)|^2 \le \delta \sum_{h'\in\mathcal{H}^*_\Lambda} |u(t-h')|^2, \tag{2.9}$$

where $\delta = \sum_{h'\in\mathcal{H}^*_\Lambda} \delta^2_{h'}$.

Note that for every $h'\in\mathcal{H}^*_\Lambda$ we have $2\ell_1 \le h' \le 2(\ell_2+\cdots+\ell_m)$, and therefore,

$$T_\Lambda + 2(n+1)\ell_1 - h' \le T_\Lambda + 2n\ell_1 \quad\text{and}\quad T_\Lambda + 2n\ell_1 - h' \ge 2(n+1)\ell_1 \ge 0.$$

This fact implies that

$$\int_{T_\Lambda+2n\ell_1-h'}^{T_\Lambda+2(n+1)\ell_1-h'} |u(t)|^2dt \le \int_0^{T_\Lambda+2n\ell_1} |u(t)|^2dt. \tag{2.10}$$

On the other hand, from (2.9) it follows that

$$\begin{aligned}\int_{T_\Lambda+2n\ell_1}^{T_\Lambda+2(n+1)\ell_1} |u(t)|^2dt &\le \delta \sum_{h'\in\mathcal{H}^*_\Lambda} \int_{T_\Lambda+2n\ell_1}^{T_\Lambda+2(n+1)\ell_1} |u(t-h')|^2dt\\ &= \delta \sum_{h'\in\mathcal{H}^*_\Lambda} \int_{T_\Lambda+2n\ell_1-h'}^{T_\Lambda+2(n+1)\ell_1-h'} |u(t)|^2dt.\end{aligned}$$

Now, taking into account (2.10), the previous inequality becomes

$$\int_{T_\Lambda+2n\ell_1}^{T_\Lambda+2(n+1)\ell_1} |u(t)|^2dt \le (2^m-1)\delta \int_0^{T_\Lambda+2n\ell_1} |u(t)|^2dt.$$

From this latter inequality we obtain

$$\begin{aligned}\int_0^{T_\Lambda+2(n+1)\ell_1} |u(t)|^2dt &= \int_0^{T_\Lambda+2n\ell_1} |u(t)|^2dt + \int_{T_\Lambda+2n\ell_1}^{T_\Lambda+2(n+1)\ell_1} |u(t)|^2dt\\ &\le \int_0^{T_\Lambda+2n\ell_1} |u(t)|^2dt + (2^m-1)\delta \int_0^{T_\Lambda+2n\ell_1} |u(t)|^2dt\\ &\le (1+(2^m-1)\delta) \int_0^{T_\Lambda+2n\ell_1} |u(t)|^2dt,\end{aligned}$$

which proves inequality (2.7) with $\gamma = 1+(2^m-1)\delta$. □

Remark 2. If $\mathcal{B}$ is an operator of type S, there exists a unique function $b(\lambda)$ such that $\mathcal{B}e^{i\lambda t} = b(\lambda)e^{i\lambda t}$. The function $b(\lambda)$ is obtained by replacing in the expression of $\mathcal{B}$ given by Proposition 2.3(i) the operators ℓ_i^+ and ℓ_i^- by the functions $\cos\lambda t$ and $i\sin\lambda t$, respectively.

2.3. Construction of $\mathcal{P}$ and $\mathcal{Q}$ in the general case. The construction of $\mathcal{P}$ and $\mathcal{Q}$ will be done by induction. We note that such operators have already been constructed for a network consisting of a single string.

We shall denote by Λ_i the set of all the lengths of the strings of the subtree $\mathcal{A}_i$, and by $\Lambda_{\mathcal{A}}$ that of all the lengths of the tree $\mathcal{A}$. Suppose that for the subtrees $\mathcal{A}_i$, $i = 1, \dots, m$, we have already constructed the operators $\mathcal{P}_i, \mathcal{Q}_i$ that belong to $S(\Lambda_i)$ and verify

$$\mathcal{P}_i G_i + \mathcal{Q}_i F_i = 0, \tag{2.11}$$

where G_i and F_i are the velocity and the tension at the root of the subtree $\mathcal{A}_i$, i.e., at the vertex $\mathcal{O}$ of $\mathcal{A}$.

We define the operators

$$\mathcal{P} = \ell^+ \sum_{i=1}^{m} \mathcal{P}_i \prod_{j\neq i} \mathcal{Q}_j + \ell^- \prod_{j=1}^{m} \mathcal{Q}_j, \tag{2.12}$$

$$\mathcal{Q} = \ell^- \sum_{i=1}^{m} \mathcal{P}_i \prod_{j\neq i} \mathcal{Q}_j + \ell^+ \prod_{j=1}^{m} \mathcal{Q}_j \tag{2.13}$$

(here the products denote the composition of operators).

PROPOSITION 2.6. *The operators $\mathcal{P}$ and $\mathcal{Q}$ defined by* (2.12)–(2.13) *belong to $S(\Lambda_{\mathcal{A}})$. If $\bar{u}$ is a solution of* (N), *then $\mathcal{P}G + \mathcal{Q}F = 0$.*

Proof. To prove that $\mathcal{P}, \mathcal{Q} \in S(\Lambda_{\mathcal{A}})$, it suffices to observe that, according to Proposition 2.3, all the terms of the sums in (2.12) and (2.13) belong to $S(\{\ell\} \sqcup \Lambda_1 \sqcup \cdots \sqcup \Lambda_m) = S(\Lambda_{\mathcal{A}})$. Using (2.4)–(2.5), the coupling conditions (1.12) between the strings may be expressed as

$$\sum_{i=1}^{m} F_i = \ell^- G + \ell^+ F, \qquad G_i = \ell^+ G + \ell^- F, \quad i = 1, \dots, m. \tag{2.14}$$

From (2.12)–(2.13) we have

$$\begin{aligned}\mathcal{P}G + \mathcal{Q}F &= \sum_{i=1}^{m} \left(\mathcal{P}_i \prod_{j\neq i} \mathcal{Q}_j \right) \ell^+ G + \prod_{j=1}^{m} \mathcal{Q}_j \ell^- G + \sum_{i=1}^{m} \left(\mathcal{P}_i \prod_{j\neq i} \mathcal{Q}_j \right) \ell^- F + \prod_{j=1}^{m} \mathcal{Q}_j \ell^+ F \\ &= \sum_{i=1}^{m} \left(\mathcal{P}_i \prod_{j\neq i} \mathcal{Q}_j \right) (\ell^+ G + \ell^- F) + \prod_{j=1}^{m} \mathcal{Q}_j (\ell^- G + \ell^+ F).\end{aligned}$$

Then, using formulas (2.14),

$$\mathcal{P}G + \mathcal{Q}F = \sum_{i=1}^{m} \left(\mathcal{P}_i \prod_{j\neq i} \mathcal{Q}_j \right) G_i + \sum_{i=1}^{m} \left(\prod_{j=1}^{m} \mathcal{Q}_j \right) F_i = \sum_{i=1}^{m} \left(\prod_{j\neq i} \mathcal{Q}_j \right) (\mathcal{P}_i G_i + \mathcal{Q}_i F_i) = 0,$$

where the last equality follows from the hypotheses (2.11). Thus, $\mathcal{P}$ and $\mathcal{Q}$, defined by (2.12)–(2.13), satisfy (2.1). ∎

Remark 3. From the definition, an $S(\Lambda)$-operator $\mathcal{B}$ may be written in the form

$$\mathcal{B} = \sum_{h \in \mathcal{H}_\Lambda} c_h \tau_h. \tag{2.15}$$

In general, this representation is not unique, since some elements of $\mathcal{H}_\Lambda$ may coincide. However, the coefficient $c_{s(\mathcal{B})} = c_{\ell_1 + \cdots + \ell_m}$, corresponding to the largest value of h, is determined in a unique way, as $\ell_1 + \cdots + \ell_m$ cannot be equal to another element of $\mathcal{H}_\Lambda$. Besides, it is easy to see that $c_{s(\mathcal{B})}$ is a multiplicative function; i.e., if $\mathcal{B}_1$ and $\mathcal{B}_2$ are S-operators with $s(\mathcal{B}_1) = s_1$ and $s(\mathcal{B}_2) = s_2$, then $c_{s_1+s_2}(\mathcal{B}_1\mathcal{B}_2) = c_{s_1}(\mathcal{B}_1)c_{s_2}(\mathcal{B}_2)$. In the next proposition we study this coefficient for the operators $\mathcal{P}$ and $\mathcal{Q}$.

PROPOSITION 2.7. *Let $c_{L_\mathcal{A}}(\mathcal{B})$ denote the coefficient corresponding to $h = s(\Lambda_\mathcal{A}) = L_\mathcal{A} \in \mathcal{H}_{\Lambda_\mathcal{A}}$ in the expansion* (2.15) *of an $S(\Lambda_\mathcal{A})$-operator $\mathcal{B}$. Then $c_{L_\mathcal{A}}(\mathcal{P}) = c_{L_\mathcal{A}}(\mathcal{Q}) > 0$.*

Proof. We proceed by induction. For a string, $\mathcal{P} = \ell^+ = \frac{\tau_h + \tau_{-h}}{2}$ and $\mathcal{Q} = \ell^+ = \frac{\tau_h - \tau_{-h}}{2}$. This implies $c_\ell(\mathcal{P}) = c_\ell(\mathcal{Q}) = \frac{1}{2}$.

Now assume the assertion is true for the subtrees $\mathcal{A}_1, \ldots, \mathcal{A}_m$. This means that

$$c_{L_i}(\mathcal{P}) = c_{L_i}(\mathcal{Q}) > 0, \quad i = 1, \ldots, m, \tag{2.16}$$

where, as above, L_i is the sum of the lengths of all the strings of the subtree $\mathcal{A}_i$.

Then, from formula (2.12) and the assumption (2.16),

$$\begin{aligned}
c_{L_\mathcal{A}}(\mathcal{P}) &= c_{L_\mathcal{A}} \left(\ell^+ \sum_{i=1}^m \mathcal{P}_i \prod_{j \neq i} \mathcal{Q}_j + \ell^- \prod_{j=1}^m \mathcal{Q}_j \right) \\
&= c_{L_\mathcal{A}} \left(\ell^+ \sum_{i=1}^m \mathcal{P}_i \prod_{j \neq i} \mathcal{Q}_j \right) + c_{L_\mathcal{A}} \left(\ell^- \prod_{j=1}^m \mathcal{Q}_j \right) \\
&= c_\ell(\ell^+) \sum_{i=1}^m c_{L_i}(\mathcal{P}_i) \prod_{j \neq i} c_{L_j}(\mathcal{Q}_j) + c_\ell(\ell^-) \prod_{j=1}^m c_{L_j}(\mathcal{Q}_j) \\
&= \frac{1}{2}(m+1) \prod_{j=1}^m c_{L_j}(\mathcal{Q}_j) > 0.
\end{aligned}$$

In the same way it may be proved that

$$c_{L_\mathcal{A}}(\mathcal{Q}) = \frac{1}{2}(m+1) \prod_{j=1}^m c_{L_j}(\mathcal{Q}_j),$$

which completes the proof. □

2.4. The action of $\mathcal{P}$ and $\mathcal{Q}$ on the tensions and velocities at the interior nodes. For the index $\bar{\alpha} = (\alpha_1, \ldots, \alpha_k) \in \mathcal{I}$ we denote

$$\tilde{\Lambda}_{\bar{\alpha}} := \{\ell, \ell_{\alpha_1}, \ell_{\alpha_1,\alpha_2}, \ldots, \ell_{\alpha_1,\alpha_2,\ldots,\alpha_{k-1}}\}.$$

Observe that $\tilde{\Lambda}_{\bar{\alpha}}$ is the set of the lengths of the strings forming the unique simple path that connects the root $\mathcal{R}$ with the subtree $\mathcal{A}_{\bar{\alpha}}$. For completeness we take for the empty index $\tilde{\Lambda} = \emptyset$. The following proposition gives information on how the operators

$\mathcal{P}$ and $\mathcal{Q}$ act on traces of the components of a solution at the interior nodes of the network.

PROPOSITION 2.8. *For any $\bar{\alpha} \in \mathcal{I}$ there exist operators $\mathcal{L}_{\bar{\alpha}} \in S(\Lambda_{\mathcal{A}} \sqcup \tilde{\Lambda}_{\bar{\alpha}})$ such that, for any solution of* (N),

$$\mathcal{Q}F_{\bar{\alpha}} = \mathcal{L}_{\bar{\alpha}}G, \qquad \mathcal{P}F_{\bar{\alpha}} = -\mathcal{L}_{\bar{\alpha}}F.$$

Proof. We proceed by induction. Note that from the relation $\mathcal{P}G + \mathcal{Q}F = 0$ it follows that when $\bar{\alpha}$ is the empty multi-index, the property is true with $\mathcal{L} = -\mathcal{P} \in S(\Lambda_{\mathcal{A}}) = S(\Lambda_{\mathcal{A}} \sqcup \tilde{\Lambda})$. In particular, for a single string the assertion of the proposition holds.

Suppose now that the operators $\mathcal{L}_{\bar{\alpha}}$ have already been constructed for the subtrees $\mathcal{A}_i$, $i = 1, \dots, m$, of $\mathcal{A}$. This means that we have for $i = 1, \dots, m$ the operators $\mathcal{L}^i_{\bar{\alpha}} \in S(\Lambda_i \sqcup \tilde{\Lambda}^i_{\bar{\alpha}})$ such that

$$\mathcal{P}_i F_{i\circ\bar{\alpha}} = -\mathcal{L}^i_{\bar{\alpha}} F_i, \qquad \mathcal{Q}_i F_{i\circ\bar{\alpha}} = \mathcal{L}^i_{\bar{\alpha}} G_i,$$

where $\tilde{\Lambda}^i_{\bar{\alpha}}$ is the set defined as $\tilde{\Lambda}_{\bar{\alpha}}$ for the subtree $\mathcal{A}_i$ and $\mathcal{P}_i$, $\mathcal{Q}_i$ are the operators $\mathcal{P}$, $\mathcal{Q}$ corresponding to that subtree.

Then, using relation (2.13),

$$\begin{aligned}
\mathcal{Q}F_{i\circ\bar{\alpha}} &= \ell^- \left(\sum_{j=1}^{m} \mathcal{P}_j \prod_{k\neq j} \mathcal{Q}_k \right) F_{i\circ\bar{\alpha}} + \ell^+ \left(\prod_{k=1}^{m} \mathcal{Q}_k \right) F_{i\circ\bar{\alpha}} \\
&= \ell^- \left(\sum_{\substack{j=1 \\ j\neq i}}^{m} \mathcal{P}_j \prod_{\substack{k\neq j \\ k\neq i}} \mathcal{Q}_k \right) \mathcal{Q}_i F_{i\circ\bar{\alpha}} + \ell^- \left(\mathcal{P}_i \prod_{k\neq i} \mathcal{Q}_k \right) F_{i\circ\bar{\alpha}} + \ell^+ \left(\prod_{k=1}^{m} \mathcal{Q}_k \right) F_{i\circ\bar{\alpha}} \\
&= \mathcal{L}^i_{\bar{\alpha}} \left(\ell^- \left(\sum_{\substack{i=1 \\ j\neq i}}^{m} \mathcal{P}_j \prod_{\substack{k\neq j \\ k\neq i}} \mathcal{Q}_k \right) G_i - \ell^- \left(\prod_{k\neq i} \mathcal{Q}_k \right) F_i + \ell^+ \left(\prod_{k=1}^{m} \mathcal{Q}_k \right) G_i \right) \\
&= \mathcal{L}^i_{\bar{\alpha}} \left(\ell^- \sum_{j\neq i} \left(\prod_{\substack{k\neq j \\ k\neq i}} \mathcal{Q}_k \right) (\mathcal{P}_j G_i + \mathcal{Q}_j \widehat{F}) - \ell^- \left(\prod_{k\neq i} \mathcal{Q}_k \right) \widehat{F} + \ell^+ \left(\prod_{k\neq i} \mathcal{Q}_k \right) G_i \right) \\
&= \mathcal{L}^i_{\bar{\alpha}} \left(\prod_{k\neq i} \mathcal{Q}_k \right) (\ell^+ \widehat{G} - \ell^- \widehat{F}) = \mathcal{L}^i_{\bar{\alpha}} \left(\prod_{k\neq i} \mathcal{Q}_k \right) \left(\ell^+ (\ell^- F + \ell^+ G) - \ell^- (\ell^+ F + \ell^- G) \right) \\
&= \mathcal{L}^i_{\bar{\alpha}} \left(\prod_{k\neq i} \mathcal{Q}_k \right) \left((\ell^+)^2 - (\ell^-)^2 \right) G.
\end{aligned}$$

In a similar way, it may be obtained that

$$\mathcal{P}F_{i\circ\bar{\alpha}} = -\mathcal{L}^i_{\bar{\alpha}} \prod_{k\neq i} \mathcal{Q}_k \left((\ell^+)^2 - (\ell^-)^2 \right) F.$$

Thus, we arrive at the recursive formula

$$\mathcal{L}_{i\circ\bar{\alpha}} = \mathcal{L}^i_{\bar{\alpha}} \prod_{k\neq i} \mathcal{Q}_k \left((\ell^+)^2 - (\ell^-)^2 \right),$$

from which, in particular, according to Proposition 2.3, it holds that the operators $\mathcal{L}_{i\circ\bar\alpha}$ belong to $S(\Lambda_i \sqcup \tilde\Lambda^i_{\bar\alpha} \sqcup \{\ell,\ell\}) = S(\Lambda_i \sqcup \{\ell\} \sqcup \tilde\Lambda^i_{\bar\alpha} \sqcup \{\ell\}) = S(\Lambda_{\mathcal{A}} \sqcup \tilde\Lambda_{i\circ\bar\alpha})$. This proves the proposition. □

The action of $\mathcal{P}$ and $\mathcal{Q}$ on the velocities $G_{\bar\alpha}$ may be described in a similar way, as follows.

PROPOSITION 2.9. *For any $\bar\alpha \in \mathcal{I}$ there exist operators $\mathcal{K}_{\bar\alpha}$, $\widehat{\mathcal{K}}_{\bar\alpha} \in S(\Lambda_{\mathcal{A}} \sqcup \tilde\Lambda_{\bar\alpha})$ such that, for any solution of* (N),

$$\mathcal{Q}G_{\bar\alpha} = \mathcal{K}_{\bar\alpha}G, \qquad \mathcal{P}G_{\bar\alpha} = \widehat{\mathcal{K}}_{\bar\alpha}F.$$

Proof. From the relation $\mathcal{P}G + \mathcal{Q}F = 0$, it follows that, for the empty multi-index, $\mathcal{K} = \mathcal{Q}$ and $\widehat{\mathcal{K}} = -\mathcal{Q}$. For the remaining indices the operators $\mathcal{K}_{\bar\alpha}$ and $\widehat{\mathcal{K}}_{\bar\alpha}$ are constructed by recurrence. Assume that for the index $\bar\alpha$ the operators $\mathcal{K}_{\bar\alpha}$ and $\widehat{\mathcal{K}}_{\bar\alpha}$, verifying the conditions of the proposition, have been already constructed.

Then, for the indices $\bar\alpha \circ i$ with $i = 1, \dots, m_{\bar\alpha}$ we have that

$$\mathcal{Q}G_{\bar\alpha\circ i} = \mathcal{Q}\widehat{G}_{\bar\alpha} = \ell^+_{\bar\alpha}\mathcal{Q}G_{\bar\alpha} + \ell^-_{\bar\alpha}\mathcal{Q}F_{\bar\alpha} = \left(\ell^+_{\bar\alpha}\mathcal{K}_{\bar\alpha} + \ell^-_{\bar\alpha}\mathcal{L}_{\bar\alpha}\right)G,$$

where $\mathcal{L}_{\bar\alpha}$ is the operator constructed in the previous proposition.

In an analogous way it may be obtained that

$$\mathcal{P}G_{\bar\alpha\circ i} = (\ell^+_{\bar\alpha}\widehat{\mathcal{K}}_{\bar\alpha} - \ell^-_{\bar\alpha}\mathcal{L}_{\bar\alpha})F.$$

Then the needed operators may be constructed by the rules

$$\mathcal{K}_{\bar\alpha\circ i} = \ell^+_{\bar\alpha}\mathcal{K}_{\bar\alpha} + \ell^-_{\bar\alpha}\mathcal{L}_{\bar\alpha}, \tag{2.17}$$

$$\widehat{\mathcal{K}}_{\bar\alpha\circ i} = \ell^+_{\bar\alpha}\widehat{\mathcal{K}}_{\bar\alpha} - \ell^-_{\bar\alpha}\mathcal{L}_{\bar\alpha}. \tag{2.18}$$

As in the proof of the Proposition 2.8, from the relations (2.17)–(2.18) it holds, in particular, that the operators $\mathcal{K}_{\bar\alpha\circ i}$ and $\widehat{\mathcal{K}}_{\bar\alpha\circ i}$ belong to $S(\Lambda_{\mathcal{A}} \sqcup \tilde\Lambda_{\bar\alpha\circ i})$. □

2.5. Action of $\mathcal{P}$ and $\mathcal{Q}$ on the solution. If $\bar u$ is a solution of (N) and $\mathcal{B}$ is an operator of type S, then, due to the linearity of $\mathcal{B}$ and (N), $\mathcal{B}\bar u$ is also a solution of (N). Moreover, if $G^{\mathcal{B}\bar u}_{\bar\alpha}$ and $F^{\mathcal{B}\bar u}_{\bar\alpha}$, $\bar\alpha \in \mathcal{I}$, denote the velocity and strength traces of the strings at the vertices of the network for the solution $\mathcal{B}\bar u$, then

$$G^{\mathcal{B}\bar u}_{\bar\alpha} = \mathcal{B}G_{\bar\alpha}, \qquad F^{\mathcal{B}\bar u}_{\bar\alpha} = \mathcal{B}F_{\bar\alpha}.$$

That is true, in particular, when $\mathcal{B}$ is one of the operators $\mathcal{P}$ or $\mathcal{Q}$. The following lemma contains a fundamental technical step in our construction.

LEMMA 2.10. *There exists a constant C, independent of $\bar u$, such that*

$$\mathbf{E}_{\mathcal{P}\bar u}(t) \le C\int_{T^*-2L_{\mathcal{A}}}^{T^*+2L_{\mathcal{A}}} |F(t)|^2 dt, \qquad \mathbf{E}_{\mathcal{Q}\bar u}(t) \le C\int_{T^*-2L_{\mathcal{A}}}^{T^*+2L_{\mathcal{A}}} |G(t)|^2 dt \tag{2.19}$$

for every $T^* \in \mathbb{R}$ and $t \in [T^* - L_{\mathcal{A}}, T^* + L_{\mathcal{A}}]$.

Proof. (i) Fix $T^* \in \mathbb{R}$. We shall prove first that

$$\mathbf{E}_{\mathcal{P}\bar u}(T^*) \le C\int_{T^*-2L_{\mathcal{A}}}^{T^*+2L_{\mathcal{A}}} |F(t)|^2 dt, \qquad \mathbf{E}_{\mathcal{Q}\bar u}(T^*) \le C\int_{T^*-2L_{\mathcal{A}}}^{T^*+2L_{\mathcal{A}}} |G(t)|^2 dt. \tag{2.20}$$

As a consequence of Propositions 2.8 and 2.9 we have

$$\mathcal{Q}F_{\bar\alpha} = \mathcal{L}_{\bar\alpha}G, \qquad \mathcal{Q}G_{\bar\alpha} = \mathcal{K}_{\bar\alpha}G,$$
$$\mathcal{P}F_{\bar\alpha} = -\mathcal{L}_{\bar\alpha}F, \qquad \mathcal{P}G_{\bar\alpha} = \widehat{\mathcal{K}}_{\bar\alpha}F$$

for $\bar\alpha \in \mathcal{I}$. Then, from Propositions 2.1 and 2.4(i) it follows that

$$E^{\bar\alpha}_{\mathcal{Q}\bar u}(T^*) \le C\int_{T^*-\ell_{\bar\alpha}}^{T^*+\ell_{\bar\alpha}} \left(|\mathcal{L}_{\bar\alpha}G(t)|^2 + |\mathcal{K}_{\bar\alpha}G(t)|^2\right) dt \le C\int_{T^*-2L_{\mathcal{A}}}^{T^*+2L_{\mathcal{A}}} |G(t)|^2 dt,$$

$$E^{\bar\alpha}_{\mathcal{P}\bar u}(T^*) \le C\int_{T^*-\ell_{\bar\alpha}}^{T^*+\ell_{\bar\alpha}} \left(|\mathcal{L}_{\bar\alpha}F(t)|^2 + |\widehat{\mathcal{K}}_{\bar\alpha}F(t)|^2\right) dt \le C\int_{T^*-2L_{\mathcal{A}}}^{T^*+2L_{\mathcal{A}}} |F(t)|^2 dt,$$

where, as above, $E^{\bar\alpha}$ is the energy of the solution in the string $\mathbf{e}_{\bar\alpha}$. It suffices to note that $\mathbf{E} = \sum_{\bar\alpha\in\mathcal{I}} E^{\bar\alpha}$ to obtain the inequalities (2.20).

(ii) Now we prove that these inequalities remain true for all $t \in [T^* - L_{\mathcal{A}}, L^* + T_{\mathcal{A}}]$. Indeed, if $t \in [T^* - L_{\mathcal{A}}, T^* + L_{\mathcal{A}}]$, from Proposition 1.1 we have

$$\begin{aligned}
\mathbf{E}_{\mathcal{P}\bar u}(t) &= \mathbf{E}_{\mathcal{P}\bar u}(T^*) - \int_{T^*}^{t} F^{\mathcal{P}\bar u}(\tau)G^{\mathcal{P}\bar u}(\tau)dt \\
&\le \mathbf{E}_{\mathcal{P}\bar u}(T^*) + \left|\int_{T^*}^{t} (|F^{\mathcal{P}\bar u}(\tau)|^2 + |G^{\mathcal{P}\bar u}(\tau)|^2)dt\right| \\
&\le \mathbf{E}_{\mathcal{P}\bar u}(T^*) + \int_{T^*-L_{\mathcal{A}}}^{T^*+L_{\mathcal{A}}} (|\mathcal{P}F(\tau)|^2 + |\mathcal{P}G(\tau)|^2)dt \\
&\le \mathbf{E}_{\mathcal{P}\bar u}(T^*) + \int_{T^*-L_{\mathcal{A}}}^{T^*+L_{\mathcal{A}}} (|\mathcal{P}F(\tau)|^2 + |\mathcal{Q}F(\tau)|^2)dt \le C\int_{T^*-2L_{\mathcal{A}}}^{T^*+2L_{\mathcal{A}}} |F(\tau)|^2 dt.
\end{aligned}$$

(In the last step we have used Proposition 2.4(i) and the result of (i).) For the operator $\mathcal{Q}$ the proof is similar. □

Remark 4. When $\bar u$ is a solution of (1.1)–(1.5) (i.e., $G \equiv 0$), Lemma 2.10 gives $\mathbf{E}_{\mathcal{Q}\bar u}(t) = 0$. This implies that $\mathcal{Q}\bar u(t) = 0$. This relation may be viewed as a generalization of the time periodicity property of the solutions of the 1-d wave equation with homogeneous Dirichlet boundary conditions, which with our notation may be written as $\ell^- u(t) = 0$. As we have shown in Proposition 2.5, this generalized periodicity implies that all the essential L^2 information on $\bar u$ is contained in an interval of length $2L_{\mathcal{A}}$.

3. The main observability theorem. In this section we prove our main observability result for the solutions of the homogeneous system (1.1)–(1.5).

For every nonempty multi-index $\bar\alpha = (\alpha_1, \dots, \alpha_k) \in \mathcal{I}$ we define the operator $\mathcal{D}_{\bar\alpha}$ by

$$\mathcal{D}_{\bar\alpha} := \left(\prod_{i=1,\ i\ne\alpha_1}^{m} \mathcal{Q}_i\right)\left(\prod_{i=1,\ i\ne\alpha_2}^{m_{\alpha_1}} \mathcal{Q}_{\alpha_1,i}\right)\cdots\left(\prod_{i=1,\ i\ne\alpha_{k-1}}^{m_{\alpha_1,\dots,\alpha_{k-1}}} \mathcal{Q}_{\alpha_1,\dots,\alpha_{k-1},i}\right), \tag{3.1}$$

and for the empty index $\mathcal{D}$ is the identity operator. We recall that $\mathcal{Q}_{\bar\beta}$ is the operator constructed in the previous section for the subtree $\mathcal{A}_{\bar\beta}$ and that the products in (3.1) denote the composition of operators.

Note that for every $\bar\alpha \in \mathcal{I}$ the operator $\mathcal{D}_{\bar\alpha}$ is of type S with $s(\mathcal{D}_{\bar\alpha}) < L_{\mathcal{A}}$.

Now the following result holds.

LEMMA 3.1. *There exists a positive constant C such that for every $\bar\alpha \in \mathcal{I}_{\mathcal{S}}$ and every solution $\bar u$ of* (N)

$$\mathbf{E}_{\mathcal{D}_{\bar\alpha}\bar u}(t) \le C \int_{t-2L_{\mathcal{A}}}^{t+2L_{\mathcal{A}}} \left(|F(\tau)|^2 + |G(\tau)|^2\right) d\tau$$

for any $t \in \mathbb{R}$.

Proof. We proceed by induction. For the case of a single string the assertion is an immediate consequence of Proposition 2.1.

Now let $\bar\alpha = (\alpha_1, \dots, \alpha_k) \in \mathcal{I}_{\mathcal{S}}$ and assume that the assertion of the theorem is true for the subtree $\mathcal{A}_{\alpha_1}$. That implies that

$$\mathbf{E}^{\alpha_1}_{\mathcal{D}^{\alpha_1}_{\bar\alpha}\bar u}(t) \le C \int_{t-2L_{\alpha_1}}^{t+2L_{\alpha_1}} \left(|F_{\alpha_1}(\tau)|^2 + |G_{\alpha_1}(\tau)|^2\right) d\tau \tag{3.2}$$

for any solution $\bar u$ of (N), where

$$\mathcal{D}^{\alpha_1}_{\bar\alpha} := \left(\prod_{i=1,\ i\neq\alpha_2}^{m_{\alpha_1}} \mathcal{Q}_{\alpha_1,i}\right) \cdots \left(\prod_{i=1,\ i\neq\alpha_{k-1}}^{m_{\alpha_1,\dots,\alpha_{k-1}}} \mathcal{Q}_{\alpha_1,\dots,\alpha_{k-1},i}\right) \tag{3.3}$$

is the operator $\mathcal{D}_{\bar\alpha}$ for the subtree $\mathcal{A}_{\alpha_1}$ with $\bar\alpha = (\alpha_2, \dots, \alpha_k)$.

First, we estimate the energy $\mathbf{E}^{\alpha_1}_{\mathcal{D}_{\bar\alpha}\bar u}$ of $\mathcal{D}_{\bar\alpha}\bar u$ on the subtree $\mathcal{A}_{\alpha_1}$. To do this, we set

$$\bar\omega := \left(\prod_{j=1,\ j\neq\alpha_1}^{m} \mathcal{Q}_j\right) \bar u, \qquad \bar\omega_i := \left(\prod_{\substack{j=1,\ j\neq\alpha_1\\ j\neq i}}^{m} \mathcal{Q}_j\right) \bar u, \quad i = 1, \dots, m. \tag{3.4}$$

Note that these functions are also solutions of (N). They verify $\bar\omega = \mathcal{Q}_i\bar\omega_i$ and

$$\mathcal{D}_{\bar\alpha}\bar u = \mathcal{D}^{\alpha_1}_{\bar\alpha}\bar\omega. \tag{3.5}$$

Additionally, from (3.2),

$$\mathbf{E}^{\alpha_1}_{\mathcal{D}^{\alpha_1}_{\bar\alpha}\bar\omega}(t) \le C \int_{t-2L_{\alpha_1}}^{t+2L_{\alpha_1}} \left(|F^{\bar\omega}_{\alpha_1}(\tau)|^2 + |G^{\bar\omega}_{\alpha_1}(\tau)|^2\right) d\tau. \tag{3.6}$$

However, from the coupling formulas (1.12) we obtain that

$$\sum_{i=1}^{m} F^{\bar\omega}_i = \widehat{F}^{\bar\omega}, \qquad G^{\bar\omega}_i = \widehat{G}^{\bar\omega}, \tag{3.7}$$

so that

$$F^{\bar\omega}_{\alpha_1} = \widehat{F}^{\bar\omega} - \sum_{i=1,\ i\neq\alpha_1}^{m} \mathcal{Q}_i F^{\bar\omega_i}_i = \widehat{F}^{\bar\omega} + \sum_{i=1,\ i\neq\alpha_1}^{m} \mathcal{P}_i \widehat{G}^{\bar\omega_i}, \qquad G^{\bar\omega}_{\alpha_1} = \hat{G}^{\bar\omega}, \tag{3.8}$$

holds. Then, using the equalities (3.8), (3.6) gives

$$\mathbf{E}^{\alpha_1}_{\mathcal{D}^{\alpha_1}_{\bar\alpha}\bar\omega}(t) \le C\int_{t-2L_{\alpha_1}}^{t+2L_{\alpha_1}} \left(\left| \widehat{F}^{\bar\omega}(\tau) + \sum_{i=1,\, i\neq\alpha_1}^{m} \mathcal{P}_i \widehat{G}^{\bar\omega_i}(\tau) \right|^2 + |\hat{G}^{\bar\omega}(\tau)|^2 \right) d\tau,$$

and this implies

$$\mathbf{E}^{\alpha_1}_{\mathcal{D}^{\alpha_1}_{\bar\alpha}\bar\omega}(t) \le C\int_{t-2L_{\alpha_1}}^{t+2L_{\alpha_1}} \left(|\widehat{F}^{\bar\omega}(\tau)|^2 + \sum_{i=1,\, i\neq\alpha_1}^{m} |\mathcal{P}_i \widehat{G}^{\bar\omega_i}(\tau)|^2 + |\hat{G}^{\bar\omega}(\tau)|^2 \right) d\tau. \tag{3.9}$$

Now, from the definition of $\bar\omega$ and the formulas (2.4), (2.5) we have

$$\widehat{F}^{\bar\omega} = \left(\prod_{j=1,\, j\neq\alpha_1}^{m} \mathcal{Q}_j \right) \widehat{F} = \left(\prod_{j=1,\, j\neq\alpha_1}^{m} \mathcal{Q}_j \right) \ell^+ F + \left(\prod_{j=1,\, j\neq\alpha_1}^{m} \mathcal{Q}_j \right) \ell^- G$$

and consequently

$$\int_{t-2L_{\alpha_1}}^{t+2L_{\alpha_1}} |\widehat{F}^{\bar\omega}|^2 d\tau \le 2\int_{t-2L_{\alpha_1}}^{t+2L_{\alpha_1}} \left(\left| \left(\prod_{j=1,\, j\neq\alpha_1}^{m} \mathcal{Q}_j \right) \ell^+ F \right|^2 + \left| \left(\prod_{j=1,\, j\neq\alpha_1}^{m} \mathcal{Q}_j \right) \ell^- G \right|^2 \right) d\tau.$$

Observe that the operators $(\prod_{j=1,\, j\neq\alpha_1}^{m} \mathcal{Q}_j)\ell^+$ and $(\prod_{j=1,\, j\neq\alpha_1}^{m} \mathcal{Q}_j)\ell^-$ are of type S with $s < L_{\mathcal{A}} - L_{\alpha_1}$, so that the latter inequality combined with Proposition 2.4 provides

$$\int_{t-2L_{\alpha_1}}^{t+2L_{\alpha_1}} |\widehat{F}^{\bar\omega}(\tau)|^2 d\tau \le C\int_{t-2L_{\mathbb{A}}}^{t+2L_{\mathcal{A}}} (|F(\tau)|^2 + |G(\tau)|^2) d\tau.$$

In a similar way it may be proved that

$$\int_{t-2L_{\alpha_1}}^{t+2L_{\alpha_1}} |\mathcal{P}_i \widehat{G}^{\bar\omega_i}(\tau)|^2 d\tau \le C\int_{t-2L_{\mathbb{A}}}^{t+2L_{\mathcal{A}}} (|F(\tau)|^2 + |G(\tau)|^2) d\tau$$

and

$$\int_{t-2L_{\alpha_1}}^{t+2L_{\alpha_1}} |\hat{G}^{\bar\omega}(\tau)|^2 d\tau \le C\int_{t-2L_{\mathbb{A}}}^{t+2L_{\mathcal{A}}} (|F(\tau)|^2 + |G(\tau)|^2) d\tau.$$

Therefore, these three inequalities together with (3.5) and (3.9) give

$$\mathbf{E}^{\alpha_1}_{\mathcal{D}_{\bar\alpha}\bar u}(t) \le C\int_{t-2L_{\mathbb{A}}}^{t+2L_{\mathcal{A}}} (|F(\tau)|^2 + |G(\tau)|^2) d\tau. \tag{3.10}$$

Now we proceed to estimate the energies $\mathbf{E}^{i}_{\mathcal{D}_{\bar\alpha}\bar u}$ of $\mathcal{D}_{\bar\alpha}\bar u$ on the remaining subtrees $\mathcal{A}_i$ (i.e., for $i \neq \alpha_1$). According to Lemma 2.10, applied to $\bar\omega_i$ in the subtree $\mathcal{A}_i$, it holds that for every t' in $[t - L_i, t + L_i]$

$$\mathbf{E}^{i}_{\bar\omega}(t') = \mathbf{E}^{i}_{\mathcal{Q}_i\bar\omega_i}(t') \le C\int_{t-2L_i}^{t+2L_i} |G_i^{\bar\omega_i}(\tau)|^2 d\tau \tag{3.11}$$

for $i = 1, \dots, m$. Taking into account that

$$G_i^{\bar\omega_i} = \left(\prod_{\substack{j=1,\ j\neq\alpha_1\\ j\neq i}}^{m} \mathcal{Q}_j \right) \widehat{G} = \left(\prod_{\substack{j=1,\ j\neq\alpha_1\\ j\neq i}}^{m} \mathcal{Q}_j \right) \ell^+ F + \left(\prod_{\substack{j=1,\ j\neq\alpha_1\\ j\neq i}}^{m} \mathcal{Q}_j \right) \ell^- G, \tag{3.12}$$

we get from (3.11) and Proposition 2.4(i)

$$\mathbf{E}^i_{\bar\omega}(t') \leq C \int_{t-L_{\mathcal{A}}-L_i+L_{\alpha_1}}^{t+L_{\mathcal{A}}+L_i-L_{\alpha_1}} (|F(\tau)|^2 + |G(\tau)|^2) d\tau \leq C \int_{t-2L_{\mathcal{A}}}^{t+2L_{\mathcal{A}}} (|F(\tau)|^2 + |G(\tau)|^2) d\tau. \tag{3.13}$$

(Here we have used the fact that the operators applied to F and G in the right-hand term of (3.12) are of type S with $s = L_{\mathcal{A}} - L_{\alpha_1} - L_i$.)

Now, if we apply Proposition 2.4(ii) with $\mathcal{B} = \mathcal{D}_{\bar\alpha}^{\alpha_1}$ to (3.13) (recall that $s(\mathcal{D}_{\bar\alpha}^{\alpha_1}) < L_{\alpha_1}$), we obtain, after choosing $t' = t$,

$$\mathbf{E}^i_{\mathcal{D}_{\bar\alpha}} \bar u(t) = \mathbf{E}^i_{\mathcal{D}_{\bar\alpha}^{\alpha_1}\bar\omega}(t') \leq \ C \int_{t-2L_{\mathcal{A}}}^{t+2L_{\mathcal{A}}} (|F(\tau)|^2 + |G(\tau)|^2) d\tau. \tag{3.14}$$

Finally, from Proposition 2.1 we obtain that the component u of $\bar u$ verifies, for every $t' \in [t - L_{\mathcal{A}}, t + L_{\mathcal{A}}]$,

$$E_u(t') \leq \ C \int_{t-\ell-L_{\mathcal{A}}}^{t+\ell+L_{\mathcal{A}}} (|F(\tau)|^2 + |G(\tau)|^2) d\tau.$$

Thus, using Proposition 2.4(ii), it holds that

$$E_{\mathcal{D}_{\bar\alpha}\bar u}(t') \leq \ C \int_{t-\ell-L_{\mathcal{A}}}^{t+\ell+L_{\mathcal{A}}} (|F(\tau)|^2 + |G(\tau)|^2) d\tau$$

for every $t' \in [t - L_{\mathcal{A}} + s(\mathcal{D}_{\bar\alpha}), t + L_{\mathcal{A}} - s(\mathcal{D}_{\bar\alpha})]$, and, since $s(\mathcal{D}_{\bar\alpha}) < L_{\mathcal{A}}$, this is true in particular for $t' = t$. Therefore,

$$E_{\mathcal{D}_{\bar\alpha}\bar u}(t) \leq \ C \int_{t-2L_{\mathcal{A}}}^{t+2L_{\mathcal{A}}} (|F(\tau)|^2 + |G(\tau)|^2) d\tau. \tag{3.15}$$

Now it suffices to combine (3.10), (3.14), (3.15) and the fact that $\mathbf{E}_{\mathcal{D}_{\bar\alpha}\bar u} = E_{\mathcal{D}_{\bar\alpha}\bar u} + \sum_{i=1}^m \mathbf{E}^i_{\mathcal{D}_{\bar\alpha}\bar u}$ to conclude the proof. □

With the help of Lemma 3.1 we obtain the following important property of the solutions of the system (1.1)–(1.6).

THEOREM 3.2. *There exists a constant C such that*

$$E_{\mathcal{D}_{\bar\alpha}\bar u}(0) = E_{\mathcal{D}_{\bar\alpha}\bar u}(t) \leq \ C \int_0^{2L_{\mathcal{A}}} |F(\tau)|^2 d\tau$$

for every solution $\bar u$ of (1.1)–(1.5) *and any $\bar\alpha \in \mathcal{I}_{\mathcal{S}}$.*

Proof. If $\bar u$ is a solution of (1.1)–(1.6), so is $\mathcal{D}_{\bar\alpha}\bar u$. In particular, the energy of $\mathcal{D}_{\bar\alpha}\bar u$ is conserved. Then, taking into account that $G \equiv 0$ for the solutions of (1.1)–(1.5), from Lemma 3.1 it holds that

$$\mathbf{E}_{\mathcal{D}_{\bar\alpha}\bar u}(0) = \mathbf{E}_{\mathcal{D}_{\bar\alpha}\bar u}(2T_{\mathcal{A}}) \leq \ C \int_0^{4L_{\mathcal{A}}} |F(\tau)|^2 d\tau. \tag{3.16}$$

On the other hand, in this case $\mathcal{Q}F \equiv 0$, and then, using Proposition 2.5 (which may be applied to $\mathcal{Q}$ on the basis of Proposition 2.7), we have

$$\int_0^{4L_{\mathcal{A}}} |F(\tau)|^2 d\tau \le C \int_0^{2L_{\mathcal{A}}} |F(\tau)|^2 d\tau.$$

With this, the assertion of the theorem follows from (3.16). ∎

4. Relation between $\mathcal{P}$ and $\mathcal{Q}$ and the eigenvalues.

4.1. The eigenvalue problem. We consider the eigenvalue problem for the Laplace operator on the network associated with the hyperbolic problem (1.1)–(1.5):

$$-\theta^{\bar\alpha}_{xx}(x) = \mu\ \theta^{\bar\alpha}(x) \qquad \text{in } [0,\ell_{\bar\alpha}], \quad \bar\alpha \in \mathcal{I}, \tag{4.1}$$

$$\theta^{\bar\alpha\circ\beta}(0) = \theta^{\bar\alpha}(\ell_{\bar\alpha}), \qquad \bar\alpha \in \mathcal{I}_{\mathcal{M}}, \quad \beta = 1,\dots,m_{\bar\alpha}, \tag{4.2}$$

$$\sum_{\beta=1}^{m_{\bar\alpha}} \theta^{\bar\alpha\circ\beta}_x(0) = \theta^{\bar\alpha}_x(\ell_{\bar\alpha}), \qquad \bar\alpha \in \mathcal{I}_{\mathcal{M}}, \tag{4.3}$$

$$\theta^{\bar\alpha}(\ell_{\bar\alpha}) = 0, \qquad \bar\alpha \in \mathcal{I}_{\mathcal{S}}, \tag{4.4}$$

$$\theta(0) = 0 \qquad \text{at the root } \mathcal{R}. \tag{4.5}$$

It is well known (see, e.g., [8]) that the spectrum of this problem is formed by a positive, increasing sequence $\{\mu_k\}_{k\in\mathbb{Z}_+}$ of eigenvalues. We call it *spectrum of* $\mathcal{A}$ and denote it by $\sigma_{\mathcal{A}}$.

Clearly, we may consider the problem (4.1)–(4.5) for each subtree $\mathcal{A}_{\bar\alpha}$ of $\mathcal{A}$. The corresponding spectrum is called the *spectrum of* $\mathcal{A}_{\bar\alpha}$ and is denoted by $\sigma_{\bar\alpha}$.

For technical reasons, as we did for the system (1.1)–(1.5), we will also consider smooth solutions of (4.1), which verify the boundary conditions (4.2)–(4.4) but not necessarily (4.5). For brevity, they are simply called *solutions of* (N_E) *corresponding to* μ.

PROPOSITION 4.1. *If μ is a common eigenvalue of two subtrees $\mathcal{A}_{\bar\alpha\circ i}$, $\mathcal{A}_{\bar\alpha\circ j}$ $(i \ne j)$ with the same root $\mathcal{O}_{\bar\alpha}$, then μ is also an eigenvalue of $\mathcal{A}$. Moreover, there exists a nonzero eigenfunction $\bar\theta$ associated with μ such that $\theta(0) = \theta_x(0) = 0$.*

Proof. Let $\bar\theta^{\bar\alpha\circ i}$, $\bar\theta^{\bar\alpha\circ j}$ be nonzero eigenfunctions corresponding to the eigenvalue μ for the subtrees $\mathcal{A}_{\bar\alpha\circ i}$ and $\mathcal{A}_{\bar\alpha\circ j}$, respectively. These functions are defined in the corresponding subtrees, but it will be sufficient to paste them conveniently to build up an eigenfunction of $\mathcal{A}$.

We may assume that the numbers $\theta^{\bar\alpha\circ i}_x(0)$, $\theta^{\bar\alpha\circ j}_x(0)$ are both different from zero. Indeed, if one of them, say $\theta^{\bar\alpha\circ i}_x(0)$, vanishes, then the relations $\theta^{\bar\alpha\circ i}(0) = \theta^{\bar\alpha\circ i}_x(0) = 0$ ensure that the function $\bar\theta$, obtained by extending by zero the function $\bar\theta^{\bar\alpha\circ i}$ to the whole tree $\mathcal{A}$, is an eigenfunction of $\mathcal{A}$.

Now define the function $\bar\theta$ by

$$\theta_{\bar\alpha'} = \begin{cases} \theta^{\bar\alpha\circ j}_x(0)\ \theta^{\bar\alpha\circ i}_{\bar\beta,x} & \text{if } \bar\alpha' = \bar\alpha\circ i\circ\bar\beta, \\ -\theta^{\bar\alpha\circ i}_x(0)\ \theta^{\bar\alpha\circ j}_{\bar\beta,x} & \text{if } \bar\alpha' = \bar\alpha\circ j\circ\bar\beta, \\ 0 & \text{otherwise;} \end{cases}$$

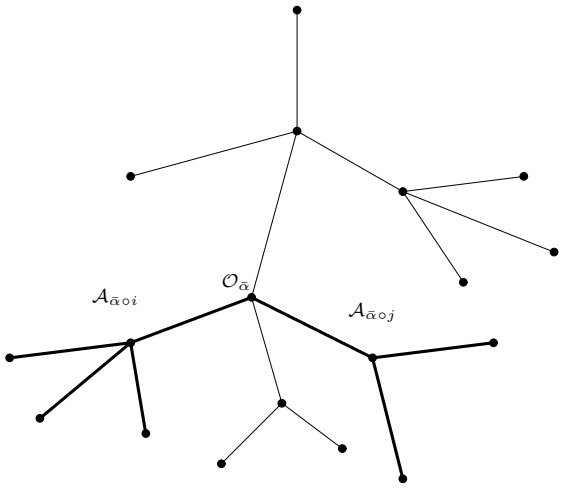


FIG. 4.1. *The subtree* $\mathcal{A}_{\bar\alpha\circ i} \vee \mathcal{A}_{\bar\alpha\circ j}$.

i.e., $\bar\theta$ coincides in the subtree $\mathcal{A}_{\bar\alpha\circ i}$ with $\theta_x^{\bar\alpha\circ j}(0)\bar\theta^{\bar\alpha\circ i}$, in $\mathcal{A}_{\bar\alpha\circ j}$ with $-\theta_x^{\bar\alpha\circ i}(0)\bar\theta^{\bar\alpha\circ j}$, and vanishes outside those subtrees. It is easy to see that $\bar\theta$ satisfies the boundary conditions (4.2)–(4.3) at $\mathcal{O}_{\bar\alpha}$:

$$\sum_{k=1}^{m_{\bar\alpha}} \theta_x^{\bar\alpha\circ k}(0) = \theta_x^{\bar\alpha\circ j}(0)\theta_x^{\bar\alpha\circ i}(0) - \theta_x^{\bar\alpha\circ i}(0)\theta_x^{\bar\alpha\circ j}(0) = 0 = \theta_x^{\bar\alpha}(\ell_{\bar\alpha}).$$

At the other nodes the coupling conditions are obviously satisfied, and therefore $\bar\theta$ is an eigenfunction of $\mathcal{A}$.

Finally, observe that in both cases the eigenfunction $\bar\theta$ constructed here is such that $\theta(0) = \theta_x(0) = 0$ (and thus, $\theta \equiv 0$; i.e., $\bar\theta$ vanishes at the whole string containing the root). □

Remark 5. Note that the eigenfunction obtained in the proof of the previous proposition vanishes everywhere outside the subtrees $\mathcal{A}_{\bar\alpha\circ i}$, $\mathcal{A}_{\bar\alpha\circ j}$. If we denote by $\mathcal{A}_{\bar\alpha\circ i} \vee \mathcal{A}_{\bar\alpha\circ j}$ the tree formed by $\mathcal{A}_{\bar\alpha\circ i}$ and $\mathcal{A}_{\bar\alpha\circ j}$ in which the node $\mathcal{O}_{\bar\alpha}$ is considered as an interior point of a string of length $\ell_{\bar\alpha\circ i} + \ell_{\bar\alpha\circ j}$ (see Figure 4.1), we obtain that these subtrees have a common eigenvalue if and only if there exists an eigenfunction of $\mathcal{A}_{\bar\alpha\circ i} \vee \mathcal{A}_{\bar\alpha\circ j}$ that vanishes at the point $\mathcal{O}_{\bar\alpha}$.

As has been shown above, the operators $\mathcal{P}$ and $\mathcal{Q}$ are of type S with $s(\mathcal{P}) = s(\mathcal{Q}) = L_{\mathcal{A}}$. According to Remark 2, there exist functions p and q such that

$$\mathcal{P}e^{i\lambda t} = p(\lambda)e^{i\lambda t}, \qquad \mathcal{Q}e^{i\lambda t} = q(\lambda)e^{i\lambda t}. \tag{4.6}$$

PROPOSITION 4.2. *Let* $\lambda \in \mathbb{R} \setminus \{0\}$ *and* $f, g \in \mathbb{C}$ *such that*

$$q(\lambda)f + i\lambda p(\lambda)g = 0. \tag{4.7}$$

If the tree $\mathcal{A}$ *satisfies the property*

$$|q_{\bar\alpha\circ i}(\lambda)| + |q_{\bar\alpha\circ j}(\lambda)| \neq 0 \quad \text{for any } \bar\alpha \in \mathcal{I}_{\mathcal{M}},\ i, j = 1, \dots, m_{\bar\alpha},\ i \neq j, \tag{4.8}$$

then there exists a unique solution $\bar{\theta}$ *of* (N_E) *corresponding to the value* $\mu = \lambda^2$ *such that*

$$\theta(0) = g \quad \textit{and} \quad \theta_x(0) = f. \tag{4.9}$$

Proof. First we construct the component θ of $\bar{\theta}$ (the one corresponding to the string $\mathbf{e}$). We set

$$\theta(x) = g\cos\lambda x + \frac{f}{\lambda}\sin\lambda x, \tag{4.10}$$

which clearly satisfies (4.9).

If the network consists of a single string of length ℓ, then $p(\lambda) = \cos\lambda\ell$, $q(\lambda) = i\sin\lambda\ell$, and condition (4.7) becomes $if\sin\lambda\ell + ig\lambda\cos\lambda\ell = 0$. This implies that $\theta(\ell) = g\cos\lambda\ell + \frac{f}{\lambda}\sin\lambda\ell = 0$, which means that θ is a solution of (N_E), and so the assertion is true in this case.

In the general case the remaining components of $\bar{\theta}$ are constructed by induction. Assume that the proposition is true for the subtrees $\mathcal{A}_1, \dots, \mathcal{A}_m$.

If we were able to choose numbers $f_1, \dots, f_m$ verifying

$$\sum_{k=1}^{m} f_k = \theta_x(\ell) \ \text{ and } \ q_k(\lambda)f_k + i\lambda p_k(\lambda)\theta(\ell) = 0 \quad \text{for } k = 1, \dots, m, \tag{4.11}$$

then, according to the induction assumption, we could find solutions $\bar{\theta}^1, \dots, \bar{\theta}^m$, defined on the subtrees $\mathcal{A}_1, \dots, \mathcal{A}_m$, respectively, such that

$$\theta^k(0) = \theta(\ell), \quad \theta^k_x(0) = f_k \text{ for } k = 1, \dots, m.$$

This would imply that

$$\sum_{k=1}^{m} \theta^k_x(0) = \theta(\ell) \ \text{ and } \quad \theta^k(0) = \theta(\ell) \ \text{ for } k = 1, \dots, m.$$

Therefore, the function $\bar{\theta}$ defined on the tree by $\theta_{k\circ\bar{\alpha}} = \theta^k_{\bar{\alpha}}$ would be the solution of (N_E), whose existence is asserted in the proposition. Consequently, it remains to prove the possibility of the decomposition (4.11).

We remark that from the definition of p and q and formulas (2.12), (2.13) it follows that

$$p = \cos\lambda\ell \sum_{k=1}^{m} p_k \prod_{j\neq k} q_j + i\sin\lambda\ell \prod_{j=1}^{m} q_j, \tag{4.12}$$

$$q = i\sin\lambda\ell \sum_{k=1}^{m} p_k \prod_{j\neq k} q_j + \cos\lambda\ell \prod_{j=1}^{m} q_j. \tag{4.13}$$

Note that condition (4.8) implies that among the numbers $q_k(\lambda)$, $k = 1, \dots, m$, at most one may be equal to zero. Thus, we consider two cases: (a) all the numbers $q_k(\lambda)$, $k = 1, \dots, m$, are different from zero and (b) exactly one of those numbers, say, $q_1(\lambda)$, is equal to zero.

Case (a). If we take

$$f_k = \frac{-i\lambda p_k(\lambda)\theta(\ell)}{q_k(\lambda)},$$

then

$$\sum_{k=1}^{m} f_k = -i\lambda\theta(\ell)\sum_{k=1}^{m}\frac{p_k}{q_k} = -i\lambda\left(g\cos\lambda\ell + \frac{f}{\lambda}\sin\lambda\ell\right)\frac{\sum_{k=1}^{m} p_k\prod_{j\neq k} q_j}{\prod_{j=1}^{m} q_j}.$$

This equality, taking into account (4.12) and (4.13), gives

$$\begin{aligned}\sum_{k=1}^{m} f_k &= -i\lambda g\left(\frac{p}{\prod_{j=1}^{m} q_j} - i\sin\lambda\ell\right) - f\left(\frac{q}{\prod_{j=1}^{m} q_j} - \cos\lambda\ell\right)\\ &= -\lambda g\sin\lambda\ell + f\cos\lambda\ell - \frac{i\lambda pg + qf}{\prod_{j=1}^{m} q_j} = -\lambda g\sin\lambda\ell + f\cos\lambda\ell = \theta_x(\ell).\end{aligned}$$

Thus, the numbers $f_1,\dots,f_m$ satisfy (4.11).

Case (b). The relations (4.12), (4.13) together with $q_1(\lambda) = 0$ give

$$p(\lambda) = \cos\lambda\ell p_1(\lambda)\prod_{j\neq 1} q_j(\lambda), \qquad q(\lambda) = i\sin\lambda\ell p_1(\lambda)\prod_{j\neq 1} q_j(\lambda),$$

and from (4.7) we obtain

$$0 = q(\lambda)f + i\lambda p(\lambda)g = i\lambda\left(g\cos\lambda\ell + \frac{g}{\lambda}\sin\lambda\ell\right)p_1(\lambda)\prod_{j\neq 1} q_j(\lambda) = i\lambda\theta(\ell)p_1(\lambda)\prod_{j\neq 1} q_j(\lambda).$$

However, $\prod_{j\neq 1} q_j(\lambda) \neq 0$ and then, necessarily, $\theta(\ell)p_1(\lambda) = 0$. This means that if we choose $f_1 = \theta_x(\ell)$ and $f_2,\dots,f_m$ verifying

$$\sum_{k=2}^{m} f_k = 0 \ \text{ and } \ q_k(\lambda)f_k + i\lambda p_k(\lambda)\theta(\ell) = 0 \quad \text{for } k = 2,\dots,m,$$

as in the previous case, then the condition (4.11) is satisfied.

So far, we have proved the existence of a solution. It turns out that for the solutions satisfying (4.2) we can give an explicit formula. Indeed, if we apply Propositions 2.8 and 2.9 to the solution $\bar\theta(t,x) = e^{i\lambda t}\bar\theta(x)$ of (N), we obtain

$$i\lambda p(\lambda)\theta^{\bar\alpha}(0) = \hat k(\lambda)\theta_x(0) = \hat k(\lambda)f, \qquad p(\lambda)\theta_x^{\bar\alpha}(0) = -l(\lambda)\theta_x(0) = -l(\lambda)f, \tag{4.14}$$

$$q(\lambda)\theta^{\bar\alpha}(0) = k(\lambda)\theta(0) = k(\lambda)g, \qquad q(\lambda)\theta_x^{\bar\alpha}(0) = i\lambda l(\lambda)\theta(0) = i\lambda l(\lambda)g, \tag{4.15}$$

where k, $\hat k$, l, and r are the functions associated with the operators $\mathcal{K}$, $\hat{\mathcal{K}}$, $\mathcal{L}$, and $\mathcal{R}$, respectively, according to Remark 2.

On the other hand, the condition (4.8) implies that at least one of the numbers $p(\lambda)$ or $q(\lambda)$ is different from zero (see Proposition 4.4 below). Therefore, one of the equalities (4.14), (4.15) provides us with an explicit formula for the values of $\theta^{\bar\alpha}(0)$ and $\theta_x^{\bar\alpha}(0)$ for any $\bar\alpha \in \mathcal{I}_{\mathcal{M}}$ and thus for the solution $\bar\theta$. In particular, if $f = g = 0$, the corresponding solution vanishes identically on $\mathcal{A}$, which clearly implies the uniqueness of the solution for arbitrary values of f and g. □

Remark 6. The converse assertion is also true, even if the condition (4.8) is not fulfilled. Indeed, if $\bar\theta$ is a solution of ($\mathrm{N_E}$), then $\bar\theta(t,x) = e^{i\lambda t}\bar\theta(x)$ is a solution of (N) and $\theta_t(t,0) = i\lambda e^{i\lambda t}\theta(0)$, $\theta_x(t,0) = e^{i\lambda t}\theta_x(0)$. Then, from the relations (2.1) and (4.6) it follows that

$$0 = \mathcal{P}\theta_t(t,0) + \mathcal{Q}\theta_x(t,0) = (ip\lambda\theta(0) + q\theta_x(0))e^{i\lambda t}$$

for every $t \in \mathbb{R}$. Thus, (4.7) holds.

Now we are ready to prove the following basic property.

PROPOSITION 4.3. *Let $0 \neq \lambda \in \mathbb{R}$. Then λ^2 is an eigenvalue of $\mathcal{A}$ if and only if $q(\lambda) = 0$.*

Proof. First we prove that $q(\lambda) = 0$ implies that λ^2 is an eigenvalue, i.e., that there exists a nonzero solution of (4.1)–(4.5) for that value of λ. If the tree verifies (4.8), then this fact follows immediately from Proposition 4.2, choosing $g = 0$, $f \neq 0$. Note that the condition $0 \neq f = \theta_x(0)$ guarantees that $\bar\theta$ is not identically equal to zero. In particular, the assertion is true for a string, as it always verifies (4.8).

In the general case when the condition (4.8) may fail, we follow an induction argument: we suppose that the assertion has been proved for all the subtrees $\mathcal{A}_{\bar\alpha}$ with nonempty $\bar\alpha$.

If $q_{\bar\alpha\circ i}(\lambda) = q_{\bar\alpha\circ j}(\lambda) = 0$ for some $\bar\alpha \in \mathcal{I}_{\mathcal{M}}$, $i \neq j$, then, according to the induction hypothesis, λ^2 is an eigenvalue of both $\mathcal{A}_{\bar\alpha\circ i}$ and $\mathcal{A}_{\bar\alpha\circ j}$. Then from Proposition 4.1 it follows that λ^2 is also an eigenvalue of $\mathcal{A}$.

Let us see now the converse assertion. Let $\bar\theta$ be a nonzero eigenfunction corresponding to the eigenvalue λ^2. Then the function $\bar u(t,x) = e^{i\lambda t}\bar\theta(x)$ is a solution of (N). Choose $\bar\alpha \in \mathcal{I}$ such that one of the numbers $\theta^{\bar\alpha}(0)$ or $\theta^{\bar\alpha}_x(0)$ is different from zero (that is possible since, otherwise, it would be $\bar\theta \equiv 0$). For this solution of (N) we have for every $\bar\alpha \in \mathcal{I}$

$$F_{\bar\alpha}(t) = e^{i\lambda t}\theta^{\bar\alpha}_x(0), \qquad G_{\bar\alpha}(t) = i\lambda e^{i\lambda t}\theta^{\bar\alpha}(0),$$

and in particular, $G \equiv 0$. Then, from the Propositions 2.8 and 2.9 it follows that

$$0 = \mathcal{L}_{\bar\alpha}G = \mathcal{Q}F_{\bar\alpha} = \mathcal{Q}e^{i\lambda t}\theta^{\bar\alpha}_x(0) = q(\lambda)\theta^{\bar\alpha}_x(0),$$

$$0 = \mathcal{K}_{\bar\alpha}G = \mathcal{Q}G_{\bar\alpha} = \mathcal{Q}e^{i\lambda t}\theta^{\bar\alpha}(0) = i\lambda q(\lambda)\theta^{\bar\alpha}(0),$$

and therefore, necessarily, $q(\lambda) = 0$. ∎

4.2. Further properties of p and q. Now we give some additional properties of the functions p and q, which are used in the next section.

PROPOSITION 4.4. *For every tree $\mathcal{A}$ the following properties hold:*

(i) *one of the functions p, q is even and the other is odd;*

(ii) *there exists $\lambda_0 \in \mathbb{R}$ such that $p(\lambda_0) = q(\lambda_0) = 0$ if and only if there exist two subtrees $\mathcal{A}_{\bar\alpha\circ i}$, $\mathcal{A}_{\bar\alpha\circ j}$, $i \neq j$, with common root $\mathcal{O}_{\bar\alpha}$ such that $q_{\bar\alpha\circ i}(\lambda_0) = q_{\bar\alpha\circ j}(\lambda_0) = 0$.*

Proof. We proceed by induction. For a single string

$$p(\lambda) = \cos\lambda\ell, \quad q(\lambda) = i\sin\lambda\ell.$$

In this case (i) is trivial. Assertion (ii) follows from the fact that $|p|^2 + |q|^2 = 1$.

Suppose now that (i), (ii) are true for the subtrees $\mathcal{A}_1, \dots, \mathcal{A}_m$. Let h be a function, which is either even or odd. Denote

$$\rho(h) = \begin{cases} 1 & \text{if } h \text{ is even,} \\ -1 & \text{if } h \text{ is odd.} \end{cases}$$

The function ρ is multiplicative: $\rho(h_1h_2) = \rho(h_1)\rho(h_2)$.

According to the definitions of p and q and the formulas (2.12), (2.13) we have that

$$q(\lambda) = i \sin \lambda\ell \sum_{i=1}^{m} p_i(\lambda) \prod_{j \neq i} q_j(\lambda) + \cos \lambda\ell \prod_{i=1}^{m} q_i(\lambda), \tag{4.16}$$

$$p(\lambda) = \cos \lambda\ell \sum_{i=1}^{m} p_i(\lambda) \prod_{j \neq i} q_j(\lambda) + i \sin \lambda\ell \prod_{i=1}^{m} q_i(\lambda). \tag{4.17}$$

The hypotheses with respect to the subtrees imply that $\rho(p_i) = -\rho(q_i)$, $i = 1, \ldots, m$. Then,

$$\rho\left(i \sin \lambda\ell \; p_i(\lambda) \prod_{j \neq i} q_j\right) = \prod_{i=1}^{m} \rho(q_i), \qquad \rho\left(\cos \lambda\ell \prod_{i=1}^{m} q_i\right) = \prod_{i=1}^{m} \rho(q_i),$$

and consequently, $\rho(q) = \prod_{i=1}^{m} \rho(q_i)$. In an analogous way it is proved that $\rho(p) = -\prod_{i=1}^{m} \rho(q_i)$, and then $\rho(p) = -\rho(q)$. This proves the property (i).

We now prove (ii). If $p(\lambda_0) = q(\lambda_0) = 0$, then from (4.16), (4.17) it follows that

$$0 = q(\lambda_0) = i \sin \lambda_0\ell \sum_{i=1}^{m} p_i(\lambda_0) \prod_{j \neq i} q_j(\lambda_0) + \cos \lambda_0\ell \prod_{i=1}^{m} q_i(\lambda_0),$$

$$0 = p(\lambda_0) = \cos \lambda_0\ell \sum_{i=1}^{m} p_i(\lambda_0) \prod_{j \neq i} q_j(\lambda_0) + i \sin \lambda_0\ell \prod_{i=1}^{m} q_i(\lambda_0).$$

This implies that

$$\sum_{i=1}^{m} p_i(\lambda_0) \prod_{j \neq i} q_j(\lambda_0) = 0, \tag{4.18}$$

$$\prod_{i=1}^{m} q_i(\lambda_0) = 0. \tag{4.19}$$

These equalities are verified if and only if for some i_0

$$q_{i_0}(\lambda_0) = 0, \qquad p_{i_0}(\lambda_0) \prod_{j \neq i_0} q_j(\lambda_0) = 0,$$

and this is equivalent to the fact that one of the following assertions is true:

(a) there exists $i_1 \neq i_0$ such that $q_{i_1}(\lambda_0) = 0$;

(b) $p_{i_0}(\lambda_0) = 0$.

In the first case assertion (ii) follows immediately. In (b), according to the induction assumption, there exist subtrees of $\mathcal{A}_{i_0}$, and consequently also of $\mathcal{A}$, that verify condition (ii). □

With the aid of the previous proposition it is possible to calculate how the operator $\mathcal{Q}$ acts on the functions $\sin \lambda t$ and $\cos \lambda t$.

COROLLARY 4.5. *The following equalities are verified:*

$$\mathcal{Q}\sin\lambda t = \begin{cases} q(\lambda)\sin\lambda t & \text{if } q \text{ is even,} \\ -iq(\lambda)\cos\lambda t & \text{if } q \text{ is odd,} \end{cases}$$

$$\mathcal{Q}\cos\lambda t = \begin{cases} q(\lambda)\cos\lambda t & \text{if } q \text{ is even,} \\ iq(\lambda)\sin\lambda t & \text{if } q \text{ is odd.} \end{cases}$$

Remark 7. As a consequence of the previous formulas, when q is an even function, then it is real valued, while when it is odd, then iq is real valued.

5. Observability results. In this section we express the inequalities from Theorem 3.2 in terms of the initial data of the solution $\bar{u}$. This allows us to obtain weighted observability inequalities, with explicit weights on the Fourier coefficient of the initial data of the solution. Further, we study under what conditions those weights are different from zero.

5.1. Weighted observability inequalities. As stated above, a solution $\bar{u}$ of (1.1)–(1.5) is expressed in terms of the initial data $\bar{u}_0$, $\bar{u}_1$ by the formula

$$\bar{u}(t) = \sum_{k\in\mathbf{Z}_+} \left(u_{0,k}\cos\lambda_k t + \frac{u_{1,k}}{\lambda_k}\sin\lambda_k t \right)\bar{\theta}_k, \tag{5.1}$$

where $\{u_{0,k}\}$, $\{u_{1,k}\}$ are the sequences of Fourier coefficients of $\bar{u}_0$, $\bar{u}_1$ with respect to the orthonormal basis of eigenfunctions $\{\bar{\theta}_k\}_{k\in\mathbb{Z}_+}$ and $\lambda_k = \sqrt{\mu_k}$.

Furthermore, the energy of the solution $\bar{u}$ is given by

$$\mathbf{E}_{\bar{u}} = \frac{1}{2}\sum_{k\in\mathbf{Z}_+} \left(\lambda_k^2 u_{0,k}^2 + u_{1,k}^2\right). \tag{5.2}$$

The operators $\mathcal{D}_{\bar{\alpha}}$ defined in section 3 are of type S. Then, according to Remark 2, there exist functions $d_{\bar{\alpha}}$ such that

$$\mathcal{D}_{\bar{\alpha}} e^{i\lambda t} = d_{\bar{\alpha}}(\lambda) e^{i\lambda t}.$$

In particular, when $\bar{\alpha}$ is the empty index, we have $d(\lambda) \equiv 1$.

These functions, in view of (3.1), are expressed as

$$d_{\bar{\alpha}} := \left(\prod_{i=1,\, i\neq\alpha_1}^{m} q_i\right)\left(\prod_{i=1,\, i\neq\alpha_2}^{m_{\alpha_1}} q_{\alpha_1,i}\right)\cdots\left(\prod_{i=1,\, i\neq\alpha_{k-1}}^{m_{\alpha_1,\ldots,\alpha_{k-1}}} q_{\alpha_1,\ldots,\alpha_{k-1},i}\right), \tag{5.3}$$

and then Proposition 4.4 ensures that, for every $\bar{\alpha}\in\mathcal{I}$, $d_{\bar{\alpha}}$ is an even or odd function. Moreover, from Corollary 4.5 we have the equalities

$$\begin{aligned} \mathcal{D}_{\bar{\alpha}}\sin\lambda t &= \begin{cases} d_{\bar{\alpha}}(\lambda)\sin\lambda t & \text{for } d_{\bar{\alpha}} \text{ even,} \\ -id_{\bar{\alpha}}(\lambda)\cos\lambda t & \text{for } d_{\bar{\alpha}} \text{ odd,} \end{cases} \\ \mathcal{D}_{\bar{\alpha}}\cos\lambda t &= \begin{cases} d_{\bar{\alpha}}(\lambda)\cos\lambda t & \text{for } d_{\bar{\alpha}} \text{ even,} \\ id_{\bar{\alpha}}(\lambda)\sin\lambda t & \text{for } d_{\bar{\alpha}} \text{ odd.} \end{cases} \end{aligned} \tag{5.4}$$

Now fix $\bar{\alpha} \in \mathcal{I}_M$ and denote $\bar{\omega} = \mathcal{D}_{\bar{\alpha}}\bar{u}$. The function $\bar{\omega}$ is also a solution of (1.1)–(1.5) and, from (5.1),

$$\bar{\omega}(t) = \mathcal{D}_{\bar{\alpha}}\bar{u}(t) = \sum_{k\in\mathbf{Z}_+} \left(u_{0,k}\, \mathcal{D}_{\bar{\alpha}} \cos\lambda_k t + \frac{u_{1,k}}{\lambda_k}\, \mathcal{D}_{\bar{\alpha}} \sin\lambda_k t \right) \bar{\theta}_k.$$

Then, from (5.4) it follows that

$$\bar{\omega}(t) = \sum_{k\in\mathbf{Z}_+} d_{\bar{\alpha}}(\lambda_k) \left(u_{0,k} \cos\lambda_k t + \frac{u_{1,k}}{\lambda_k} \sin\lambda_k t \right) \bar{\theta}_k \qquad \text{if } d_{\bar{\alpha}} \text{ is even,}$$

$$\bar{\omega}(t) = \sum_{k\in\mathbf{Z}_+} i d_{\bar{\alpha}}(\lambda_k) \left(u_{0,k} \sin\lambda_k t - \frac{u_{1,k}}{\lambda_k} \cos\lambda_k t \right) \bar{\theta}_k \qquad \text{if } d_{\bar{\alpha}} \text{ is odd.}$$

Thus, in both cases, the energy of $\bar{\omega}$ computed by the formula (5.2) is given by

$$\mathbf{E}_{\bar{\omega}} = \frac{1}{2} \sum_{k\in\mathbf{Z}_+} |d_{\bar{\alpha}}(\lambda_k)|^2 \left(\lambda_k^2 u_{0,k}^2 + u_{1,k}^2 \right). \tag{5.5}$$

With this, the inequality of Theorem 3.2 may be written in terms of the initial data of the solution $\bar{u}$ as

$$\sum_{k\in\mathbf{Z}_+} |d_{\bar{\alpha}}(\lambda_k)|^2 \left(\lambda_k^2 u_{0,k}^2 + u_{1,k}^2 \right) \le C \int_0^{2T_{\mathcal{A}}} |F(t)|^2 dt = C \int_0^{2T_{\mathcal{A}}} |u_x(t,0)|^2 dt. \tag{5.6}$$

Consequently, if we define

$$c_k = \max_{\bar{\alpha}\in\mathcal{I}_{\mathcal{S}}} |d_{\bar{\alpha}}(\lambda_k)|, \tag{5.7}$$

we obtain the following result.

THEOREM 5.1. *There exists a positive constant C such that*

$$\sum_{k\in\mathbf{Z}_+} c_k^2 \left(\lambda_k^2 u_{0,k}^2 + u_{1,k}^2 \right) \le C \int_0^{2T_{\mathcal{A}}} |u_x(t,0)|^2 dt, \tag{5.8}$$

for every solution $\bar{u}$ with initial data $(\bar{u}_0,\ \bar{u}_1) \in V \times H$.

Remark 8. It is easy to prove, using, e.g., formula (5.1) for the solutions, that if inequality (5.8) holds, then for every $\alpha, T \in \mathbb{R}$,

$$\sum_{k\in\mathbf{Z}_+} c_k^2 \left(\lambda_k^2 u_{0,k}^2(T) + u_{1,k}^2(T) \right) \le C \int_\alpha^{\alpha+2T_{\mathcal{A}}} |u_x(t,0)|^2 dt, \tag{5.9}$$

where $u_{0,k}(T)$ and $u_{1,k}(T)$ are the Fourier coefficients of $\bar{u}|_{t=T}$ and $\bar{u}_t|_{t=T}$, respectively, in the basis $\{\bar{\theta}_k\}_{k\in\mathbb{Z}_+}$.

5.2. Nondegenerate trees. In general, some of the coefficients c_k in the inequality (5.8) may vanish. That is why we consider a special class of trees for which all those numbers are different from zero.

DEFINITION 5.2. *A tree $\mathcal{A}$ is said to be nondegenerate if the numbers c_k, defined for that tree by* (5.7)*, are different from zero for every $k \in \mathbb{Z}_+$. Otherwise, the tree is said to be degenerate.*

The following proposition provides us with a more transparent characterization of nondegenerate trees.

PROPOSITION 5.3. *The tree $\mathcal{A}$ is nondegenerate if and only if the spectra $\sigma_{\bar{\alpha}\circ i}$, $\sigma_{\bar{\alpha}\circ j}$ of any two subtrees $\mathcal{A}_{\bar{\alpha}\circ i}$, $\mathcal{A}_{\bar{\alpha}\circ j}$ of $\mathcal{A}$ with common $\mathcal{O}_{\bar{\alpha}}$ root are disjoint.*

Proof. Note that the following general fact is true: an inequality like (5.8) with different-from-zero coefficients c_k (not necessarily given by (5.7)) is impossible for a tree having two subtrees with common root that share an eigenvalue μ. Indeed, in such a case, with the help of Proposition 4.1 we can construct a nonzero solution $\bar{u}$ of (1.1)–(1.5) such that $u_x(t,0) \equiv 0$. With this, a (5.8)-like inequality would give

$$\sum_{k\in\mathbf{Z}_+} c_k^2 \left(\lambda_k^2 u_{0,k}^2 + u_{1,k}^2\right) \le 0,$$

which is false, since $\bar{u}$ is not identically equal to zero.

For the converse assertion we argue by contradiction. We will prove that if $c_k = 0$ for some $k \in \mathbb{Z}_+$ and any two subtrees of $\mathcal{A}$ with common root have disjoint spectra, then $d_{\bar{\alpha}}(\lambda_k) = 0$ for any $\bar{\alpha} \in \mathcal{I}$. In particular, $d(\lambda_k) = 0$, which would contradict the fact that $d(\lambda_k) = 1$.

Note first that the property is immediate for exterior nodes, since $c_k \ge |d_{\bar{\alpha}}(\lambda_k)|$ for $\bar{\alpha} \in \mathcal{I}_{\mathcal{S}}$.

For the interior nodes we follow a recursive argument: if $\bar{\alpha} \in \mathcal{I}_{\mathcal{M}}$ and $d_{\bar{\alpha}\circ\beta}(\lambda_k) = 0$ for all $\beta = 1, \dots, m_{\bar{\alpha}}$, then $d_{\bar{\alpha}}(\lambda_k) = 0$.

Indeed, we have that, for every $\beta = 1, \dots, m_{\bar{\alpha}}$,

$$d_{\bar{\alpha}\circ\beta} = d_{\bar{\alpha}} \prod_{i\neq\beta} q_{\bar{\alpha}\circ i}. \tag{5.10}$$

Assume that $d_{\bar{\alpha}} \neq 0$. Then (5.10) implies that $\prod_{i\neq 1} q_{\bar{\alpha}\circ i} = 0$, and thus there exists $i^* \neq 1$ such that

$$q_{\bar{\alpha}\circ i^*} = 0. \tag{5.11}$$

But then, from the equalities $d_{\bar{\alpha}\circ i^*} = 0$ and (5.10) it follows that there exists $j^* \neq i^*$ satisfying

$$q_{\bar{\alpha}\circ j^*} = 0. \tag{5.12}$$

However, the equalities (5.11) and (5.12) ensure, according to Proposition 4.3, that $\mu_k = \lambda_k^2$ is a common eigenvalue of the subtrees $\mathcal{A}_{\bar{\alpha}\circ i^*}$ and $\mathcal{A}_{\bar{\alpha}\circ j^*}$. But that is impossible for the tree $\mathcal{A}$. Thus, $d_{\bar{\alpha}}(\mu_k) = 0$. This completes the proof of the proposition. □

Remark 9. According to the previous proposition, if the spectra of some two subtrees of $\mathcal{A}$ have a nonvoid intersection, inequality (5.8) degenerates. However, as indicated in the proof, this fact is not due to the technique used to obtain the inequality, since for degenerate trees no (5.8)-like inequality holds with all the coefficients c_k being different from zero. Thus, Theorem 5.1 is sharp in the following sense: *it provides inequality* (5.8) *whenever such an inequality exists.*

COROLLARY 5.4 (unique continuation property). *If the tree $\mathcal{A}$ is nondegenerate and $\bar{u}$ is a solution of (1.1)–(1.5) such that $u_x(t,0) = 0$ for $t \in [0, 2L_{\mathcal{A}}]$, then $\bar{u} \equiv 0$.*

Remark 10. Combining Propositions 4.4(ii) and 5.3, we obtain an alternative characterization of the nondegenerate trees: $\mathcal{A}$ is nondegenerate if and only if $|p(\lambda)|^2 + |q(\lambda)|^2 \neq 0$ for every $\lambda \in \mathbb{R}$.

PROPOSITION 5.5. *If the tree $\mathcal{A}$ is nondegenerate, then all its eigenvalues are simple.*

Proof. If λ_k^2 is an eigenvalue of a nondegenerate tree, then, according to Proposition 4.3 and Remark 10, $q(\lambda_k) = 0$ and $p(\lambda_k) \neq 0$. Consequently, if $\bar{\theta_k}$ is an eigenfunction of $\mathcal{A}$ corresponding to λ_k^2, formula (4.14) gives

$$\theta^{\bar{\alpha}}(0) = \frac{\hat{k}(\lambda_k)}{i\lambda_k p(\lambda_k)}\theta_x(0), \qquad \theta_x^{\bar{\alpha}}(0) = \frac{-l(\lambda_k)}{p(\lambda_k)}\theta_x(0),$$

Thus, $\bar{\theta_k}$ is determined, up to the constant factor $\theta_x(0)$, in a unique way. □

Remark 11. Let $\{\tilde{\mu}_k\}_{n\in\mathbb{Z}_+}$ be the strictly increasing sequence of the eigenvalues μ_k of a tree without taking into account their multiplicity. It may be shown that $\tilde{\mu}_k$ verifies $\mu_k^N \leq \tilde{\mu}_k \leq \mu_k^D$, for $k \in \mathbb{Z}_+$, where $\{\mu_k^D\}_{k\in\mathbb{Z}_+}$ and $\{\mu_k^N\}_{k\in\mathbb{Z}_+}$ are the ordered sequences formed by the distinct eigenvalues of the strings with Dirichlet or Neumann homogeneous boundary conditions, respectively (see [11]). This fact allows us to prove that the complete radius of the sequence $\lambda_k = \sqrt{\mu_k}$ is equal to $L_{\mathcal{A}}$ and therefore that a (5.8)-like inequality is impossible for $T < 2L_{\mathcal{A}}$.

Besides, it may be also proved that, if the network contains more than one string and (5.8) holds, then necessarily $\liminf c_k = 0$. This fact implies that the whole space $H \times V'$ is not controllable; i.e., there exist initial states in $H \times V'$, which cannot be driven to rest under the action of L^2-controls.

5.3. On the size of the set of nondegenerate trees. Now we give some information on the size of the set of degenerate trees.

We shall say that two trees are topologically equivalent if their edges can be numbered with the same set of multi-indices. This means that they may differ only in the lengths of their edges. In particular, two equivalent trees have the same number of edges and vertices. The classes of topologically equivalent trees are called topological configurations.

Fix a topological configuration Σ with d edges. We assume that in the set of indices $\mathcal{I}$ for the elements of the trees belonging to Σ a criterion of ordering has been defined, and we use the notation $\langle\mathcal{A}\rangle$ for the corresponding ordered set of the lengths of the edges of $\mathcal{A} \in \Sigma$.

Then Σ may be identified with $(\mathbb{R}_+)^d$ by means of the mapping $\pi : \Sigma \to \mathbb{R}^d$ defined by

$$\pi(\mathcal{A}) = \langle\mathcal{A}\rangle \in \mathbb{R}^d.$$

Let μ_Σ be the measure induced in Σ by the Lebesgue measure of $\mathbb{R}^d$ through the mapping π. That is, if $B \subset \Sigma$, then $\mu_\Sigma(B) = m_d(\pi(B))$, where m_d is the usual Lebesgue measure in $\mathbb{R}^d$.

We now have the following result.

PROPOSITION 5.6. *Given a topological configuration Σ, almost every tree (in the sense of the measure μ_Σ) with that topological configuration is nondegenerate.*

Proof. Let $D_{\bar{\alpha}}^{i,j} \subset \Sigma$ denote the set of those trees $\mathcal{A}$ such that the subtrees $\mathcal{A}_{\bar{\alpha}\circ i}$ and $\mathcal{A}_{\bar{\alpha}\circ j}$ are nondegenerate and have a common eigenvalue. Then the set $\Sigma_{deg} \subset \Sigma$

of degenerate trees may be decomposed as

$$\Sigma_{deg} = \bigcup_{\bar\alpha\in\mathcal{I}_{\mathcal{M}}} \bigcup_{i,j=1\ i\neq j}^{m_{\bar\alpha}} D_{\bar\alpha}^{i,j}. \tag{5.13}$$

We will prove that $\mu_\Sigma(D_{\bar\alpha}^{i,j}) = 0$ for every $\bar\alpha \in \mathcal{I}_{\mathcal{M}}$, $i, j = 1, \dots, m_{\bar\alpha}$, $i \neq j$. This fact, in view of (5.13), will imply $\mu_\Sigma(\Sigma_{deg}) = 0$. In what follows we consider that $\bar\alpha$, i, and j are fixed.

The idea of the proof is simple. We fix a tree $\mathcal{B}$ having the structure of $\mathcal{A}_{\bar\alpha\circ i} \vee \mathcal{A}_{\bar\alpha\circ j}$ (defined as in Remark 5) and extend it to a tree $\mathcal{A} \in D_{\bar\alpha}^{i,j}$. According to Remark 5, that is equivalent to choosing the node $\mathcal{O}_{\bar\alpha}$ of $\mathcal{A} \in \Sigma$ in a point of a string of $\mathcal{B}$ (precisely, of that string where it should be located to agree with the structure of Σ) where some eigenfunction of $\mathcal{B}$ vanishes. Once $\mathcal{O}_{\bar\alpha}$ has been chosen, the lengths of the remaining strings of $\mathcal{A}$ may be taken arbitrarily.

Observe that we may assume that no (nonidentically zero) eigenfunction of $\mathcal{B}$ vanishes identically on the string that contains $\mathcal{O}_{\bar\alpha}$, since, otherwise, one of the subtrees of $\mathcal{A}_{\bar\alpha\circ i}$ or $\mathcal{A}_{\bar\alpha\circ j}$ of the tree $\mathcal{A}$ obtained with this procedure would be degenerate and thus $\mathcal{A} \notin D_{\bar\alpha}^{i,j}$. This assumption implies that all the eigenfunctions of $\mathcal{B}$ are simple, and then the node $\mathcal{O}_{\bar\alpha}$ should be chosen in a set of points, which is at most denumerable.

Thus, we have obtained, after some reordering if needed, that the set $\pi(D_{\bar\alpha}^{i,j})$ is contained in a set of the form

$$\left\{(h_1, h_2, \dots, h_d) \in (\mathbb{R}_+)^d :\ h_1 + h_2 = h,\ h_1 \in \mathbf{N}(h, h_3, \dots, h_d)\right\}, \tag{5.14}$$

where $\mathbf{N}(h, h_3, \dots, h_d)$ is a denumerable set depending on h and $h_3, \dots, h_d$.

It is easy to see, using, e.g., Fubini's theorem, that a set defined by (5.14) has d-dimensional Lebesgue measure equal to zero. Thus, the same is true of $\pi(D_{\bar\alpha}^{i,j})$, and then $\mu_\Sigma(D_{\bar\alpha}^{i,j}) = 0$. This completes the proof. □

COROLLARY 5.7. *The set $\Sigma \setminus \Sigma_{deg}$ of nondegenerate trees is dense in Σ provided with the metrics induced in Σ by the usual metrics of $\mathbb{R}^d$ through π.*

Remark 12. The set Σ_{deg}, even though is small in the sense of μ_Σ, is dense in Σ. Indeed, it suffices to see that, if two edges of a tree with rationally dependent lengths have a common vertex and their other vertices are exterior, then the tree is degenerate.

6. Consequences concerning controllability. Gathering the facts of the previous sections, we obtain the following characterization of the controllability properties of trees.

THEOREM 6.1. *Let $\mathcal{A}$ be a tree and $T > 0$; then*

(a) *If $T \geq 2L_{\mathcal{A}}$, the properties*

– *$\mathcal{A}$ is spectrally controllable in time T,*

– *$\mathcal{A}$ is approximately controllable in time T,*

– *$\mathcal{A}$ is nondegenerate,*

– *any two subtrees of $\mathcal{A}$ with common root have disjoint spectra*

are equivalent, and when they are true, all the initial data in the space $\mathcal{W}$, defined by (1.9)*, are controllable. Moreover, these properties hold for almost every tree topologically equivalent to $\mathcal{A}$.*

(b) *If $T < 2L_{\mathcal{A}}$, the properties of spectral and approximate controllability are false, regardless of whether $\mathcal{A}$ is degenerate or not.*

7. Simultaneous observability and controllability of networks. The results of the previous sections allow us to consider the one-node control problem for several (a finite number of) tree-shaped networks when the same control function is used to control all of them, i.e., when they are controlled simultaneously.

Let $\mathcal{A}^1, \dots, \mathcal{A}^R$ be the trees associated with the controlled networks. For the elements of the network whose graph is $\mathcal{A}^r$, we will use the same notation as in the preceeding sections but adding the superscript r to them. Thus, the solution of (1.1)–(1.6) for the tree $\mathcal{A}^r$ (in what follows we shall briefly refer to this problem as $(1.1)_r$–$(1.6)_r$) is denoted by $\bar{u}^r$, and the spaces V and H constructed for that tree are denoted by V^r and H^r.

We define the space

$$\mathbf{W} = \prod_{r=1}^{R} V^r \times H^r,$$

endowed with the product Hilbert structure. The elements of $\mathbf{W}$ are called *simultaneous states*.

We shall say that the simultaneous state $\bar{w} \in \mathbf{W}$ is *controllable in time* T if it is possible to find a control function $v \in L^2(0,T)$ such that the solutions $\bar{u}^r$ of $(1.1)_r$–$(1.6)_r$ with initial states $(\bar{u}_0^r, \bar{u}_1^r)$ (the components of $\bar{w}$) and $v^r = v$ verify

$$\bar{u}^r(T,x) = \bar{u}_t^r(T,x) = \bar{0}$$

for every $i = 1, \dots, R$.

Once again using HUM, the problem of characterizing the controllable simultaneous states is reduced to the study of observability inequalities for the corresponding homogeneous systems. Indeed, assume that there exist nonzero numbers c_n^k, $n \in \mathbb{Z}_+$, $k = 1, \dots, R$, such that for every k the inequality

$$\int_0^T \left| \sum_{r=1}^{R} u_x^r(0,t) \right|^2 dt \geq \sum_{n \in \mathbf{Z}_+} (c_n^k)^2 \left(\mu_n^k |u_{0,n}^k|^2 + |u_{1,n}^k|^2 \right) \tag{7.1}$$

holds for all the initial simultaneous states $\bar{w} \in \mathbf{W}$, where $\{u_{0,n}^r\}$ and $\{u_{1,n}^r\}$ are the sequences of Fourier coefficients of the components $\bar{u}_0^r$ and $\bar{u}_1^r$, respectively, of the initial state in the bases $\{\bar{\theta}_n^r\}$ of H^r, and $\bar{u}^r$ is the solution of $(1.1)_r$–$(1.6)_r$ with $v^r = 0$. Define the sets

(7.2)

$$\mathcal{W}^r = \left\{ (\bar{u}_0^r, \bar{u}_1^r) \in V^r \times (H^r)' : \ \|(\bar{u}_0^r, \bar{u}_1^r)\|_r^2 = \sum_{n \in \mathbb{Z}_+} \frac{1}{(c_n^r)^2} \left(|u_{0,n}^r|^2 + \frac{1}{\mu_n^r} |u_{1,n}^r|^2 \right) < \infty \right\}.$$

Then all the initial simultaneous states $\bar{w} \in \mathcal{W} = \prod_{i=1}^{r} \mathcal{W}^r$ are controllable in time T.

In particular, if the inequalities (7.1) hold, then the initial simultaneous states $\bar{w} \in \prod_{1=1}^{r} Z^r \times Z^r$ are controllable. (Recall that Z^r is the set of all finite linear combinations of the eigenfunctions $\bar{\theta}_n^r$.) In this case, the networks are said to be *simultaneously spectrally controllable*.

Moreover, the set of controllable simultaneous states in time T is dense in $\mathbf{W}$ (when that holds, the networks are said to be *simultaneously approximately control-*

lable in time T) if and only if the following unique continuation property holds:

$$\sum_{r=1}^{R} u_x^r(0,t) = 0 \ \text{ in } \ L^2(0,T) \ \text{ implies } \ (\bar{u}_0^r, \bar{u}_1^r) = \bar{0} \ \text{ for every } \ r = 1, \dots, R. \tag{7.3}$$

It is clear that if a simultaneous state is controllable, then each of its components is also controllable for the corresponding network. This implies that if we expect at least the approximate controllability to hold, then we need to assume that all the trees supporting the networks are nondegenerate.

On the other hand, if two of the trees, say $\mathcal{A}^1$ and $\mathcal{A}^2$, have a common eigenvalue, then, using the pasting procedure described in the proof of Proposition 4.1, we can construct nonzero solutions of $(1.1)_r$–$(1.6)_r$, $r = 1, 2$, such that

$$u_x^1(t,0) + u_x^2(t,0) = 0, \qquad t \in \mathbb{R}.$$

Therefore, choosing zero initial states for all the remaining trees $\mathcal{A}^r$, $r = 3, \dots, R$, we obtain a simultaneous initial state in $\mathbf{W}$ for which inequalities (7.1) are impossible and, moreover, for which the unique continuation property (7.3) fails.

Thus, the conditions that the trees $\mathcal{A}^r$, $r = 1, \dots, R$, be nondegenerate and their spectra be pairwise disjoint are necessary for simultaneous approximate controllability, and then for spectral controllability. As we shall see, these conditions are also sufficient.

Set $T^* = \sum_{i=1}^{r} L^i$. For every $k = 1, \dots, R$ we define the operator

$$\widehat{\mathcal{Q}}_k := \prod_{r=1,\ r\neq k}^{R} \mathcal{Q}^r,$$

where $\mathcal{Q}^r$ is the operator $\mathcal{Q}$ for the tree $\mathcal{A}^r$. Note that $\widehat{\mathcal{Q}}_k$ is an S-operator with $s(\widehat{\mathcal{Q}}_k) = T^* - L^k$.

Let $\widehat{q}_k$ be the function associated with $\widehat{\mathcal{Q}}_k$ according to Remark 2. Then

$$\widehat{q}_k = \prod_{r=1,\ r\neq k}^{R} q^r, \tag{7.4}$$

where q^r is the function corresponding to $\mathcal{Q}^r$.

PROPOSITION 7.1. *If for a given k there exist numbers c_n, $n \in \mathbb{Z}_+$, such that every solution of $(1.1)_k$–$(1.6)_k$ with $v_k = 0$ and initial state*

$$(\bar{u}_0^k, \bar{u}_1^k) = \left(\sum_{n\in\mathbb{Z}_+} u_{0,n}^k \bar{\theta}_n^k, \sum_{n\in\mathbb{Z}_+} u_{1,n}^k \bar{\theta}_n^k \right) \in V^k \times H^k$$

satisfies

$$\int_0^{2L_k} |u_x^k(0,t)|^2 dt \geq \sum_{n\in\mathbf{Z}_+} c_n^2 \left(\mu_n^k |u_{0,n}^k|^2 + |u_{1,n}^k|^2 \right), \tag{7.5}$$

then

$$\int_0^{2T^*} \left| \sum_{r=1}^{R} u_x^r(0,t) \right|^2 dt \geq \sum_{n\in\mathbf{Z}_+} c_n^2 |\widehat{q}_k(\lambda_n^k)|^2 \left(\mu_n^k |u_{0,n}^k|^2 + |u_{1,n}^k|^2 \right) \tag{7.6}$$

for every $(\bar{u}_0^r, \bar{u}_1^r) \in V^r \times H^r$, $r = 1, \dots, R$.

Proof. As $\widehat{\mathcal{Q}}_k$ is an S-operator with $s(\widehat{\mathcal{Q}}_k) = T^* - L^k$, using Proposition 2.4(i), we get

$$\int_0^{2T^*} \left| \sum_{r=1}^{R} u_x^i(0,t) \right|^2 dt \geq \int_{T^*-L^k}^{T^*+L^k} \left| \widehat{\mathcal{Q}}_k \sum_{r=1}^{R} u_x^i(0,t) \right|^2 dt. \tag{7.7}$$

However, as $\mathcal{Q}^r u_x^r(0,t) = 0$, then $\widehat{\mathcal{Q}}_k \bar{u}^k$ if $r \neq k$. Thus, inequality (7.7) becomes

$$\int_0^{2T^*} \left| \sum_{r=1}^{R} u_x^i(0,t) \right|^2 dt \geq \int_{T^*-L^k}^{T^*+L^k} |\widehat{\mathcal{Q}}_k u_x^k(0,t)|^2 dt. \tag{7.8}$$

Now we consider the function $\bar{\omega} = \widehat{\mathcal{Q}}_k \bar{u}$. As $\bar{\omega}$ is clearly a solution of $(1.1)_k$–$(1.5)_k$, then, according to (7.5) and Remark 5.9 we have

$$\int_{T^*-L^k}^{T^*+L^k} |\omega_x(0,t)|^2 dt \geq \sum_{n\in\mathbb{Z}_+} c_n^2 \left(\mu_n^k \omega_{0,n}^2 + \omega_{1,n}^2\right). \tag{7.9}$$

On the other hand, it is simple to prove that the Fourier coefficients of the initial data of $\bar{u}$ and $\bar{\omega}$ are related by

$$\mu_n^k \omega_{0,n}^2 + \omega_{1,n}^2 = q_k^2(\lambda_n^k) \left(\mu_n^k u_{0,n}^2 + u_{1,n}^2\right). \tag{7.10}$$

Finally, combining (7.8)–(7.10) and the fact that $\omega_x(0,t) = \widehat{\mathcal{Q}}_k u_x^k(0,t)$, the inequality (7.6) is obtained. □

Now, if the trees $\mathcal{A}^1, \ldots, \mathcal{A}^R$ are nondegenerate, then we have for every $r = 1, \ldots, R$ inequalities (7.5) with nonzero coefficients c_n (depending on r), which are explicitly computed by formulas (5.7). Therefore, according to the previous proposition, we shall also have inequalities (7.1) with explicitly computed coefficients $c_n^r = |\widehat{q}_r(\lambda_n^r)| c_n$, which are all different from zero whenever the spectra of any two of the trees $\mathcal{A}^r$ are disjoint, since $\widehat{q}_r(\lambda_n^r) \neq 0$ for all $r = 1, \ldots, R$ and $n \in \mathbb{Z}_+$. Indeed, if $\widehat{q}_r(\lambda_n^r) = 0$ for some r and n, then equality (7.4) would imply that $q^i(\lambda_n^r) = 0$ for some $i \neq r$, and thus, from Proposition (4.3), μ_n^r would be a common eigenvalue of the trees $\mathcal{A}^r$ and $\mathcal{A}^i$.

Consequently, we are able to construct, under those assumptions, a space $\mathcal{W} = \prod_{r=1}^{R} \mathcal{W}^r$, where $\mathcal{W}^r$ are defined by (7.2), of controllable simultaneous states in time $2T^*$. In particular, we obtain the following.

COROLLARY 7.2. *The trees $\mathcal{A}^1, \ldots, \mathcal{A}^R$ are simultaneously spectrally controllable in some time T (and then in time $2T^*$) if and only if they are spectrally controllable and their spectra are pairwise disjoint.*

Acknowledgment. The author would like to thank E. Zuazua for suggesting that we study this problem, as well as for his valuable help.

REFERENCES

[1] S. A. AVDONIN AND W. MORAN, *Simultaneous control problems for systems of elastic strings and beams*, Systems Control Lett., 44 (2001), pp. 147–155.

[2] S. A. AVDONIN AND M. TUCSNAK, *Simultaneous controllability in sharp time for two elastic strings*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 259–273.

[3] C. BAIOCCHI, V. KOMORNIK, AND P. LORETI, *Ingham-Beurling type theorems with weakened gap conditions*, Acta Math. Hungar., 97 (2002), pp. 55–95.

[4] R. Dáger and E. Zuazua, *Controllability of star-shaped networks of strings*, in Proceedings of the Fifth International Conference on Mathematical and Numerical Aspects of Wave Propagation, Santiago de Compostela, Spain, A. Bermúdez, D. Gómez, C. Hazard, P. Joly, and J. E. Roberts, eds., Proceedings in Appl. Math. 102, SIAM, Philadelphia, PA, 2000, pp. 1006–1010.

[5] R. Dáger and E. Zuazua, *Controllability of star-shaped networks of strings*, C. R. Acad. Sci. Paris Sér. I, 332 (2001), pp. 621–626.

[6] R. Dáger and E. Zuazua, *Controllability of tree-shaped networks of strings*, C. R. Acad. Sci. Paris Sér. I, 332 (2001), pp. 1087–1092.

[7] B. Dekoninck and S. Nicaise, *Control of networks of Euler-Bernoulli beams*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 57–81.

[8] J. E. Lagnese, G. Leugering, and E. J. Schmidt, *Modelling, Analysis and Control of Dynamic Elastic Multi-Link Structures*, Systems Control Found. Appl., Birkhäuser, Basel, 1994.

[9] G. Leugering and E. Zuazua, *On exact controllability of generic trees*, in Proceedings of the meeting Control of Systems Governed by Partial Differential Equations, Nancy, France, 1999, ESAIM Proc. 8, Soc. Math. Appl. Indust., Paris, 2000, pp. 95–105.

[10] J.-L. Lions, *Contrôlabilité Exacte Perturbations et Stabilisation de Systèmes Distribués*, Vol. 1, Masson, Paris, 1988.

[11] S. Nicaise, *Spectre des réseaux topologiques finis*, Bull. Sci. Math., 111 (1987), pp. 401–413.

[12] S. Rolewicz, *On controllability of systems of strings*, Studia Math., 36 (1970), pp. 105–110.

[13] E. J. P. G. Schimidt, *On the modelling and exact controllability of networks of vibrating strings*, SIAM J. Control Optim., 30 (1992), pp. 229–245.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 624–642

NASH EQUILIBRIUM PAYOFFS FOR NONZERO-SUM STOCHASTIC DIFFERENTIAL GAMES*

RAINER BUCKDAHN†, PIERRE CARDALIAGUET†, AND CATHERINE RAINER†

Abstract. Existence and characterization of Nash equilibrium payoffs are proved for stochastic nonzero-sum differential games.

Key words. nonzero-sum differential game, stochastic differential game, Nash equilibrium

AMS subject classifications. 91A23, 91A15

DOI. 10.1137/S0363012902411556

1. Introduction. In this paper, we investigate the notion of Nash equilibrium payoff for nonzero-sum stochastic two-players differential games. Our main result is an existence theorem for such equilibrium payoffs. We also give a characterization of these payoffs.

Let us denote by $X_s^{t,x,u,v}$ the solution of the following equation:

$$dX_s = f(s, X_s, u_s, v_s)ds + \sigma(s, X_s, u_s, v_s)dB_s, \qquad t \le s,$$

with initial condition

$$X_t = x.$$

Here $B_\cdot$ is a d-dimensional standard Brownian motion, and $u_\cdot$ and $v_\cdot$ are stochastic processes taking values in some compact subsets U and V of some finite dimensional spaces. Precise assumptions on $f : [0,T] \times \mathbb{R}^n \times U \times V \to \mathbb{R}^n$ and on $\sigma : [0,T] \times \mathbb{R}^n \times U \times V \to \mathbb{R}^{n\times d}$ are given in the next section.

The payoff of the players is a terminal payoff, given by $J_1(t,x,u,v) = E[g_1(X_T^{t,x,u,v})]$ for Player I and by $J_2(t,x,u,v) = E[g_2(X_T^{t,x,u,v})]$ for Player II. Loosely speaking, Player I aims to maximize $J_1(t,x,u,v)$ while the goal of Player II is to maximize $J_2(t,x,u,v)$. As usual in differential game theory, the players play not time-measurable controls but *strategies*. In order to avoid for the moment the technical details, we postpone the definition of the strategies to the next section. Here we only need to assume that for any strategy α of Player I and any strategy β of Player II one can define a payoff $J_1(t,x,\alpha,\beta)$ for Player I and a payoff $J_2(t,x,\alpha,\beta)$ for Player II.

A particularly important notion for investigating nonzero-sum games is given by Nash equilibria. In our framework, a Nash equilibrium is a pair $(\bar\alpha, \bar\beta)$ of strategies such that, for any other pair (α,β) of strategies, we have

$$J_1(t,x,\bar\alpha,\bar\beta) \ge J_1(t,x,\alpha,\bar\beta) \quad \text{and} \quad J_2(t,x,\bar\alpha,\bar\beta) \ge J_2(t,x,\bar\alpha,\beta). \tag{1.1}$$

The couple $(J_1(t,x,\bar\alpha,\bar\beta), J_2(t,x,\bar\alpha,\bar\beta))$ is called a Nash equilibrium payoff. In general, we do not expect Nash equilibria to exist but only Nash equilibrium payoffs

*Received by the editors July 19, 2002; accepted for publication (in revised form) November 26, 2003; published electronically July 23, 2004.
http://www.siam.org/journals/sicon/43-2/41155.html

†Département de Mathématiques, Université de Bretagne Occidentale, 6, avenue Victor-le-Gorgeu, B.P. 809, 29285 Brest cedex, France (Rainer.Buckdahn@univ-brest.fr, Pierre.Cardaliaguet@univ-brest.fr, Catherine.Rainer@univ-brest.fr).

(e_1, e_2) which can be approximated by the payoffs of strategies $(\bar{\alpha}_\epsilon, \bar{\beta}_\epsilon)$ for which the inequalities (1.1) hold true only up to some $\epsilon > 0$ for any (α, β). (Note also that, in general, Nash equilibrium payoffs are not unique.)

The main result of this paper (Theorem 2.9) states that Nash equilibrium payoffs exist for any initial position. Moreover, we characterize these Nash equilibrium payoffs. To explain this characterization, we have to introduce the zero-sum games associated with the payoffs J_1 and J_2. Under the Isaacs condition (see (2.2)), Fleming and Souganidis [2] (see also [7]) have proved that the zero-sum game, where Player I wants to maximize J_1 and Player II wants to minimize J_1, has a value, which will be denoted here by W_1:

$$W_1(t,x) = \inf_\beta \sup_\alpha J_1(t,x,\alpha,\beta) = \sup_\alpha \inf_\beta J_1(t,x,\alpha,\beta).$$

In the same way, the zero-sum game, in which Player I aims at minimizing the payoff J_2 and Player II aims at maximizing it, has also a value, denoted W_2:

$$W_2(t,x) = \sup_\beta \inf_\alpha J_2(t,x,\alpha,\beta) = \inf_\alpha \sup_\beta J_2(t,x,\alpha,\beta).$$

Our characterization result Theorem 2.10 loosely states (up to technical details) that a pair $(e_1, e_2) \in \mathbb{R}^2$ is a Nash equilibrium payoff for the initial position (t,x) if and only if there is some pair $(u_\cdot, v_\cdot) : [t,T] \to U \times V$ of adapted controls such that

(i) for $j = 1, 2$, $E[g_j(X_T^{t,x,u,v})|\mathcal{F}_{t,s}] \geq W_j(s, X_s^{t,x,u,v})$ a.s. for any $s \in [t,T]$, where $\mathcal{F}_{t,s}$ is the σ-algebra generated by $\{B_u - B_t,\ u \in [t,s]\}$;

(ii) for $j = 1, 2$, $e_j = J_j(t,x,u,v)$.

(In practice, the existence of such $(u_\cdot, v_\cdot)$ is out of reach, and we prove only the existence, for any $\epsilon > 0$, of some adapted controls $(u_\cdot^\epsilon, v_\cdot^\epsilon)$ for which (i) holds true up to ϵ with a probability larger than $1 - \epsilon$ and (ii) holds true up to ϵ. However, this is enough for characterizing the Nash equilibrium payoffs.)

The controls $u_\cdot$ and $v_\cdot$ can be interpreted as follows. The players agree at the beginning of the game to play, respectively, $u_\cdot$ and $v_\cdot$. Condition (ii) then guarantees that their payoff is (e_1, e_2) if they indeed play $u_\cdot$ and $v_\cdot$ up to the terminal time T. If on the contrary one of the players (say, Player II) deviates at some time $t' \in (t,T)$, i.e., does not play $v_\cdot$ on $[t',T]$, then Player I punishes Player II by playing some strategy that minimizes the expected payoff of Player II. Condition (i) guarantees that such a strategy exists and that the resulting payoff of Player II is not larger than e_2. So Player II gains nothing by deviating.

In the deterministic case, the results presented in this paper have already been established by Kononenko in [6] and by Kleimenov in [5] in the framework of positional strategies, and by Tolwinski, Haurie, and Leitmann in [9] in the framework of Friedman strategies. Let us point out that the generalization to the stochastic case is far from being straightforward for at least two reasons: first, because of measurability issues, already encountered by Fleming and Souganidis when generalizing the existence of a value (and the dynamic programming) from deterministic zero-sum differential games to zero-sum stochastic differential games, and second, because the method used by Kononenko and Kleimenov—which makes an extensive use of the extremal aiming and of the existence of quasi-optimal positional strategies for some associated zero-sum differential games—does not apply to stochastic differential games.

Let us finally recall another approach for the existence problem of Nash equilibrium payoffs—the dynamic programming approach. The idea is to find the Nash equilibrium payoff (e_1, e_2) as a function of the initial position (t,x): $(e_1, e_2) =$

$(e_1(t,x), e_2(t,x))$. This function can be constructed as a solution of some system of parabolic PDE (as in [1], for instance) or by using backward or backward-forward stochastic differential equations (as in [3], [4]). Both methods rely heavily on a nondegeneracy assumption on σ. In fact, it can be proved (see [8]) that a payoff $(e_1(t,x), e_2(t,x))$ given by such a method is a Nash equilibrium payoff in our sense.

The paper is organized as follows. We first state the assumptions, notation, and main results of the paper. Then we prove the characterization theorem, from which we derive the existence result. We complete the paper with some remarks on the notions of strategies.

2. Statements of the main results. Let $T > 0$ be a fixed finite time horizon. For $t \in [0,T]$, we consider the following doubly controlled stochastic system:

$$\begin{aligned} &dX_s = f(s, X_s, u_s, v_s)ds + \sigma(s, X_s, u_s, v_s)dB_s, \qquad s \in [t,T], \\ &X_t = x, \end{aligned} \tag{2.1}$$

where B is a d-dimensional standard Brownian motion on the canonical Wiener space $(\Omega, \mathcal{F}, P)$, i.e., Ω is the set of continuous functions from $[0,T]$ to $\mathbb{R}^d$ issued from 0, $\mathcal{F}$ is the completed Borell σ-algebra over Ω, P is the Wiener measure, and B is the canonical process: $B_s(\omega) = \omega(s), s \in [0,T]$. The processes u and v are assumed to take their values in some compact metric spaces U and V, respectively. We suppose that the functions $f : [0,T] \times \mathbb{R}^n \times U \times V \to \mathbb{R}^n$ and $\sigma : [0,T] \times \mathbb{R}^n \times U \times V \to \mathbb{R}^{n \times d}$ are measurable and satisfy the assumption (H):

(H) f and σ are bounded and Lipschitz continuous with respect to (t,x), uniformly in $(u,v) \in U \times V$.

We should also assume that the Isaacs condition, i.e., that for all $(t,x) \in [0,T] \times \mathbb{R}^n$, $p \in \mathbb{R}^n$, and all $A \in \mathcal{S}_n$ (where $\mathcal{S}_n$ is the set of symmetric $n \times n$ matrices), holds:

$$\begin{aligned} &\inf_u \sup_v \{\langle f(t,x,u,v), p\rangle + \tfrac{1}{2} Tr(A\sigma(t,x,u,v)\sigma^*(t,x,u,v))\} \\ &\qquad = \sup_v \inf_u \{\langle f(t,x,u,v), p\rangle + \tfrac{1}{2} Tr(A\sigma(t,x,u,v)\sigma^*(t,x,u,v))\}. \end{aligned} \tag{2.2}$$

We define the sets of admissible controls as follows.

DEFINITION 2.1. *An admissible control process u for Player* I *(resp.,* II*) on $[t,T]$ is a process taking values in U (resp., V), progressively measurable with respect to the filtration $(\mathcal{F}_{t,s}, s \geq t)$, where*

$$\mathcal{F}_{t,s} = \sigma\{B_r - B_t, r \in [t,s]\}, \qquad s \in [t,T],$$

augmented by all null-sets of P.

The set of admissible controls for Player I *(resp.,* II*) on $[t,T]$ is denoted by $\mathcal{U}(t)$ (resp., $\mathcal{V}(t)$).*

We identify two processes u and $\bar{u}$ in $\mathcal{U}(t)$ and write $u \equiv \bar{u}$, if $P\{u = \bar{u}$ a.e. in $[t,s]\} = 1$.

Under assumption (H), for all $(t,x) \in [0,T] \times \mathbb{R}^n$ and $(u,v) \in \mathcal{U}(t) \times \mathcal{V}(t)$, there exists a unique solution to (2.1) that we denote by $X_\cdot^{t,x,u,v}$.

Now we have to define strategies. Let us first recall the definition of nonanticipative strategies.

DEFINITION 2.2. *A nonanticipative strategy for Player* I *on $[t,T]$ is a mapping $\alpha : \mathcal{V}(t) \to \mathcal{U}(t)$ such that, for any $s \in [t,T]$ and for any $v_1, v_2 \in \mathcal{V}(t)$, if $v_1 \equiv v_2$ on $[t,s]$, then $\alpha(v_1) \equiv \alpha(v_2)$ on $[t,s]$.*

Nonanticipative strategies for Player II are defined symmetrically.

For several reasons explained below, nonanticipative strategies are not the proper ones for nonzero-sum differential games. We merely use the notion of admissible strategies, whose definition needs some preliminary remarks, as follows.

For all $\bar{t}, t \in [0,T]$ with $\bar{t} \le t$, let $\Omega_{\bar{t},t}$ be the set of continuous functions from $[\bar{t},t]$ to $\mathbb{R}^d$, issued from 0 and $P_{\bar{t},t}$ the Wiener measure on $\Omega_{\bar{t},t}$ (in particular $\Omega_{0,T} = \Omega$ and $P_{0,T} = P$). If, for fixed $0 \le \bar{t} \le t \le T$, and $\omega \in \Omega_{\bar{t},T}$, we define $\pi(\omega) = (\omega_1, \omega_2)$ by

$$\begin{aligned} \omega_1 &= \omega|_{[\bar{t},t]}, \\ \omega_2 &= (\omega - \omega(t))|_{[t,T]}, \end{aligned}$$

we can identify $\Omega_{\bar{t},T}$ with $\Omega_{\bar{t},t} \times \Omega_{t,T}$, and we have $P_{\bar{t},T} = P_{\bar{t},t} \otimes P_{t,T}$.

Furthermore, to every random variable Y on $\Omega_{\bar{t},T}$ and all $\omega_1 \in \Omega_{\bar{t},t}$, we can associate a random variable $(Y(\omega_1))(\cdot)$ on $\Omega_{t,T}$, by setting $(Y(\omega_1))(\omega_2) = Y(\omega)$.

We also remark that, for all $(\mathcal{F}_{\bar{t},s}, s \ge \bar{t})$-progressively measurable processes $(Y_s, s \ge \bar{t})$ and almost every $\omega_1 \in \Omega_{\bar{t},t}$, the process $(Y(\omega_1)_s, s \ge t)$ is $(\mathcal{F}_{t,s}, s \ge t)$-progressively measurable. This allows us to apply a nonanticipative strategy α defined on $[t,T]$ to controls living on the larger time interval $[\bar{t},T]$: for $v \in \mathcal{V}(\bar{t})$, we define $\alpha(v|_{[t,T]})$ by

$$\alpha(v|_{[t,T]})(\omega)_s = (\alpha(v(\omega_1)))(\omega_2)_s, \qquad s \in [t,T] \tag{2.3}$$

(with a symmetric notation for strategies β for Player II).

DEFINITION 2.3. *An admissible strategy for Player* I *on* $[t,T]$ *is a mapping* $\alpha : \mathcal{V}(t) \to \mathcal{U}(t)$ *such that the following hold:*

(i) α *is a strongly nonanticipative strategy. Namely, for any* $(\mathcal{F}_{t,s})_{s\in[t,T]}$*-stopping time* S *and any* $v, \tilde{v} \in \mathcal{V}(t)$*, if* $v \equiv \tilde{v}$ *on* $[\![t,S]\!]$*, then* $\alpha(v) \equiv \alpha(\tilde{v})$ *on* $[\![t,S]\!]$ *(with the notation* $[\![t,S]\!] = \{(s,\omega) \in [0,T] \times \Omega, t \le s \le S(\omega)\}$*).*

(ii) α *is a nonanticipative strategy with delay. Namely, there is some partition* $t = t_0 < t_1 < \cdots < t_m = T$ *such that for all* $v, \tilde{v} \in \mathcal{V}(t)$*, we have* $\alpha(v) = \alpha(\tilde{v})$ *on* $[t,t_1]$ *and for any* $i < n$*, if* $v \equiv \tilde{v}$ *on* $[t,t_i]$*, then* $\alpha(v) \equiv \alpha(\tilde{v})$ *on* $[t,t_{i+1}]$*.*

(iii) α *is an r-strategy. Namely, for every* $0 \le \bar{t} < t$ *and* $v \in \mathcal{V}(\bar{t})$ *the process* $\alpha(v|_{[t,T]})$ *is* $(\mathcal{F}_{\bar{t},s}, s \ge t)$*-progressively measurable.*

The set of all admissible strategies for Player I *on* $[t,T]$ *is denoted by* $\mathcal{A}(t)$*. The set of admissible strategies* $\beta : \mathcal{U}(t) \to \mathcal{V}(t)$ *for Player* II*, which are defined symmetrically, is denoted by* $\mathcal{B}(t)$*.*

The r-strategies were introduced in [2], motivated by technical problems related to measurability issues. Not surprisingly, we have encountered the same kind of difficulties, hence the requirement for the admissible strategies to be r-strategies.

To the best of our knowledge, the notion of *strongly* nonanticipative strategies has never been introduced before. However, it is, in our opinion, much more natural for stochastic differential games than that of standard nonanticipative strategies. Indeed strongly nonanticipative strategies formalize the fact that a player is allowed to take into account only the control of his opponent he observes in the present state of the world. In other words, if we want to make rigorous the requirement "If for some ω, there exists a time $s \ge 0$ such that $v_1(\omega) = v_2(\omega)$ before s, then $\alpha(v_1(\omega)) = \alpha(v_2(\omega))$ before s," it becomes clear that s depends on ω and thus that the nonanticipativity has to involve random times.

The main reason for introducing nonanticipative strategies *with delay* is the following lemma.

LEMMA 2.4. *Let $\alpha \in \mathcal{A}(t)$ be an admissible strategy and $\beta : \mathcal{U}(t) \to \mathcal{V}(t)$ be nonanticipative. There is a unique control-pair $(u, v) \in \mathcal{U}(t) \times \mathcal{V}(t)$ such that*

$$\alpha(v) = u \quad \text{and} \quad \beta(u) = v. \tag{2.4}$$

Of course, a symmetric result holds if α is nonanticipative and β is admissible, or if both α and β are admissible. Let us point out that one cannot omit one of the strategies to be with delay.

Proof of Lemma 2.4. Let $t = t_0 < t_1 < \cdots < t_m = T$ be a partition associated with the admissible strategy α. We construct the controls $(u_\cdot, v_\cdot)$ by induction on the interval $[t_k, t_{k+1})$.

For $k = 0$, we know that, for any $v' \in \mathcal{V}(t)$, the restriction of $u = \alpha(v')$ to the interval $[t, t_1)$ is independent of v' since α admissible. Let us set $v = \beta(u)$ on $[t, t_1)$, which depends only on the values of u on $[t, t_1)$ since β is nonanticipative. Let us point out that this procedure uniquely defines (u, v) on $[t, t_1)$.

Let us now assume that u and v are uniquely defined on $[t, t_k)$. Then the restriction of $u = \alpha(v')$ to the interval $[t_k, t_{k+1}]$ does not depend on the values of v' on $[t_k, t_{k+1})$ provided that $v' = v$ on $[t, t_k)$, because α is admissible. This defines u on $[t, t_{k+1})$. Then v is uniquely defined on $[t, t_{k+1})$ by $v = \beta(u)$ since β is nonanticipative.

This completes the proof by induction. □

We give several remarks and comments on strategies later, in the appendix. Here is an example of admissible strategy. This example is borrowed from [2].

Example 2.5. Let $t_0 = t < t_1 < \cdots < t_m = T$ be a fixed partition of $[t, T]$, and, for $1 \le j \le m$, let $(O_{ij})_{i \in \mathbb{N}}$ be a Borel partition of $\mathbb{R}^n$ and $u_{ij} \in U$, for $i \in \mathbb{N}$, be fixed. For any control $v \in \mathcal{V}(t)$, we define $\alpha(v)$ by induction on $[t, t_j)$ by setting

$$\alpha(v)(s) = u_{10} \text{ on } [t, t_1),$$

and, if $\alpha(v)$ is built on $[t, t_j)$, we set

$$\alpha(v)(s) = \sum_i u_{ij} 1_{\{X_{t_j}^{t,x,\alpha(v),v} \in O_{ij}\}} \text{ on } [t_j, t_{j+1}).$$

Then α is an admissible strategy.

The main result of [2] is that zero-sum stochastic games have a value when the players play nonanticipative strategies. A careful examination of the proof of [2] shows the following result.

THEOREM 2.6. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be bounded and Lipschitz continuous and set*

$$\forall (u, v) \in \mathcal{U}(t) \times \mathcal{V}(t), \quad J(t, x, u, v) = E[g(X_T^{t,x,u,v})].$$

Let f and σ satisfy the assumption (H) *and the Isaacs condition* (2.2). *Then*

$$\inf_{\beta \in \mathcal{B}(t)} \sup_{u \in \mathcal{U}(t)} J(t, x, u, \beta(u)) = \sup_{\alpha \in \mathcal{A}(t)} \inf_{v \in \mathcal{V}(t)} J(t, x, \alpha(v), v).$$

We briefly explain this result, which is a straightforward consequence of several results of [2], in the following proof.

Proof of Theorem 2.6. Let us set

$$W^\sharp = \sup_{\alpha \in \mathcal{A}(t)} \inf_{v \in \mathcal{V}(t)} J(t, x, \alpha(v), v) \quad \text{and} \quad W^\flat = \inf_{\beta \in \mathcal{B}(t)} \sup_{u \in \mathcal{U}(t)} J(t, x, u, \beta(u)).$$

For any $\epsilon > 0$, let us choose some admissible strategies α and β such that

$$W^\sharp \le \inf_{v\in\mathcal{V}(t)} J(t,x,\alpha(v),v) + \epsilon \quad \text{and} \quad W^\flat \ge \sup_{u\in\mathcal{U}(t)} J(t,x,u,\beta(u)) - \epsilon.$$

Using Lemma 2.4, there is some control-pair $(u,v) \in \mathcal{U}(t)\times\mathcal{V}(t)$ such that $\alpha(v) = u$ and $\beta(u) = v$. Hence

$$W^\sharp \le J(t,x,\alpha(v),v) + \epsilon = J(t,x,u,v) + \epsilon = J(t,x,u,\beta(u)) + \epsilon \le W^\flat + 2\epsilon.$$

Therefore we have proved that $W^\sharp \le W^\flat$.

For proving the reverse inequality, let us set

$$V^\flat = \inf_{\substack{\beta\,:\,\mathcal{U}(t)\to\mathcal{V}(t)\\ \text{nonanticipative}}} \sup_{u\in\mathcal{U}(t)} J(t,x,u,\beta(u)).$$

Combining formula (2.4), Proposition 2.5, and Theorem 2.6 of [2] yields the existence of some nonanticipative strategy α such that

$$V^\flat \le \inf_{v\in\mathcal{V}(t)} J(t,x,\alpha(v),v) + \epsilon.$$

A careful examination of the proof of (2.4) also shows that the strategy α can be chosen from $\mathcal{A}(t)$. Indeed it is actually of the form of Example 2.5. This proves that $V^\flat \le W^\sharp$. Using symmetric argument, one can prove that $V^\sharp \ge W^\flat$, where

$$V^\sharp = \sup_{\substack{\alpha\,:\,\mathcal{V}(t)\to\mathcal{U}(t)\\ \text{nonanticipative}}} \inf_{v\in\mathcal{V}(t)} J(t,x,\alpha(v),v).$$

Hence we have already proved that

$$V^\flat \le W^\sharp \le W^\flat \le V^\sharp.$$

Since, under the Isaacs condition (2.2), the game has a value, i.e., $V^\sharp = V^\flat$ (Theorem 2.6 of [2]), equality $W^\sharp = W^\flat$ holds. □

Now let $g_1 : \mathbb{R}^n \to \mathbb{R}$ and $g_2 : \mathbb{R}^n \to \mathbb{R}$ be two Lipschitz continuous functions bounded by some $C > 0$. For $(t,x) \in [0,T]\times\mathbb{R}^n, (u,v) \in \mathcal{U}(t)\times\mathcal{V}(t)$, set

$$J_1(t,x,u,v) = E[g_1(X_T^{t,x,u,v})] \quad \text{and} \quad J_2(t,x,u,v) = E[g_2(X_T^{t,x,u,v})].$$

In what follows, for all couples of a nonanticipative and an admissible strategy (α,β), we will also use the following notation:

$$J_j(t,x,\alpha,\beta) = J_j(t,x,u,v) \qquad (\text{for } j=1 \text{ or } j=2),$$

where (u,v) are associated to (α,β) by (2.4).

Recall that Player I wants to maximize $J_1(t,x,\alpha,\beta)$, while Player II wants to maximize $J_2(t,x,\alpha,\beta)$.

DEFINITION 2.7. *We say that a couple $(e_1,e_2) \in \mathbb{R}^2$ is a Nash equilibrium payoff at the point (t,x) if, for any $\epsilon > 0$, there exist $(\alpha_\epsilon,\beta_\epsilon) \in \mathcal{A}(t)\times\mathcal{B}(t)$ such that*

(2.5) $\forall \quad (\alpha,\beta) \in \mathcal{A}(t)\times\mathcal{B}(t)$ *it holds that*

$$J_1(t,x,\alpha_\epsilon,\beta_\epsilon) \ge J_1(t,x,\alpha,\beta_\epsilon) - \epsilon \quad \text{and} \quad J_2(t,x,\alpha_\epsilon,\beta_\epsilon) \ge J_2(t,x,\alpha_\epsilon,\beta) - \epsilon$$

and

(2.6) $$\text{for } j=1,2, \quad |J_j(t,x,\alpha_\epsilon,\beta_\epsilon) - e_j| \le \epsilon.$$

Remarks.

1. Condition (2.5) means that if one of the players deviates from his strategy (α_ϵ or β_ϵ), then he cannot expect to get much more (less than ϵ) than what he would have had by keeping his strategy.
2. The definition still makes sense if one uses the notion of nonanticipative strategies with delay instead of admissible strategies.

In what follows, we shall often use an equivalent formulation of Condition (2.5) given by the following lemma.

LEMMA 2.8. *Let $\epsilon > 0$ and $(\alpha_\epsilon, \beta_\epsilon) \in \mathcal{A}(t) \times \mathcal{B}(t)$. Condition* (2.5) *holds if and only if*

$$\begin{aligned} &\text{for any } (u,v) \in \mathcal{U}(t) \times \mathcal{V}(t), \\ &J_1(t,x,\alpha_\epsilon,\beta_\epsilon) \geq J_1(t,x,u,\beta_\epsilon(u)) - \epsilon \quad \text{and} \quad J_2(t,x,\alpha_\epsilon,\beta_\epsilon) \geq J_2(t,x,\alpha_\epsilon(v),v) - \epsilon. \end{aligned} \tag{2.7}$$

Proof of Lemma 2.8. Suppose that (2.7) holds and let $\alpha \in \mathcal{A}(t)$. By Lemma 2.4, there exists $(u,v) \in \mathcal{U}(t) \times \mathcal{V}(t)$ such that $\alpha(v) = u$ and $\beta_\epsilon(u) = v$. By (2.7) applied to this couple (u,v), we get

$$J_1(t,x,\alpha_\epsilon,\beta_\epsilon) \geq J_1(t,x,u,\beta_\epsilon(u)) - \epsilon = J_1(t,x,\alpha,\beta_\epsilon) - \epsilon.$$

Repeating the same argument for some $\beta \in \mathcal{B}(t)$, we get Condition (2.5).

Conversely, for any fixed $u \in \mathcal{U}(t)$, we can define a strategy $\alpha \in \mathcal{A}(t)$ by setting $\alpha(v) = u$ for all $v \in \mathcal{V}(t)$. In particular, for $v = \beta_\epsilon(u)$, we have again $u = \alpha(v)$ and $v = \beta_\epsilon(u)$. It is then easy to deduce (2.7) from (2.5). □

From now on, we denote by W_1 and W_2 the value functions of the zero-sum games where Player I (resp., Player II) aims at maximizing g_1 (resp., g_2). According to Theorem 2.6, this means that

$$W_1(t,x) = \inf_{\beta \in \mathcal{B}(t)} \sup_{u \in \mathcal{U}(t)} J_1(t,x,u,\beta(u)) = \sup_{\alpha \in \mathcal{A}(t)} \inf_{v \in \mathcal{V}(t)} J_1(t,x,\alpha(v),v)$$

and

$$W_2(t,x) = \sup_{\beta \in \mathcal{B}(t)} \inf_{u \in \mathcal{U}(t)} J_2(t,x,u,\beta(u)) = \inf_{\alpha \in \mathcal{A}(t)} \sup_{v \in \mathcal{V}(t)} J_2(t,x,\alpha(v),v).$$

Now, once all notation and assumptions are stated, we are able to announce the two results of this paper, as follows.

THEOREM 2.9 (existence). *Under the Isaacs condition* (2.2)*, for any initial position $(t,x) \in [0,T] \times \mathbb{R}^n$, there is some Nash equilibrium payoff at (t,x).*

THEOREM 2.10 (characterization). *Assume the Isaacs condition holds. Then a couple $(e_1,e_2) \in \mathbb{R}^2$ is a Nash equilibrium payoff at a point (t,x) if and only if for any $\epsilon > 0$ there exists $(u^\epsilon, v^\epsilon) \in \mathcal{U}(t) \times \mathcal{V}(t)$ such that*

(i) *for any $s \in [t,T]$ and $j = 1,2$,*

$$P\left\{E[g_j(X_T^\epsilon)|\mathcal{F}_{t,s}] \geq W_j(s,X_s^\epsilon) - \epsilon\right\} \;\geq\; 1 - \epsilon,$$

where $X_\cdot^\epsilon = X_\cdot^{t,x,u^\epsilon,v^\epsilon}$,

(ii) *and*

$$\text{for } \; j = 1,2, \quad |E[g_j(X_T^\epsilon)] - e_j| \leq \epsilon.$$

Remarks.

1. For proving that the conditions in Theorem 2.10 are necessary, we do not need the notion of admissible strategies; in fact, we need only that the strategies defining the Nash equilibrium payoff are nonanticipative with delay and satisfy some condition (**C**) introduced in the appendix. The fact that the strategy is a strongly nonanticipative r-strategy is not needed. However, although the notion of Nash equilibrium payoff still could be defined by using the bigger class of nonanticipative strategies with delay, we do not know whether the characterization result remains true if one removes the requirement that the strategies satisfy condition (**C**). In other words, we do not know if this characterization holds if one allows the players to use the knowledge of the full control of his or her opponent (in any state of the world).
2. The generalization to the case of more than two players is not difficult. (The idea is that each player can act as if he would be confronted by only one opponent which activates the set of all the other controls.) But to write it down properly, it needs to reintroduce the whole definition and notation and several preliminary results. We will not do it here.

The sketch of the proof of the two previous results is the following: We first show the equivalence in Theorem 2.10, and, using this equivalence, we finally prove Theorem 2.9.

3. Proof of the characterization: Sufficient condition. The object of this section is the proof of the sufficient condition of Theorem 2.10.

We first point out a technical lemma, which will also be used in the proof of the other results.

LEMMA 3.1. *Fix* $(t,x) \in [0,T] \times \mathbb{R}^n$ *and* $u \in \mathcal{U}(t)$.

(a) *For all* $\theta \in [t,T]$ *and* $\epsilon > 0$, *there exists a strongly nonanticipative r-strategy* $\alpha : \mathcal{V}(t) \to \mathcal{U}(t)$ *such that, for any* $v \in \mathcal{V}(t)$,

$$\begin{aligned} &\alpha(v) \equiv u \quad \text{on } [t,\theta], \\ &E[g_2(X_T^{t,x,\alpha(v),v})|\mathcal{F}_{t,\theta}] \le W_2(\theta, X_\theta^{t,x,\alpha(v),v}) + \epsilon, \qquad P \text{ a.s.} \end{aligned} \tag{3.1}$$

(b) *Let* B *be a compact subset of* $\mathbb{R}^n$. *For all* $\theta \in [t,T]$ *and* $\epsilon > 0$, *there exists an admissible strategy* $\alpha \in \mathcal{A}(t)$ *such that, for any* $v \in \mathcal{V}(t)$,

$$\begin{aligned} &\alpha(v) \equiv u \quad \text{on } [t,\theta], \\ &E[g_2(X_T^{t,x,\alpha(v),v})|\mathcal{F}_{t,\theta}] \le W_2(\theta, X_\theta^{t,x,\alpha(v),v}) + \epsilon, \qquad P \text{ a.s. on } \{X_\theta^{t,x,\alpha(v),v} \in B\}. \end{aligned} \tag{3.2}$$

Remark. It can be proved similarly that, for all $\theta \in [t,T]$ and $\epsilon > 0$, there exists a strongly nonanticipative r-strategy $\alpha : \mathcal{V}(t) \to \mathcal{U}(t)$ such that, for any $v \in \mathcal{V}(t)$,

$$\begin{aligned} &\alpha(v) \equiv u \quad \text{on } [t,\theta], \\ &E[g_1(X_T^{t,x,\alpha(v),v})|\mathcal{F}_{t,\theta}] \ge W_1(\theta, X_\theta^{t,x,\alpha(v),v}) - \epsilon, \qquad P \text{ a.s.} \end{aligned}$$

Proof of Lemma 3.1.

(a) From the definition of the value function W_2, for any $y \in \mathbb{R}^n$, there is some admissible strategy $\alpha_y \in \mathcal{A}(\theta)$ such that

$$\sup_{v\in\mathcal{V}(\theta)} E[g_2(X_T^{\theta,y,\alpha_y(v),v})] \le W_2(\theta,y) + \epsilon/2.$$

Since $z \to W_2(\theta,z)$ and $z \to \sup_{v\in\mathcal{V}(\theta)} E[g_2(X_T^{\theta,z,\alpha_y(v),v})]$ are continuous, one can find a Borelian partition $(O_i,\ i=1,2,\ldots)$ of $\mathbb{R}^n$ such that, for any i, there is some $y_i \in O_i$ with

$$\forall z \in O_i,\ \sup_{v\in\mathcal{V}(\theta)} E[g_2(X_T^{\theta,z,\alpha_{y_i}(v),v})] \le W_2(\theta,z) + \epsilon. \tag{3.3}$$

Now we define the following strategy α:

$$\forall v \in \mathcal{V}(t), \alpha(v)_s = \begin{cases} u_s & \text{for } s \in [t,\theta], \\ \alpha_{y_i}(v|_{[\theta,T]})_s & \text{for } s \in (\theta,T] \qquad \text{on } \{X_\theta^{t,x,u,v} \in O_i\}, \end{cases} \tag{3.4}$$

where the notation $\alpha_{y_i}(v|_{[\theta,T]})_s$ is defined by (2.3). By a tiresome but straightforward proof, we get that α is a nonanticipative r-strategy. The fact that it is strongly nonanticipative is proved in Lemma 6.1. It is clear that $\alpha(v) \equiv u$ on $[t,\theta]$.

Further, we obviously have (see also Lemma 1.11 in [2]), if we set $X = X^{t,x,\alpha(v),v}$,

$$\begin{aligned} E[g_2(X_T)|\mathcal{F}_{t,\theta}](\omega_1,\cdot) &= E_{\theta,T}[g_2(X_T^{\theta,X_\theta(\omega_1),\alpha(v(\omega_1)),v(\omega_1)})] \\ &= \textstyle\sum_{i\in\mathbb{N}} 1_{\{X_\theta(\omega_1)\in O_i\}} E_{\theta,T}[g_2(X_T^{\theta,X_\theta(\omega_1),\alpha_{y_i}(v(\omega_1)),v(\omega_1)})], \\ &\qquad\qquad P_{t,\theta}(d\omega_1) \text{ a.s.} \end{aligned} \tag{3.5}$$

Now (3.1) follows from (3.3).

(b) Let $(O_i)_{i\in\{0,\ldots,m\}}$ be a finite Borelian partition of $\mathbb{R}^n$ and $y_i \in O_i, i \in \{0,\ldots,k\}$, be such that $O_0 = B^c$ and, for $i \in \{1,\ldots,m\}$, (3.3) holds. (Indeed, since B is compact, a finite partition of B is sufficient to get (3.3).) Note that there is no condition on $y_0 \in B^c$.

Now let α be built like in (3.4). We already know that $\alpha(v) \equiv u$ on $[t,\theta]$ and that α is a strongly nonanticipative r-strategy. Let us prove that α is nonanticipative with delay. It is easy to see that α has delays corresponding to a partition that, from θ on, is a partition for all strategies $\alpha_{y_i}, i \in \{0,\ldots,m\}$. Since the number of strategies α_{y_i} involved in the construction of α is finite, this partition is also finite. It follows from the construction and from Proposition 6.3 that $\alpha \in \mathcal{A}(t)$.

Using again (3.3) and (3.5) and the choice of $O_1,\ldots,O_m$, one has that

$$1_{\{X_\theta\in B\}} E[g_2(X_T)|\mathcal{F}_{t,\theta}] \le 1_{\{X_\theta\in B\}}(W_2(\theta,X_\theta)+\epsilon)$$

for any $v \in \mathcal{V}(t)$, and relation (3.2) follows evidently. □

Now let us assume that (e_1,e_2) satisfies conditions (i) and (ii) of Theorem 2.10. For any $\epsilon > 0$, there is some control pair $(u^\epsilon, v^\epsilon) \in \mathcal{U}(t) \times \mathcal{V}(t)$ such that

(i) for any $s \in [t,T]$ and $j = 1,2$,

$$P\left\{E[g_j(X_T^\epsilon)|\mathcal{F}_{t,s}] \ge W_j(s,X_s^\epsilon) - \epsilon\right\} \ \ge\ 1-\epsilon, \tag{3.6}$$

where $X_\cdot^\epsilon = X_\cdot^{t,x,u^\epsilon,v^\epsilon}$, and

(ii)

$$\text{for } j=1,2, \qquad |E[g_j(X_T^{\epsilon})]-e_j|\le\epsilon. \tag{3.7}$$

We have to prove that (e_1,e_2) is a Nash equilibrium payoff for the initial position (t,x).

To do this, we are going to define, for any $\epsilon>0$, some strategies $(\alpha_\epsilon,\beta_\epsilon)\in\mathcal{A}(t)\times\mathcal{B}(t)$ satisfying (2.5) and (2.6). We explain only the construction of α_ϵ, the construction of β_ϵ being symmetric.

To simplify the notation, we assume throughout the proof that $g_j\ge 0$ for $j=1,2$. Let us point out that we can make this assumption without loss of generality since this just adds some constant to the functions g_j. In particular, this assumption entails that $W_j\ge 0$. Recall that there exists also some $C>0$ such that $|g_j|\le C$.

Since the dynamic is bounded, it is easy to prove that, for all $(u,v),(u',v')\in\mathcal{U}(t)\times\mathcal{V}(t)$, for all $(\mathcal{F}_{t,s},s\ge t)$-stopping times S with $P[S\le T]=1$ and such that $X_S^{t,x,u,v}=X_S^{t,x,u',v'}$, P a.s., and for all $\tau\ge 0$,

$$E\left[\sup_{0\le s\le\tau}|X^{t,x,u,v}_{(S+s)\wedge T}-X^{t,x,u',v'}_{(S+s)\wedge T}|^2\right]\le C_0\tau,$$

where the constant $C_0>0$ depends only on the dynamic. Thus, since $W_2(s,\cdot)$ is Lipschitz, uniformly in s, we can choose $\tau>0$ such that, for all $(u,v),(u',v')\in\mathcal{U}(t)\times\mathcal{V}(t)$, for all $(\mathcal{F}_{t,s},s\ge t)$-stopping times S with $P[S\le T]=1$, and such that $X_S^{t,x,u,v}=X_S^{t,x,u',v'}$, P a.s.,

$$E\left[\sup_{0\le s\le\tau}|W_2((S+s)\wedge T,X^{t,x,u,v}_{(S+s)\wedge T})-W_2((S+s)\wedge T,X^{t,x,u',v'}_{(S+s)\wedge T})|^2\right]\le(\epsilon/4)^2. \tag{3.8}$$

Let us fix some partition $t_0=t<t_1<\cdots<t_m=T$ which satisfies $\sup_i|t_{i+1}-t_i|\le\tau$.

We also fix some M large enough such that

$$\sup_{u\in\mathcal{U}(t)}\sup_{v\in\mathcal{V}(t)}P\left(\sup_{t\le s\le T}|X_s^{t,x,u,v}|>M\right)\le\epsilon/(4C). \tag{3.9}$$

Let us set

$$\epsilon_0=\frac{\epsilon}{4(2+mC)} \tag{3.10}$$

and let $(\bar u,\bar v)=(u^{\epsilon_0},v^{\epsilon_0})$ satisfy (3.6) and (3.7) for $\epsilon=\epsilon_0$.

By Lemma 3.1(b) applied to the closed ball B in $\mathbb{R}^n$ with center 0 and radius M, to the control $\bar u$ and to $\theta=t_1,\dots,t_m$, we get m admissible strategies $\alpha_1\in\mathcal{A}(t),\dots,\alpha_m\in\mathcal{A}(t)$ such that, for any $v\in\mathcal{V}(t)$, for any $l\in\{1,\dots,m\}$, $\alpha_j(v)\equiv\bar u$ on $[t,t_j]$ and

$$1_{\{X^{\alpha_j}_{t_j}\in B\}}E[g_2(X_T^{\alpha_j})|\mathcal{F}_{t,t_j}]\le 1_{\{X^{\alpha_j}_{t_j}\in B\}}W_2(t_j,X^{\alpha_j}_{t_j})+\epsilon/4, \tag{3.11}$$

where we have set $X_\cdot^{\alpha_j}=X_\cdot^{t,x,\alpha_j(v),v}$.

For all $v\in\mathcal{V}(t)$, we introduce the stopping times

$$S^v=\inf\{s\ge t,v_s\ne\bar v_s\}\qquad\text{and}\qquad t^v=\inf\{t_i>S^v,\ i\ge 1\},$$

with the convention $t^v = T$ if $v_s = \bar{v}_s$ on $[t, T]$.

We are now ready to define the admissible strategy α_ϵ by setting

$$\forall v \in \mathcal{V}(t), \alpha_\epsilon(v) = \begin{cases} \bar{u} & \text{on } [\![t, t^v]\!], \\ \alpha_j(v) & \text{on } (t_j, T] \times \{t^v = t_j\}. \end{cases} \tag{3.12}$$

It is easy to check that α_ϵ is an admissible strategy.

Let $v \in \mathcal{V}(t)$ be fixed and let us set $X_\cdot = X_\cdot^{t,x,\alpha_\epsilon(v),v}$. Let us notice that $\alpha_\epsilon(v) \equiv \bar{u}$ on $[\![t, t^v]\!]$ and that

$$X = \begin{cases} X^{t,x,\bar{u},v} & \text{on } [\![t, t^v]\!] \ P \text{ a.s.}, \\ \sum_j X^{t,x,\alpha_j(v),v} 1_{t^v = t_j} & \text{on } [\![t^v, T]\!] \ P \text{ a.s.} \end{cases}$$

Then using (3.11) we get

$$1_{\{X_{t^v} \in B\}} E[g_2(X_T) | \mathcal{F}_{t,t^v}] \le 1_{\{X_{t^v} \in B\}} W_2(t^v, X_{t^v}) + \epsilon/4. \tag{3.13}$$

We now claim that

$$\forall v \in \mathcal{V}(t), \ J_2(t, x, \alpha_\epsilon(v), v) \le e_2 + \epsilon \qquad \text{and} \qquad \alpha_\epsilon(\bar{v}) = \bar{u}. \tag{3.14}$$

Indeed, it follows from (3.13) that for any $v \in \mathcal{V}(t)$, if we set $X_\cdot = X_\cdot^{t,x,\alpha_\epsilon(v),v}$, we have

$$\begin{aligned} J_2(t, x, \alpha_\epsilon(v), v) &\le E[g_2(X_T) 1_{\{|X_{t^v}| > M\}}] + E[W_2(t^v, X_{t^v}) 1_{\{|X_{t^v}| \le M\}}] + \epsilon/4 \\ &\le E[W_2(t^v, X_{t^v})] + \epsilon/2, \end{aligned} \tag{3.15}$$

where the last inequality comes from the choice of M in (3.9).

Now set $\overline{X} = X^{t,x,\bar{u},\bar{v}}$. Recall that, by the definition of S^v and the fact that $\alpha_\epsilon(v) = \bar{u}$ on $[\![t, t^v]\!]$, we have $\overline{X}_s = X_s$ on $\{s \le S^v\}$.

Further, we have $S^v \le t^v \le S^v + \tau$; thus, by (3.8), we get

$$\|W_2(t^v, X_{t^v}) - W_2(t^v, \overline{X}_{t^v})\|_2 \le \epsilon/4.$$

Accordingly,

$$E[W_2(t^v, X_{t^v})] \le E[W_2(t^v, \overline{X}_{t^v})] + \epsilon/4. \tag{3.16}$$

Let us now denote by Ω_s (for $s \in [t, T]$) the set

$$\Omega_s = \left\{ E[g_2(\overline{X}_T) | \mathcal{F}_{t,s}] \ge W_2(s, \overline{X}_s) - \epsilon_0 \right\}.$$

We recall that $P(\Omega_s) \ge 1 - \epsilon_0$ thanks to (3.6). Thus

$$\begin{aligned} E[W_2(t^v, \overline{X}_{t^v})] &= \textstyle\sum_{i=1}^m E[W_2(t_i, \overline{X}_{t_i}) 1_{t^v = t_i} 1_{\Omega_{t_i}}] + \sum_{i=1}^m E[W_2(t_i, \overline{X}_{t_i}) 1_{t^v = t_i} 1_{\Omega_{t_i}^c}] \\ &\le \textstyle\sum_{i=1}^m E[(E[g_2(\overline{X}_T) | \mathcal{F}_{t,t_i}] + \epsilon_0) 1_{t^v = t_i} 1_{\Omega_{t_i}}] \\ &+ \textstyle\sum_{i=1}^m CP(\Omega_{t_i}^c \cap \{t^v = t^i\}) \\ &\le E[g_2(\overline{X}_T)] + \epsilon_0 + \textstyle\sum_{i=1}^m CP(\Omega_{t_i}^c) \\ &\le (e_2 + 2\epsilon_0) + mC\epsilon_0, \end{aligned} \tag{3.17}$$

where we have used in the last inequality on the one hand the fact that $g_2 \ge 0$ and (3.7) and, on the other hand, the fact that $P(\Omega_{t_i}^c) \le \epsilon_0$ for any i.

Putting (3.15), (3.16), and (3.17) together yields

$$J_2(t,x,\alpha_\epsilon(v),v) \le (e_2+2\epsilon_0) + mC\epsilon_0 + \epsilon/4 + \epsilon/2 \le e_2+\epsilon$$

from the choice of ϵ_0. The last assertion of (3.14) being obvious, (3.14) is proved.

In the same way one can build an admissible strategy $\beta_\epsilon \in \mathcal{B}(t)$ such that

$$\forall u \in \mathcal{U}(t),\ J_1(t,x,u,\beta_\epsilon(u)) \le e_1+\epsilon \qquad \text{and} \qquad \beta_\epsilon(\bar{u}) = \bar{v}.$$

Combining (3.14), the previous assertion, and Lemma 2.8 implies that $(\alpha_\epsilon,\beta_\epsilon)$ satisfies the two inequalities of the definition of a Nash equilibrium payoff. □

4. Proof of the characterization: Necessary condition. Suppose that there is a Nash equilibrium payoff $(e_1,e_2) \in \mathbb{R}^2$ at a point (t,x). For some fixed $\epsilon > 0$, let $(\alpha_\epsilon,\beta_\epsilon) \in \mathcal{A}(t) \times \mathcal{B}(t)$ be such that for any $(\alpha,\beta) \in \mathcal{A}(t) \times \mathcal{B}(t)$, the following inequalities hold:

$$J_1(t,x,\alpha_\epsilon,\beta_\epsilon) \ge J_1(t,x,\alpha,\beta_\epsilon) - \epsilon^2/2 \ \text{ and } \ J_2(t,x,\alpha_\epsilon,\beta_\epsilon) \ge J_2(t,x,\alpha_\epsilon,\beta) - \epsilon^2/2 \tag{4.1}$$

and

$$\text{for } j=1,2,\ |J_j(t,x,\alpha_\epsilon,\beta_\epsilon) - e_j| \le \epsilon^2/2.$$

Thanks to Lemma 2.4, there exists a unique couple of controls $(u^\epsilon,v^\epsilon) \in \mathcal{U}(t) \times \mathcal{V}(t)$ such that

$$\alpha_\epsilon(v^\epsilon) = u^\epsilon \text{ and } \beta_\epsilon(u^\epsilon) = v^\epsilon.$$

Let us set $X_\cdot^\epsilon = X_\cdot^{t,x,u^\epsilon,v^\epsilon}$.

We argue by contradiction and assume that there is some time $\theta \in [t,T)$ and some $j=1,2$ (say, $j=1$ to fix the idea) such that

$$P\left\{E[g_1(X_T^\epsilon)|\mathcal{F}_{t,\theta}] < W_1(\theta,X_\theta^\epsilon) - \epsilon\right\} > \epsilon.$$

We set

$$A = \left\{E[g_1(X_T^\epsilon)|\mathcal{F}_{t,\theta}] < W_1(\theta,X_\theta^\epsilon) - \epsilon\right\}. \tag{4.2}$$

By Lemma 3.1 (and the remark following the lemma) applied to θ and to the control u^ϵ, there exists a nonanticipative strategy $\tilde{\alpha} : \mathcal{V}(t) \to \mathcal{U}(t)$ such that, for any $v \in \mathcal{V}(t)$, $\tilde{\alpha}(v) = u^\epsilon$ on $[t,\theta]$ and P a.s.,

$$E[g_1(X_T^{t,x,\tilde{\alpha}(v),v})|\mathcal{F}_{t,\theta}] \ge W_1(\theta,X_\theta^{t,x,\tilde{\alpha}(v),v}) - \epsilon/2. \tag{4.3}$$

Let (u,v) be the unique couple associated with $(\tilde{\alpha},\beta_\epsilon)$. Let us notice that $u \equiv u^\epsilon$ on $[t,\theta]$. We define a control $\bar{u}$ in the following way:

$$\bar{u} = u^\epsilon \text{ on } ([t,\theta)\times\Omega) \cup ([\theta,T]\times A^c), \qquad \bar{u} = u \text{ on } [\theta,T]\times A.$$

Since β_ϵ is *strongly* nonanticipative, Corollary 6.4 states that

$$\beta_\epsilon(\bar{u}) \equiv v^\epsilon \text{ on } [t,\theta) \text{ and } \beta_\epsilon(\bar{u})_s = \begin{cases} v_s & \text{on } A \\ v_s^\epsilon & \text{on } A^c \end{cases} \quad \text{for } s \in [\theta,T].$$

Hence, we have

$$X^{t,x,\bar{u},\beta_\epsilon(\bar{u})} \equiv X^\epsilon \text{ on } [t,\theta] \qquad \text{and } X_s^{t,x,\bar{u},\beta_\epsilon(\bar{u})} = \begin{cases} X_s^{t,x,\tilde{\alpha}(v),v} & \text{on } A \\ X_s^\epsilon & \text{on } A^c \end{cases} \text{ for } s \in [\theta,T].$$

Accordingly, by (4.3), we have

$$\begin{aligned} J_1(t,x,\bar{u},\beta_\epsilon(\bar{u})) &= E\left[g_1(X_T^\epsilon)1_{A^c}\right] + E\left[E[g_1(X_T^{t,x,\tilde{\alpha}(v),v})|\mathcal{F}_{t,\theta}]1_A\right] \\ &\geq E\left[g_1(X_T^\epsilon)1_{A^c}\right] + E\left[W_1(\theta,X_\theta^\epsilon)1_A\right] - \tfrac{\epsilon}{2}P[A]. \end{aligned} \tag{4.4}$$

It follows from the definition (4.2) of A that

$$J_1(t,x,\bar{u},\beta_\epsilon(\bar{u})) > E\left[g_1(X_T^\epsilon)\right] + \frac{\epsilon}{2}P(A) \geq J_1(t,x,\alpha_\epsilon,\beta_\epsilon) + \epsilon^2/2.$$

This is in contradiction with (4.1) and the proof is complete. □

5. Proof of the existence. For proving Theorem 2.9, it is enough to show that, for any $\epsilon > 0$, there are some controls u^ϵ and v^ϵ satisfying the conditions (i) and (ii) of Theorem 2.10. In fact, we give a slightly stronger result, as follows.

PROPOSITION 5.1. *Suppose the assumptions of Theorem* 2.9 *hold. Then, for any* $\epsilon > 0$*, there is a control-pair* $(u^\epsilon, v^\epsilon) \in \mathcal{U}(t) \times \mathcal{V}(t)$ *such that, for any* $t \leq s_1 \leq s_2 \leq T$ *and* $j = 1, 2$*,*

$$P\left\{E[W_j(s_2, X_{s_2}^\epsilon)|\mathcal{F}_{t,s_1}] \geq W_j(s_1, X_{s_1}^\epsilon) - \epsilon\right\} \geq 1 - \epsilon,$$

where $X_\cdot^\epsilon = X_\cdot^{t,x,u^\epsilon,v^\epsilon}$.

Proof of Theorem 2.9. Combining the above proposition applied to $s_1 = s$ and $s_2 = T$ with Theorem 2.10 gives the result for any (e_1, e_2) which is an accumulation point of the payoff $(J_1(t,x,u^\epsilon,v^\epsilon), J_2(t,x,u^\epsilon,v^\epsilon))$ as $\epsilon \to 0^+$. □

The proof of Proposition 5.1 is split into several lemmas.

LEMMA 5.2. *For any* $\epsilon > 0$*, there is a couple* $(u^\epsilon, v^\epsilon) \in \mathcal{U}(t) \times \mathcal{V}(t)$ *such that, for any* $t \leq s \leq T$ *and* $j = 1, 2$*,*

$$E[W_j(s, X_s^\epsilon)] \geq W_j(t,x) - \epsilon,$$

where $X_\cdot^\epsilon = X_\cdot^{t,x,u^\epsilon,v^\epsilon}$.

Proof. Let us choose $\alpha_\epsilon \in \mathcal{A}(t)$ and $\beta_\epsilon \in \mathcal{B}(t)$ such that α_ϵ is $\epsilon/2$-optimal for $W_1(t,x)$ while β_ϵ is $\epsilon/2$-optimal for $W_2(t,x)$. Namely,

$$W_1(t,x) \leq \inf_{v\in\mathcal{V}(t)} J_1(t,x,\alpha_\epsilon(v),v) + \epsilon/2 \quad \text{and} \quad W_2(t,x) \leq \inf_{u\in\mathcal{U}(t)} J_2(t,x,u,\beta_\epsilon(u)) + \epsilon/2. \tag{5.1}$$

Let (u^ϵ, v^ϵ) be the unique pair of controls such that

$$\alpha_\epsilon(v^\epsilon) = u^\epsilon \text{ and } \beta_\epsilon(u^\epsilon) = v^\epsilon.$$

We intend to prove that the couple (u^ϵ, v^ϵ) satisfies the conclusion of the lemma. For this, we argue by contradiction and assume that there is some $\theta \in (t,T]$ and some $j = 1, 2$ (to fix the ideas, we suppose $j = 2$) such that

$$E[W_2(\theta, X_\theta^\epsilon)] < W_2(t,x) - \epsilon. \tag{5.2}$$

By Lemma 3.1 applied to θ and to the control u^ϵ, there exists a nonanticipative strategy $\alpha : \mathcal{U}(t) \to \mathcal{V}(t)$ such that, for any $v \in \mathcal{V}(t)$, $\alpha(v) \equiv u^\epsilon$ on $[t, \theta]$ and

$$E[g_2(X_T^{t,x,\alpha(v),v})|\mathcal{F}_{t,\theta}] \le W_2(\theta, X_\theta^{t,x,\alpha(v),v}) + \epsilon/2, \qquad P \text{ a.s.} \tag{5.3}$$

By Lemma 2.4, there exists a unique couple of controls $(u, v) \in \mathcal{U}(t) \times \mathcal{V}(t)$ such that, P a.s.,

$$\alpha(v) = u \text{ and } \beta_\epsilon(u) = v.$$

Since α is nonanticipative and β_ϵ is admissible, and since $\alpha(v^\epsilon) \equiv u^\epsilon$ and $\beta_\epsilon(u^\epsilon) \equiv v^\epsilon$ on $[t, \theta]$, it is easy to check that

$$u \equiv u^\epsilon \text{ and } v \equiv v^\epsilon \text{ on } [t, \theta].$$

Thus $X_\theta^{t,x,u,v} = X_\theta^{t,x,\alpha(v),v} = X_\theta^\epsilon$. It follows then from (5.2) and (5.3) that

$$\begin{aligned} J_2(t,x,u,\beta_\epsilon(u)) = J_2(t,x,\alpha(v),v) &= E[E(g_2(X_T^{t,x,\alpha(v),v})|\mathcal{F}_{t,\theta})] \\ &\le E[W_2(\theta, X_\theta^{t,x,\alpha(v),v})] + \epsilon/2 \\ &< W_2(t,x) - \epsilon/2, \end{aligned}$$

which is in contradiction with (5.1). □

LEMMA 5.3. *Let us fix $\epsilon > 0$ and $t_0 = t < t_1 < \cdots < t_m = T$. Then there is some $(u^\epsilon, v^\epsilon) \in \mathcal{U}(t) \times \mathcal{V}(t)$ such that, for any $i = 0, \ldots, (m-1)$, for $j = 1, 2$, P a.s.,*

$$E[W_j(t_{i+1}, X_{t_{i+1}}^\epsilon)|\mathcal{F}_{t,t_i}] \ge W_j(t_i, X_{t_i}^\epsilon) - \epsilon,$$

where $X_\cdot^\epsilon = X_\cdot^{t,x,u^\epsilon,v^\epsilon}$.

Proof. We construct (u^ϵ, v^ϵ) by induction on the interval $[t_i, t_{i+1})$. Let us first notice that the result for $i = 0$ is given by Lemma 5.2.

Let us now assume that (u^ϵ, v^ϵ) is constructed on $[t_0, t_i)$ and let us define it on $[t_i, t_{i+1})$. From Lemma 5.2, for any $y \in \mathbb{R}^n$, there is some $(u^y, v^y) \in \mathcal{U}(t_i) \times \mathcal{V}(t_i)$ such that for any $s \in [t_i, T]$ and for $j = 1, 2$,

$$E[W_j(s, X_s^{t_i,y,u^y,v^y})] \ge W_j(t_i, y) - \epsilon/2.$$

Using the continuity of $W_j(t_i, \cdot)$ and of $E[W_j(s, X_s^{t_i,\cdot,u^y,v^y})]$, we can find a Borel partition $(O_l \; i = 1, 2, \ldots)$ of $\mathbb{R}^n$ such that, for any l, there is some $y_l \in O_l$ with, for $j = 1, 2$,

$$\forall z \in O_l, \; E[W_j(s, X_s^{t_i,z,u^{y_l},v^{y_l}})] \ge W_j(t_i, z) - \epsilon.$$

Then we define, for any $z \in \mathbb{R}^n$, the control pair $(\tilde{u}(z), \tilde{v}(z))$ by

$$\forall s \ge t, \; \tilde{u}(z)_s = \sum_l 1_{O_l}(z) u_s^{y_l} \text{ and } \tilde{v}(z)_s = \sum_l 1_{O_l}(z) v_s^{y_l},$$

and we set

$$(u^\epsilon, v^\epsilon) = (\tilde{u}(X_{t_i}^{t,x,u^\epsilon,v^\epsilon}), \tilde{v}(X_{t_i}^{t,x,u^\epsilon,v^\epsilon})) \text{ on } [t_i, t_{i+1}).$$

We have, for all $s \ge t_i$, P a.s.,

$$E[W_j(s, X_s^\epsilon)|\mathcal{F}_{t,t_i}] \ge W_j(t_i, X_{t_i}^\epsilon) - \epsilon,$$

where, as usual, $X^{\epsilon}_{\cdot} = X^{t,x,u^{\epsilon},v^{\epsilon}}_{\cdot}$. Using the above inequality for $s = t_{i+1}$ completes the proof by induction. □

We are now ready to prove Proposition 5.1. Let us choose a partition $t_0 = t < t_1 < \cdots < t_m = T$ and let us set

$$\tau = \sup_i |t_{i+1} - t_i|.$$

Since $W_j(\cdot, y)$ are uniformly Hölder continuous and $W_j(s, \cdot)$ are uniformly Lipschitz continuous and since the dynamic is bounded, we can choose τ sufficiently small in such a way that, for all $k \in \{0, \dots, m-1\}$ and all $s \in [t_k, t_{k+1})$,

$$\|W_j(t_{k+1}, X^{t,x,u,v}_{t_{k+1}}) - W_j(s, X^{t,x,u,v}_{s})\|_2 \le \gamma, \tag{5.4}$$

where $\gamma = \epsilon^{\frac{3}{2}}/4$.

Let $(u, v) \in \mathcal{U}(t) \times \mathcal{V}(t)$ be defined by Lemma 5.3 for $\epsilon = \epsilon/(2m)$: for any $i = 0, \dots, (m-1)$, for $j = 1, 2$, P a.s.,

$$E[W_j(t_{i+1}, X_{t_{i+1}})|\mathcal{F}_{t,t_i}] \ge W_j(t_i, X_{t_i}) - \epsilon/(2m), \tag{5.5}$$

where $X_{\cdot} = X^{t,x,u,v}_{\cdot}$. Let us now fix $t \le s_1 < s_2 \le T$. Let also t_i and t_k be such that $t_{i-1} \le s_1 < t_i$ and $t_k < s_2 \le t_{k+1}$. Then we have, thanks to (5.5), for $j = 1, 2$, P a.s.,

$$E[W_j(t_{k+1}, X_{t_{k+1}})|\mathcal{F}_{t,t_i}] \ge W_j(t_i, X_{t_i}) - \epsilon/2.$$

Taking the conditional expectation with respect to $\mathcal{F}_{t,s_1}$ gives, since $t_i \ge s_1$, P a.s.,

$$E[W_j(t_{k+1}, X_{t_{k+1}})|\mathcal{F}_{t,s_1}] \ge E[W_j(t_i, X_{t_i})|\mathcal{F}_{t,s_1}] - \epsilon/2. \tag{5.6}$$

Let us set

$$Z_1 = E[W_j(t_{k+1}, X_{t_{k+1}})|\mathcal{F}_{t,s_1}] - E[W_j(t_i, X_{t_i})|\mathcal{F}_{t,s_1}] + \epsilon/2$$

and

$$Z_2 = E[W_j(s_2, X_{s_2})|\mathcal{F}_{t,s_1}] - W_j(s_1, X_{s_1}) + \epsilon/2.$$

From (5.4) and (5.6), we have $Z_1 \ge 0$, P a.s., and $\|Z_2 - Z_1\|_2 \le 2\gamma$. Therefore,

$$P[Z_2 < -\epsilon/2] \le P[|Z_2 - Z_1| > \epsilon/2] \le 4\gamma^2/(\epsilon/2)^2 = \epsilon,$$

i.e.,

$$P\left\{E[W_j(s_2, X_{s_2})|\mathcal{F}_{t,s_1}] \ge W_j(s_1, X_{s_1}) - \epsilon\right\} \ge 1 - \epsilon.$$

Therefore the proof of Proposition 5.1 is complete. □

6. Appendix: On strategies. In the appendix we bring together several technical facts on strongly nonanticipative strategies. Some of them (Lemma 6.1 and Corollary 6.4) are used in the proofs of Theorems 2.9 and 2.10, and some others have seemed to us of general interest for a better understanding of the different notions of strategies.

In the following lemma we prove that the strategy built in the proof of Lemma 3.1 is strongly nonanticipative. Let us first recall the construction of the strategy α.

Let $\theta \in (t,T)$ be a fixed time, $(O_i)_{i\in\mathbb{N}}$ be a Borel partition of $\mathbb{R}^n$, and, for any $i \in \mathbb{N}$, $\alpha_i \in \mathcal{A}(\theta)$ be an admissible strategy. We also fix some control $u \in \mathcal{U}(t)$. The strategy α is defined by setting

$$\forall v \in \mathcal{V}(t), \alpha(v)_s = \begin{cases} u_s & \text{for } s \in [t,\theta], \\ \alpha_i(v|_{[\theta,T]})_s & \text{for } s \in (\theta,T], \text{ on } \{X_\theta^{t,x,u,v} \in O_i\}. \end{cases}$$

LEMMA 6.1. *The strategy α is strongly nonanticipative.*

Proof. Let S be an $(\mathcal{F}_{t,s})_{s\in[t,T]}$-stopping time and let $v_1, v_2 \in \mathcal{V}(t)$ be such that $v_1 \equiv v_2$ on $[\![t,S]\!]$. We have to prove that $\alpha(v_1) \equiv \alpha(v_2)$ on $[\![t,S]\!]$.

Let us notice that the result is obvious on $\{S \le \theta\}$, because we have $\alpha(v_1) = \alpha(v_2) = u$ on $[t,\theta]$. Let us now set $B = \{S > \theta\}$ and $B_i = B \cap \{X_\theta^{t,x,u,v} \in O_i\}$. We can assume that B is not neglectable, since otherwise there is nothing to prove. Let us now prove that $\alpha(v_1) \equiv \alpha(v_2)$ on $[\![\theta,S]\!] \cap ([t,T] \times B)$.

For this, let us first recall the identifications $\Omega_{t,T} = \Omega_{t,\theta} \times \Omega_{\theta,T}$ and $P_{t,T} = P_{t,\theta} \otimes P_{\theta,T}$. We know that, for almost all $\omega_1 \in \Omega_{t,\theta}$, the controls $v_1(\omega_1)$ and $v_2(\omega_1)$ belong to $\mathcal{V}(\theta)$. Let us finally introduce, for any $0 \le r \le s \le T$, $\mathcal{F}^0_{r,s} = \sigma\{B_\tau - B_r,\ \tau \in [r,s]\}$ (let us underline that $\mathcal{F}_{r,s}$ is nothing but $\mathcal{F}^0_{r,s}$ augmented by all null-sets of P). Using the identification $\mathcal{F}^0_{t,T} = \mathcal{F}^0_{t,\theta} \otimes \mathcal{F}^0_{\theta,T}$ and the fact that the section $N(\omega_1) = \{\omega_2 \in \Omega_{\theta,T},\ (\omega_1,\omega_2) \in N\}$ of a $P_{t,T}$-neglectable set $N \in \mathcal{F}_{t,T}$ is $P_{\theta,T}$-neglectable for almost all $\omega_1 \in \Omega_{t,\theta}$, it can also be proved that, for almost all $\omega_1 \in \Omega_{t,\theta}$, $S(\omega_1)$ is an $(\mathcal{F}_{\theta,s})_{s\in[\theta,T]}$-stopping time. From the definition of the strategy α, we have, for any $i \in \mathbb{N}$ and for almost all $\omega_1 \in B_i$, $\alpha(v_1)(\omega_1,\cdot) = \alpha_i(v_1(\omega_1)_{[\theta,T]})$ and $\alpha(v_2)(\omega_1,\cdot) = \alpha_i(v_2(\omega_1)_{[\theta,T]})$ on $[\theta,T]$. Since α_i is strongly nonanticipative and since $v_1(\omega_1)_{[\theta,T]} \equiv v_2(\omega_1)_{[\theta,T]}$ on $[\![\theta,S(\omega_1)]\!]$ for almost all $\omega_1 \in B_i$, we have, for almost all $\omega_1 \in B_i$,

$$\alpha_i(v_1(\omega_1)_{[\theta,T]}) \equiv \alpha_i(v_2(\omega_1)_{[\theta,T]}) \qquad \text{on } [\![\theta,S(\omega_1)]\!].$$

Therefore, $\alpha(v_1)(\omega_1,\cdot) \equiv \alpha(v_2)(\omega_1,\cdot)$ on $[\![\theta,S(\omega_1)]\!]$ for almost all $\omega_1 \in B$. This completes the proof of the lemma. □

Let us now introduce the following notion, which appears naturally in the proof of Theorem 2.10.

DEFINITION 6.2 (condition (**C**)). *Let $\alpha : \mathcal{V}(t) \to \mathcal{U}(t)$ be a nonanticipative strategy. We say that α satisfies condition* (**C**) *if for any (deterministic) $\theta \in [t,T)$, for any control $v, \bar{v} \in \mathcal{V}(t)$ with $v \equiv \bar{v}$ on $[t,\theta]$ and for any $A \in \mathcal{F}_{t,\theta}$, we have*

$$\alpha(v\mathbf{1}_A + \bar{v}\mathbf{1}_{A^c}) \equiv \alpha(v)\mathbf{1}_A + \alpha(\bar{v})\mathbf{1}_{A^c},$$

where $v\mathbf{1}_A + \bar{v}\mathbf{1}_{A^c}$ denotes for simplicity the control equal to $v \equiv \bar{v}$ on $[t,\theta] \times \Omega$, to v on $[\theta,T] \times A$, and to $\bar{v}$ on $[\theta,T] \times A^c$.

Remark. Although condition (**C**) is weaker than the assumption of being strongly nonanticipative for a strategy (cf. Corollary 6.4), the main results of this paper, Theorems 2.9 and 2.10, remain unchanged if condition (i) in the definition of admissible strategies is replaced by condition (**C**). However, as we shall see, condition (**C**) does not seem to be the right definition for modelizing the fact that a player can take into account only the observation of the control played by his opponent.

An equivalent formulation of condition (**C**) follows.

PROPOSITION 6.3. *Let $\alpha : \mathcal{V}(t) \to \mathcal{U}(t)$ be a nonanticipative strategy. The following assertions are equivalent:*

1. *α satisfies condition* (**C**).

2. *For all stopping times S taking a finite number of values in $[t,T]$, if $v \equiv \tilde{v}$ on $[\![t,S]\!]$, then $\alpha(v) \equiv \alpha(\tilde{v})$ on $[\![t,S]\!]$.*

This result implies that condition (**C**) is weaker than the notion of strongly nonanticipative strategies.

COROLLARY 6.4. *A strongly nonanticipative strategy α satisfies condition* (**C**).

Indeed, a strongly nonanticipative strategy α obviously satisfies the second condition of the proposition.

Proof of Proposition 6.3. Let us first assume that α satisfies the second condition. For proving that α satisfies condition (**C**), let us fix $\theta \in [t,T)$, $A \in \mathcal{F}_{t,\theta}$, and $v, \bar{v}$ in $\mathcal{V}(t)$ such that $v \equiv \bar{v}$ on $[t,\theta]$. We set $v_1 = v1_A + \bar{v}1_{A^c}$. We want to prove that $\alpha(v_1) = \alpha(v)1_A + \alpha(\bar{v})1_{A^c}$ on $[\theta,T]$.

For this let us introduce the stopping time S defined by

$$S = \theta \text{ on } A^c \text{ and } S = T \text{ on } A.$$

Then $v_1 \equiv v$ on $[\![t,S]\!]$. Since α satisfies 2, $\alpha(v_1) \equiv \alpha(v)$ on $[\![t,S]\!]$. This implies that $\alpha(v_1) \equiv \alpha(v)$ on $[\theta,T] \times A$. We can prove in the same way (using the stopping time $S' = \theta$ on A and $S' = T$ on A^c) that $\alpha(v_1) \equiv \alpha(\bar{v})$ on $[\theta,T] \times A^c$. Therefore, α satisfies condition (**C**).

Let us now assume that α satisfies condition (**C**). Let S be a stopping time taking values in $t_0 = t < t_1 < \cdots < t_m = T$, and let $v, \bar{v}$ be such that $v \equiv \bar{v}$ on $[\![t,S]\!]$. We have to prove that $\alpha(v) \equiv \alpha(\bar{v})$ on $[\![t,S]\!]$. For this, let us set, for any $j \in \{0,\dots,m\}$,

$$A_j = \{S \le t_j\} \text{ and } v_j = \begin{cases} v & \text{in } ([t,t_j) \times \Omega) \cup ([t_j,T] \times A_j), \\ \bar{v} & \text{in } [t_j,T] \times A_j^c. \end{cases}$$

We are going to prove by induction that

$$\alpha(v_j) \equiv \alpha(\bar{v}) \quad \text{in } [t_j,T] \times A_j^c \quad \text{and } \alpha(v) \equiv \alpha(\bar{v}) \quad \text{in } [\![t, S \wedge t_j]\!]. \tag{6.1}$$

For $j = m$, this shows the desired result: $\alpha(v) = \alpha(\bar{v})$ on $[\![t,S]\!]$.

For $j = 0$, we have $A_0 = \{S = t\} \in \mathcal{F}_{t,t} = \{\emptyset, \Omega\}$ P a.s., and thus either $v_0 = v$ a.s. (if $A_0 = \Omega$) or $v_0 = \bar{v}$ a.s. (if $A_0 = \emptyset$). In both cases, equalities in (6.1) are clear.

Let us assume that (6.1) holds for some j. Let us first prove that $\alpha(\bar{v}) = \alpha(v)$ in $[\![t, S \wedge t_{j+1}]\!]$. For that purpose, let us notice that $v_j \equiv v$ on $[t,t_{j+1})$, because $v \equiv \bar{v}$ on $[t_j,t_{j+1}) \times A_j^c$, since $A_j^c = \{S \ge t_{j+1}\}$ and $v \equiv \bar{v}$ on $[\![t,S]\!]$. Since α is nonanticipative, we have therefore that $\alpha(v_j) \equiv \alpha(v)$ on $[t,t_{j+1})$. Using assumption (6.1), which states that $\alpha(v_j) \equiv \alpha(\bar{v})$ in $[t_j,T] \times A_j^c$, we deduce that $\alpha(\bar{v}) = \alpha(v)$ in $[t_j,t_{j+1}] \times A_j^c$. Let us now notice that

$$[\![t, S \wedge t_{j+1}]\!] = [\![t, S \wedge t_j]\!] \cup ([t_j,t_{j+1}] \times A_j^c).$$

From assumption (6.1) we know that $\alpha(\bar{v}) = \alpha(v)$ in $[\![t, S \wedge t_j]\!]$ and we have just proved that $\alpha(\bar{v}) = \alpha(v)$ in $[t_j,t_{j+1}] \times A_j^c$. Therefore, we have established that $\alpha(\bar{v}) = \alpha(v)$ in $[\![t, S \wedge t_{j+1}]\!]$.

It remains to show that $\alpha(v_{j+1}) \equiv \alpha(\bar{v})$ in $[t_{j+1},T] \times A_{j+1}^c$. Let us first notice that $v_{j+1} = v_j 1_{A_{j+1}^c} + v1_{A_{j+1}}$. Then, since $v_j \equiv v$ on $[t,t_{j+1}]$ and since $A_{j+1} \in \mathcal{F}_{t,t_{j+1}}$, we have, from assumption (**C**),

$$\alpha(v_{j+1}) = \alpha(v_j 1_{A_{j+1}^c} + v1_{A_{j+1}}) \equiv \alpha(v_j)1_{A_{j+1}^c} + \alpha(v)1_{A_{j+1}} \quad \text{on } [t,T]. \tag{6.2}$$

In particular, $\alpha(v_{j+1}) \equiv \alpha(v_j)$ on $[t,T] \times A^c_{j+1}$. Moreover, from (6.1), $\alpha(v_j) \equiv \alpha(\bar{v})$ on $[t_{j+1},T] \times A^c_{j+1}$, because $A^c_{j+1} \subset A^c_j$. Thus

$$\alpha(v_{j+1}) \equiv \alpha(v_j) \equiv \alpha(\bar{v}) \text{ on } [t_{j+1},T] \times A^c_{j+1}.$$

By induction the proof is now complete. □

We complete this appendix by showing that condition (**C**) is not equivalent to the notion of strongly nonanticipative strategies. More precisely, we build a strategy that is nonanticipative with delay, is an r-strategy, and satisfies condition (**C**) but is not strongly nonanticipative.

Let us suppose that each of the spaces in which the controls take their values has only two elements: $U = \{u_1, u_2\}$ and $V = \{v_1, v_2\}$. Let $t \in [0,T]$. We define the strategy $\alpha : \mathcal{V}(t) \to \mathcal{U}(t)$ in the following way.

Let $t = t_0 < t_1 < t_2 = T$ be a partition of $[t,T]$. For any $v \in \mathcal{V}(t)$, we set $\alpha(v)_s = u_1$ for $s \in [t,T]$ if v satisfies the following property:

$$\exists \epsilon > 0 \text{ such that } v \equiv v_1 \text{ on } [t, t+\epsilon]. \tag{6.3}$$

Otherwise we set $\alpha(v)_s = u_1$ for $s \in [t,t_1]$ and $\alpha(v)_s = u_2$ for $s \in (t_1,T]$.

It is easy to check that α is a nonanticipative strategy with delay. Let us now prove that α is an r-strategy. If $t > 0$, for $\bar{t} \in [0,t)$ and $v \in \mathcal{V}(\bar{t})$, define the process $(\tilde{u}_s = \alpha(v|_{[t,T]})_s, s \in [t,T])$ (using the notation of section 2). This process satisfies $\tilde{u} \equiv u_1$ on $[t,t_1]$ and then is constant on $(t_1,T]$ equal to u_1 or u_2, with

$$\{\tilde{u}_s = u_1, s \in (t_1,T]\} = \bigcap_{\epsilon > 0 \text{ rational}} \{v \equiv v_1 \text{ in } [t, t+\epsilon]\} \in \mathcal{F}_{\bar{t},t+} \subset \mathcal{F}_{\bar{t},t_1}.$$

Thus $\tilde{u}$ is adapted to the filtration $(\mathcal{F}_{\bar{t},s})_{s \geq t}$. Since, moreover, its paths are left continuous, it follows that $\tilde{u}$ is $(\mathcal{F}_{\bar{t},s})_{s \geq t}$-progressively measurable.

Let us now prove that α satisfies the condition (**C**). Let us point out that $\mathcal{F}_{t,t} = \{\emptyset, \Omega\}$ P a.s. Thus (**C**) is trivially satisfied for $\theta = t$. Further, by the construction of α, if for $v, \bar{v} \in \mathcal{V}(t)$, we have $v \equiv \bar{v}$ on some time interval $[t,\theta], \theta > t$, then $\alpha(v) = \alpha(\bar{v})$ on $[t,T]$. In particular, α satisfies (**C**).

We finally show that α is not strongly anticipative. Let S be an $(\mathcal{F}_{t,s})_{s \in [t,T]}$-stopping time such that $P[S > t_1] > 0$ and, for all $\epsilon > 0, P[S \leq t+\epsilon] > 0$, say, $S = \inf\{s \geq t, B_s - B_t = 1\} \wedge T$. We define v and $\bar{v} \in \mathcal{V}(t)$ by

$$v \equiv v_1 \text{ on } [t,T],$$
$$\bar{v} \equiv v_1 \text{ on } [\![t,S]\!] \text{ and } \bar{v} \equiv v_2 \text{ on }]\!]S,T]\!].$$

It holds that $v \equiv \bar{v}$ on $[\![t,S]\!]$. But the strategy α applied to the two controls gives

$$\alpha(v) \equiv u_1 \text{ and } \alpha(\bar{v}) \equiv u_2 \text{ on } (t_1,T].$$

Remark 6.5. Note also that in assertion 2 of Proposition 6.3, we have found an example of a property that holds for every stopping time taking a finite number of values but that cannot be generalized to all stopping times.

REFERENCES

[1] A. Bensoussan and J. Frehse, *Stochastic games for N players,* J. Optim. Theory Appl., 105 (2000), pp. 543–565.

[2] W. H. Fleming and P. E. Souganidis, *On the existence of value functions of two-player, zero-sum stochastic differential games,* Indiana Univ. Math. J., 38 (1989), pp. 293–314.
[3] S. Hamadène, J.-P. Lepeltier, and S. Peng, *BSDEs with continuous coefficients and stochastic differential games,* in Backward Stochastic Differential Equations, N. El Karoui et al., eds., Pitman Res. Notes Math. Ser. 364, Longman, London, 1997, pp. 115–128.
[4] S. Hamadène, *Nonzero sum linear-quadratic stochastic differential games and backward-forward equations,* Stochastic Anal. Appl., 17 (1999), pp. 117–130.
[5] A. F. Kleimenov, *Nonantagonist Differential Games*, Nauka Uralprime skoje Otdelenie, Ekaterinburg, 1993 (in Russian).
[6] A. F. Kononenko, *On equilibrium positional strategies in nonantagonistic differential games,* Dokl. Akad. Nauk SSSR, 231 (1976), pp. 285–288.
[7] M. Nisio, *Stochastic differential games and viscosity solutions of Isaacs equations,* Nagoya Math. J., 110 (1988), pp. 163–184.
[8] C. Rainer, *On Feedback Controls for Nonzero-Sum Stochastic Differential Games*, manuscript, 2003.
[9] B. Tolwinski, A. Haurie, and G. Leitmann, *Cooperative equilibria in differential games,* J. Math. Anal. Appl., 119 (1986), pp. 182–202.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 643–669

ASYMPTOTIC STABILITY OF THE WONHAM FILTER: ERGODIC AND NONERGODIC SIGNALS*

PETER BAXENDALE†, PAVEL CHIGANSKY‡, AND ROBERT LIPTSER‡

Abstract. The stability problem of the Wonham filter with respect to initial conditions is addressed. The case of ergodic signals is revisited in view of a gap in the classic work of H. Kunita (1971). We give new bounds for the exponential stability rates, which do not depend on the observations. In the nonergodic case, the stability is implied by identifiability conditions, formulated explicitly in terms of the transition intensities matrix and the observation structure.

Key words. nonlinear filtering, stability, Wonham filter

AMS subject classifications. 93E11, 60J57

DOI. 10.1137/S0363012902416924

1. Introduction.

The optimal filtering estimate of a signal from the record of noisy observations is usually generated by a nonlinear recursive equation subject to the signal a priori distribution. If the latter is unknown and the filtering equation is initialized by an arbitrary initial distribution, the obtained estimate is suboptimal in general. From an applications point of view, it is important to know whether such an estimate becomes close to the optimal one at least after enough time elapses. This property of filters to forget the initial conditions is far from being obvious and in fact generally remains an open and challenging problem.

In this paper, we consider the filtering setting for signals with a finite state space. Specifically, let $X = (X_t)_{t\geq 0}$ be a continuous time homogeneous Markov chain observed via

$$Y_t = \int_0^t h(X_s)ds + \sigma W_t \tag{1.1}$$

with the Wiener process $W = (W_t)_{t\geq 0}$, independent of X, some bounded function h, and $\sigma \neq 0$.

We assume that X_t takes values in the finite alphabet $\mathbb{S} = \{a_1, \dots, a_n\}$ and admits several ergodic classes. Namely,

$$\mathbb{S} = \Big\{ \underbrace{a_1^1, \dots, a_{n_1}^1}_{\mathbb{S}_1}, \dots, \underbrace{a_1^m, \dots, a_{n_m}^m}_{\mathbb{S}_m} \Big\},$$

where the subalphabets $\mathbb{S}_1, \dots, \mathbb{S}_m$ are noncommunicating in the sense that for any $i \neq j$ and $t \geq s$

$$P\big(X_t \in \mathbb{S}_j | X_s \in \mathbb{S}_i\big) = 0. \tag{1.2}$$

*Received by the editors November 1, 2002; accepted for publication (in revised form) October 16, 2003; published electronically July 23, 2004.
http://www.siam.org/journals/sicon/43-2/41692.html

†Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113 (baxendal@math.usc.edu). The research of this author was supported by ONR grant N00014-96-1-0413.

‡Department of Electrical Engineering Systems, Tel Aviv University, 69978 Tel Aviv, Israel (pavelm@eng.tau.ac.il, liptser@eng.tau.ac.il).

So, unless $m = 1$, X_t is a compound Markov chain with the transition intensities matrix

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 & 0 \\ 0 & \Lambda_2 & 0 \\ \dots & \dots & \dots \\ 0 & 0 & \Lambda_m \end{pmatrix} \tag{1.3}$$

of m ergodic classes and is not ergodic itself.

The filtering problem consists in computation of the conditional distribution,

$$\pi_t^\nu(1) = P(X_t^\nu = a_1 | \mathscr{Y}_{[0,t]}^\nu), \dots, \pi_t^\nu(n) = P(X_t^\nu = a_n | \mathscr{Y}_{[0,t]}^\nu),$$

where $\mathscr{Y}_{[0,t]}^\nu$ is the filtration, generated by $\{Y_s^\nu, 0 \le s \le t\}$ satisfying the usual conditions (henceforth, the superscript ν is used to emphasize that the distribution of X_0 is ν).

The vector-valued random process π_t^ν with entries $\pi_t^\nu(1), \dots, \pi_t^\nu(n)$ is generated by the Wonham filter [45] (see also [29, Chap. 9])

$$\begin{aligned} \pi_0^\nu &= \nu, \\ d\pi_t^\nu &= \Lambda^* \pi_t^\nu dt + \sigma^{-2}\big(\operatorname{diag}(\pi_t^\nu) - \pi_t^\nu (\pi_t^\nu)^*\big) h (dY_t^\nu - h^* \pi_t^\nu dt), \end{aligned} \tag{1.4}$$

where $\operatorname{diag}(x)$ is the scalar matrix with the diagonal $x \in \mathbb{R}^n$, h is the column vector with entries $h(a_1), \dots, h(a_n)$, and * is the transposition symbol. If ν is unknown and some other distribution β (on $\mathbb{S}$) is used to initialize the filter, the "wrong" conditional distribution $\pi_t^{\beta\nu}$ is obtained:

$$\begin{aligned} \pi_0^{\beta\nu} &= \beta, \\ d\pi_t^{\beta\nu} &= \Lambda^* \pi_t^{\beta\nu} dt + \sigma^{-2}\big(\operatorname{diag}(\pi_t^{\beta\nu}) - \pi_t^{\beta\nu} (\pi_t^{\beta\nu})^*\big) h (dY_t^\nu - h^* \pi_t^{\beta\nu} dt). \end{aligned} \tag{1.5}$$

According to the intuitive notion of stability, given at the beginning of this section, the filter defined in (1.5) is said to be asymptotically stable if

$$\lim_{t\to\infty} E\|\pi_t^\nu - \pi_t^{\beta\nu}\| = 0, \tag{1.6}$$

where $\|\cdot\|$ is the total variation norm.

If the state space of the Markov chain X consists of one ergodic class ($m = 1$), our setting is in the framework studied by Ocone and Pardoux [35]. In this case, there exists the unique invariant distribution μ, so that

$$\lim_{t\to\infty} \|S_t\gamma - \mu\| = 0, \tag{1.7}$$

where S_t is the semigroup corresponding to X and γ is an arbitrary probability distribution on $\mathbb{S}$. Moreover,

$$\lim_{t\to\infty} \int_{\mathbb{S}} |\mathcal{S}_t f(x) - \mu(f)| d\mu(x) = 0 \tag{1.8}$$

holds for any bounded $f : \mathbb{S} \mapsto \mathbb{R}$. So, it may seem that it remains only to assume

$$\nu \ll \beta \tag{1.9}$$

and allude to [35]. However, the proof of (1.6) given in [35] uses as its central argument the uniqueness theorem for the stationary measure of the filtering process π_t^ν which appeared in the work of H. Kunita [22]. Unfortunately, the proof of this theorem (Theorem 3.3 in [22]) contains a serious gap, as elaborated in the next section.

A different approach to the stability analysis of the filters for ergodic signals was initiated by Delyon and Zeitouni [19]. The authors studied the top Lyapunov exponent of the filtering equation

$$\gamma_\sigma(\beta',\beta'') = \overline{\lim}_{t\to\infty} \frac{1}{t} \log \bigl\|\pi_t^{\beta'\nu} - \pi_t^{\beta''\nu}\bigr\|, \quad \beta' \text{ and } \beta'' \text{ distributions on } \mathbb{S},$$

and showed that $\gamma_\sigma(\beta',\beta'') < 0$ too when Λ and h satisfy certain conditions. Moreover, the filter is found to be stable in the low signal-to-noise regime: $\overline{\lim}_{\sigma\to\infty} \gamma_\sigma(\beta',\beta'') \le \Re\bigl[\lambda^{\max}(\Lambda)\bigr]$ with $\lambda^{\max}(\Lambda)$ being the eigenvalue of Λ with the largest nonzero real part.

These results were further extended by Atar and Zeitouni [3], where it is shown that uniformly in $\sigma > 0$ and h

$$\gamma_\sigma(\beta',\beta'') \le -2 \min_{p\neq q} \sqrt{\lambda_{pq}\lambda_{qp}}, \text{ a.s.}, \tag{1.10}$$

and the high signal-to-noise asymptotics are obtained:

$$\overline{\lim}_{\sigma\to 0} \sigma^2 \gamma_\sigma \le -\frac{1}{2} \sum_{i=1}^d \mu_i \min_{j\neq i} \bigl[h(a_i) - h(a_j)\bigr]^2,$$

$$\underline{\lim}_{\sigma\to 0} \sigma^2 \gamma_\sigma \ge -\frac{1}{2} \sum_{i=1}^d \mu_i \sum_{j=1}^d \bigl[h(a_i) - h(a_j)\bigr]^2,$$

where μ is the ergodic measure of X.

The method in [3] (and its full development in [2]) does not rely on [22] and is based on the analysis of the Zakai equation, corresponding to (1.4) (see (5.2) below). The analysis is carried out by means of the Hilbert projective metric and the Birkhoff inequality, etc.; see section 5 for more details. This approach proved out its efficiency in several filtering scenarios (see [1], [9], [11]).

Other results and methods related to the filtering stability can be found in [4], [10], [12], [13], [14], [16], [17], [18], [15], [24], [25], [26], [27], [36], [37]. The linear Kalman–Bucy case, being the most understood, is extensively treated by several authors: [5], [32], [33], [19], [35], [28], [30] (sections 14.6 and 16.2).

In the present paper, we consider both ergodic and nonergodic signals. Applying the technique from Atar and Zeitouni [2], we show that in the ergodic case the asymptotic stability holds true without any additional assumptions. In other words, the conclusion of H. Kunita [22] is valid in the specific case under consideration.

In view of the counterexample given in section 3, it is clear that in general γ_σ may vanish at $\sigma = 0$. So, it is interesting to find out which ergodic properties of the signal are inherited by the filter regardless of the specific observation structure. In this connection we prove the inequality

$$\overline{\lim}_{t\to\infty} \frac{1}{t} \log \|\pi_t^{\beta\nu} - \pi_t^\nu\| \le -\sum_{r=1}^n \mu_r \min_{i\neq r} \lambda_{ri}.$$

Since μ is the positive measure on $\mathbb{S}$, unlike (1.10), this bound remains negative if at least one row of Λ has all nonzero entries.

Also we give the nonasymptotic bound (compare with (1.10))

$$\|\pi_t^\nu - \pi_t^{\beta\nu}\| \le C \exp\Big(- 2t \min_{p\neq q} \sqrt{\lambda_{pq}\lambda_{qp}}\Big)$$

with some positive constant C depending on ν and β only.

For the discrete time case, related results can be found in Del Moral and Guionnet [18] and Le Gland and Mevel [24]. For example, in [24] the positiveness assumption for all transition probabilities is relaxed under certain constraints on the observation process noise density.

In the case of nonergodic signal, $m > 1$, we show that the filtering stability holds true if the ergodic classes can be identified via observations and the filter matched to each class is stable. We formulate explicit sufficient identifiability conditions in terms of Λ and h.

The paper is organized as follows. In section 2, we introduce the necessary notations and clarify the role of condition $\nu \ll \beta$ in the filtering stability (Proposition 2.1). This section also gives a link to the gap in Kunita's proof [22], while in section 3 the filtering setting is described for which the stability fails and the gap becomes evident.

The main results are formulated in section 4 and proved in sections 5 and 6.

2. Preliminaries and connection to the gap in [22].

2.1. Notations. Throughout, $\nu \ll \beta$ is assumed.

In order to explain our approach, let us consider a general setting when (X, Y) is a Markov process with paths from the Skorokhod space $\mathbb{D} = \mathbb{D}_{[0,\infty)}(\mathbb{R}^2)$ of right continuous functions having limits to the left functions. Moreover, the signal component X is a Markov process itself.

We introduce a measurable space $(\mathbb{D}, \mathscr{D})$, where $\mathscr{D} = \sigma\{(x_s, y_s), s \ge 0\}$ is the Borel σ-algebra on $\mathbb{D}$. Let $D = (\mathscr{D}_t)_{t\ge 0}$ be the filtration of $\mathscr{D}_t = \sigma\{(x_s, y_s), s \le t\}$ and let $D^y = (\mathscr{D}_t^y)_{t\ge 0}$ be the filtration of $\mathscr{D}_t^y = \sigma\{y_s, s \le t\}$.

As before, we write (X_t^ν, Y_t^ν) and (X_t^β, Y_t^β), when the distribution of X_0 is ν or β, respectively, meaning that both pairs are defined on the same probability space, have the same transition semigroup, but different initial distributions.

For a bounded measurable function f, we introduce $\pi_t^\nu(f) := E(f(X_t^\nu)|\mathscr{Y}_{[0,t]}^\nu)$ and $\pi_t^\beta(f) := E(f(X_t^\beta)|\mathscr{Y}_{[0,t]}^\beta)$. Since $\pi_t^\nu(f)$ and $\pi_t^\beta(f)$ are $\mathscr{Y}_{[0,t]}^\nu$- and $\mathscr{Y}_{[0,t]}^\beta$-measurable random variables, respectively, it is convenient to identify $\pi_t^\nu(f)$ and $\pi_t^\beta(f)$ with some $\mathscr{D}_t^y$-measurable functionals of trajectories $Y_{[0,t]}^\nu = \{Y_s^\nu, s \le t\}$ and $Y_{[0,t]}^\beta = \{Y_s^\beta, s \le t\}$.

For this purpose, let Q^ν and Q^β denote the distributions of (X^ν, Y^ν) and (X^β, Y^β) on $(\mathbb{D}, \mathscr{D})$, respectively, and Q_t^ν and Q_t^β be their restrictions on $[0, t]$, so that Q_0^ν, Q_0^β are the distributions of $(X_0^\nu, Y_0^\nu), (X_0^\beta, Y_0^\beta)$. We also assume that

$$\frac{dQ_0^\nu}{dQ_0^\beta}(x, y) = \frac{d\nu}{d\beta}(x_0). \tag{2.1}$$

Since (X_t^ν, Y_t^ν) and (X_t^β, Y_t^β) have the same transition law, we have $Q^\nu \ll Q^\beta$ with

$$\frac{dQ^\nu}{dQ^\beta}(x, y) = \frac{d\nu}{d\beta}(x_0).$$

Without loss of generality, we assume that the filtrations D and D^y satisfy the general conditions with respect to $(Q^\nu + Q^\beta)/2$.

For fixed t, let $H_t^\beta(y)$ be a $\mathscr{D}_t^y$-measurable functional so that $H_t^\beta(Y^\beta) = \pi_t^\beta(f)$ a.s. Moreover, due to $Q^\nu \ll Q^\beta$, a version of $H_t^\beta(y)$ can be chosen such that the random variable $H_t^\beta(Y^\nu)$ is well defined. Then we identify $\pi_t^{\beta\nu}(f)$ with $H_t^\beta(Y^\nu)$.

We do not assume that $\beta \ll \nu$ (and thus $Q^\beta \not\ll Q^\nu$), so this construction fails for $\pi_t^{\nu\beta}(f)$. Nevertheless, a version of $H_t^\nu(y)$ can be chosen such that $H_t^\nu(Y^\nu) = \pi_t^\nu(f)$ a.s. and used for the definition of $\pi_t^{\nu\beta}(f)$. Indeed, let $\overline{Q}^\beta$ and $\overline{Q}^\nu$ be the distributions of Y^ν and Y^β, respectively, i.e., the marginal distributions of Q^β and Q^ν, obviously, $\overline{Q}^\nu \ll \overline{Q}^\beta$ as well as $\overline{Q}_t^\nu \ll \overline{Q}_t^\beta$; the restrictions of $\overline{Q}^\nu$ and $\overline{Q}^\beta$ on the interval $[0,t]$. Moreover, $\frac{d\overline{Q}_t^\nu}{d\overline{Q}_t^\beta}(Y^\beta) = E(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}_{[0,t]}^\beta)$. Now define

$$\pi_t^{\nu\beta}(f) := H_t^\nu(Y^\beta) I\left(\frac{d\overline{Q}_t^\nu}{d\overline{Q}_t^\beta}(Y^\beta) > 0\right).$$

We introduce the decreasing filtration $\mathscr{X}_{[t,\infty)}^\beta = \sigma\{X_s^\beta, s \ge t\}$, the tail σ-algebra

$$\mathscr{T}(X^\beta) = \bigcap_{t\ge 0} \mathscr{X}_{[t,\infty)}^\beta, \tag{2.2}$$

and σ-algebras $\mathscr{X}_t^\beta = \sigma\{X_t^\beta\}$, $\quad \mathscr{Y}_{[0,\infty)}^\beta = \bigvee_{t\ge 0} \mathscr{Y}_{[0,t]}^\beta$.

Set

$$\pi_t^{\beta_0}(f) = E\big(f(X_t^\beta)|\mathscr{Y}_{[0,t]}^\beta \vee \mathscr{X}_0^\beta\big). \tag{2.3}$$

2.2. Filter stability. For bounded and measurable f, the estimate $\pi_t^\nu(f)$ is asymptotically stable with respect to β if

$$\lim_{t\to\infty} E\big|\pi_t^\nu(f) - \pi_t^{\beta\nu}(f)\big| = 0. \tag{2.4}$$

Note that, when the signal process takes values in a finite alphabet and (2.4) holds for any bounded f, then (2.4) and (1.6) are equivalent.

We establish below that (2.4) holds if for large values of t the additional measurement X_0^β is useless for estimation of $f(X_t^\beta)$ via $Y_{[0,t]}^\beta$ or, analogously, if the additional measurement X_t^β is useless for estimation of $\frac{d\nu}{d\beta}(X_0^\beta)$ via $Y_{[0,\infty)}^\beta$.

PROPOSITION 2.1. *Assume $\nu \ll \beta$. Then, any of the conditions*

1.

$$\lim_{t\to\infty} E\big|\pi_t^\beta(f) - \pi_t^{\beta_0}(f)\big| = 0, \tag{2.5}$$

2.

$$E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}_{[0,\infty)}^\beta\Big) = \lim_{t\to\infty} E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}_{[0,\infty)}^\beta \vee \mathscr{X}_{[t,\infty)}^\beta\Big) \tag{2.6}$$

provides (2.4).

Proof. Let us first show that, under $\nu \ll \beta$, for any bounded f

$$\begin{aligned} &E\big|\pi_t^{\beta\nu}(f) - \pi_t^\nu(f)\big| \\ &= E\Big|E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}_{[0,t]}^\beta\Big)E\Big(f(X_t^\beta)|\mathscr{Y}_{[0,t]}^\beta\Big) - E\Big(\frac{d\nu}{d\beta}(X_0^\beta)f(X_t^\beta)\big|\mathscr{Y}_{[0,t]}^\beta\Big)\Big|. \end{aligned} \tag{2.7}$$

Write

$$E\big|\pi_t^{\beta\nu}(f)-\pi_t^{\nu}(f)\big| = E\frac{d\nu}{d\beta}(X_0^\beta)\big|\pi_t^{\beta}(f)-\pi_t^{\nu\beta}(f)\big|$$
$$= EE\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big)\big|\pi_t^{\beta}(f)-\pi_t^{\nu\beta}(f)\big| = E\Big|E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big)\big(\pi_t^{\beta}(f)-\pi_t^{\nu\beta}(f)\big)\Big|$$
$$= E\Big|E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big)E\Big(f(X_t^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big) - E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\pi_t^{\nu\beta}(f)\big|\mathscr{Y}^\beta_{[0,t]}\Big)\Big|.$$

So, it remains to show

$$E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\pi_t^{\nu\beta}(f)\big|\mathscr{Y}^\beta_{[0,t]}\Big) = E\Big(\frac{d\nu}{d\beta}(X_0^\beta)f(X_t^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big). \tag{2.8}$$

With $\mathscr{Y}_t^y$-measurable and bounded function $\Psi_t(y)$ we get

$$E\Big\{\Psi_t(Y^\beta)E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\pi_t^{\nu\beta}(f)\big|\mathscr{Y}^\beta_{[0,t]}\Big)\Big\} = E\Big(\Psi_t(Y^\beta)\frac{d\nu}{d\beta}(X_0^\beta)\pi_t^{\nu\beta}(f)\Big)$$
$$= E\Big(\Psi_t(Y^\nu)\pi_t^{\nu}(f)\Big) = E\Big(\Psi_t(Y^\nu)f(X_t^\nu)\Big) = E\Big(\Psi_t(Y^\beta)\frac{d\nu}{d\beta}(X_0^\beta)f(X_t^\beta)\Big)$$

and notice that (2.8) is valid by the arbitrariness of Ψ_t.

The proof of (2.5)⇒(2.4). Using (2.7) and

$$E\Big(\frac{d\nu}{d\beta}(X_0^\beta)f(X_t^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big) = E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\pi_t^{\beta_0}(f)\big|\mathscr{Y}^\beta_{[0,t]}\Big),$$

we derive

$$E\big|\pi_t^{\beta\nu}(f)-\pi_t^{\nu}(f)\big| = E\Big|E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big)\pi_t^{\beta}(f) - E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\pi^{\beta_0}(f)\big|\mathscr{Y}^\beta_{[0,t]}\Big)\Big|$$
$$= E\Big|E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big(\pi_t^{\beta}(f)-\pi^{\beta_0}(f)\big)\big|\mathscr{Y}^\beta_{[0,t]}\Big)\Big| \le E\frac{d\nu}{d\beta}(X_0^\beta)\big|\pi_t^{\beta}(f)-\pi_t^{\beta_0}(f)\big|,$$

where the Jensen inequality has been used. Let for definiteness $|f|\le K$ with some constant K. Then $\pi_t^\beta(f)$, $\pi_t^{\beta_0}(f)$ can also be chosen such that $|\pi_t^\beta(f)|$ and $|\pi_t^{\beta_0}(f)|$ are bounded by K. Hence, for any $C>0$, we have

$$E\big|\pi_t^{\beta\nu}(f)-\pi_t^{\nu}(f)\big| \le CE\big|\pi_t^{\beta}(f)-\pi_t^{\beta_0}(f)\big| + 2KP\Big(\frac{d\nu}{d\beta}(X_0^\beta)>C\Big).$$

Therefore, $\overline{\lim}_{t\to\infty}E\big|\pi_t^{\beta\nu}(f)-\pi_t^{\nu}(f)\big| \le 2KP\big(\frac{d\nu}{d\beta}(X_0^\beta)>C\big)$ and by the Chebyshev inequality $P\big(\frac{d\nu}{d\beta}(X_0^\beta)>C\big)\le C^{-1}\to 0,\ C\to\infty$.

The proof of (2.6)⇒(2.4). By (2.7)

$$E\big|\pi_t^{\beta\nu}(f)-\pi_t^{\nu}(f)\big|$$
$$= E\Big|E\Big(f(X_t^\beta)E\Big[\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big]\Big|\mathscr{Y}^\beta_{[0,t]}\Big) - E\Big(f(X_t^\beta)\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big)\Big|.$$

Notice also

$$E\Big(f(X_t^\beta)\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big) = E\Big(f(X_t^\beta)E\Big[\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)}\vee\mathscr{X}^\beta_{[t,\infty)}\Big]\big|\mathscr{Y}^\beta_{[0,t]}\Big).$$

Since $|f| \le K$, by the Jensen inequality we have

$$E\big|\pi_t^{\beta\nu}(f) - \pi_t^{\nu}(f)\big| \le KE\Big|E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big) - E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{X}^\beta_{[t,\infty)}\Big)\Big|. \tag{2.9}$$

Both random processes $E(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]})$ and $E(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{X}^\beta_{[t,\infty)})$ are uniformly integrable forward and backward martingales with respect to the filtrations $(\mathscr{Y}^\beta_{[0,t]})_{t\ge0}$ and $(\mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{X}^\beta_{[t,\infty)})_{t\ge0}$. Therefore, they admit limits a.s. in $t \to \infty$: $E(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)})$ and $\lim_{t\to\infty} E(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{X}^\beta_{[t,\infty)})$, respectively. By (2.6)

$$\lim_{t\to\infty}\Big|E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big) - E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{X}^\beta_{[t,\infty)}\Big)\Big| = 0.$$

We show also that

$$\lim_{t\to\infty} E\Big|E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big) - E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{X}^\beta_{[t,\infty)}\Big)\Big| = 0. \tag{2.10}$$

Denote by α_t any of $E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,t]}\Big)$ and $E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{X}^\beta_{[t,\infty)}\Big)$ and

$$\alpha_\infty = \lim_{t\to\infty} \alpha_t.$$

It is clear that (2.10) holds true if $\lim_{t\to\infty} E|\alpha_t - \alpha_\infty| = 0$. Since $\lim_{t\to\infty} \alpha_t = \alpha_\infty$, $\alpha_t \ge 0$, and $E\alpha_t \equiv E\alpha_\infty = 1$, by the Scheffe theorem we get the desired property.

Thus the right-hand side of (2.9) converges to zero and the result follows. □

2.3. Connection to the gap in [22]. In [22], H. Kunita studies[1] ergodic properties of the filtering process π_t^ν. He considers π_t^ν as a Markov process with values in the space of probability measures and claims (in Theorem 3.3) that there exists the unique invariant measure being "limit point" of marginal distributions of π_t^ν, $t \nearrow \infty$. As was later shown in [35], this result is the key to the stability analysis under (1.8).

Below we demonstrate that the main argument, used in the proof of Theorem 3.3 of [22], cannot be taken for granted. We discuss this issue in the context of Proposition 2.1. Suppose the Markov process X is ergodic in the sense of (1.7) and (1.8). It is well known that its tail σ-algebra $\mathscr{T}(X^\beta)$ (see (2.2) for definition) is empty a.s. It is very tempting in this case to change the order of intersection and supremum as follows:

$$\bigcap_{t\ge0} \mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{X}^\beta_{[t,\infty)} = \mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{T}(X^\beta) \quad \text{a.s.} \tag{2.11}$$

Then, the right-hand side of (2.6) is transformed to

$$\lim_{t\to\infty} E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{X}^\beta_{[t,\infty)}\Big) = E\left(\frac{d\nu}{d\beta}(X_0^\beta)\Big|\bigcap_{t\ge0}\Big\{\mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{X}^\beta_{[t,\infty)}\Big\}\right)$$
$$= E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)} \vee \mathscr{T}(X^\beta)\Big) = E\Big(\frac{d\nu}{d\beta}(X_0^\beta)\big|\mathscr{Y}^\beta_{[0,\infty)}\Big)$$

[1] The notations of this paper are used here.

and (2.6) would be correct, regardless (!) of any other ingredients of the problem (e.g., with $\sigma = 0$ in (1.1)).

In [22], the relation of (2.11) type plays the key role in verification of the uniqueness for the invariant measure corresponding to $\pi^{\nu}_t, t \geq 0$. However, the validity of (2.11) is far from being obvious. According to Williams [44], it "...tripped up even Kolmogorov and Wiener" (see Sinai [39, p. 837] for some details). The reader can find a discussion concerning (2.11) in von Weizsäcker [43]; unfortunately, the counterexample there is incorrect. A proper counterexample to (2.11) is given in Exercise 4.12 in Williams [44], which, however, seems somewhat artificial in the filtering context. It turns out that the example, considered by Delyon and Zeitouni in [19] (see [21] by Kaijser for its earlier discrete time version), is nothing but another case when (2.11) fails.

For the reader's convenience, we give below a detailed analysis of this example.

It is important to note that the counterexamples mentioned above do not fit exactly into the setup considered by Kunita. They merely indicate that (2.11) is not evident and so the claim of Theorem 3.3 in [22] remains a conjecture.

Generally, the stability of nonlinear filters for ergodic Markov processes remains an open problem, and some results [23], [40], [41], [6], [8], [7], [35] based on [22] have to be revised.

3. Counterexample. Below we give a detailed discussion of one counterexample to (2.11). Consider Markov process X with values in $\mathbb{S} = \{1, 2, 3, 4\}$, with the initial distribution ν and the transition intensities matrix

$$\Lambda = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix}. \tag{3.1}$$

All states of Λ communicate, and so X is an ergodic Markov process (see, e.g., [34]) with the unique invariant measure $\mu = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$. Let $h(x) = I(x = 1) + I(x = 3)$, that is,

$$Y_t = \int_0^t \big[I(X_s = 2) + I(X_s = 3)\big]\, ds + \sigma W_t.$$

By Theorem 4.1 below, the filter is stable in this case for any $\sigma > 0$.

3.1. Noiseless observation. Consider the case $\sigma = 0$.

It will be convenient to redefine the observation process as follows:

$$Y_t = [I(X_t = 1) + I(X_t = 3)].$$

We assume $\nu \ll \beta$ and notice that (2.1) holds true. We omit the superscripts ν and β when the initial condition does not play a significant role. Since X is an ergodic Markov process, satisfying (1.8), $\mathscr{T}(X) = (\Omega, \varnothing)$ a.s.

Proposition 3.1.

$$\bigcap_{t\geq 0} \Big(\mathscr{Y}_{[0,\infty)} \vee \mathscr{X}_{[t,\infty)}\Big) \supsetneq \mathscr{Y}_{[0,\infty)} \ \ a.s. \tag{3.2}$$

Proof. It suffices to show that X_0 is a $\bigcap_{t\geq 0}(\mathscr{Y}_{[0,\infty)} \vee \mathscr{X}_{[t,\infty)})$-measurable random variable and at the same time $X_0 \notin \mathscr{Y}_{[0,\infty)}$.

TABLE 3.1
Typical trajectory of π_t for $Y_0 = 1$.

t	$[0,\tau_1)$	$[\tau_1,\tau_2)$	$[\tau_2,\tau_3)$	$[\tau_3,\tau_4)$	$[\tau_4,\tau_5)$	$\dots$
Y_t	1	0	1	0	1	$\dots$
$\pi_t(1)$	$\frac{\nu_1}{\nu_1+\nu_3}$	0	$\frac{\nu_3}{\nu_1+\nu_3}$	0	$\frac{\nu_1}{\nu_1+\nu_3}$	$\dots$
$\pi_t(2)$	0	$\frac{\nu_1}{\nu_1+\nu_3}$	0	$\frac{\nu_3}{\nu_1+\nu_3}$	0	$\dots$

The structure of matrix Λ admits only cyclic transitions in the following order:

$$\cdots \to \{3\} \to \{4\} \to \{1\} \to \{2\} \to \{3\} \to \cdots .$$

So, since Y and X jump simultaneously, X_0 can be recovered exactly from the trajectory $Y_s, s \le t$, and X_t for any $t > 0$, i.e., X_0 is $\mathscr{X}_t \vee \mathscr{Y}_{[0,t]}$-measurable. Owing to $\mathscr{X}_t \vee \mathscr{Y}_{[0,t]} \subset \mathscr{X}_{[t,\infty)} \vee \mathscr{Y}_{[0,\infty)}$, X_0 is measurable with respect to

$$\bigcap_{t\ge 0} \Big(\mathscr{Y}_{[0,\infty)} \vee \mathscr{X}_{[t,\infty)} \Big).$$

Denote by $(\tau_i)_{i\ge 1}$ the time moments where Y jumps. It is not hard to check that $(\tau_i)_{i\ge 0}$ is independent of (X_0, Y_0) and, moreover,

$$\mathscr{Y}_{[0,t]} = \bigvee_{i\ge 0} \sigma\{\tau_i \le t\} \vee \sigma\{Y_0\}.$$

Thus for any $t \ge 0$

$$\begin{aligned} P\big(X_0 = 1 | \mathscr{Y}_{[0,t]}\big) &= P\left(X_0 = 1 | \bigvee_{i\ge 0} \sigma\{\tau_i \le t\} \vee \sigma\{Y_0\} \right) \\ &= P\big(X_0 = 1 | Y_0\big) = \frac{\nu_1}{\nu_1+\nu_3} Y_0. \end{aligned} \tag{3.3}$$

Since (3.3) is valid for any $t \ge 0$, we conclude that

$$P\big(X_0 = 1 | \mathscr{Y}_{[0,\infty)}\big) = \frac{\nu_1}{\nu_1+\nu_3} Y_0.$$

Obviously $I(X_0 = 1) \ne \frac{\nu_1}{\nu_1+\nu_3} Y_0$ and thus X_0 is not $\mathscr{Y}_{[0,\infty)}$-measurable. □

3.2. Invariant measures of π_t and the filter instability. Since $I_t(2) + I_t(4) = 1 - Y_t$ and $I_t(1) + I_t(3) = Y_t$, only $I_t(1)$ and $I_t(2)$ have to be filtered while $\pi_t(3) = Y_t - \pi_t(1)$ and $\pi_t(4) = (1 - Y_t) - \pi_t(2)$. The derivation of the filtering equations is sketched in the appendix.

PROPOSITION 3.2. *The optimal filtering estimate satisfies*

$$\begin{aligned} d\pi_t(1) &= \big(1 - \pi_{t-}(2)\big)(1 - Y_{t-})dY_t + \pi_{t-}(1) Y_{t-} dY_t, \\ d\pi_t(2) &= -\pi_{t-}(2)(1 - Y_{t-})dY_t - \pi_{t-}(1) Y_{t-} dY_t \end{aligned}$$

subject to $\pi_0(1) = \frac{\nu_1}{\nu_1+\nu_3} Y_0$, $\pi_0(2) = \frac{\nu_2}{\nu_2+\nu_4}(1 - Y_0)$.

Let us examine the behavior of the filter from Proposition 3.2. A pair of typical trajectories are given in Table 3.1 (for $Y_0 = 1$) and Table 3.2 (for $Y_0 = 0$).

It is not hard to see that Y is itself a Markov chain with values in $\{0, 1\}$ and the transition intensities matrix $\left(\begin{smallmatrix} -1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$, and thus its invariant measure is $\mu' =$

TABLE 3.2
Typical trajectory of π_t for $Y_0 = 0$.

t	$[0,\tau_1)$	$[\tau_1,\tau_2)$	$[\tau_2,\tau_3)$	$[\tau_3,\tau_4)$	$[\tau_4,\tau_5)$	$\dots$
Y_t	0	1	0	1	0	$\dots$
$\pi_t(1)$	0	$\frac{\nu_2}{\nu_2+\nu_4}$	0	$\frac{\nu_4}{\nu_2+\nu_4}$	0	$\dots$
$\pi_t(2)$	$\frac{\nu_2}{\nu_2+\nu_4}$	0	$\frac{\nu_4}{\nu_2+\nu_4}$	0	$\frac{\nu_2}{\nu_2+\nu_4}$	$\dots$

$\begin{pmatrix}1/2 & 1/2\end{pmatrix}$. Hence, the invariant measure Φ of the filtering process $(\pi_t(1),\pi_t(2))$ is concentrated on eight vectors

$$\phi_1=\begin{pmatrix}\frac{\nu_1}{\nu_1+\nu_3}\\0\end{pmatrix},\quad \phi_2=\begin{pmatrix}0\\\frac{\nu_1}{\nu_1+\nu_3}\end{pmatrix},\quad \phi_3=\begin{pmatrix}\frac{\nu_3}{\nu_1+\nu_3}\\0\end{pmatrix},\quad \phi_4=\begin{pmatrix}0\\\frac{\nu_3}{\nu_1+\nu_3}\end{pmatrix},$$
$$\phi_5=\begin{pmatrix}\frac{\nu_2}{\nu_2+\nu_4}\\0\end{pmatrix},\quad \phi_6=\begin{pmatrix}0\\\frac{\nu_2}{\nu_2+\nu_4}\end{pmatrix},\quad \phi_7=\begin{pmatrix}\frac{\nu_4}{\nu_2+\nu_4}\\0\end{pmatrix},\quad \phi_8=\begin{pmatrix}0\\\frac{\nu_4}{\nu_2+\nu_4}\end{pmatrix}$$

with

$$\Phi(\phi_i)=(\nu_1+\nu_3)/4,\quad i=1,2,3,4,$$
$$\Phi(\phi_i)=(\nu_2+\nu_4)/4,\quad i=5,6,7,8,$$

and, consequently, Φ is not unique. Moreover, the optimal filter is not stable in the sense of (1.6). In fact, for different initial conditions, the filtering distribution $\pi_t, t>0$, can "sit" on different vectors!

4. Main results.

4.1. Ergodic case. Markov chain X is ergodic if and only if all entries of its transition intensities matrix Λ *communicate*, i.e., for any pair of indices i and j, a string of indices $\{\ell_1,\dots,\ell_m\}$ can be found so that $\lambda_{i\ell_1}\lambda_{\ell_1\ell_2}\dots\lambda_{\ell_m j}\neq 0$ (see, e.g., [34]). In this case, the distribution of X_t converges to the positive invariant distribution μ being the unique solution of $\Lambda^*\mu=0$ in the class of vectors with positive entries the sum of which is equal to one.

THEOREM 4.1. *If all states of Λ communicate, then there exists a positive constant c such for any ν and β*

$$\overline{\lim}_{t\to\infty}\frac{1}{t}\log\|\pi_t^{\beta\nu}-\pi_t^{\nu}\|<-c\ \ a.s.$$

Remark 1. Clearly, Theorem 4.1 provides (1.6). Also it allows us to conclude that $\lim_{t\to\infty}\|\pi_t^{\beta\nu}-\pi_t^{\nu}\|=0$ a.s. for β concentrated in a single state of $\mathbb{S}$. Then, in particular, we have

$$\lim_{t\to\infty}\|\pi_t^{\mu_0}-\pi_t^{\mu}\|=0$$

which is the main argument in the proof of existence of the unique invariant measure for the process $(\pi_t)_{t\geq 0}$. This fact corroborates Kunita's result from [22] in the finite state space setup of Theorem 4.1.

Actually, Theorem 4.1 verifies the logarithmic rate in $t\to\infty$ which is in general a function of Λ, h and σ. However, stronger assumptions on Λ guarantee exponential or logarithmic rates, regardless of h and σ (σ is only required to be nonzero).

THEOREM 4.2. *Assume all states of* Λ *communicate. Then*

$$\overline{\lim}_{t\to\infty}\frac{1}{t}\log\|\pi_t^{\beta\nu}-\pi_t^{\nu}\| \le -\sum_{r=1}^{n}\mu_r\min_{i\neq r}\lambda_{ri}. \tag{4.1}$$

Remark 2. The bound (4.1) is negative if at least one row of Λ has all nonzero entries.

THEOREM 4.3. *Assume all entries of* Λ *are nonzero.*

1. *If* $\nu \ll \beta$, *then*

$$E\|\pi_t^{\beta\nu}-\pi_t^{\nu}\| \le n\sum_{j=1}^{n}\frac{d\nu}{d\beta}(a_j)\exp\Big(-2t\min_{p\neq q}\sqrt{\lambda_{pq}\lambda_{qp}}\Big),\ t>0. \tag{4.2}$$

2. *If* $\nu \sim \beta$, *then*

$$\|\pi_t^{\beta\nu}-\pi_t^{\nu}\| \le n^2\max_j\frac{d\nu}{d\beta}(a_j)\max_j\frac{d\beta}{d\nu}(a_j)\exp\Big(-2t\min_{p\neq q}\sqrt{\lambda_{pq}\lambda_{qp}}\Big),\ t>0. \tag{4.3}$$

4.2. Nonergodic case. Let $m \ge 2$ and Λ be given in (1.3). If $X_0 \in \mathbb{S}_j$, then X is a Markov process with values in $\mathbb{S}_j$ with transition intensities matrix Λ_j. We denote this process by X^j. In addition to h, introduce column vectors h_j, $j = 1, \dots, m$, with entries $h(a_1^j), \dots, h(a_{n_j}^j)$, respectively.

THEOREM 4.4. *Assume the following.*

A-1. For any j, all states of Λ_j communicate.

A-2. For each j, k with $j \neq k$, either

$$h_j^*\mu^j \neq h_k^*\mu^k$$

or

$$h_j^*\,\mathrm{diag}(\mu^j)\Lambda_j^q h_j \neq h_k^*\,\mathrm{diag}(\mu^k)\Lambda_k^q h_k,\ \textit{for some}\ \ 0 \le q \le n_j+n_k-1.$$

Then the asymptotic stability (1.6) *holds true.*

The condition A-1 is inherited from Theorem 4.1 to ensure the stability within each ergodic class, while under A-2, $\mathscr{Y}_{[0,\infty)}$ completely identifies the class in which X actually resides.

5. Proofs for the ergodic case. Recall that under $m = 1$, X is a homogeneous ergodic Markov chain with values in the finite alphabet $\mathbb{S} = \{a_1, \dots, a_n\}$ with the transition intensities matrix Λ. The unique invariant measure $\mu = (\mu_1, \dots, \mu_n)$ is the positive distribution on $\mathbb{S}$. Let ν be the distribution of X_0 and β a probability measure on $\mathbb{S}$. The observation process Y is defined in (1.1). Recall that the entries of π_t^{ν} and $\pi_t^{\beta\nu}$ are the true and "wrong" conditional probabilities, respectively, as defined in the introduction.

5.1. The proof of Theorem 4.1. We use the method proposed by Atar and Zeitouni in [2], which is elaborated for the considered filtering setup for the reader's convenience.

Recall the following facts from the theory of nonnegative matrices. For a pair (p, q) of nonnegative measures on $\mathbb{S}$ (i.e., vectors with nonnegative entries), the Hilbert projective metric $H(p, q)$ is defined as the following (see, e.g., [38]):

$$H(p,q) = \begin{cases} \log\dfrac{\max\limits_{j:q_j>0}(p_j/q_j)}{\min\limits_{i:q_i>0}(p_i/q_i)}, & p \sim q, \\ \infty, & p \not\sim q. \end{cases} \tag{5.1}$$

The Hilbert metric is known to satisfy the following properties:

1. $H(c_1p, c_2q) = H(p,q)$ for any positive constants c_1 and c_2.
2. For matrix A with nonnegative entries (A_{ij}),

$$H\big(Ap, Aq\big) \le \tau(A) H\big(p,q\big) \quad \text{(see, e.g., [38])},$$

where $\tau(A) = \frac{1-\sqrt{\psi(A)}}{1+\sqrt{\psi(A)}}$ is the Birkhoff contraction coefficient with

$$\psi(A) = \min_{i,j,k,\ell} \frac{A_{ik}A_{j\ell}}{A_{i\ell}A_{jk}}.$$

3. $\|p-q\| \le \frac{2}{\log 3} H(p,q)$ (Lemma 1 in [2]).

Returning to the filtering problem, let us first consider the special case when $\nu = \mu$, and thus the signal X^μ is the stationary Markov chain. It is well known that $\pi_t^\mu = \eta_t^\mu / \langle \mathbf{1}, \eta_t^\mu \rangle$, where $\mathbf{1}$ denotes the vector with unit entries, $\langle \cdot, \cdot \rangle$ is the usual inner product, and η_t^μ solves the Zakai equation

$$d\eta_t^\mu = \Lambda^* \eta_t^\mu dt + \sigma^{-2} \operatorname{diag}(h) \eta_t^\mu dY_t^\mu \tag{5.2}$$

subject to $\eta_0^\mu = \mu$. Similarly, $\pi_t^{\beta\mu} = \eta_t^{\beta\mu} / \langle \mathbf{1}, \eta_t^{\beta\mu} \rangle$, where $\eta_t^{\beta\mu}$ is the solution of (5.2) subject to $\eta_0^{\beta\mu} = \beta$.

The Zakai equation possesses the unique strong solution which is linear with respect to the initial condition. Hence, $\eta_t^\mu = J_{[0,t]}\mu$ and $\eta_t^{\beta\mu} = J_{[0,t]}\beta$, $t > 0$, where $J_{[0,t]}$ is the random Cauchy matrix corresponding to (5.2).

The matrix $J_{[0,t]}$ can be factored (here $\lfloor t \rfloor$ is the integer part of t):

$$J_{[0,t]} = J_{[\lfloor t \rfloor, t]} \left(\prod_{n=2}^{\lfloor t \rfloor} J_{[n-1,n]} \right) J_{[0,1]}.$$

The properties of the Hilbert metric, listed above, provide

$$\begin{aligned} \big\| \pi_t^\mu - \pi_t^{\beta\mu} \big\| &\le \frac{2}{\log 3} H\big(\pi_t^\mu, \pi_t^{\beta\mu}\big) = \frac{2}{\log 3} H\big(J_{[0,t]}\mu, J_{[0,t]}\beta\big) \\ &\le \frac{2}{\log 3} \tau\big(J_{[\lfloor t \rfloor, t]}\big) \prod_{n=2}^{\lfloor t \rfloor} \tau\big(J_{[n-1,n]}\big) H\big(J_{[0,1]}\mu, J_{[0,1]}\beta\big). \end{aligned}$$

Assume for a moment that $H\big(J_{[0,1]}\mu, J_{[0,1]}\beta\big) < \infty$ a.s. Then

$$\begin{aligned} \overline{\lim}_{t\to\infty} \frac{1}{t} \log \big\| \pi_t^\mu - \pi_t^{\beta\mu} \big\| &\le \overline{\lim}_{t\to\infty} \frac{1}{\lfloor t \rfloor} \sum_{n=2}^{\lfloor t \rfloor} \log \tau\big(J_{[n-1,n]}\big) \\ &\le \overline{\lim}_{t\to\infty} \frac{1}{\lfloor t \rfloor} \sum_{n=2}^{\lfloor t \rfloor} \big\{ -1 \vee \log \tau\big(J_{[n-1,n]}\big) \big\} = E\big[-1 \vee \log \tau\big(J_{[0,1]}\big) \big] \le 0. \end{aligned} \tag{5.3}$$

The equality is implied by the law of large numbers, which is valid since $-1 \le \big\{ -1 \vee \log \tau\big(J_{[n-1,n]}\big) \big\} \le 0$ and $\log \tau\big(J_{[n-1,n]}\big)$ is generated by

$$\{ X_s^\mu - X_{n-1}^\mu,\ W_s - W_{n-1} \},\ n-1 \le s < n,$$

where the processes X^μ and W are independent and X^μ is an ergodic Markov chain.

Let $J^{\nu}_{[n-1,n]}$ be the matrices defined similarly to $J_{[n-1,n]}$ with Y^{μ} replaced by Y^{ν}. Recall that μ is the positive measure on $\mathbb{S}$, so that $\nu \ll \mu$ and, in turn, $\overline{Q}^{\nu} \ll \overline{Q}^{\mu}$ (here $\overline{Q}^{\mu}$ is the distribution of Y^{μ}).

Since (5.3) holds $\overline{Q}^{\mu}$-a.s., it also holds $\overline{Q}^{\nu}$-a.s., i.e., with $J_{[n-1,n]}$ replaced by $J^{\nu}_{[n-1,n]}$ which gives the following theorem.

THEOREM 5.1. (version of Theorem 1(a) in Atar and Zeitouni [2]). *Assume that all states of* Λ *communicate, i.e.,* X *is an ergodic Markov chain. Assume* $J_{[0,1]}\beta$ *and* $J_{[0,1]}\nu$ *have positive entries a.s. Then,*

$$\overline{\lim}_{t\to\infty} \frac{1}{t} \log \left\| \pi^{\nu}_t - \pi^{\beta\nu}_t \right\| \le E\big[-1 \vee \log \tau\big(J_{[0,1]}\big)\big] \ \text{a.s.} \tag{5.4}$$

Now the statement of Theorem 4.1 follows from the lemma below.

LEMMA 5.2. *The right-hand side of* (5.4) *is strictly negative.*

Proof. It suffices to show that all entries of $J_{[0,1]}$ are positive a.s. For fixed i, j, we have

$$J_{[0,t]}(i,j) = \delta_{ij} + \int_0^t J_{[0,s]}(i,j)\big[\lambda_{ii}ds + \sigma^{-2}h(a_i)dY^{\mu}_s\big] + \int_0^t \sum_{r\ne i} \lambda_{ri} J_{[0,s]}(r,j)ds.$$

With the help of the Itô formula and with

$$\phi_t(i) = \exp\big\{\lambda_{ii}t + \sigma^{-2}h(a_i)Y^{\mu}_t - (1/2)\sigma^{-2}h^2(a_i)t\big\}$$

we derive

$$\begin{aligned} J_{[0,t]}(j,j) &= \phi_t(j)\Big(1 + \int_0^t \phi_s^{-1}(j) \sum_{r\ne j} \lambda_{rj} J_{[0,s]}(r,j)ds\Big), \\ J_{[0,t]}(i,j) &= \phi_t(i) \int_0^t \phi_s^{-1}(i) \sum_{r\ne i} \lambda_{ri} J_{[0,s]}(r,j)ds, \quad i \ne j. \end{aligned} \tag{5.5}$$

Also notice that the entries of $J_{[0,t]}$ are unnormalized conditional probabilities and so nonnegative a.s. Since all states of Λ communicate, for a pair of indices (i,j) there is a string of indexes $j = i_\ell, \dots, i_1 = i$ such that $\lambda_{i_\ell i_{\ell-1}}, \dots, \lambda_{i_2 i_1} > 0$. So from (5.5), it follows that a.s.

$$\begin{aligned} &J_{[0,t]}(i_\ell, i_\ell) \ge \phi_t(i_\ell) > 0, \\ &J_{[0,t]}(i_{\ell-1}, i_\ell) \ge \phi_t(i_{\ell-1}) \int_0^t \phi_s^{-1}(i_{\ell-1}) \lambda_{i_\ell i_{\ell-1}} J_{[0,s]}(i_\ell, i_\ell)ds > 0, \\ &J_{[0,t]}(i_{\ell-2}, i_\ell) \ge \phi_t(i_{\ell-2}) \int_0^t \phi_s^{-1}(i_{\ell-2}) \lambda_{i_{\ell-1} i_{\ell-2}} J_{[0,s]}(i_{\ell-1}, i_\ell)ds > 0 \end{aligned}$$

for any $t > 0$, and so on until we get $J_{[0,t]}(i_1, i_\ell) > 0$, $t > 0$. □

5.2. The proof of Theorem 4.2. Denote $\rho_{ji}(t) = P\big(X^{\beta}_0 = a_j | \mathscr{Y}^{\beta}_{[0,t]}, X^{\beta}_t = a_i\big)$. If β is a positive distribution, then by Lemma 9.5 in [29, Chap. 9] we have

$$\begin{aligned} \rho_{ji}(0) &= \begin{cases} 1, & j = i, \\ 0, & j \ne i, \end{cases} \\ \frac{d\rho_{ji}(t)}{dt} &= \sum_{r\ne i} \frac{\lambda_{ri}\pi^{\beta}_t(r)}{\pi^{\beta}_t(i)} \big(\rho_{jr}(t) - \rho_{ji}(t)\big), \quad i = 1, \dots, n. \end{aligned} \tag{5.6}$$

Remark 3. By the arguments used in the proof of Lemma 5.2, it can be readily shown that $\pi^\beta_t(i) > 0$ a.s., $i = 1, \dots, n$, for any $t > 0$. Then (5.6) remain valid for $t > t_0$ for any $t_0 > 0$ initialized by

$$\rho_{ji}(t_0) = P\big(X^\beta_0 = a_j | \mathscr{Y}^\beta_{[0,t_0]}, X^\beta_{t_0} = a_i\big).$$

Set $i^\diamond(t) = \operatorname{argmax}_{i\in\mathbb{S}} \rho_{ji}(t)$ and $i_\diamond(t) = \operatorname{argmin}_{i\in\mathbb{S}} \rho_{ji}(t)$ (if the maximum or the minimum is attained at several indices, the lowest one is taken by convention). Set

$$\rho^\diamond(t) := \rho_{ji^\diamond(t)}(t) \quad \text{and} \quad \rho_\diamond(t) := \rho_{ji_\diamond(t)}(t). \tag{5.7}$$

LEMMA 5.3. *The processes $\rho^\diamond(t)$ and $\rho_\diamond(t)$ have absolutely continuous paths with*

$$\begin{aligned} d\rho^\diamond(t) &= \sum_{i=1}^n I(i^\diamond(t) = i)\dot\rho_{ji}(t)dt, \\ d\rho_\diamond(t) &= \sum_{i=1}^n I(i_\diamond(t) = i)\dot\rho_{ji}(t)dt. \end{aligned} \tag{5.8}$$

The proof of this lemma uses two results formulated in Propositions 5.4 and 5.5 below.

PROPOSITION 5.4 (Theorem A.6.3 in Dupuis and Ellis [20]). *Let $g = g(t)$ be an absolutely continuous function mapping of $[0,1]$ into $\mathbb{R}$. Then for each real number a the set $\{t : g(t) = a, \dot g(t) \neq 0\}$ has Lebesgue measure 0.*

PROPOSITION 5.5. *Let $X(t,\omega)$ be a random process with absolutely continuous paths with respect to dt in the sense that there exists a measurable random process $x(t,\omega)$ such that $\int_0^t |x(s,\omega)|ds < \infty$ a.s., $t > 0$, and*

$$X(t,\omega) = X(0,\omega) + \int_0^t x(s,\omega)ds. \tag{5.9}$$

Then

$$|X(t,\omega)| = |X(0,\omega)| + \int_0^t \operatorname{sign}(X(s,\omega))x(s,\omega)ds,$$

where $\operatorname{sign}(0) = 0$.

Proof. Set $V_t(\omega) = \int_0^t |x(s,\omega)|ds$ and notice that for any $t' \le t''$ it holds that

$$\big||X(t'',\omega)| - |X(t',\omega)|\big| \le |X(t'',\omega) - X(t',\omega)| \le (V_{t''}(\omega) - V_{t'}(\omega)).$$

Hence, for fixed ω, the function $|X(t,\omega)|$ possesses bounded total variation for any finite time interval. Denote by $U_t(\omega)$ this total variation corresponding to $[0,t]$. Obviously, $dU_t(\omega) \ll dV_t(\omega) \ll dt$. Recall that $U_t(\omega) = U'_t(\omega) + U''_t(\omega)$, where $U'_t(\omega)$, $U''_t(\omega)$ are increasing continuous in t functions such that for any $t > 0$ and measurable set A from $\mathbb{R}_+$, $\int_{A\cap[0,t]} dU''_s(\omega) = 0$ and $\int_{(\mathbb{R}_+\setminus A)\cap[0,t]} dU'_s(\omega) = 0$, and at the same time $|X(t,\omega)| = U''_t(\omega) - U'_t(\omega)$. Since $dU'_t \ll dU_t(\omega)$, $dU''_t \ll dU_t(\omega)$, it follows that

$d|X(t,\omega)| \ll dU_t(\omega) \ll dV_t(\omega) \ll dt$ and so that

$$|X(t,\omega)| = |X(0,\omega)| + \int_0^t g(s,\omega)ds \tag{5.10}$$

though we may not claim that $g(t,\omega)$ is measurable in (t,ω).

Now, we show that $\operatorname{sign}(X(s,\omega))x(s,\omega)$ is a measurable version of $g(s,\omega)$. By (5.9), we have $X^2(t,\omega) = X^2(0,\omega)+2\int_0^t X(s,\omega)x(s,\omega)ds$. At the same time, by (5.10) it holds that $|X(t,\omega)|^2 = |X(0,\omega)|^2 + 2\int_0^t |X(s,\omega)|g(s,\omega)ds$. Hence, the following identity is valid: For any $t \geq 0$

$$\int_0^t |X(s,\omega)|g(s,\omega)ds \equiv \int_0^t X(s,\omega)x(s,\omega)ds.$$

Therefore, $|X(s,\omega)|g(s,\omega) = X(s,\omega)x(s,\omega)$ for almost all s with respect to Lebesgue measure. Consequently, we have $I(|X(s,\omega)| \neq 0)g(s,\omega) = \operatorname{sign}(X(s,\omega))x(s,\omega)$ for almost all s with respect to Lebesgue measure. It remains to show that

$$I(X(s,\omega) = 0)g(s,\omega) = 0$$

for almost all s with respect to Lebesgue measure. Taking into account (5.10), it suffices to prove that $\int_0^\infty I(X(s,\omega) = 0)d|X(s,\omega)| = 0$ a.s. On the other hand, whereas $d|X(t,\omega)| \ll dV_t(\omega)$, it suffices to show that $\int_0^\infty I(X(s,\omega) = 0)dV_s(\omega) = 0$ a.s. The latter holds by Proposition 5.4. □

Now we give the proof for Lemma 5.3.

Proof. Let us introduce $\rho^{\diamond,i}(t) = \rho_{j1} \vee \rho_{j2} \vee \cdots \vee \rho_{ji}$ and $\rho_{\diamond,i}(t) = \rho_{j1} \wedge \rho_{j2} \wedge \cdots \wedge \rho_{ji}$ and notice that $\rho^{\diamond,n}(t) = \rho^{\diamond}(t)$, $\rho_{\diamond,n}(t) = \rho_{\diamond}(t)$.

The use of obvious identities

$$\begin{aligned}
\rho^{\diamond,2}(t) + \rho_{\diamond,2}(t) &= \rho_{j1}(t) + \rho_{j2}(t),\\
\rho^{\diamond,2}(t) - \rho_{\diamond,2}(t) &= |\rho_{j1}(t) - \rho_{j2}(t)|
\end{aligned}$$

and the fact, provided by Proposition 5.5, that $d|\rho_{j1}(t) - \rho_{j2}(t)| = p(t,\omega)dt$ with measurable derivative $p(\omega,t)$, allow us to claim that $\rho^{\diamond,2}(t)$ and $\rho_{\diamond,2}(t)$ are absolutely continuous with respect to dt with measurable derivatives.

Further, taking into account $\rho^{\diamond,i}(t) = \rho^{\diamond,i-1}(t) \vee \rho_{ji}$ and $\rho_{\diamond,i}(t) = \rho_{\diamond,i-1}(t) \wedge \rho_{ji}(t)$ and consequent identities

$$\begin{aligned}
\rho^{\diamond,i}(t) + \rho^{\diamond,i-1}(t) \wedge \rho_{ji}(t) &= \rho^{\diamond,i-1}(t) + \rho_{ji}(t),\\
\rho^{\diamond,i}(t) - \rho^{\diamond,i-1}(t) \wedge \rho_{ji}(t) &= |\rho^{\diamond,i-1}(t) - \rho_{ji}(t)|,\\
\rho_{\diamond,i-1}(t) \vee \rho_{ji}(t) + \rho_{\diamond,i}(t) &= \rho_{\diamond,i-1}(t) + \rho_{ji}(t),\\
\rho_{\diamond,i-1}(t) \vee \rho_{ji}(t) - \rho_{\diamond,i}(t) &= |\rho_{\diamond,i-1}(t) - \rho_{ji}(t)|,
\end{aligned}$$

absolute continuity for $\rho^{\diamond}(t)$ and $\rho_{\diamond}(t)$ is verified by the induction method.

Thus, $d\rho^{\diamond}(t) = u(t)dt$ with some density $u(t)$ such that $\int_0^t |u(s)|ds < \infty$ a.s., $t > 0$. On the other hand, since $\sum_{i=1}^n I(i^{\diamond}(t) = i) = 1$, we have

$$\rho^{\diamond}(t) = \rho^{\diamond}(0) + \int_0^t \sum_{i=1}^n I(i^{\diamond}(s) = i)u(s)ds.$$

So, it suffices to show that for any $t > 0$ and any $i = 1, 2, \dots, n$

$$\int_0^t I(i^\diamond(s) = i)|u(s) - \dot\rho_{ji}(s)|ds = 0 \text{ a.s.}$$

The latter holds true by Proposition 5.4, since

$$\begin{aligned}
&\int_0^t I(i^\diamond(s) = i)|u(s) - \dot\rho_{ji}(s)|ds \\
&= \int_0^t I(\rho^\diamond(s) - \rho_{ji}(s) = 0)|u(s) - \dot\rho_{ji}(s)|ds \\
&= \int_0^t I\big(\rho^\diamond(s) - \rho_{ji}(s) = 0, u(s) - \dot\rho_{ji}(s) \neq 0\big)|u(s) - \dot\rho_{ji}(s)|ds = 0. \qquad \square
\end{aligned}$$

LEMMA 5.6. *Under the assumptions of Theorem* 4.2,

$$\overline{\lim}_{t\to\infty} \frac{1}{t} \log \max_{1\le j,k,\ell\le n} \big|\rho_{jk}(t) - \rho_{j\ell}(t)\big| \le -\sum_{r=1}^n \mu_r \min_{i\neq r} \lambda_{ri}. \tag{5.11}$$

Proof. By (5.6) and (5.8), we have[2]

$$\begin{aligned}
\frac{d\rho_\diamond(t)}{dt} &= \sum_{r\neq i_\diamond(t)} \frac{\lambda_{ri_\diamond(t)}\pi_t^\beta(r)}{\pi_t^\beta(i_\diamond(t))}\big(\rho_{jr}(t) - \rho_\diamond(t)\big), \\
\frac{d\rho^\diamond(t)}{dt} &= \sum_{r\neq i^\diamond(t)} \frac{\lambda_{ri^\diamond(t)}\pi_t^\beta(r)}{\pi_t^\beta(i^\diamond(t))}\big(\rho_{jr}(t) - \rho^\diamond(t)\big).
\end{aligned} \tag{5.12}$$

In what follows, we will omit the time variable in $i_\diamond(t)$ and $i^\diamond(t)$ for brevity. Set $\triangle_t = \rho^\diamond(t) - \rho_\diamond(t)$. By (5.12) we have

$$\begin{aligned}
\frac{d\triangle_t}{dt} &= -\sum_{r\neq i^\diamond} \frac{\lambda_{ri^\diamond}\pi_t^\beta(r)}{\pi_t^\beta(i^\diamond)}\big(\rho^\diamond(t) - \rho_{jr}(t)\big) - \sum_{r\neq i_\diamond} \frac{\lambda_{ri_\diamond}\pi_t^\beta(r)}{\pi_t^\beta(i_\diamond)}\big(\rho_{jr}(t) - \rho_\diamond(t)\big) \\
&= -\triangle_t\left(\frac{\lambda_{i_\diamond i^\diamond}\pi_t^\beta(i_\diamond)}{\pi_t^\beta(i^\diamond)} + \frac{\lambda_{i^\diamond i_\diamond}\pi_t^\beta(i^\diamond)}{\pi_t^\beta(i_\diamond)}\right) \\
&\quad -\triangle_t\left(\sum_{\substack{r\neq i^\diamond(t)\\ r\neq i_\diamond(t)}}\left[\frac{\lambda_{ri^\diamond}\pi_t^\beta(r)}{\pi_t^\beta(i^\diamond)}\left(\frac{\rho^\diamond(t) - \rho_{jr}(t)}{\triangle_t}\right) + \frac{\lambda_{ri_\diamond}\pi_t^\beta(r)}{\pi_t^\beta(i_\diamond)}\left(\frac{\rho_{jr}(t) - \rho_\diamond(t)}{\triangle_t}\right)\right]\right).
\end{aligned} \tag{5.13}$$

Letting $0/0 = 1/2$, set $\alpha_r(t) = \frac{\rho^\diamond(t) - \rho_{jr}(t)}{\triangle_t}$. Then, we get $1 - \alpha_r(t) = \frac{\rho_{jr}(t) - \rho_\diamond(t)}{\triangle_t}$

[2] In (5.12)–(5.14) we use for brevity a form of differential equalities (inequalities) which are valid for any ω and almost all t with respect to Lebesgue measure.

and $0 \le \alpha_r(t) \le 1$ and (5.13) implies

$$
\begin{aligned}
\frac{d\triangle_t}{dt} &= -\triangle_t \left(\frac{\lambda_{i_\diamond i^\diamond} \pi_t^\beta(i_\diamond)}{\pi_t^\beta(i^\diamond)} + \frac{\lambda_{i^\diamond i_\diamond} \pi_t^\beta(i^\diamond)}{\pi_t^\beta(i_\diamond)} \right) \\
&\quad - \triangle_t \left(\sum_{\substack{r \neq i^\diamond(t) \\ r \neq i_\diamond(t)}} \left[\alpha_r(t) \frac{\lambda_{ri^\diamond} \pi_t^\beta(r)}{\pi_t^\beta(i^\diamond)} + \big(1 - \alpha_r(t)\big) \frac{\lambda_{ri_\diamond} \pi_t^\beta(r)}{\pi_t^\beta(i_\diamond)} \right] \right) \\
&\le -\triangle_t \Big(\lambda_{i_\diamond i^\diamond} \pi_t^\beta(i_\diamond) + \lambda_{i^\diamond i_\diamond} \pi_t^\beta(i^\diamond) \Big) \\
&\quad - \triangle_t \left(\sum_{\substack{r \neq i^\diamond(t) \\ r \neq i_\diamond(t)}} \Big[\alpha_r(t) \lambda_{ri^\diamond} + \big(1 - \alpha_r(t)\big) \lambda_{ri_\diamond} \Big] \pi_t^\beta(r) \right) \\
&\le -\triangle_t \left(\lambda_{i_\diamond i^\diamond} \pi_t^\beta(i_\diamond) + \lambda_{i^\diamond i_\diamond} \pi_t^\beta(i^\diamond) + \sum_{\substack{r \neq i^\diamond(t) \\ r \neq i_\diamond(t)}} \Big[\lambda_{ri^\diamond} \wedge \lambda_{ri_\diamond} \Big] \pi_t^\beta(r) \right).
\end{aligned} \tag{5.14}
$$

Recall that all offdiagonal entries of Λ are nonnegative and $\sum_{r=1}^n \lambda_{ir} = 0$ for any i. Then, $\big|\lambda_{i_\diamond i^\diamond}\big| \wedge \big|\lambda_{i_\diamond i_\diamond}\big| \ge \lambda_{i_\diamond i^\diamond}$, $\big|\lambda_{i^\diamond i^\diamond}\big| \wedge \big|\lambda_{i^\diamond i_\diamond}\big| \ge \lambda_{i^\diamond i_\diamond}$, and (5.14) provides

$$
\begin{aligned}
\frac{d\triangle_t}{dt} &\le -\triangle_t \sum_{r=1}^n \Big(|\lambda_{ri^\diamond}| \wedge |\lambda_{ri_\diamond}| \Big) \pi_t^\beta(r) \le -\triangle_t \sum_{r=1}^n \min_{1 \le i \le n} |\lambda_{ri}| \pi_t^\beta(r) \\
&= -\triangle_t \sum_{r=1}^n \pi_t^\beta(r) \min_{i \neq r} \lambda_{ri}.
\end{aligned}
$$

Since the derivative $\frac{d\triangle_t}{dt}$ is defined for each ω and almost everywhere (a.e.) in t with respect to dt, the above inequality $\frac{d\triangle_t}{dt} \le -\triangle_t \sum_{r=1}^n \pi_t^\beta(r) \min_{i \neq r} \lambda_{ri}$ is also valid a.e. So, it allows us to define a.e. the function

$$
H(t) = -\triangle_t \sum_{r=1}^n \pi_t^\beta(r) \min_{i \neq r} \lambda_{ri} - \frac{d\triangle_t}{dt}.
$$

Moreover, for definiteness, we may redefine $H(t)$ everywhere so as $H(t) \ge 0$. Then we have

$$
d\triangle_t = -\left[\triangle_t \sum_{r=1}^n \pi_t^\beta(r) \min_{i \neq r} \lambda_{ri} + H(t) \right] dt.
$$

Notice also that $\int_0^t |H(s)| ds < \infty$ a.s. for any $t > 0$ and recall that $\triangle_0 = 1$. Then, we get

$$
\triangle_t = \exp\left(-\int_0^t \sum_{r=1}^n \pi_s^\beta(r) \min_{i \neq r} \lambda_{ri} ds \right) - \int_0^t \exp\left(-\int_s^t \sum_{r=1}^n \pi_u^\beta(r) \min_{i \neq r} \lambda_{ri} du \right) H(s) ds
$$

and in turn

$$
\frac{1}{t} \log \triangle_t \le -\sum_{r=1}^n \Big(\min_{i \neq r} \lambda_{ri} \Big) \frac{1}{t} \int_0^t \pi_s^\beta(r) ds.
$$

So, it is left to verify that

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t \pi_s^\beta(r)ds=\mu_r \quad \text{a.s.} \tag{5.15}$$

Similarly to (1.4), π_t^β satisfies

$$\begin{aligned}\pi_0^\beta =&\beta,\\ d\pi_t^\beta =&\Lambda^*\pi_t^\beta dt+\sigma^{-2}\big(\operatorname{diag}(\pi_t^\beta)-\pi_t^\beta(\pi_t^\beta)^*\big)h(dY_t^\beta-h^*\pi_t^\beta dt).\end{aligned}$$

Recall that $\sigma^{-1}\big(Y_t^\beta-\int_0^t h^*\pi_s^\beta ds\big)$ is the innovation Wiener process (see, e.g., Theorem 9.1 in Chapter 10 in [30]). Hence $M_t=\int_0^t\big(\operatorname{diag}(\pi_s^\beta)-\pi_s^\beta(\pi_s^\beta)^*\big)h(dY_s^\beta-h^*\pi_s^\beta ds)$ is a vector-valued continuous martingale. Its entries $M_t(i)$, $i=1,\dots,n$, have predictable quadratic variation processes $\langle M(i)\rangle_t$ with the following property: For some positive constant c, $d\langle M(i)\rangle_t\le cdt$. Then by Theorem 10, Chapter 3 in [31], $\lim_{t\to\infty}\frac{1}{t}M_t(i)=0$ a.s. This fact and the boundedness of π_t^β provide $\Lambda^*\lim_{t\to\infty}\frac{1}{t}\int_0^t\pi_s^\beta ds=0$. The vector $Z_t=\frac{1}{t}\int_0^t\pi_s^\beta ds$ has nonnegative entries, whose sum equals 1. Therefore the limit vector Z_∞, obeying the same property, is the unique solution of the linear algebraic equation $\Lambda^*Z_\infty=0$, i.e., $Z_\infty=\mu$. □

To prove Theorem 4.2, without loss generality, due to Remark 3, we may assume that $\nu\sim\beta$. Then, we show that for any $t\ge 0$ and $i=1,\dots,n$

$$\big|\pi_t^\nu(i)-\pi_t^{\beta\nu}(i)\big|\le n\max_j\frac{d\nu}{d\beta}(a_j)\max_j\frac{d\beta}{d\nu}(a_j)\max_{1\le i,j,k\le d}\big|\rho_{ji}(t)-\rho_{jk}(t)\big|. \tag{5.16}$$

Recall that Q^ν and Q^β are distributions of (X^ν,Y^ν) and (X^β,Y^β), respectively, which are equivalent, by virtue of $\nu\sim\beta$, with

$$\frac{dQ^\beta}{dQ^\nu}(X^\nu,Y^\nu)\equiv\frac{d\beta}{d\nu}(X_0^\nu)\quad\text{and}\quad\frac{dQ^\nu}{dQ^\beta}(X^\beta,Y^\beta)\equiv\frac{d\nu}{d\beta}(X_0^\beta).$$

Now, we show that for any $i=1,\dots,d$ and $t>0$, Q^ν- and Q^β-a.s.

$$\pi_t^{\beta\nu}(i)=\frac{\sum_{j=1}^n\Big(\frac{d\beta}{d\nu}(a_j)P\big(X_0^\nu=a_j),X_t^\nu=a_i|\mathscr{Y}_{[0,t]}^\nu\Big)}{E\Big(\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}_{[0,t]}^\nu\Big)}. \tag{5.17}$$

To this end, with any bounded $\mathscr{D}_t^y$-measurable function $\psi_t(y)$, write

$$\begin{aligned}&E\psi_t(Y^\nu)\pi_t^{\beta\nu}(i)E\Big(\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}_{[0,t]}^\nu\Big)=E\psi_t(Y^\nu)\pi_t^{\beta\nu}(i)\frac{d\beta}{d\nu}(X_0^\nu)\\ &=E\psi_t(Y^\nu)\pi_t^{\beta\nu}(i)\frac{dQ^\beta}{dQ^\nu}(X^\nu,Y^\nu)=E\psi_t(Y^\beta)\pi_t^\beta(i)\\ &=E\psi_t(Y^\beta)I(X_t^\beta=a_i)=E\psi_t(Y^\nu)I(X_t^\nu=a_i)\frac{dQ^\beta}{dQ^\nu}(X^\nu,Y^\nu)\\ &=E\psi_t(Y^\nu)I(X_t^\nu=a_i)\frac{d\beta}{d\nu}(X_0^\nu)=E\psi_t(Y^\nu)E\Big(I(X_t^\nu=a_i)\frac{d\beta}{d\nu}(X_0^\nu)\big|\mathscr{Y}_{[0,t]}^\nu\Big).\end{aligned}$$

Hence, by the arbitrariness of $\psi_t(y)$,

$$\pi_t^{\beta\nu}(i)E\Big(\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}_{[0,t]}^\nu\Big)=E\left(I(X_t^\nu=a_i)\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}_{[0,t]}^\nu\right).$$

Further, $Q^\nu \sim Q^\beta$ provides $E\big(\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}^\nu_{[0,t]}\big) > 0$, Q^ν- and Q^β-a.s., so that

$$\pi_t^{\beta\nu}(i) = \frac{E\big(I(X_t^\nu = a_i)\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}^\nu_{[0,t]}\big)}{E\Big(\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}^\nu_{[0,t]}\Big)}$$

and it remains to notice that

$$E\left(I(X_t^\nu = a_i)\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}^\nu_{[0,t]}\right) = \sum_{j=1}^n \frac{d\beta}{d\nu}(a_j)P\big(X_t^\nu = a_i, X_0^\nu = a_j|\mathscr{Y}^\nu_{[0,t]}\big).$$

Taking into consideration (5.17), we find

$$\begin{aligned}\big|\pi_t^\nu(i) - \pi_t^{\beta\nu}(i)\big| &= \left|\pi_t^\nu(i) - \frac{\sum_{j=1}^n \Big(\frac{d\beta}{d\nu}(a_j)P\big(X_0^\nu = a_j, X_t^\nu = a_i|\mathscr{Y}^\nu_{[0,t]}\big)\Big)}{E\Big(\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}^\nu_{[0,t]}\Big)}\right| \\ &= \frac{\Big|\sum_{j=1}^n \frac{d\beta}{d\nu}(a_j)\Big(\pi_t^\nu(i)P\big(X_0^\nu = a_j|\mathscr{Y}^\nu_{[0,t]}\big) - P\big(X_0^\nu = a_j, X_t^\nu = a_i|\mathscr{Y}^\nu_{[0,t]}\big)\Big)\Big|}{E\Big(\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}^\nu_{[0,t]}\Big)}.\end{aligned}$$

Then, since by the Jensen inequality $1/E\big(\frac{d\beta}{d\nu}(X_0^\nu)|\mathscr{Y}^\nu_{[0,t]}\big) \le E\big(\frac{d\nu}{d\beta}(X_0^\nu)|\mathscr{Y}^\nu_{[0,t]}\big)$, we get the chain of estimates

$$\begin{aligned}\big|\pi_t^\nu(i) - \pi_t^{\beta\nu}(i)\big| &\le \max_{a_j\in\mathbb{S}}\frac{d\beta}{d\nu}(a_j)\max_{a_j\in\mathbb{S}}\frac{d\nu}{d\beta}(a_j) \\ &\quad\times\left|\sum_{j=1}^n \pi_t^\nu(i)\Big(P\big(X_0^\nu = a_j|\mathscr{Y}^\nu_{[0,t]}\big) - P\big(X_0^\nu = a_j\big|X_t^\nu = a_i, \mathscr{Y}^\nu_{[0,t]}\big)\Big)\right| \\ &\le \max_{a_j\in\mathbb{S}}\frac{d\beta}{d\nu}(a_j)\max_{a_j\in\mathbb{S}}\frac{d\nu}{d\beta}(a_j) \\ &\quad\times\sum_{j=1}^n \pi_t^\nu(i)\Big|P\big(X_0^\nu = a_j|\mathscr{Y}^\nu_{[0,t]}\big) - P\big(X_0^\nu = a_j|X_t^\nu = a_i, \mathscr{Y}^\nu_{[0,t]}\big)\Big| \qquad (5.18) \\ &\le \max_{a_j\in\mathbb{S}}\frac{d\beta}{d\nu}(a_j)\max_{j\in\mathbb{S}}\frac{d\nu}{d\beta}(a_j) \\ &\quad\times\sum_{j=1}^n \Big|P\big(X_0^\nu = a_j|\mathscr{Y}^\nu_{[0,t]}\big) - P\big(X_0^\nu = a_j|X_t^\nu = a_i, \mathscr{Y}^\nu_{[0,t]}\big)\Big| \\ &= \max_{a_j\in\mathbb{S}}\frac{d\beta}{d\nu}(a_j)\max_{a_j\in\mathbb{S}}\frac{d\nu}{d\beta}(a_j)\sum_{j=1}^n \Big|P\big(X_0^\nu = a_j|\mathscr{Y}^\nu_{[0,t]}\big) - \rho_{ji}(t)\Big|.\end{aligned}$$

The obvious formula $P\big(X_0^\nu = a_j|\mathscr{Y}^\nu_{[0,t]}\big) = \sum_{k=1}^n \pi_t^\nu(k)\rho_{jk}(t)$, and (5.18) provide

$$\begin{aligned}\big|\pi_t^\nu(i) - \pi_t^{\beta\nu}(i)\big| &\le \max_{a_j\in\mathbb{S}}\frac{d\beta}{d\nu}(a_j)\max_{a_j\in\mathbb{S}}\frac{d\nu}{d\beta}(a_j)\sum_{j=1}^n \Big|\sum_{k=1}^n \pi_t^\nu(k)\rho_{jk}(t) - \rho_{ji}(t)\Big| \\ &\le \max_{a_j\in\mathbb{S}}\frac{d\beta}{d\nu}(a_j)\max_{a_j\in\mathbb{S}}\frac{d\nu}{d\beta}(a_j)\sum_{j=1}^n\sum_{k=1}^n \pi_t^\nu(k)\big|\rho_{jk}(t) - \rho_{ji}(t)\big| \qquad (5.19)\end{aligned}$$

and (5.16). Thus, by Lemma 5.6, the desired statement (4.1) holds true.

5.3. The proof of Theorem 4.3. We start with the following lemma.

LEMMA 5.7. *Under the assumptions of Theorem* 4.3*, for any* $t > 0$

$$\max_{1\le j,k,\ell\le n}\big|\rho_{jk}(t)-\rho_{j\ell}(t)\big| \le \exp\Big(-2t\min_{p\neq q}\sqrt{\lambda_{pq}\lambda_{qp}}\Big). \tag{5.20}$$

Proof. Here we follow the notations from Lemma 5.6. From (5.14), it follows that

$$\frac{d\triangle_t}{dt} \le -\triangle_t\left(\frac{\lambda_{i_\diamond i^\diamond}\pi^\beta_t(i_\diamond)}{\pi^\beta_t(i^\diamond)}+\frac{\lambda_{i^\diamond i_\diamond}\pi^\beta_t(i^\diamond)}{\pi^\beta_t(i_\diamond)}\right) \tag{5.21}$$

subject to $\triangle_0 = 1$. Set $\tau = \inf\{t : i^\diamond(t) = i_\diamond(t)\}$. Since $\triangle_t$ is a nonincreasing function, $\triangle_t \equiv 0$ for $t \ge \tau$, and (5.20) holds trivially. For $t < \tau$, as previously we find

$$\begin{aligned}\triangle_t &\le \exp\left\{-\int_0^t\left(\frac{\lambda_{i_\diamond i^\diamond}\pi^\beta_s(i_\diamond)}{\pi^\beta_s(i^\diamond)}+\frac{\lambda_{i^\diamond i_\diamond}\pi^\beta_s(i^\diamond)}{\pi^\beta_s(i_\diamond)}\right)ds\right\}\\ &\le \exp\left\{-\int_0^t\min_{x\ge0}\left(\lambda_{i_\diamond i^\diamond}x+\lambda_{i^\diamond i_\diamond}\frac{1}{x}\right)ds\right\}\\ &= \exp\left\{-\int_0^t 2\sqrt{\lambda_{i_\diamond i^\diamond}\lambda_{i^\diamond i_\diamond}}ds\right\} \le \exp\Big(-2t\min_{p\neq q}\sqrt{\lambda_{pq}\lambda_{qp}}\Big),\end{aligned}$$

and (5.20) follows. □

To prove the first statement of the theorem, taking into account $\nu \ll \beta$ we replicate a fragment from the proof of Proposition 2.1.

Using the notations introduced in section 2.1, write $\pi^\nu_t(i) := \pi^\nu_t(f)$ and $\pi^{\beta\nu}_t(i) := \pi^{\beta\nu}_t(f)$ for $f(x) = I(x = a_i)$. Then,

$$E\big|\pi^{\beta\nu}_t(i)-\pi^\nu_t(i)\big| \le E\Big|E\Big(\frac{d\nu}{d\beta}(X^\beta_0)\big|\mathscr{Y}^\beta_{[0,t]}\Big)-E\Big(\frac{d\nu}{d\beta}(X^\beta_0)\big|\mathscr{Y}^\beta_{[0,\infty)}\vee\mathscr{X}^\beta_{[t,\infty)}\Big)\Big| \tag{5.22}$$

and, since (X^β, Y^β) is a Markov process,

$$E\Big(\frac{d\nu}{d\beta}(X^\beta_0)\big|\mathscr{Y}^\beta_{[0,\infty)}\vee\mathscr{X}^\beta_{[t,\infty)}\Big) = E\Big(\frac{d\nu}{d\beta}(X^\beta_0)\big|\mathscr{Y}^\beta_{[0,t]}\vee\mathscr{X}^\beta_t\Big).$$

Then,

$$\begin{aligned}&E\Big(\frac{d\nu}{d\beta}(X^\beta_0)\big|\mathscr{Y}^\beta_{[0,t]}\Big)-E\Big(\frac{d\nu}{d\beta}(X^\beta_0)\big|\mathscr{Y}^\beta_{[0,\infty)}\vee\mathscr{X}^\beta_{[t,\infty)}\Big)\\ &=\sum_{j=1}^n\frac{d\nu}{d\beta}(a_j)\Big(P\big(X^\beta_0=a_j|\mathscr{Y}^\beta_{[0,t]}\big)-P\big(X^\beta_0=a_j|\mathscr{Y}^\beta_{[0,t]}\vee\mathscr{X}^\beta_t\big)\Big)\\ &=\sum_{j=1}^n\sum_{\ell=1}^n I(X^\beta_t=a_\ell)\frac{d\nu}{d\beta}(a_j)\Big(P\big(X^\beta_0=a_j|\mathscr{Y}^\beta_{[0,t]}\big)-\rho_{j\ell}(t)\Big) \qquad (5.23)\\ &=\sum_{j=1}^n\sum_{\ell=1}^n\sum_{k=1}^n\pi^\beta_t(k)I(X^\beta_t=a_\ell)\frac{d\nu}{d\beta}(a_j)\big(\rho_{jk}(t)-\rho_{j\ell}(t)\big)\\ &\le\max_{1\le j,k,\ell\le n}\big|\rho_{jk}(t)-\rho_{j\ell}(t)\big|\sum_{j=1}^n\frac{d\nu}{d\beta}(a_j).\end{aligned}$$

The first statement of Theorem 4.3 follows from (5.22), (5.23), and Lemma 5.7. The second statement follows from (5.16) and Lemma 5.7.

6. Proofs for the nonergodic case. Recall that in the nonergodic setting under consideration

$$\mathbb{S} = \Big\{ \underbrace{a_1^1, \dots, a_{n_1}^1}_{\mathbb{S}_1}, \dots, \underbrace{a_1^m, \dots, a_{n_m}^m}_{\mathbb{S}_m} \Big\}, \quad m \geq 2,$$

with subalphabets $\mathbb{S}_1, \dots, \mathbb{S}_m$ noncommunicating in the sense of (1.2).

6.1. Auxiliary lemmas. In this subsection, $\widetilde{X}_t^j$ is an independent copy of X_t^j with the initial distribution μ^j, defined on some auxiliary probability space $(\widetilde{\Omega}, \widetilde{\mathscr{F}}, \widetilde{P})$, and $\widetilde{E}$ is the expectation with respect to $\widetilde{P}$. Recall that μ^j is the invariant measure, so that $\widetilde{X}_t^j$ is a stationary process.

LEMMA 6.1. *Fix $r > 0$ and define $Z_n = \sum_{i=1}^n \big(Y_{ir}^\beta - Y_{(i-1)r}^\beta\big)^2$. Then with $n \to \infty$*

$$\frac{1}{n} Z_n \to r + \sum_{j=1}^m I(X_0^\beta \in \mathbb{S}_j) \widetilde{E} \left(\int_0^r h(\widetilde{X}_s^j)\, ds \right)^2.$$

Proof. Define

$$F(i) = E \left[\left(\int_0^r h(X_s^\beta)\, ds \right)^2 \Big| X_0^\beta = a_i \right]$$

and $\mathscr{G}_n = \sigma\{Y_{[0,nr]}\} \vee \sigma\{X_{[0,nr]}\}$. Then $E[(Y_{(n+1)r}^\beta - Y_{nr}^\beta)^2 | \mathscr{G}_n] = r + F(X_{nr}^\beta)$ so that the sequence $M_n = Z_n - nr - \sum_{i=0}^{n-1} F(X_{ir}^\beta)$ is a martingale with respect to the filtration $(\mathscr{G}_n)_{n \geq 1}$. It is easy to verify that there exists $K < \infty$ such that for all n we have $E(M_{n+1} - M_n)^2 \leq K$. It follows that $(1/n)M_n \to 0$ a.s. as $n \to \infty$ (see, e.g., Chapter VII, Section 5, Theorem 4 in [42]).

Now consider $(1/n) \sum_{i=0}^{n-1} F(X_{ir}^\beta)$. If $X_0 \in \mathbb{S}_j$, then $X_t \in \mathbb{S}_j$ for all $t \geq 0$ and the process is ergodic in $\mathbb{S}_j$ with stationary distribution μ^j. Applying the ergodic theorem for each class $\mathbb{S}_j$ we obtain

$$\frac{1}{n} \sum_{i=0}^{n-1} F(X_{ir}^\beta) \to \sum_{j=1}^m \widetilde{E}(F(\widetilde{X}_0)) I(X_0 \in \mathbb{S}_j) = \sum_{j=1}^m \widetilde{E} \left(\int_0^r h(\widetilde{X}_s^j)\, ds \right)^2 I(X_0^\beta \in \mathbb{S}_j)$$

as $n \to \infty$ a.s. Finally

$$\begin{aligned} \lim_{n \to \infty} \frac{1}{n} Z_n &= \lim_{n \to \infty} \frac{1}{n} M_n + r + \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} F(X_{ir}^\beta) \\ &= r + \sum_{j=1}^m \widetilde{E} \left(\int_0^r h(\widetilde{X}_s^j)\, ds \right)^2 I(X_0^\beta \in \mathbb{S}_j) \end{aligned}$$

and we are done. □

With $\widetilde{X}_t^j$ defined as in Lemma 6.1 and $r \geq 0$ let $d_j(r) = \widetilde{E}(\int_0^r h(\widetilde{X}_s^j)\, ds)^2$.

LEMMA 6.2. *For any $k \neq j$ the following are equivalent:*

i. *$d_k(r) = d_j(r)$ for all $r \geq 0$;*

ii. *$h_k^* \operatorname{diag}(\mu_k) \Lambda_k^q h_k = h_j^* \operatorname{diag}(\mu_j) \Lambda_j^q h_j$ for all $0 \leq q \leq n_i + n_j - 1$.*

Proof. Notice first that

$$d_j(r) = 2\widetilde{E}\int_0^r\int_0^s h(\widetilde{X}^j_u)h(\widetilde{X}^j_s)\,du\,ds = 2\int_0^r\int_0^s \widetilde{E}h(\widetilde{X}^j_u)h(\widetilde{X}^j_s)\,du\,ds$$
$$= 2\int_0^r\int_0^s \widetilde{E}h(\widetilde{X}^j_0)h(\widetilde{X}^j_{s-u})\,du\,ds = 2\int_0^r\int_0^s \widetilde{E}h(\widetilde{X}^j_0)h(\widetilde{X}^j_v)\,dv\,ds.$$

Now, introduce the vector $\widetilde{I}^j_t$ with entries $I(\widetilde{X}^j_t = a^j_1), \dots, I(\widetilde{X}^j_t = a^j_{n_j})$ and notice also that

$$\widetilde{E}h(\widetilde{X}^j_0)h(\widetilde{X}^j_v) = \widetilde{E}h^*_j\widetilde{I}^j_0(\widetilde{I}^j_v)^*h_j = \widetilde{E}h^*_j\widetilde{I}^j_0(\widetilde{I}^j_0)^*e^{\Lambda_j v}h_j$$
$$= h^*_j\widetilde{E}\,\mathrm{diag}(\widetilde{I}^j_0)e^{\Lambda_j v}h_j = h^*_j\,\mathrm{diag}(\mu^j)e^{\Lambda_j v}h_j.$$

Therefore $d_j(r) = 2\int_0^r\int_0^s h^*_j\,\mathrm{diag}(\mu^j)e^{\Lambda_j v}h_j dv ds$, so $d_j(0) = d_j{}'(0) = 0$ and

$$d_j{}''(r) = 2h^*_j\,\mathrm{diag}(\mu^j)e^{\Lambda_j r}h_j.$$

Differentiating with respect to r a further q times and then putting $r = 0$ we get

$$d_j^{(2+q)}(0) = 2h^*_j\,\mathrm{diag}(\mu^j)\Lambda^q_j h_j.$$

It follows immediately that if $d_k(r) = d_j(r)$ for all $r \geq 0$, then

$$h^*_k\,\mathrm{diag}(\mu^k)\Lambda^q_k h_k = h^*_j\,\mathrm{diag}(\mu^j)\Lambda^q_j h_j$$

for all $q \geq 0$ and so in particular for all $0 \leq q \leq n_k + n_j - 1$.

Suppose conversely that $h^*_j\,\mathrm{diag}(\mu^j)\Lambda^q_j h_j = h^*_k\,\mathrm{diag}(\mu^k)\Lambda^q_k h_k$ for all $0 \leq q \leq n_k + n_j - 1$. The Cayley–Hamilton theorem applied to the $(n_k+n_j)\times(n_k+n_j)$ block diagonal matrix $\left(\begin{smallmatrix}\Lambda_k & 0\\ 0 & \Lambda_j\end{smallmatrix}\right)$ gives constants $c_0, c_1, \dots, c_{n_k+n_j-1}$ so that

$$\Lambda^{n_k+n_j}_k = \sum_{q=0}^{n_k+n_j-1} c_q\Lambda^q_k \quad\text{and}\quad \Lambda^{n_i+n_j}_j = \sum_{q=0}^{n_k+n_j-1} c_q\Lambda^q_j.$$

Therefore we have $h^*_k\,\mathrm{diag}(\mu^k)\Lambda^q_k h_k = h^*_j\,\mathrm{diag}(\mu^j)\Lambda^q_j h_j$ for all $q > n_j + n_k - 1$ as well. Using the fact that $e^{\Lambda_j r} = \sum_{q=0}^\infty \frac{r^q\Lambda^q_j}{q!}$, we see that $d_k{}''(r) = d_j{}''(r)$ for all $r \geq 0$, and hence $d_k(r) = d_j(r)$ for all $r \geq 0$. ∎

LEMMA 6.3. *Assume* A-2. *For any* β

$$\lim_{t\to\infty} E\Big|P\big(X^\beta_0 \in \mathbb{S}_j|\mathscr{Y}^\beta_{[0,t]}\big) - I(X^\beta_0 \in \mathbb{S}_j)\Big| = 0,\ j \geq 1.$$

Proof. We use the notation $Z^{(r)}_n$ to express the dependence on r of the function Z_n in Lemma 6.1. We have $\frac{1}{n}Y^\beta_n \to \sum_{j=1}^m h^*_j\mu^j I(X^\beta_0 \in \mathbb{S}_j)$ and

$$\frac{1}{n}Z^{(r)}_n \to r + \sum_{j=1}^m d_j(r)I(X^\beta_0 \in \mathbb{S}_j)$$

as $n \to \infty$ a.s. Using assumption A-2 and Lemma 6.2 we can find an integer ℓ and numbers $r_i > 0, i = 1, \dots, \ell$, and construct a random variable of the form $V_n = (Y^\beta_n, Z^{(r_1)}_n - nr_1, \dots, Z^{(r_\ell)}_n - nr_\ell)$ so that $\frac{1}{n}V_n \to \sum_{j=1}^m v_j I(X^\beta_0 \in \mathbb{S}_j)$ as $n \to \infty$, P-a.s., where the $v_1, \dots, v_m$ are distinct vectors in $\mathbb{R}^{\ell+1}$. Therefore $\{X^\beta_0 \in \mathbb{S}_j\}$ is $Y^\beta_{[0,\infty)}$-measurable a.s. and the result follows immediately. ∎

6.2. The proof of Theorem 4.4. By Proposition 2.1, it suffices to show that

$$\lim_{t\to\infty} E\big\|\pi_t^{\beta} - \pi_t^{\beta_0}\big\| = 0.$$

We introduce a new filter, intermediate between π_t^{β} and $\pi_t^{\beta_0}$. Define the random variable U by $U = j$ on the set $\{X_0^{\beta} \in \mathbb{S}_j\}$, and then define

$$\pi_t^{\beta,U}(i) = P(X_t^{\beta} = a_i | \mathscr{Y}_{[0,t]}^{\beta}, U).$$

Then

$$\begin{aligned}
\big\|\pi_t^{\beta} - \pi_t^{\beta,U}\big\| &= \sum_{i=1}^{n} \Big| P(X_t^{\beta} = a_i | \mathscr{Y}_{[0,t]}^{\beta}) - P(X_t^{\beta} = a_i | \mathscr{Y}_{[0,t]}^{\beta}, U) \Big| \\
&= \sum_{i=1}^{n} \left| \sum_{j=1}^{m} P(X_t^{\beta} = a_i | \mathscr{Y}_{[0,t]}^{\beta}, U = j) \Big(P(U = j | \mathscr{Y}_{[0,t]}^{\beta}) - I(U = j) \Big) \right| \\
&\le \sum_{j=1}^{m} \Big| P(U = j | \mathscr{Y}_{[0,t]}^{\beta}) - I(U = j) \Big|
\end{aligned}$$

and

$$\begin{aligned}
&\big\|\pi_t^{\beta,U} - \pi_t^{\beta_0}\big\| = \sum_{i=1}^{n} \Big| P(X_t^{\beta} = a_i | \mathscr{Y}_{[0,t]}^{\beta}, U) - P(X_t^{\beta} = a_i | \mathscr{Y}_{[0,t]}^{\beta}, X_0^{\beta}) \Big| \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} I(U = j) \Big| P(X_t^{\beta} = a_i | \mathscr{Y}_{[0,t]}^{\beta}, U = j) - P(X_t^{\beta} = a_i | \mathscr{Y}_{[0,t]}^{\beta}, U = j, X_0^{\beta}) \Big| \\
&= \sum_{j=1}^{m} I(U = j) \big\|\pi_t^{\beta^j} - \pi_t^{\beta_0^j}\big\|,
\end{aligned}$$

where β^j denotes the conditional distribution of β restricted to the subalphabet $\mathbb{S}_j$. By Lemma 6.3,

$$\sum_{j=1}^{m} \Big| P(U = j | \mathscr{Y}_{[0,t]}^{\beta}) - I(U = j) \Big| \xrightarrow[t\to\infty]{P} 0$$

while $\sum_{j=1}^{m} I(U = j) \big\|\pi_t^{\beta^j} - \pi_t^{\beta_0^j}\big\| \xrightarrow[t\to\infty]{\mathbb{L}_1} 0$ by applying Theorem 4.1 to each $\mathbb{S}_j$.

Appendix. Proof of Proposition 3.2.

Proof (sketch). We use the following construction for X. Let X_0 be a random variable with values in $\mathbb{S} = \{1, 2, 3, 4\}$ and $P(X_0 = j) = \nu_j$, $j = 1, \dots, 4$. Introduce independent of X_0 the matrix-valued process

$$\mathcal{N}_t = \begin{pmatrix} -N_{12}(t) & N_{12}(t) & 0 & 0 \\ 0 & -N_{23}(t) & N_{23}(t) & 0 \\ 0 & 0 & -N_{34}(t) & N_{34}(t) \\ N_{41}(t) & 0 & 0 & -N_{41}(t) \end{pmatrix}, \tag{A.1}$$

where $N_{ij}(t)$ are independent copies of the Poisson process with the unit rate. Let us consider the Itô equation

$$I_t = I_0 + \int_0^t d\mathcal{N}_s^* I_{s-} \tag{A.2}$$

with I_0 the vector with entries $I_0(j) = I(X_0 = j)$, $j = 1, \dots, 4$. Since the jumps of Poisson processes $N_{ij}(t)$'s are disjoint, for any $t > 0$ the vector I_t has only one nonzero entry. Moreover, whereas the increments of $\mathcal{N}_t$ are independent for nonoverlapping intervals, I_t is a Markov process. It is readily checked that, with the row vector $g = \begin{pmatrix} 1 & 2 & 3 & 4 \end{pmatrix}$, $X_t = gI_t$ is a Markov process with values in $\mathbb{S}$ and the transition intensities matrix Λ and $I_t(j) = I(X_t = j)$, $j = 1, \dots, 4$.

We will follow Theorem 4.10.1 from [31]. The random process Y has piecewise constant paths with jumps of two magnitudes, $+1$ and -1. Due to (A.2), its saltus measure $p(dt, dy)$ is completely described by

$$\begin{aligned} p(dt, \{1\}) &= \big\{ I_{t-}(4) dN_{41}(t) + I_{t-}(2) dN_{23}(t) \big\}, \\ p(dt, \{-1\}) &= \big\{ I_{t-}(1) dN_{12}(t) + I_{t-}(3) dN_{34}(t) \big\}. \end{aligned}$$

So, the compensator $\overline{q}(dt, dy)$ of $p(dt, dy)$ with respect to the filtration $(\mathscr{Y}_{[0,t]})_{t \geq 0}$ is defined as

$$\begin{aligned} \overline{q}(dt, \{1\}) &= \big(\pi_{t-}(4) + \pi_{t-}(2) \big) dt = (1 - Y_{t-}) dt, \\ \overline{q}(dt, \{-1\}) &= \big(\pi_{t-}(1) + \pi_{t-}(3) \big) dt = Y_{t-} dt. \end{aligned} \tag{A.3}$$

Notice also that

$$p(dt, \{1\}) = (1 - Y_{t-}) dY_t \quad \text{and} \quad p(dt, \{-1\}) = -Y_{t-} dY_t. \tag{A.4}$$

Equation (A.2) also gives "drift+martingale" presentation for $I_1(t)$, $I_2(t)$:

$$\begin{aligned} dI_t(1) &= \big(- I_t(1) + I_t(4) \big) dt + dM_1(t), \\ dI_t(2) &= \big(I_t(1) - I_t(2) \big) dt + dM_2(t) \end{aligned} \tag{A.5}$$

with martingales

$$\begin{aligned} M_1(t) &= \int_0^t \Big(- I_{s-}(1) d(N_{12}(s) - s) + I_{s-}(4) d(N_{41}(s) - s) \Big), \\ M_2(t) &= \int_0^t \Big(I_{s-}(1) d(N_{12} - s) - I_{s-}(2) d(N_{23}(s) - s) \Big). \end{aligned}$$

Then, by Theorem 4.10.1 in [31], adapted to the case considered, we have

$$\begin{aligned} d\pi_1(t) &= \big(- \pi_t(1) + \pi_t(4) \big) dt + \int H_1(\omega, t, y) \big[p(dt, dy) - \overline{q}(dt, dy) \big], \\ d\pi_2(t) &= \big(\pi_t(1) - \pi_t(2) \big) dt + \int H_2(\omega, t, y) \big[p(dt, dy) - \overline{q}(dt, dy) \big], \end{aligned} \tag{A.6}$$

where $H_i(\omega, t, y)$, $i = 1, 2$, are $\mathscr{P}(Y) \otimes \mathscr{B}(\mathbb{R})$-measurable functions (here $\mathscr{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$ and $\mathscr{P}(Y)$ is the predictable σ-algebra on $\Omega \times \mathbb{R}_+$ with respect to the filtration $(\mathscr{Y}_{[0,t]})_{t \geq 0}$). Moreover

$$H_i(\omega, t, y) = \mathsf{M}\big(\triangle M_i + I_-(i) | \mathscr{P}(\mathcal{Y}) \otimes \mathscr{B}(\mathbb{R}) \big)(\omega, t, y) - \pi_{t-}(i),$$

where $\triangle M_i$ and $I_-(i)$ are the processes $M_i(t) - M_i(t-)$ and $I_{t-}(i)$, respectively, and $\mathsf{M}\big(\,\cdot\,|\mathscr{P}(Y)\otimes\mathscr{B}(\mathbb{R})\big)$ is the conditional expectation with respect to the measure $\mathsf{M}(d\omega, dt, dy) = P(d\omega)p(dt, dy)$ given $\mathscr{P}(Y)\otimes\mathscr{B}(\mathbb{R})$.

By (A.5), $\triangle M_i(t) + I_{t-}(i) = I_t(i)$ and the structure of compensator $\overline{q}$ provides (here $\triangle I_t(i) = I_t(i) - I_{t-}(i)$)

$$\mathsf{M}\big(I(i)|\mathscr{P}(Y)\otimes\mathscr{B}(\mathbb{R})\big) - \pi_{t-}(i) = \mathsf{M}\big(\triangle I(i)|\mathscr{P}(Y)\otimes\mathscr{B}(\mathbb{R})\big).$$

The desired conditional expectation is determined uniquely from the following identity: For any bounded, compactly supported in t and $\mathscr{P}(\mathcal{Y})\otimes\mathscr{B}(\mathbb{R})$-measurable function $\phi(\omega, t, y)$

$$E\int_0^\infty\int \phi(\omega,t,y)\triangle I_t(i)p(dt,dy) = E\int_0^\infty\int \phi(\omega,t,y)\mathsf{M}\big(\triangle I(i)|\mathscr{P}(Y)\otimes\mathscr{B}(\mathbb{R})\big)(\omega,t,y)\overline{q}(dt,dy).$$

By (A.2)

$$\begin{aligned}\triangle I_t(1) &= -I_{t-}(1)\triangle N_{12}(t) + I_{t-}(4)\triangle N_{41}(t),\\ \triangle I_t(2) &= I_{t-}(1)\triangle N_{12}(t) - I_{t-}(2)\triangle N_{23}(t),\end{aligned}$$

and so

$$\begin{aligned}\triangle I_t(1)p(dt,\{1\}) &= I_{t-}(4)dN_{41}(t),\\ \triangle I_t(1)p(dt,\{-1\}) &= -I_{t-}(1)dN_{12}(t),\\ \triangle I_t(2)p(dt,\{1\}) &= -I_{t-}(2)dN_{23}(t),\\ \triangle I_t(2)p(dt,\{-1\}) &= I_{t-}(1)dN_{12}(t).\end{aligned}$$

Owing to the obvious relations

$$\begin{aligned}I_4(t) \equiv I_4(t)(1-Y_t),\quad I_2(t) \equiv I_2(t)(1-Y_t),\\ I_1(t) \equiv I_1(t)Y_t,\quad I_3(t) \equiv I_3(t)Y_t\end{aligned}$$

we have

$$\begin{aligned}\pi_{t-}(2)dt = \pi_{t-}(2)(1-Y_{t-})dt,\quad \pi_{t-}(2)dt = \pi_{t-}(2)(1-Y_{t-})dt,\\ \pi_{t-}(1)dt = \pi_{t-}(1)Y_{t-}dt,\quad \pi_{t-}(3)dt = \pi_{t-}(3)Y_{t-}dt.\end{aligned}\tag{A.7}$$

Taking into account (A.3), we find

$$H_1(\omega,t,y) = \begin{cases}\pi_{t-}(4), & y=1,\\ -\pi_{t-}(1), & y=-1,\end{cases}$$

$$H_2(\omega,t,y) = \begin{cases}-\pi_{t-}(2), & y=1,\\ \pi_{t-}(1), & y=-1.\end{cases}$$

In accordance with (A.3), (A.4), the formulae for H_1, H_2, and (A.7), we transform (A.6) to

$$\begin{aligned}d\pi_1(t) &= \big(-\pi_t(1)+\pi_t(4)\big)dt + \pi_{t-}(4)(1-Y_{t-})(dY_t - dt) + \pi_{t-}(1)Y_{t-}(dY_t + dt)\\ &= \pi_{t-}(4)(1-Y_{t-})dY_t + \pi_{t-}(1)Y_{t-}dY_t\\ &= \big(1-\pi_{t-}(2)\big)(1-Y_{t-})dY_t + \pi_{t-}(1)Y_{t-}dY_t,\\ d\pi_2(t) &= \big(\pi_t(1)-\pi_t(2)\big)dt - \pi_{t-}(2)(1-Y_{t-})(dY_t - dt) - \pi_{t-}(1)Y_{t-}(dY_t + dt)\\ &= -\pi_{t-}(2)(1-Y_{t-})dY_t - \pi_{t-}(1)Y_{t-}dY_t.\qquad\square\end{aligned}$$

Acknowledgments. The authors gratefully acknowledge Boris Tsirelson for bringing [43] and the example in [44] to their attention, Rami Atar for suggesting use of Theorem 1 in [2] for the proof of Theorem 4.1, and the anonymous referees whose comments and advice allowed us to improve the paper significantly.

REFERENCES

[1] R. Atar, *Exponential stability for nonlinear filtering of diffusion processes in a noncompact domain*, Ann. Probab., 26 (1998), pp. 1552–1574.

[2] R. Atar and O. Zeitouni, *Exponential stability for nonlinear filtering*, Ann. Inst. H. Poincaré Probab. Statist., 33 (1997), pp. 697–725.

[3] R. Atar and O. Zeitouni, *Lyapunov exponents for finite state nonlinear filtering*, SIAM J. Control Optim., 35 (1997), pp. 36–55.

[4] R. Atar, F. Viens, and O. Zeitouni, *Robustness of Zakai's equation via Feynman-Kac representations*, in Stochastic Analysis, Control, Optimization, and Applications, W. M. McEneaney, G. Yin, and Q. Zhang, eds., Birkhäuser Boston, Boston, 1998, pp. 339–352.

[5] V. E. Beneš and I. Karatzas, *Estimation and control for linear, partially observable systems with non-Gaussian initial distribution*, Stochastic Process Appl., 14 (1983), pp. 233–248.

[6] A. G. Bhatt, A. Budhiraja, and R. L. Karandikar, *Markov property and ergodicity of the nonlinear filter*, SIAM J. Control Optim., 39 (2000), pp. 928–949.

[7] A. Budhiraja, *Ergodic properties of the nonlinear filter*, Stochastic Process Appl., 95 (2001), pp. 1–24.

[8] A. Budhiraja, *On invariant measures of discrete time filters in the correlated signal-noise case*, Ann. Appl. Probab., 12 (2002), pp. 1096–1113.

[9] A. Budhiraja and D. Ocone, *Exponential stability of discrete-time filters for bounded observation noise*, Systems Control Lett., 30 (1997), pp. 185–193.

[10] A. Budhiraja and D. Ocone, *Exponential stability in discrete-time filtering for nonergodic signals*, Stochastic Process. Appl., 82 (1999), pp. 245–257.

[11] A. Budhiraja and H. J. Kushner, *Robustness of nonlinear filters over the infinite time interval*, SIAM J. Control Optim., 36 (1998), pp. 1618–1637.

[12] A. Budhiraja and H. J. Kushner, *Approximation and limit results for nonlinear filters over an infinite time interval*, SIAM J. Control Optim., 37 (1999), pp. 1946–1979.

[13] A. Budhiraja and H. J Kushner, *Approximation and limit results for nonlinear filters over an infinite time interval.* II. *Random sampling algorithms*, SIAM J. Control Optim., 38 (2000), pp. 1874–1908.

[14] F. Cerou, *Long time behavior for some dynamical noise free nonlinear filtering problems*, SIAM J. Control Optim., 38 (2000), pp. 1086–1101.

[15] J. M. C. Clark, D. L. Ocone, and C. Coumarbatch, *Relative entropy and error bounds for filtering of Markov process*, Math. Control Signals Systems, 12 (1999), pp. 346–360.

[16] G. Da Prato, M. Fuhrman, and P. Malliavin, *Asymptotic ergodicity of the process of conditional law in some problem of non-linear filtering*, J. Funct. Anal., 164 (1999), pp. 356–377.

[17] P. Del Moral and A. Guionnet, *On the stability of measure valued processes with applications to filtering*, C. R. Acad. Sci. Paris Ser. I Math., 329 (1999), pp. 429–434.

[18] P. Del Moral and A. Guionnet, *On the stability of interacting processes with applications to filtering and genetic algorithms*, Ann. Inst. H. Poincaré Probab. Statist., 37 (2001), pp. 155–194.

[19] B. Delyon and O. Zeitouni, *Lyapunov exponents for filtering problem*, in Applied Stochastic Analysis, M. H. A. Davis and R. J.Elliot, eds., Gordon & Breach, New York, 1991, pp. 511–521.

[20] P. Dupuis and R. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley & Sons, New York, 1997.

[21] T. Kaijser, *A limit theorem for partially observed Markov chains*, Ann. Probab., 3 (1975), pp. 677–696.

[22] H. Kunita, *Asymptotic behavior of the nonlinear filtering errors of Markov processes*, J. Multivariate Anal., 1 (1971), pp. 365–393.

[23] H. Kunita, *Ergodic properties of nonlinear filtering processes*, in Spatial Stochastic Processes, Progr. Probab. 19, Birkhäuser Boston, Boston, 1991, pp. 233–256.

[24] F. Le Gland and L. Mevel, *Exponential forgetting and geometric ergodicity in hidden Markov models*, Math. Control Signals Systems, 13 (2000), pp. 63–93.

[25] P. Del Moral and L. Miclo, *On the stability of nonlinear Feynman-Kac semigroups*, Ann.

Fac. Sci. Toulouse Math.6 , 11 (2002), pp. 135–175.
[26] F. Le Gland and N. Oudjane, *Stability and Uniform Approximation of Nonlinear Filters Using the Hilbert Metric, and Application to Particle Filters*, RR-4215, INRIA, LeChesnay, France, 2001. Available online at http://www.inria.fr/rrrt/rr-4215.html.
[27] F. Le Gland and N. Oudjane, *A Robustification Approach to Stability and to Uniform Particle Approximation of Nonlinear Filters: The Example of Pseudo-mixing Signals*, 4431, INRIA, 2002.
[28] A. Le Breton and M. Roubaud, *Asymptotic optimality of approximate filters in stochastic systems with colored noises*, SIAM J. Control Optim., 39 (2000), pp. 917–927.
[29] R. Sh. Liptser and A. N. Shiryaev, *Statistics of Random Processes* I, 2nd ed., Springer-Verlag, Berlin, 2001.
[30] R. Sh. Liptser and A. N. Shiryaev, *Statistics of Random Processes* II, 2nd ed., Springer-Verlag, Berlin, 2001.
[31] R. Sh. Liptser and A. N. Shiryaev, *Theory of Martingales*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.
[32] A. M. Makowski, *Filtering formula for partially observed linear systems with non-Gaussian initial conditions*, Stochastics, 16 (1986), pp. 1–24.
[33] A. M. Makowski and R. B. Sowers, *Discrete-time filtering for linear systems with non-Gaussian initial conditions: Asymptotic behaviors of the difference between the MMSE and the LMSE estimates*, IEEE Trans. Automat. Control, 37 (1992), pp. 114–120.
[34] J. R. Norris, *Markov Chains*, Cambridge University Press, Cambridge, UK, 1997.
[35] D. Ocone and E. Pardoux, *Asymptotic stability of the optimal filter with respect to its initial condition*, SIAM J. Control Optim., 34 (1996), pp. 226–243.
[36] D. Ocone, *Asymptotic stability of Beneš filters*, Stochastic Anal. Appl., 17 (1999), pp. 1053–1074.
[37] D. Ocone, *Entropy inequalities and entropy dynamics in nonlinear filtering of diffusion processes*, in Stochastic Analysis, Control, Optimization and Applications, Systems Control Found. Appl., Birkhäuser Boston, Boston, 1999, pp. 477–496.
[38] E. Seneta, *Nonnegative Matrices and Markov Chains*, 2nd ed., Springer-Verlag, New York, 1981.
[39] Ya. D. Sinai, *Kolmogorov's work on ergodic theory*, Ann. Probab., 17 (1989), pp. 833–839.
[40] L. Stettner, *On invariant measure of filtering processes*, in Stochastic Differential Systems, Lecture Notes in Control and Inform. Sci. 126, Springer-Verlag, Berlin, 1989, pp. 279–292.
[41] L. Stettner, *Invariant measures of the pair: State, approximate filtering process*, Colloq. Math., 62 (1991), pp. 347–351.
[42] A. N. Shiryaev, *Probability*, 2nd ed., Springer-Verlag, New York, 1996.
[43] H. von Weizsäcker, *Exchanging the order of taking suprema and countable intersections of σ-algebras*, Ann. Inst. H. Poincaré Sect. B, 19 (1983), pp. 91–100.
[44] D. Williams, *Probability with Martingales*, Cambridge University Press, Cambridge, UK, 1991.
[45] W. M. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering*, J. Soc. Indust. Appl. Math. Ser. A Control, 2 (1965), pp. 347–369.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 670–684

SLIDING MODE DESIGN VIA QUADRATIC PERFORMANCE OPTIMIZATION WITH POLE-CLUSTERING CONSTRAINT*

KYUNG-SOO KIM† AND YOUNGJIN PARK‡

Abstract. We consider a novel method for designing the sliding mode that minimizes the quadratic performance while keeping a pole-clustering constraint. Our approach is based on the manipulations of linear matrix inequalities (LMIs) imposed by the design objectives. For this purpose, we newly propose LMI conditions for the quadratic performance optimization and the pole-clustering problem, respectively, in a full order state. Then they are combined in the LMI framework that is typically devised for the sliding mode design in the convex form. An effort is made to reduce the generic conservatism by allowing different Lyapunov matrices. In addition, a class of polytope uncertain systems is addressed to illustrate the advantages of the proposed method.

Key words. sliding mode, linear matrix inequalities (LMIs), Lyapunov matrix, quadratic performance optimization, pole-clustering problem, polytopic uncertain system

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/S0363012901388476

1. Introduction.

Sliding mode control has been one of the major concerns in control theory thanks to the robustness against matched uncertainties (or disturbances) through the sliding mode behavior. Once the system reaches the sliding manifold, the system dynamics becomes invariant to the matched disturbances. This typical characteristic has drawn much attention to many practical problems which are hard to solve by using only the linear control methods. For example, the robot manipulators with friction or stabilized head mirror systems can be effectively dealt with by the sliding mode control. Refer to DeCarlo, Zak, and Matthews [2] and the references therein for more details.

In the literature, much effort has been made to design sliding modes that satisfy the desired performance criteria. The well-known criteria include quadratic performance optimization [1], guaranteed H_2 cost minimization [14], the eigen-structure assignment including pole clustering [3, 4, 6], the robustness to parametric uncertainties [5, 7, 13], etc. It is noted that most of the design methods developed so far have considered only the single design objective. In fact, there have been few researches that address multiple design objectives in the sliding mode. Recently, in [15], the sliding mode design with multiple constraints has been introduced based on the linear matrix inequality (LMI) approach by employing the linear full state feedback design. On the other hand, remarkable progress has been made in linear control theory for solving optimal problems with multiple constraints based on LMIs (e.g., see [10] and [18]). The basic idea of the multiobjective approach based on LMIs is to seek a *common* Lyapunov matrix that satisfies different parametric constraints imposed by the design performances simultaneously. Assuming the common variables may cause the conservatism, however, it does provide the flexibility of the control design with multiple objectives, and the ease of synthesis in the parameter space based on LMIs

*Received by the editors April 22, 2001; accepted for publication (in revised form) December 4, 2003; published electronically August 4, 2004.
http://www.siam.org/journals/sicon/43-2/38847.html
†Multimedia Technical Center, STMicroelectronics Co. Ltd., 16 Tao Hua Road, Futian Free Trade Zone, Shenzen 518045, People's Republic of China (kimkyungsoo@ieee.org).
‡Department of Mechanical Engineering, KAIST, 373-1 Kusong-dong, Yusong-gu, Taejon 305-701, Korea (yjpark@kaist.ac.kr).

[8]. Also, there have been notable results in the literature in reducing the design conservatism (e.g., see [11] and [16]).

The paper is devoted to establishing an LMI framework for the sliding mode design which effectively solves the constrained optimization problems by adopting the common Lyapunov matrix idea of the linear control theory. To this end, attention is paid especially to quadratic performance optimization with the pole-clustering constraint in the sliding mode. First, we newly present the parametric constraints in LMIs for describing quadratic performance optimization and the pole-clustering condition, respectively. Then they are combined in an LMI-based optimization problem of special structure. It will be shown that the proposed approach for the sliding mode design should be different from the standard full state feedback problem by Chilali and Gahinet [10] in linear control theory. Furthermore, the proposed approach is further extended to reduce the generic conservatism by adopting the bounding technique in Shimomura and Fujii [16].

The basic idea of the paper starts from the derivation of LMI conditions in the full order state (not in the reduced order) by adopting and generalizing the parameterization technique in Kim, Park, and Oh [7] to allow all the feasible linear sliding modes to be represented explicitly. Any feasible sliding modes can be obtained immediately by combining the partitions of the Lyapunov matrix in an explicit manner. From this point of view, the proposed approach generalizes the previous results using the full order Lyapunov (or Riccati-type) approaches (see [5, 6]), which motivated the study. Note that the approach in the full order state has the advantage of simplicity in description and easy application of the linear full state feedback theories to the sliding mode design with least modification (which will be further discussed with a class of polytopic uncertain systems later on).

In section 2, the problem of interest is formulated with a brief explanation for the LMI scheme with multiple constraints. In section 3, we deal with quadratic performance optimization and the pole-clustering problem separately based on parametric constraints. Then in section 4, we formulate the LMI approach by accumulating parametric constraints in a single problem of convex form. Also, to reduce the design conservatism, a numerically tractable algorithm is presented. In section 5, we address the extension of the proposed scheme to a class of polytopic uncertain systems in order to show the potential advantage of the basic idea of the paper.

The notations used in the paper are fairly standard, among them, the inequality signs for matrices denote sign-definiteness for real symmetric matrices.

2. Problem formulation and preliminary. Consider the system

$$\dot{x} = Ax + B(u + Dw), \tag{2.1}$$

where $x \in \Re^n$ and $u \in \Re^m$ are the state vector and the control input, respectively, and $w \in \Re^l$ is the disturbance of which each element is bounded as $|w_j(t)| \le \overline{w}_j \ \forall j \in [0, l]$ for the known $\overline{w}_j$. The stabilizability of the pair (A, B) is assumed. And, for simplicity of description, suppose that the system is of the regular form [1, 2]

$$\begin{cases} \dot{x}_1 = A_{11}x_1 + A_{12}x_2, \\ \dot{x}_2 = A_{21}x_1 + A_{22}x_2 + B_2(u + Dw), \end{cases} \tag{2.2}$$

where $x_1 \in \Re^{n-m}$, $x_2 \in \Re^m$, and B_2 is nonsingular. Without loss of generality, consider the sliding function

$$s(t) = Sx_1 + x_2 \tag{2.3}$$

for some $S \in \Re^{m\times(n-m)}$. Suppose that a control law is employed to satisfy the reachability condition such that $\dot{s}(t)^T s(t) < 0\ \forall t > 0$, which would result in the sliding mode (i.e., $s(t) = 0\ \forall t \geq t_s$ for some $t_s > 0$). Then the system behavior would be governed by the reduced order system and the constrained state as follows:

$$\begin{cases} \dot{x}_1 = (A_{11} - A_{12}S)\, x_1, \\ x_2 = -Sx_1. \end{cases} \tag{2.4}$$

Note that in the sliding mode, the system dynamics will be determined by the choice of sliding function coefficient S. Among many criteria for selecting the matrix S, one may consider the quadratic performance optimization to make a tradeoff between the system states behaviors. In the paper, we are interested in solving the following problem.

Problem formulation. Given the real scalars c and $\rho > 0$ and a matrix $0 < Q \in \Re^{n\times n}$, find the sliding function coefficient S that minimizes the quadratic performance

$$J(S) \triangleq \int_{t_s}^{\infty} x^T Q x\ dt \tag{2.5}$$

subject to the dynamics (2.4) and the pole-clustering constraint

$$\lambda\,(A_{11} - A_{12}S) \subset \mathcal{Z}(c, \rho), \tag{2.6}$$

where $\lambda(\cdot)$ is the set of eigenvalues of the argument matrix, and $\mathcal{Z}(c, \rho)$ denotes the circular region, in the complex domain,

$$\mathcal{Z}(c, \rho) \triangleq \{z \in \mathcal{C} \mid\ |z + c| < \rho,\ \mathrm{Re}(z) < 0\}\,. \tag{2.7}$$

For the set $\mathcal{Z}(c, \rho)$ to be nonempty, it is necessary that $c > -\rho$.

Note that the problem above consists of two design objectives (i.e., the quadratic performance optimization and the pole-clustering problem). To tackle the above problem, suppose that the quadratic performance is bounded as $J < \gamma$ if there exist some Lyapunov matrices P_1 satisfying an affine matrix inequality $F_1(P_1, S) < 0$. Also, suppose that the pole-clustering constraint is satisfied if there exist some multipliers P_2 satisfying $F_2(P_2, S) < 0$. Often, each of the parametric constraints (i.e., $F_1 < 0$ and $F_2 < 0$) can be converted to be convex, but not jointly. In this case, let us consider this problem: *find a matrix S that satisfies the constraints $F_i(P_i, S) < 0$ $(i = 1, 2)$ simultaneously.* Since the constraints are not jointly convex, the problem is difficult to solve. One way to recover the convexity is to further assume that $P_1 = P_2$ regardless of the conservatism, which is one of the main approaches in the LMI-based constrained optimization (or multiobjective) theory. See [10] and [18] for more details.

In the next section, two objectives are dealt with distinctly based on the parameterization technique for the sliding mode in the LMI form. Then in section 4, they are combined in an LMI optimization approach specifically adapted to the sliding mode design and different from the standard full state feedback method [10].

3. LMI approaches to sliding mode design.

3.1. Quadratic performance optimization. In this subsection, we handle the quadratic performance optimization problem in the parameter space by manipulating the Lyapunov equation.

Given a matrix $Q > 0$, let us define the set

$$\Omega(Q) \triangleq \left\{P \mid (A - BK)^T P + P(A - BK) + Q = 0,\ P > 0,\ K \in \Re^{m\times n}\right\}, \tag{3.1}$$

which is the nonempty set of positive definite matrices as long as the pair (A, B) is stabilizable (e.g., see [17]). Then we present the following result.

THEOREM 3.1. *Suppose that a matrix $Q > 0$ is arbitrarily chosen. Then the following statements hold:*

(i) *Any stabilizing sliding function coefficient exists in the form*

$$S = P_{22}^{-1} P_{12}^T \tag{3.2}$$

for a $P \in \Omega(Q)$, where P_{ij}'s are defined as

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix} \in \begin{bmatrix} \Re^{(n-m)\times(n-m)} & \Re^{(n-m)\times m} \\ \Re^{m\times(n-m)} & \Re^{m\times m} \end{bmatrix}.$$

(ii) *For any $P \in \Omega(Q)$, the matrix S given by* (3.2) *is a stabilizing sliding function coefficient.*

Proof. Statement (i). Let S be the stabilizing sliding function coefficient which guarantees the stability of the matrix $A_s := A_{11} - A_{12}S$. Define a positive definite matrix as $Q_r = \left[I_{n-m},\ -S^T\right] Q \left[I_{n-m},\ -S^T\right]^T$. Then there should exist a $P_r > 0$ satisfying

$$\left(A_{11} - A_{12}S\right)^T P_r + P_r \left(A_{11} - A_{12}S\right) + Q_r = 0 \tag{3.3}$$

due to the stability of A_s. Now, for an arbitrary matrix $0 < P_{22} \in \Re^{m\times m}$, define the matrices

$$P_{12} \triangleq S^T P_{22}, \quad P_{11} \triangleq P_r + P_{12}^T P_{22}^{-1} P_{12}, \quad K \triangleq [K_1,\ K_2], \tag{3.4}$$

where

$$K_1 = B_2^{-1}\left\{A_{21} + P_{22}^{-1}P_{12}^T A_{11} + P_{22}^{-1}A_{12}^T P_r\right\}, \quad K_2 = B_2^{-1}\left\{A_{22} + P_{22}^{-1}P_{12}^T A_{12}\right\}.$$

Observe that $P := \left[\begin{smallmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{smallmatrix}\right]$ is positive definite since $P_{22} > 0$ and $P_{11} - P_{12}P_{22}^{-1}P_{12}^T$ $(= P_r) > 0$. With the matrices P and K above, it may be shown, through some manipulations, that

$$\begin{bmatrix} I_{n-m} & -S^T \\ 0 & I_m \end{bmatrix} \left\{(A - BK)^T P + P(A - BK) + Q\right\} \begin{bmatrix} I_{n-m} & -S^T \\ 0 & I_m \end{bmatrix}^T = 0, \tag{3.5}$$

which implies that $P \in \Omega(Q)$.

Statement (ii). For a $P \in \Omega(Q)$, define $T_r \triangleq \left[I_{n-m},\ -P_{12}P_{22}^{-1}\right]$. Then pre- and postmultiplying the Lyapunov equation in (3.1) by T_r and T_r^T, respectively, yields

$$\left(A_{11} - A_{12}P_{22}^{-1}P_{12}^T\right)^T P_r + P_r \left(A_{11} - A_{12}P_{22}^{-1}P_{12}^T\right) + T_r Q T_r^T = 0, \tag{3.6}$$

where $P_r = P_{11} - P_{12}^T P_{22}^{-1} P_{12}$, which is positive definite since $P > 0$. Then, by choosing $S = P_{22}^{-1}P_{12}^T$, the asymptotic stability of the matrix $A_{11} - A_{12}S$ is guaranteed. This completes the proof. □

Theorem 3.1 shows that all of the linear sliding modes can be represented by combining portions of the full order Lyapunov matrix (i.e., P_{12} and P_{22}). Using the result, the sliding mode can be obtained easily simply by solving the Lyapunov equation. Here, we point out that the specific choice of Q does not constrain the

range of the feasible sliding function coefficients. The matrix Q will be used to define the quadratic performance in the following.

Using Theorem 3.1, we can tackle the linear quadratic sliding (LQS) mode optimization problem in the following.

THEOREM 3.2. *Given the stabilizing sliding function coefficient S, the quadratic cost function* (2.5) *satisfies*

$$J(S) = x_1(t_s)^T P_r x_1(t_s) \tag{3.7}$$

for the $P \in \Omega(Q)$ satisfying (3.2)*, where $P_r = P_{11} - P_{12}P_{22}^{-1}P_{12}^T$.*

Proof. For the sliding function coefficient S, there exists a $P \in \Omega(Q)$ satisfying $S = P_{22}^{-1}P_{12}^T$ for some K (from result (i) of Theorem 3.1). Then it follows that

$$\begin{aligned} s(t) &= x_2 + Sx_1 \\ &= P_{22}^{-1}\left[\, P_{12}^T \quad P_{22}\,\right] x \\ &= \left(P_{22}^{-1}B_2^{-T}\right) B^T P x, \end{aligned} \tag{3.8}$$

which implies $B^T P x = 0$ on $s(t) = 0$ $(\forall t \geq t_s)$ since $B_2^T P_{22}$ is nonsingular. Note that $B^T = [0_{m\times(n-m)},\ B_2^T]$ from (2.1) and (2.2).

Let us consider the derivative of a quadratic function $V = x^T P x$ for $t \geq t_s$. Rewriting the system equation as

$$\dot{x} = (A - BK)x + B(u + Dw + Kx)$$

for the matrix K associated with the matrix P, it follows that

$$\begin{aligned} \dot{V} &= x^T\left\{(A-BK)^T P + P(A-BK)\right\} x + 2x^T PB\,(u + Dw + Kx) \\ &= -x^T Q x + 2x^T PB\,(u + Dw + Kx) \\ &= -x^T Q x, \end{aligned} \tag{3.9}$$

since $B^T P x = 0$ on $s(t) = 0$. Integrating both sides in (3.9), we have

$$\begin{aligned} \int_{t_s}^{\infty} x^T Q x\ dt &= x(t_s)^T P x(t_s) \\ &= x_1(t_s)^T\left(P_{11} - P_{12}P_{22}^{-1}P_{12}^T\right) x_1(t_s), \end{aligned} \tag{3.10}$$

since $x_2(t_s) = -P_{22}^{-1}P_{12}^T x_1(t_s)$. This completes the proof. ∎

Theorems 3.1 and 3.2 provide an important result—that the quadratic performance index is redefined in the parameter space so that the optimization can be equivalently expressed as

$$\min_S J(S) = \min_{P\in\Omega(Q)} x_1(t_s)^T P_r x_1(t_s). \tag{3.11}$$

One of the advantages of the above results is that a convex approach is allowed based on the LMI method [8] that utilizes the change of variables such as $Y := P^{-1}$ and $L := KP^{-1}$. For converting the LQS design problem into LMIs, the cost index (expressed by P_r) should be of concern when using the matrix inversion property

$$Y = P^{-1} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12}^T & Y_{22} \end{bmatrix}, \tag{3.12}$$

where Y_{ij}'s are as follows:

$$\begin{aligned} Y_{11} &= (P_{11} - P_{12}P_{22}^{-1}P_{12}^T)^{-1}, \\ Y_{12} &= -(P_{11} - P_{12}P_{22}^{-1}P_{12}^T)^{-1}P_{12}P_{22}^{-1}, \\ Y_{22} &= P_{22}^{-1} + P_{22}^{-1}P_{12}^T Y_{11} P_{12} P_{22}^{-1}. \end{aligned}$$

Observe that P_r is described by the inverse of Y_{11}. Now, we summarize the LQS design method in the following.

LQS PROBLEM. *Given $Q > 0$, minimize γ w.r.t. $Y > 0$ and L satisfying*

$$\begin{bmatrix} AY + YA^T - BL - L^T B^T & YC_Q \\ C_Q^T Y & -I_q \end{bmatrix} < 0, \tag{3.13}$$

$$\begin{bmatrix} \gamma I_{n-m} & I_{n-m} \\ I_{n-m} & Y_{11} \end{bmatrix} > 0, \tag{3.14}$$

where $Q = C_Q C_Q^T$ and $q = rank(Q)$. Then, for the optimal value of Y, the sliding function coefficient (3.2) *is determined by*

$$S = -Y_{12}^T Y_{11}^{-1}. \tag{3.15}$$

Remark 3.3. The inequality (3.14) removes the optimal solution's dependency on the initial state, which results in $\int_{t_s}^{\infty} x^T Q x dt < \gamma \|x_1(t_s)\|^2$. Since the initial state is hardly known in reality, we adopt (3.14) instead of the inequality

$$\begin{bmatrix} \gamma & x_1(t_s)^T \\ x_1(t_s) & Y_{11} \end{bmatrix} > 0 \ .$$

The quadratic performance optimization was first introduced and solved by Utkin and Yang [1] based on the manipulation of an algebraic Riccati equation of reduced order, which has been considered a standard approach in the literature (e.g., see [12]). On the other hand, the approach of the paper is derived in the full order state while preserving the generality of the sliding mode design (i.e., in terms of the optimality and the existence of stabilizing sliding surfaces). In this context, the above results can be seen as a generalization, in the full order state, of the conventional approach. Furthermore, emphasis is placed on the underlying features in the parameterization technique (introduced in Theorem 3.1) in the full order state. That is, the Lyapunov matrices that solve the full state feedback provide all the feasible sliding modes. As a matter of fact, the idea can be consistently applied not only to the quadratic performance problem but also to other issues such as the pole-clustering constraint and the polytope uncertain systems, which can be dealt with by the Lyapunov (or Riccati) approaches developed for the full state feedback design. This will be investigated later. As pointed out by Su, Drakunov, and Özgüner [5], the simplicity of the problem description (by manipulating the full order system) is one of the practical advantages.

3.2. Pole-clustering problem in the sliding mode. We consider a method to represent the pole-clustering constraint (2.6) in terms of a parametric approach. To this end, we rely on the result from Yedavalli [9] and Chilali and Gahinet [10] in the following.

LEMMA 3.4. *The pole clustering constraint* (2.6) *holds for the matrix* S *if and only if there exist some* $0 < P_r \in \Re^{(n-m)\times(n-m)}$ *satisfying*

$$\begin{bmatrix} \rho P_r & (A_{11} - A_{12}S)^T P_r + cP_r \\ P_r(A_{11} - A_{12}S) + cP_r & \rho P_r \end{bmatrix} > 0. \tag{3.16}$$

Note that the Lyapunov matrix (i.e., P_r) of reduced order may be used to parameterize the sliding function coefficient alone. However, it is not apparent how it is related to the Lyapunov matrix of full order. Regarding this issue, we present the following result.

THEOREM 3.5. *Given the region* (2.7)*, the following statements hold.*

(i) *Any sliding function coefficient satisfying* (2.6) *exists in the form* $S = P_{22}^{-1}P_{12}^T$ *for a* $0 < P \in \Re^{n\times n}$ *satisfying, for some* $K \in \Re^{m\times n}$*,*

$$\begin{bmatrix} \rho P & (A - BK)^T P + cP \\ P(A - BK) + cP & \rho P \end{bmatrix} > 0\ . \tag{3.17}$$

(ii) *For any* $P > 0$ *satisfying* (3.17)*, the matrix* $S = P_{22}^{-1}P_{12}^T$ *attains the pole-clustering property* (2.6)*.*

Proof. Statement (i). Let S be the sliding function coefficient matrix satisfying (2.6). Then we can prove that there exists $P_r > 0$ satisfying (3.16) based on the necessity condition of Lemma 3.4. Define the matrices, for an arbitrary $0 < P_{22} \in \Re^{m\times m}$, as

$$P_{12} \triangleq S^T P_{22}, \quad P_{11} \triangleq P_r + P_{12}^T P_{22}^{-1} P_{12}, \quad K \triangleq [K_1,\ K_2], \tag{3.18}$$

where

$$\begin{aligned} K_1 &= B_2^{-1}\left\{P_{22}^{-1}P_{12}^T(A_{11} + cI_{n-m}) + A_{21}\right\}, \\ K_2 &= B_2^{-1}\left\{P_{22}^{-1}P_{12}^T A_{12} + A_{22} + cI_m\right\}. \end{aligned}$$

Note that the matrix $P := \left[\begin{smallmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{smallmatrix}\right]$ is positive definite since $P_{22} > 0$ and $P_r > 0$. In the following, we want to show that P and K defined above satisfy (3.17). To this end, consider the matrices

$$H = \left[\begin{array}{cc|cc} 0_{m\times(n-m)} & I_m & 0_{m\times(n-m)} & 0_m \\ \hline 0_{m\times(n-m)} & 0_m & 0_{m\times(n-m)} & I_m \end{array}\right], \tag{3.19}$$

$$T = \left[\begin{array}{cc|cc} I_{n-m} & -P_{12}P_{22}^{-1} & 0_{n-m} & 0_{(n-m)\times m} \\ \hline 0_{n-m} & 0_{(n-m)\times m} & I_{n-m} & -P_{12}P_{22}^{-1} \end{array}\right], \tag{3.20}$$

which have the nonsingular matrix $\left[\begin{smallmatrix} H \\ T \end{smallmatrix}\right] \in \Re^{2n\times 2n}$. For brevity of notation, let $L_{(3.16)}$ and $L_{(3.17)}$ denote the left-hand sides of (3.16) and (3.17), respectively. Through some elaborate manipulations, it may be shown that

$$\begin{aligned} \begin{bmatrix} H \\ T \end{bmatrix} L_{(3.17)} \begin{bmatrix} H \\ T \end{bmatrix}^T &= \left[\begin{array}{c|c} HL_{(3.17)}H^T & HL_{(3.17)}T^T \\ \hline TL_{(3.17)}H^T & TL_{(3.17)}T^T \end{array}\right] \\ &= \left[\begin{array}{cc|c} \rho P_{22} & 0 & \\ 0 & \rho P_{22} & 0 \\ \hline & 0 & L_{(3.16)} \end{array}\right]. \end{aligned} \tag{3.21}$$

Therefore, the inequality (3.17) holds because $P_{22} > 0$ and $L_{(3.16)} > 0$.

Statement (ii). For T_r (as defined in (3.6)), define the augmented matrix

$$T = \begin{bmatrix} T_r & 0_{(n-m)\times n} \\ 0_{(n-m)\times n} & T_r \end{bmatrix}. \tag{3.22}$$

Pre- and postmultiplying (3.17) by T and T^T, respectively, one may have (3.16) for the choice of $S = P_{22}^{-1}P_{12}^T$ and $P_r = P_{11} - P_{12}^T P_{22}^{-1} P_{12}$. This completes the proof. □

Remark 3.6. Note that the inequality (3.17) is the necessary and sufficient condition for the existence of the full state feedback, which places the closed loop poles in the specified region such that $\lambda(A - BK) \subset \mathcal{Z}(c, \rho)$.

Using change of variables such as $Y := P^{-1}$ and $L := KP^{-1}$, one may have the LMI for (3.17) as follows:

$$\begin{bmatrix} \rho Y & YA^T - L^T B^T + cY \\ AY - BL + cY & \rho Y \end{bmatrix} > 0. \tag{3.23}$$

Also, the sliding function coefficient is given by $S = -Y_{12}^T Y_{11}^{-1}$ as in (3.15).

Here, we stress the common structure of the sliding function coefficient under Theorems 3.1 and 3.5. That is, once the full state feedback problem is feasible, the sliding mode can be obtained immediately by defining and combining the partitions of the Lyapunov matrix without loss of generality. Also, the sliding function coefficient is concerned only with some portions of the Lyapunov matrix.

4. The LMI-based constrained optimization. In this section, we present the (sub-)optimal solution to the problem of concern (defined in section 2) using the LMI constraints proposed in the previous sections. For the readability of the manuscript, the basic idea is illustrated first, and the typical approaches to the sliding mode design will be derived both in the convex form and in the nonconvex but numerically tractable form.

4.1. Combining the parametric constraints. To tackle the problem defined in section 2, let us consider solving the LQS problem with the matrix variables Y and L for a certain level of γ (see (3.13) and (3.14)). At the same time, assume that the matrix variables are also satisfying the inequality (3.23). Then, if the problem is feasible, it is easy to see that the pole-clustering constraint is met by the resulting sliding mode with the upper bound of the quadratic performance. Even though the assumption for the common variables may cause conservatism—the infeasibility or the overly conservative upper bound of the performance index—the approach has proven to be effective in providing flexibility in the control design requiring multiple performances with convexity (see Chilali and Gahinet [10] and the references therein).

For reducing conservatism, the assumption for the common variables can be relaxed by using scales either in scalar [11] or matrix [16] form. Particularly, we focus on the idea of the matrix scale which results in the iterative convex searches. Once an initial guess is found, each step of synthesis is convex and convergent through the iteration.

Relying on the above ideas (i.e., the convex approach and the relaxation idea), the remainder of the paper is devoted to devising novel methods for the constrained optimization problem. Then the methods will be verified by an example.

4.2. Convex formulation. To avoid confusion, let Y_{LQS} and Y_{POL} be the Lyapunov matrices in the LQS problem (i.e., (3.13) and (3.14)) and the pole-clustering constraint (i.e., (3.23)), respectively. According to the typical structure of the sliding function coefficient (i.e., $S = -Y_{12}^T Y_{11}^{-1}$), the optimal solution to the problem (defined in section 2) needs the condition

$$Y_{LQS,12}^T Y_{LQS,11}^{-1} = Y_{POL,12}^T Y_{POL,11}^{-1} \tag{4.1}$$

to produce the single optimal sliding mode. The conventional way to remove nonconvexity is to assume that $Y_{LQS} = Y_{POL}$, as in the aforementioned full state feedback design. However, we assume that the Lyapunov matrices are composed as

$$Y_{LQS} = \begin{bmatrix} M & N \\ N^T & Z_1 \end{bmatrix}, \quad Y_{POL} = \begin{bmatrix} M & N \\ N^T & Z_2 \end{bmatrix} \tag{4.2}$$

for some M, N, Z_1, and Z_2 with dimensions as partitioned in (3.2). Note that (4.1) is simply satisfied with the above structure (4.2). Also, the matrix variables L in inequalities (3.13) and (3.23) can be allowed to be independent while keeping convexity. Based on these properties, we have the following result.

THEOREM 4.1. *Given $Q > 0$ and $\mathcal{Z}(c, \rho)$, there exist some S such that* (i) $\lambda(A_{11} - A_{12}S) \subset \mathcal{Z}(c, \rho)$ *and* (ii) $\int_{t_s}^{\infty} x^T Q x dt < \gamma \|x_1(t_s)\|^2$ *if there exist some M, N, Z_i, L_i ($\forall i = 1, 2$), and γ satisfying*

$$Y_{LQS} > 0, \tag{4.3}$$

$$Y_{POL} > 0, \tag{4.4}$$

$$\begin{bmatrix} \rho Y_{POL} & Y_{POL}A^T - L_1^T B^T + cY_{POL} \\ AY_{POL} - BL_1 + cY_{POL} & \rho Y_{POL} \end{bmatrix} > 0, \tag{4.5}$$

$$\begin{bmatrix} AY_{LQS} + Y_{LQS}A^T - BL_2 - L_2^T B^T & Y_{LQS}C_Q \\ C_Q^T Y_{LQS} & -I_q \end{bmatrix} < 0, \tag{4.6}$$

$$\begin{bmatrix} \gamma I_{n-m} & I_{n-m} \\ I_{n-m} & M \end{bmatrix} > 0, \tag{4.7}$$

where Y_{LQS} and Y_{POL} have the structure in (4.2). *Then the admissible sliding mode is given by $S = -N^T M^{-1}$ for the feasible parameters.*

Since the problem is convex, the minimal γ can be computed within polynomial time using the LMI method. Note that the proposed scheme is different from the standard feedback design [10] in that matrix L and the $(2, 2)$ blocks of the Lyapunov matrices are not necessarily common.

4.3. Reducing conservatism. So far, an effort has been made to keep convexity in the analysis. Now we present a general method in order to eliminate the conservatism while losing the convexity. However, the proposed method is suited for numerical efficiency based on the iteratively convex search.

Without loss of generality, instead of starting with (4.2), we start with the Lyapunov matrices

$$Y_{LQS} = \begin{bmatrix} M_1 & M_1 M_2^{-1} N_2 \\ N_2^T M_2^{-1} M_1 & Z_1 \end{bmatrix}, \quad Y_{POL} = \begin{bmatrix} M_2 & N_2 \\ N_2^T & Z_2 \end{bmatrix} \tag{4.8}$$

for some $M_i,\ Z_i$ $(\forall i = 1, 2)$, and N_2, which guarantee the essential requirement (4.1). Note that the structure in (4.2) is a special case of (4.8) when $M_1 = M_2$.

Hereafter, we consider the case in which (4.8) is employed in Theorem 4.1 instead of (4.2). Then the convexity of inequalities (4.3) and (4.6) related to Y_{LQS} would be destroyed. Hence, attention is paid to handling them by the convex approach in this section. For this purpose, we introduce the following properties.

PROPOSITION 4.2. *Given matrices G and H and symmetric W with appropriate dimensions, it holds that*

$$G^T W^{-1} H + H^T W^{-1} G \geq \begin{pmatrix} \Pi(H+G) + (H+G)^T \Pi^T + \Pi W \Pi^T \\ - G^T W^{-1} G - H^T W^{-1} H \end{pmatrix} \tag{4.9}$$

for any Π. Moreover, the equality holds when $\Pi = (H+G)^T W^{-1}$.

Proof. Consider the inequality

$$(G^T + H^T - \Pi W) W^{-1} (G^T + H^T - \Pi W)^T \geq 0, \tag{4.10}$$

which is always true for any Π. Note that the equality holds when $\Pi = (G^T + H^T) W^{-1}$. Then the completion of the square proves Proposition 4.2. This completes the proof. □

PROPOSITION 4.3. *Given matrices G and H and symmetric W with appropriate dimensions, it holds that*

$$G^T W^{-1} H + H^T W^{-1} G \leq \begin{pmatrix} \Pi(H-G) + (H-G)^T \Pi^T + \Pi W \Pi^T \\ + G^T W^{-1} G + H^T W^{-1} H \end{pmatrix} \tag{4.11}$$

for any Π. Moreover, the equality holds when $\Pi = (G-H)^T W^{-1}$.

Proof. Similar to the proof of Proposition 4.2, it is easy to show Proposition 4.3 using the relationship

$$(G^T - H^T - \Pi W) W^{-1} (G^T - H^T - \Pi W)^T \geq 0 \tag{4.12}$$

for any Π. This completes the proof. □

The bounding idea with scale is adapted from the result in Shimomura and Fujii [16], which was applied to the dynamic output feedback synthesis. We point out that scales have an important role in converting nonconvex inequalities into convex ones through an iterative procedure. This will be further illustrated in what follows.

First, the inequality (4.3) with the structure (4.8) is investigated. For description, rewrite Y_{LQS} as

$$Y_{LQS} = Y_{BD} + G^T M_2^{-1} H + H^T M_2^{-1} G, \tag{4.13}$$

where $Y_{BD} \triangleq \mathit{blockdiag}[M_1, Z_1],\ G \triangleq [0_{n-m},\ N_2]$, and $H \triangleq [M_1,\ 0_{(n-m)\times m}]$ are linear in terms of the variables. Then, using Proposition 4.2 and the Schur complement [8], the inequality (4.3) (i.e., $Y_{LQS} > 0$) is assured if and only if the following inequality is satisfied:

$$\begin{bmatrix} Y_{BD} + \Pi_1(H+G) + (H+G)^T \Pi_1^T + \Pi_1 M_2 \Pi_1^T & \star & \star \\ G & M_2 & \star \\ H & 0_{n-m} & M_2 \end{bmatrix} > 0 \tag{4.14}$$

TABLE 4.1
Illustration for the sequentially feasible iteration procedure.

Iteration no.	Scales		Feasible variables
k	$\Pi_{1,k}$	$\rightarrow$	$(M_{1,k},\ M_{2,k},\ N_{2,k})$
		$\swarrow$	*scale update*
$k+1$	$\Pi_{1,k+1}$	$\rightarrow$	$(M_{1,k+1},\ M_{2,k+1},\ N_{2,k+1})$
		$\swarrow$	*scale update*
$k+2$	$\Pi_{1,k+2}$	$\rightarrow$	$(M_{1,k+2},\ M_{2,k+2},\ N_{2,k+2})$
$\vdots$	$\vdots$		$\vdots$

for some Π_1, where $\star$ denotes the transpose of the corresponding off-diagonal block. Note that the necessity holds when

$$\Pi_1 = (H+G)^T M_2^{-1}. \tag{4.15}$$

In order to illustrate the property of (4.14), assume that the scale is fixed with $\Pi_1 := \Pi_{1,k}$ to make (4.14) a feasible LMI. Also, given $\Pi_{1,k}$, let $(M_{1,k},\ M_{2,k},\ N_{2,k})$ be the set of feasible variables for (4.14). Then the set should guarantee that $Y_{LQS,k} > 0$ due to sufficiency. In turn, according to necessity, the updated scale

$$\Pi_{1,k+1} = (H_k + G_k)^T M_{2,k}^{-1} \tag{4.16}$$

makes the inequality (4.14) feasible again (with the fixed scale $\Pi_1 := \Pi_{1,k+1}$). The procedure is depicted in Table 4.1. Consequently, these properties—the convex form by an LMI and the successive feasibility—allow us to handle the nonconvex inequality (i.e., $Y_{LQS} > 0$) by the convex one in the iterative procedure.

In a similar manner, one may eliminate nonconvexity in the inequality (4.6) by applying Proposition 4.3. That is, inequality (4.6) (with the nonconvex Y_{LQS}) holds if and only if the following inequality is satisfied:

$$\begin{bmatrix} \Psi(\Pi_2,\Pi_3) & \star & \star & \star & \star \\ GA_e^T & -M_2 & \star & \star & \star \\ HU_e^T & 0_{n-m} & -M_2 & \star & \star \\ HA_e^T & 0_{n-m} & 0_{n-m} & -M_2 & \star \\ GU_e^T & 0_{n-m} & 0_{n-m} & 0_{n-m} & -M_2 \end{bmatrix} < 0 \tag{4.17}$$

for some Π_2 and Π_3, where $A_e^T = [A^T,\ C_Q]$, $U_e^T = [I_n,\ 0_{n\times q}]$, and

$$\begin{aligned} \Psi(\Pi_2,\Pi_3) = &\begin{bmatrix} AY_{BD} + Y_{BD}A^T - BL_2 - L_2^T B^T & Y_{BD}C_Q \\ C_Q^T Y_{BD} & -I_q \end{bmatrix} \\ &+ \Pi_2(HU_e^T - GA_e^T) + (HU_e^T - GA_e^T)^T\Pi_2^T + \Pi_2 M_2 \Pi_2^T \\ &+ \Pi_3(GU_e^T - HA_e^T) + (GU_e^T - HA_e^T)^T\Pi_3^T + \Pi_3 M_2 \Pi_3^T. \end{aligned}$$

Moreover, the necessity holds when

$$\Pi_2 = \left(A_e G^T - U_e H^T\right) M_2^{-1}, \tag{4.18}$$

$$\Pi_3 = \left(A_e H^T - U_e G^T\right) M_2^{-1}. \tag{4.19}$$

Note that (4.17) becomes an LMI as long as the scales are fixed. Based on the discussion above, we now present an iterative algorithm that is successively feasible and convergent.

ALGORITHM I.

1. Assume $\Pi_{i,0}$ ($\forall i = 1, 2, 3$) that admit the feasibility of (4.14) and (4.17). And, set as $k := 1$.
2. Given $\Pi_i := \Pi_{i,k-1}$ ($\forall i = 1, 2, 3$), minimize γ w.r.t. $M_1,\ M_2,\ N_2,\ L_1$, and L_2 subject to (4.14), (4.4), (4.5), (4.17), and

$$\begin{bmatrix} \gamma I_{n-m} & I_{n-m} \\ I_{n-m} & M_1 \end{bmatrix} > 0, \tag{4.20}$$

   where Y_{POL} has the structure in (4.8).
3. Set $\gamma_k := \gamma_*$ and update the scales as

$$\Pi_{1,k} := (H_* + G_*)^T M_{2*}^{-1}, \tag{4.21}$$

$$\Pi_{2,k} := \left(A_e G_*^T - U_e H_*^T\right) M_{2*}^{-1}, \tag{4.22}$$

$$\Pi_{3,k} := \left(A_e H_*^T - U_e G_*^T\right) M_{2*}^{-1}, \tag{4.23}$$

   where $(_*)$ denotes the optimal values from step 2.
4. If $\gamma_{k-1} - \gamma_k < \epsilon$, stop the iteration. Otherwise, increase $k := k + 1$ and go to step 2.

Observe that the typical features of the algorithm are (a) the iteration is successively feasible once the initial guesses $\Pi_{i,0}$ ($i = 1, 2, 3$) are given, and (b) the performance upper-bound γ_k is successively decreasing such that $\gamma_{LQS} \le \gamma_{k+1} \le \gamma_k$, where γ_{LQS} alone is the optimum for the LQS problem.

Remark 4.4. For initiating Algorithm I, a set of matrix scales is needed in step 1. Observe that the solution from Theorem 4.1 is a special case when $M_1 = M_2$. Hence, the initial scales can be easily computed, according to (4.21)–(4.23), by utilizing the solution of Theorem 4.1. However, in case Theorem 4.1 is infeasible, we suggest the following steps. First, solve the pole-clustering problem alone to have matrices S and Y_{POL}. Then, given the sliding function coefficient S, construct $P_{LQS}(:= Y_{LQS}^{-1})$ following the construction method in the proof for statement (i) of Theorem 3.1 (i.e., use (3.4)). Now, using matrices Y_{LQS} and Y_{POL}, compute the scales following (4.21)–(4.23). In this way, Algorithm I can always be started as long as the pole-clustering constraint is feasible.

4.4. A numerical example. Consider the system borrowed from [5]:

$$A = \begin{bmatrix} 0.2325 & -0.9285 & 0.0154 & 0.1222 \\ -0.7274 & 1.0116 & -0.0224 & 0.1576 \\ -1.6883 & 0.2214 & 0.6534 & 1.6278 \\ -0.5310 & -0.2603 & -0.0052 & 1.1025 \end{bmatrix}, \quad B = \begin{bmatrix} 3.00 & 2.00 \\ 0 & 1.00 \\ 0.50 & -2.00 \\ 1.30 & 0 \end{bmatrix}.$$

The design objectives are defined by

$$Q = I_{4\times 4}, \quad c = 2, \quad \rho = 0.5.$$

To obtain the regular form (2.2), we choose a state transformation matrix $X = [U_2,\ U_1]^T$ from the singular value decomposition of B such that $B = [U_1,\ U_2]\left[\begin{smallmatrix} \sigma \\ 0_{2\times 2} \end{smallmatrix}\right] V^T$. Then we can start the design procedure using the transformed system

$$A := XAX^{-1}, \quad B := XB, \quad Q := X^{-T} Q X^{-1}.$$

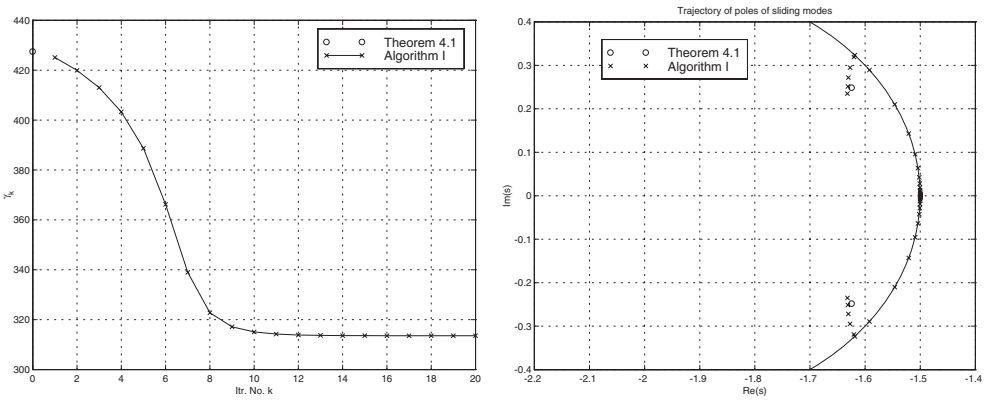


FIG. 4.1. *Convergence of γ_k (left). Trajectory of poles during iteration (right).*

First, we apply Theorem 4.1 to the above system, and the results are as follows:

$$\gamma_{cnvx} = 427.44, \quad S_{cnvx} = \begin{bmatrix} -2.8104 & -1.274 \\ -3.8506 & 36.686 \end{bmatrix}.$$

The poles of the sliding mode are placed at $\{-1.6244 \pm 0.24845i\}$, inside the constraint circle.

Now, as illustrated in Remark 4.4, we solve the algorithm using the results above (i.e., Y_{POL} and Y_{LQS} in solving Theorem 4.1) and calculate (4.21)–(4.23) in order to obtain the initial guesses for scales (i.e., $\Pi_{i,0}$). Then Algorithm I was started with them. As shown in Figure 4.1, the performance index ever decreases as the iteration proceeds. Interestingly, the poles in each iteration step move toward $s = -1.5$ following the circle boundary for the pole-clustering constraint. The algorithm shows the fast convergence and produces the results (at the 20th iteration)

$$\gamma_{20} = 313.56, \quad S_{20} = \begin{bmatrix} -1.6732 & -15.951 \\ -2.3586 & 21.139 \end{bmatrix},$$

which locates the sliding mode poles at $s = -1.5$. Observe that γ_{20} is reduced by 27% compared with γ_{cnvx}. The numerical results are summarized in Table 4.2.

TABLE 4.2
Comparison of numerical results.

Applied methods	Theorem 4.1	Algorithm I
Minimal γ	427.44	313.56

5. Further study. One of the interesting features of the proposed approach is that the sliding mode is derived from the usage of LMI conditions known for full state feedback design. In fact, this idea enables us to solve a few important issues in the sliding mode control context. Among them, let us consider polytopic uncertain systems [8] as follows:

$$(A, B) \in \mathbf{Co}\{(A^1, B^1), \ldots, (A^L, B^L)\}, \tag{5.1}$$

where $\mathbf{Co}\{\cdot\}$ denotes the convex hull and the matrices (A^i, B^i), $i = 1, \ldots, L$, are given and of the regular form. Then we have the result.

THEOREM 5.1. *There exist some stabilizing sliding function coefficients if there exist some matrices* $P > 0$ *and* $K \in \Re^{m\times n}$, *given any* $0 \le Q \in \Re^{n\times n}$, *satisfying*

$$(A^i - B^iK)^T P + P(A^i - B^iK) + Q < 0 \tag{5.2}$$

$\forall i \in [1, \ldots, L]$. *Furthermore, for the feasible matrix* P, *the matrix* $S = P_{22}^{-1}P_{12}^T$ *is a stabilizing sliding function coefficient.*

The proof is very similar to that of Theorem 3.1 (i.e., as in (3.6)). By pre- and postmultiplying the equation (5.2) by T_r and T_r^T, respectively, we have

$$\left(A_{11}^i - A_{12}^i P_{22}^{-1} P_{12}^T\right)^T P_r + P_r \left(A_{11}^i - A_{12}^i P_{22}^{-1} P_{12}^T\right) + T_r Q T_r^T < 0, \tag{5.3}$$

where $P_r = P_{11} - P_{12}^T P_{22}^{-1} P_{12}$. Since $P_r > 0$, the above implies that the matrix $P_{22}^{-1}P_{12}^T$ stabilizes the polytopic uncertain system defined in $\mathbf{Co}\{(A_{11}^1, A_{12}^1), \ldots, (A_{11}^L, A_{12}^L)\}$ in the reduced order space. Hence, with the matrix $S = P_{22}^{-1}P_{12}^T$, the sliding mode dynamics is stable in the invariant subspace $\{x \in \Re^n | x_2 + Sx_1 = 0\}$.

From the discussion above, the basic technique is expected, in a similar fashion, to be easily extended to a number of issues that have been treated by LMI-based full state feedback methods. For example, for the time-delayed systems or the parametric uncertain systems, the sliding mode can be obtained simply by solving the full state feedback problem and then combining the partitions of the corresponding Lyapunov matrix.

6. Concluding remarks. We proposed a new method for optimizing quadratic performance with the pole-clustering constraint in the sliding mode. It was explicitly shown that quadratic performance optimization and the pole-clustering problem can be solved by LMI approaches of full order without loss of generality. Using the proposed parametric constraints, the design objectives were combined under the LMI framework typically suited for sliding mode design in both convex and nonconvex forms. In the case of nonconvex formulation, a numerically tractable and efficient algorithm was presented. The effectiveness of the proposed approaches was illustrated by an example.

Finally, we illustrated the possible extension of the proposed approach to a class of polytopic uncertain systems. The sliding mode can be obtained easily by simply solving a full state feedback problem known in the literature. Note that the idea significantly simplifies sliding mode design procedures both conceptually and technically.

Acknowledgments. The authors sincerely thank the anonymous reviewers for their valuable comments and keen insights that substantially improved the study. In particular, the issue on the polytopic uncertain systems (motivated by the reviewers' suggestion) considerably extended the scope of the paper.

REFERENCES

[1] V. I. UTKIN AND K. D. YANG, *Methods for constructing discontinuity planes in multidimensional variable structure systems*, Autom. Remote Control, 39 (1979), pp. 1466–1470.

[2] R. A. DECARLO, S. H. ZAK, AND G. P. MATTHEWS, *Variable structure control of nonlinear multivariable systems: A tutorial*, Proc. IEEE, 76 (1988), pp. 212–232.

[3] O. M. E. EL-GHEZAWI, A. S. I. ZINOBER, AND S. A. BILLINGS, *Analysis and design of variable structure systems using a geometric approach*, Internat. J. Control, 38 (1983), pp. 657–671.

[4] C. M. DORLING AND A. S. I. ZINOBER, *Two approaches to hyperplane design in multivariable structure control systems*, Internat. J. Control, 44 (1986), pp. 65–82.

[5] W.-C. Su, S. V. Drakunov, and Ü. Özgüner, *Constructing discontinuity surfaces for variable structure systems: A Lyapunov approach*, Automatica J. IFAC, 32 (1996), pp. 925–928.
[6] H.-H. Choi, *An explicit formula of linear sliding surfaces for a class of uncertain dynamic systems with mismatched uncertainties*, Automatica J. IFAC, 34 (1998), pp. 1015–1020.
[7] K.-S. Kim, Y. Park, and S.-H. Oh, *Designing robust sliding hyperplanes for parametric uncertain systems: A Riccati approach*, Automatica J. IFAC, 36 (2000), pp. 1041–1048.
[8] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, PA, 1994.
[9] R. K. Yedavalli, *Robust root clustering for linear uncertain systems using generalized Lyapunov theory*, Automatica J. IFAC, 29 (1993), pp. 237–240.
[10] M. Chilali and P. Gahinet, *H_∞ design with pole placement constraints: An LMI approach*, IEEE Trans. Automat. Control, 41 (1996), pp. 358–367.
[11] K.-S. Kim and F. Jabbari, *Using scales in the multiobjective approach*, IEEE Trans. Automat. Control, 45 (2000), pp. 973–977.
[12] C. Edwards and S. K. Spurgeon, *Sliding Mode Control: Theory and Applications*, Taylor & Francis, London, 1998.
[13] K. Yasuda and Y. Nakatsuji, *Robust sliding mode control of uncertain systems*, in Proceedings of the 1996 IEEE Workshop on Variable Structure Systems, 1996, pp. 15–19.
[14] R. H. C. Takahashi and P. L. D. Peres, *H_2 guaranteed cost-switching surface design for sliding modes with nonmatching disturbances*, IEEE Trans. Automat. Control, 44 (1999), pp. 2214–2218.
[15] K.-S. Kim and Y. Park, *Using Lyapunov matrices for sliding mode design*, in Proceedings of the 39th Conference on Decision and Control, 2000, pp. 2204–2209.
[16] T. Shimomura and T. Fujii, *Multiobjective control design via successive overbounding of quadratic terms*, in Proceedings of the 39th Conference on Decision and Control, 2000, pp. 2763–2768.
[17] S. Barnett, *Polynomials and Linear Control Systems*, Monogr. Textbooks Pure Appl. Math. 77, Marcel Dekker, New York, 1983.
[18] C. Scherer, P. Gahinet, and M. Chilali, *Multi-objective output-feedback control via LMI optimization*, IEEE Trans. Automat. Control, 42 (1997), pp. 896–911.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 685–696

NULL CONTROLLABILITY OF THE FREE SURFACE OF A LIQUID IN A CONTAINER*

VILMOS KOMORNIK†, PAOLA LORETI‡, AND LEONARDO MAZZINI§

Abstract. This paper deals with the well-posedness, boundary observability, and boundary controllability of a model of lateral sloshing in moving containers. It begins with a short review of the lateral sloshing theory, then the problem is solved by combining the Fourier method and the Hilbert uniqueness method of J.-L. Lions. At the end an application is discussed.

Key words. controllability, Fourier series, fluid model, free surface

AMS subject classifications. 93B05, 93C20, 42C15, 76B10

DOI. 10.1137/S0363012902405285

1. The control of sloshing in moving containers. The sloshing equations in moving containers are used in many applications to study the motion of propellant fluids inside ships, aircraft, or space platforms; see, e.g., Silverman and Abramson [6]. The motion is often analyzed in order to demonstrate the stability of attitude control systems of the propellant carriers.

The behavior of the solutions depend very much on the proportion between the gravity forces and the surface tension terms, i.e., on the so-called Bond number defined by

$$\text{Bond number} = \frac{g\rho r^2}{\sigma}.$$

The equations we use are widely applied when the Bond number is much higher than one, as in the sloshing of water in a 1-g environment or in the sloshing of standard monomethile hydrazine propellant in accelerated rockets (with acceleration between $0,1$ g and 1 g) when the dimension (r) is larger than a half meter (as is often the case).

The model of lateral sloshing proposed in the paper has been used for many years in the sloshing of propellant inside satellite tanks in propulsed phases. The same model is used world-wide for the same purposes because the Bond number in such cases is typically higher than 10.

Today the sloshing in moving containers is best controlled by passive techniques like baffles and separators inside the tank. The control of the actual profile of the liquid phase has gotten less attention thus far, although it may become interesting in industrial applications to control the low frequency disturbances in spacecraft or ships that try to maintain a very quiet attitude. The generation of a pressure slope over a liquid phase is very far from being an impossible technical problem.

A simple implementation locates a number of valves connected to a high pressure air reservoir and a number of low pressure valves connected to a low pressure reservoir.

*Received by the editors April 8, 2002; accepted for publication (in revised form) January 14, 2004; published electronically August 4, 2004.

http://www.siam.org/journals/sicon/43-2/40528.html

†Institut de Recherche Mathématique Avancée, Université Louis Pasteur et CNRS, 7 rue René Descartes, 67084 Strasbourg Cedex, France (komornik@math.u-strasbg.fr).

‡Dipartimento di Metodi e Modelli, Matematici per le Scienze Applicate, Università di Roma "La Sapienza," Via A. Scarpa 16, 00161 Roma, Italy (loreti@dmmm.uniroma1.it).

§Alenia Spazio S.p.A., via Saccomuro 24, 00100 Roma, Italy (l.mazzini@rmd250b.roma.alespazio.it).

The valves are distributed on the top of the tank and are controlled by a computer. A set of pressure sensors located in the tank in the gas phase provides the computer with information for commanding the valves in order to obtain the desired slope of pressure.

The aim of this paper is to provide exact controllability results via the Hilbert uniqueness method (HUM) technique for the problem of controlling the liquid motion in a container via the distribution of the pressures in the container itself. This result deviates from any classic approach because it applies directly to the partial differential equations that describe the motion of the liquid without using any preliminary discretization of the problem, while most of the research on this subject uses preliminary discretization.

Active control of the liquid itself inside its container may be achieved by controlling the pressure distribution above the fluid free surface.

At the beginning of the paper, we introduce the equations for studying the linearized motion of an incompressible fluid inside a moving container, then we establish a proper mathematical setting for this problem, and finally we provide necessary and sufficient conditions for the control of this motion.

This control problem may find proper applications where the frequency of motion of the fluid strongly interferes with the controller of the attitude of the carrier, making it necessary to actively control the dynamics of the fluids. This approach is new in engineering and in control theory for this particular subject.

1.1. The sloshing equations. Our treatment shall start with the basic Navier–Stokes equations written for an accelerating frame introducing the hypothesis of fluid incompressibility and fluid irrotational motion. Following the reference [1], we introduce the following quantities:

Ω	liquid free boundary,
h	tank height,
$Q = \Omega \times (-h, 0)$	tank interior,
$\Gamma = \partial\Omega \times [-h, 0]$	tank wet boundary,
$\overrightarrow{x} = (x, y, z) \in \Omega \times [-h, 0]$	position coordinate inside the tank.

The following fields are relevant to the physics of the problem:

$\delta(x, y, t)$	liquid free surface height,
$\overrightarrow{v}(x, y, z, t) \in \mathbb{R}^3$	liquid velocity in the accelerating frame,
$\overrightarrow{u} = \int \overrightarrow{v}\, dt$	liquid displacement,
$\phi(x, y, z, t)$	scalar potential of liquid velocity $\nabla\phi = v$,
$\psi(x, y, z, t)$	scalar potential of liquid displacement $\nabla\psi = u$,
$p(x, y, z, t)$	pressure of the liquid,
$p_0(x, y, t)$	ullage pressure above the liquid surface,
$\overrightarrow{n} \in \mathbb{R}^3$	normal to the container,
$\overrightarrow{V} \in \mathbb{R}^3$	velocity of the container walls in the accelerating frame,
$\overrightarrow{w}(t)$	container rigid motion translation vector,
$\overrightarrow{\theta}(t)$	container rigid motion rotation vector;

furthermore, g denotes the acceleration $+z$ of the reference frame or equivalent gravitational acceleration in the $-z$ direction.

From the classical Navier–Stokes equations written in a gravitational field (taken, for example, from reference [7]), introducing the hypotheses of incompressibility and irrotational motion with respect to that frame, we derive that the liquid velocity can be derived as gradient of a potential that satisfies the so-called Bernoulli equation:

$$\nabla\left(\partial_t\phi + \frac{1}{2}\nabla\phi^T\nabla\phi + \frac{p}{\rho} + gz\right) = 0.$$

The velocity is derived by the following statements:

$$\Delta\phi = 0, \qquad \nabla\phi = \overrightarrow{v}.$$

The potential is not unique, for a generic time function can be added to it leaving the same physical solution, and we can add as a gauge condition that the Bernoulli equation is written as

$$\partial_t\phi + \frac{1}{2}\nabla\phi^T\nabla\phi + \frac{p}{\rho} + gz = 0.$$

The boundary conditions for such equations are

$$\partial_n\phi = \overrightarrow{V}\cdot\overrightarrow{n} \quad \text{in} \quad \Gamma$$

and

$$p(x,y,0,t) = p_0(x,y,t) \quad \text{in} \quad \Omega.$$

When the container is stationary in the accelerating frame and the ullage pressure is constant, we can derive the static equilibrium solution:

$$p = p_0 - \rho g z, \quad \phi = 0, \quad v = 0, \quad \delta = 0.$$

Now, let us consider a small rigid motion of the container and a small pressure variation in the ullage produced either by a control action or by external environmental conditions. We then have

$$\overrightarrow{V}(x,y,z,t) = \frac{d\overrightarrow{w}}{dt}(t) + \frac{d\overrightarrow{\theta}}{dt}(t)\times\overrightarrow{x}.$$

Since we consider rigid motion of the container, we have

$$\overrightarrow{u} = \nabla\psi = \nabla\int\phi dt,$$

$$\delta(x,y,t) = u_z(x,y,0,t) = \partial_z\psi|_{z=0}.$$

We can finally write for the linearized problem in the displacement potential

$$\Delta\psi = 0 \quad \text{in} \quad Q\times[0,T] \tag{1.1}$$

with boundary conditions at the container wet wall

$$\partial_n\psi = \overrightarrow{n}\cdot(\overrightarrow{w} + \overrightarrow{\theta}\times\overrightarrow{x}) \quad \text{on} \quad \Gamma\times[0,T] \tag{1.2}$$

and boundary conditions at the free surface derived by the Bernoulli equation:

$$\partial_{tt}^2\psi + g\partial_z\psi + \frac{p_0}{\rho} = 0 \quad \text{on} \quad \Omega \times [0,T]. \tag{1.3}$$

The pressure inside the liquid and at the container boundary can be derived by the Bernoulli equation

$$\partial_{tt}^2\psi + g(\partial_z\psi + z) + \frac{p}{\rho} = 0. \tag{1.4}$$

Equations (1.1) and (1.2), (1.3) in a suitable functional setting provide a well-defined set of equations, as will be shown in the following.

It is quite common in practice to divide the potential into components:

$$\psi = \overrightarrow{\psi}_w \cdot \overrightarrow{w} + \overrightarrow{\psi}_\theta \cdot \overrightarrow{\theta} + \psi_K.$$

Here $\overrightarrow{\psi}_w$ satisfies

$$\begin{aligned}
&\Delta \overrightarrow{\psi}_w = 0 \quad \text{in} \quad Q \times [0,T],\\
&\partial_n \overrightarrow{\psi}_w = \overrightarrow{n} \quad \text{in} \quad \Gamma \times [0,T],\\
&\partial_{tt}^2 \overrightarrow{\psi}_w + g\partial_z \overrightarrow{\psi}_w = 0 \quad \text{in} \quad \Omega \times [0,T],
\end{aligned}$$

while $\overrightarrow{\psi}_\theta$ satisfies

$$\begin{aligned}
&\Delta \overrightarrow{\psi}_\theta = 0 \quad \text{in} \quad Q \times [0,T],\\
&\partial_n \overrightarrow{\psi}_\theta = \overrightarrow{n} \quad \text{in} \quad \Gamma \times [0,T],\\
&\partial_{tt}^2 \overrightarrow{\psi}_\theta + g\partial_z \overrightarrow{\psi}_\theta = 0 \quad \text{in} \quad \Omega \times [0,T].
\end{aligned}$$

The last potential ψ_K has homogeneous boundary conditions at the wet container walls. So, ψ_K satisfies

$$\begin{aligned}
&\Delta\psi_K = 0 \quad \text{in} \quad Q \times [0,T],\\
&\partial_n\psi_K = 0 \quad \text{in} \quad \Gamma \times [0,T],
\end{aligned}$$

and

$$\partial_{tt}^2\psi_K + g\partial_z\psi_K + \frac{p_0}{\rho} = -\left(\overrightarrow{\psi}_w \cdot \frac{d^2\overrightarrow{w}}{dt^2} + g\partial_z \overrightarrow{\psi}_w \overrightarrow{w}\right) - \left(\overrightarrow{\psi}_\theta \cdot \frac{d^2\overrightarrow{\theta}}{dt^2} + g\partial_z \overrightarrow{\psi}_\theta \cdot \overrightarrow{\theta}\right)$$

in $\Omega \times [0,T]$.

While the first two vectorial potentials depend only on the shape of the container and can be solved a priori, the third scalar potential has homogeneous conditions at the wet boundary of the container and boundary conditions at the free surface containing all the external inputs, which are the ullage pressure distribution and the container motion variables with their second derivatives. These external inputs can be considered as control variables or external perturbations. It is to be remarked that controlling $\psi_K(x,y,z,t)$ to zero means that the motion of the fluid is rigid with the container. In the following we will focus our study on the production of a suitable controlling ullage pressure $p_0(x,y,t)$, so to produce $\psi_K(x,y,z,t) = 0$ for $t \geq T$ when the potential has been previously excited by a container motion occurring before $t = 0$, as per the classic paradigm of continuous controllability.

2. Well-posedness and boundary observability of the homogeneous problem. Let Ω be a bounded open domain in $\mathbb{R}^2$, having a sufficiently smooth boundary Γ. Given two positive numbers h and g, consider the following problem for an unknown function $\varphi(x,y,z,t)$:

$$\begin{cases} \Delta\varphi = 0 \quad \text{in} \quad \Omega\times(-h,0)\times\mathbb{R}, \\ \partial_\nu\varphi = 0 \quad \text{on} \quad \Gamma\times(-h,0)\times\mathbb{R}, \\ \partial_\nu\varphi = 0 \quad \text{on} \quad \Omega\times\{-h\}\times\mathbb{R}, \\ \frac{\partial^2\varphi}{\partial t^2} + g\frac{\partial\varphi}{\partial z} = 0 \quad \text{on} \quad \Omega\times\{0\}\times\mathbb{R}, \\ \varphi(x,y,0,0) = \varphi_0(x,y) \quad (x,y)\in\Omega, \\ \frac{\partial\varphi}{\partial t}(x,y,0,0) = \varphi_1(x,y), \quad (x,y)\in\Omega. \end{cases} \tag{2.1}$$

Let us introduce an orthonormal basis (e_n) of $L^2(\Omega)$, consisting of eigenfunctions of $-\Delta$ with homogeneous Neumann boundary conditions, corresponding to nonnegative eigenvalues λ_n, tending to ∞. Set

$$\omega_n = \sqrt{g\lambda_n \tanh(\lambda_n h)}.$$

PROPOSITION 2.1. *Given $\varphi_0 \in L^2(\Omega)$ and $\varphi_1 \in (H^1(\Omega))'$, problem (2.1) has a unique solution of the form*

$$\varphi(x,y,z,t) = \sum_{n=1}^{\infty} e_n(x,y)\cosh[\lambda_n(z+h)]\big(a_n e^{i\omega_n t} + b_n e^{-i\omega_n t}\big), \tag{2.2}$$

with suitable complex coefficients a_n and b_n. The series converges in $L^2(\Omega)$ uniformly in z and t, and its termwise derivative with respect to t,

$$\sum_{n=1}^{\infty} e_n(x,y)\cosh[\lambda_n(z+h)]i\omega_n\big(a_n e^{i\omega_n t} - b_n e^{-i\omega_n t}\big), \tag{2.3}$$

converges in $(H^1(\Omega))'$ uniformly in z and t.

Furthermore,

$$\varphi(\cdot,\cdot,0,\cdot) \in C(\mathbb{R};L^2(\Omega)) \cap C^1(\mathbb{R};(H^1(\Omega))')$$

and

$$\|\varphi(\cdot,\cdot,0,t)\|^2_{L^2(\Omega)} + \left\|\frac{\partial\varphi}{\partial t}(\cdot,\cdot,0,t)\right\|^2_{(H^1(\Omega))'} \le \alpha\big(\|\varphi_0\|^2_{L^2(\Omega)} + \|\varphi_1\|^2_{(H^1(\Omega))'}\big)$$

with some constant α, independent of $t\in\mathbb{R}$ and of the particular choice of the initial data φ_0 and φ_1.

Proof. Thanks to the choice of the eigenfunctions $e_n(x,y)$ and of the numbers ω_n, the functions of the form (2.2) satisfy the boundary conditions (2.3), regardless of the particular choice of the coefficients a_n and b_n. Setting $h=0$ and $t=0$ in (2.2) and (2.3) and using the initial conditions, we obtain the equations

$$\varphi_0(x,y) = \sum_{n=1}^{\infty} e_n(x,y)\cosh(\lambda_n h)(a_n+b_n)$$

and

$$\varphi_1(x,y)=\sum_{n=1}^{\infty} e_n(x,y)i\omega_n\cosh(\lambda_n h)(a_n-b_n).$$

Since (e_n) is an orthogonal basis in both $L^2(\Omega)$ and $(H^1(\Omega))'$, these equations determine the coefficients a_n and b_n uniquely.

On the other hand, choosing the coefficients according to these equations, a straightforward computation shows the convergence of the series as stated in the proposition.

For the proof of the inequality we first obtain by a direct computation the following estimates:

$$\|\varphi(\cdot,\cdot,0,t)\|^2_{L^2(\Omega)}\le\sum_{n=1}^{\infty}\cosh(\lambda_n h)^2\left|a_n e^{i\omega_n t}+b_n e^{-i\omega_n t}\right|^2$$

and

$$\left\|\frac{\partial\varphi}{\partial t}(\cdot,\cdot,0,t)\right\|^2_{(H^1(\Omega))'}\le\sum_{n=1}^{\infty}\cosh(\lambda_n h)^2\frac{|\omega_n|^2}{\lambda_n}\left|a_n e^{i\omega_n t}+b_n e^{-i\omega_n t}\right|^2.$$

Since ${\omega_n}^2/\lambda_n$ converges to one, it follows that

$$\begin{aligned}c_1\sum_{n=1}^{\infty}\cosh(\lambda_n h)^2\big(|a_n|^2+|b_n|^2\big)&\\ \le\|\varphi(\cdot,\cdot,0,t)\|^2_{L^2(\Omega)}+\left\|\frac{\partial\varphi}{\partial t}(\cdot,\cdot,0,t)\right\|^2_{(H^1(\Omega))'}&\\ \le c_2\sum_{n=1}^{\infty}\cosh(\lambda_n h)^2\big(|a_n|^2+|b_n|^2\big),&\end{aligned}$$

with two positive constants c_1 and c_2 which do not depend on the particular choice of t. Hence the desired inequality follows with $\alpha=c_2/c_1$. □

Fix an arbitrary positive number T and set

$$\delta:=\max_n\left|\frac{\sin(\omega_n T)}{\omega_n T}\right|.$$

Observe that $0\le\delta<1$.

We are going to establish the following inequalities.

PROPOSITION 2.2. *All solutions of* (2.1) *satisfy the inequalities*

$$\begin{aligned}(1-\delta)\sum_{n=1}^{\infty}\cosh^2(\lambda_n h)\big(|a_n|^2+|b_n|^2\big)\le\frac{1}{T}\int_0^T\int_\Omega|\varphi(x,y,0,t)|^2\,dx\,dy\,dt&\\ \le(1+\delta)\sum_{n=1}^{\infty}\cosh^2(\lambda_n h)\big(|a_n|^2+|b_n|^2\big).&\end{aligned}$$

Proof. We obtain by a straightforward computation, using the orthonormality of the sequence (e_n), the equality

$$\int_\Omega|\varphi(x,y,0,t)|^2\,dx\,dy=\sum_{n=1}^{\infty}\cosh^2(\lambda_n h)\left|a_n e^{i\omega_n t}+b_n e^{-i\omega_n t}\right|^2,$$

and then the equality

$$\int_0^T \int_\Omega |\varphi(x,y,0,t)|^2 \, dx \, dy \, dt = \sum_{n=1}^{\infty} \cosh^2(\lambda_n h) \int_0^T |a_n|^2 + |b_n|^2 + 2\Re(a_n \overline{b_n} e^{2i\omega_n t}) \, dt.$$

The proof will be completed if we show for every n the estimate

$$\left| \int_0^T 2\Re(a_n \overline{b_n} e^{2i\omega_n t}) \, dt \right| \le \delta T \big(|a_n|^2 + |b_n|^2 \big).$$

This can be shown as follows:

$$\begin{aligned} \left| \int_0^T 2\Re(a_n \overline{b_n} e^{2i\omega_n t}) \, dt \right| &= \left| 2\Re \left(a_n \overline{b_n} \int_0^T e^{2i\omega_n t} \, dt \right) \right| \\ &= \left| 2\Re \left(a_n \overline{b_n} \frac{e^{2i\omega_n T} - 1}{2i\omega_n} \right) \right| \\ &= \left| 2\Re \left(a_n \overline{b_n} e^{i\omega_n T} \frac{\sin \omega_n T}{\omega_n} \right) \right| \\ &\le 2|a_n| \cdot |b_n| \cdot \left| \frac{\sin(\omega_n T)}{\omega_n} \right| \\ &\le \delta T \big(|a_n|^2 + |b_n|^2 \big). \qquad \square \end{aligned}$$

Let us consider two examples.

Example 1. If Ω is a rectangle, say

$$\Omega = (0,a) \times (0,b),$$

then it is more convenient to arrange the eigenfunctions and eigenvalues in a double sequence

$$\begin{aligned} e_{m,n}(x,y) &= \frac{2}{\sqrt{ab}} \cos \frac{m\pi x}{a} \cdot \cos \frac{n\pi y}{b}, \\ \lambda_{m,n} &= \pi^2 \left(\frac{m^2}{a^2} + \frac{n^2}{b^2} \right), \end{aligned}$$

where m, n are integers ranging from 0 to ∞, and to set

$$\omega_{m,n} = \sqrt{g \lambda_{m,n} \tanh(\lambda_{m,n} h)}.$$

Then the formula (2.2) is replaced by

$$\begin{aligned} &\varphi(x,y,z,t) \\ &= \frac{2}{\sqrt{ab}} \sum_{m,n=0}^{\infty} \cos \frac{m\pi x}{a} \cdot \cos \frac{n\pi y}{b} \cdot \cosh[\lambda_{m,n}(z+h)] \big(a_{m,n} e^{i\omega_{m,n} t} + b_{m,n} e^{-i\omega_{m,n} t} \big), \end{aligned}$$

and the sums in the estimate of Proposition 2.1 are replaced by

$$\sum_{m,n=0}^{\infty} \cosh^2(\lambda_{m,n}h)\big(|a_{m,n}|^2+|b_{m,n}|^2\big).$$

Example 2. If Ω is a disc of radius R, then it is more convenient to use polar coordinates (r,θ) and to arrange the eigenfunctions and eigenvalues in a double sequence by using the Bessel functions, as, e.g., in [2] or [3, pp. 138–141]. Then the solutions of (2.1) are given by the series

$$\varphi(r,\theta,z,t)=\sum_{n=0}^{\infty}\sum_{k=1}^{\infty} J_n(\lambda_{nk}r)\cosh[\lambda_{nk}(z+h)]\cdot\Big(A_{nk}(\theta)e^{i\omega_{nk}t}+B_{nk}(\theta)e^{-i\omega_{nk}t}\Big),$$

where for each fixed n,

- J_n denotes the Bessel function of order n,
- $R\lambda_{n1}<R\lambda_{n2}<\cdots$ are the (strictly) positive roots of its derivative J_n',
- $\omega_{nk}=\sqrt{g\lambda_{nk}\tanh(\lambda_{nk}h)}$,
- $A_{nk}(\theta)$ and $B_{nk}(\theta)$ are suitable linear combinations of the functions $\cos(k\theta)$ and $\sin(k\theta)$, depending on the initial data.

The sums in the estimate of Proposition 2.1 are replaced by

$$\sum_{n=0}^{\infty}\sum_{k=1}^{\infty}\left(\int_0^R rJ_n(r)^2\,dr\right)\cdot\cosh^2(\lambda_{nk}h)\cdot\left(\int_0^{2\pi}|A_{nk}(\theta)|^2+|B_{nk}(\theta)|^2\,d\theta\right).$$

3. Well posedness of the nonhomogeneous problem. Now let us consider the following nonhomogeneous version of problem (2.1) for an unknown function $\psi(x,y,z,t)$:

$$\begin{cases}\Delta\psi=0 \quad \text{in} \quad \Omega\times(-h,0)\times\mathbb{R},\\ \partial_\nu\psi=0 \quad \text{on} \quad \Gamma\times(-h,0)\times\mathbb{R},\\ \partial_\nu\psi=0 \quad \text{on} \quad \Omega\times\{-h\}\times\mathbb{R},\\ \frac{\partial^2\psi}{\partial t^2}+g\frac{\partial\psi}{\partial z}=v \quad \text{on} \quad \Omega\times\{0\}\times\mathbb{R},\\ \psi(x,y,0,0)=\psi_0(x,y), \quad (x,y)\in\Omega,\\ \frac{\partial\psi}{\partial t}(x,y,0,0)=\psi_1(x,y), \quad (x,y)\in\Omega.\end{cases} \tag{3.1}$$

We are going to define the solutions of this problem by the method of transposition. Consider an arbitrary solution of (2.1). By a formal computation, we have for every real number T the following equalities:

$$\begin{aligned}0&=\int_0^T\int_{-h}^0\int_\Omega(\Delta\varphi)\psi-\varphi(\Delta\psi)\,dx\,dy\,dz\,dt\\ &=\int_0^T\int_{-h}^0\int_\Gamma(\partial_\nu\varphi)\psi-\varphi(\partial_\nu\psi)\,\Gamma\,dz\,dt\\ &\qquad+\int_0^T\int_\Omega\big((\partial_\nu\varphi)\psi-\varphi(\partial_\nu\psi)\big)(x,y,-h,t)\,dx\,dy\,dt\\ &\qquad+\int_0^T\int_\Omega\big((\partial_\nu\varphi)\psi-\varphi(\partial_\nu\psi)\big)(x,y,0,t)\,dx\,dy\,dt\\ &=\int_0^T\int_\Omega\left(\frac{\partial\varphi}{\partial z}\psi-\varphi\frac{\partial\psi}{\partial z}\right)(x,y,0,t)\,dx\,dy\,dt\end{aligned}$$

$$= \frac{1}{g}\int_0^T \int_\Omega \left(-\frac{\partial^2\varphi}{\partial t^2}\psi + \varphi\frac{\partial^2\psi}{\partial t^2} - \varphi v\right)(x,y,0,t)\ dx\ dy\ dt$$
$$= -\frac{1}{g}\int_0^T \int_\Omega (\varphi v)(x,y,0,t)\ dx\ dy\ dt + \frac{1}{g}\left[\int_\Omega \left(-\frac{\partial\varphi}{\partial t}\psi + \varphi\frac{\partial\psi}{\partial t}\right)(x,y,0,t)\ dx\ dy\right]_0^T.$$

Taking the initial conditions into account, it follows that

$$\int_\Omega \left(-\frac{\partial\varphi}{\partial t}\psi + \varphi\frac{\partial\psi}{\partial t}\right)(x,y,0,T)\ dx\ dy = \int_\Omega (-\varphi_1\psi_0 + \varphi_0\psi_1)\ dx\ dy + \int_0^T\int_\Omega (\varphi v)(x,y,0,t)\ dx\ dy\ dt.$$

Identifying $L^2(\Omega)$ with its dual $(L^2(\Omega))'$ as usual, we have the dense and continuous inclusions

$$(H^1(\Omega))' \subset (L^2(\Omega))' = L^2(\Omega) \subset H^1(\Omega).$$

Then we may rewrite the last identity in the following form:

$$\left\langle \left(\frac{\partial\psi}{\partial t}, -\psi\right)(\cdot,\cdot,0,T), \left(\varphi, \frac{\partial\varphi}{\partial t}\right)(\cdot,\cdot,0,T)\right\rangle_{L^2(\Omega)\times H^1(\Omega), L^2(\Omega)\times(H^1(\Omega))'} = \langle(\psi_1,-\psi_0),(\varphi_0,\varphi_1)\rangle_{L^2(\Omega)\times H^1(\Omega), L^2(\Omega)\times(H^1(\Omega))'} + (\varphi(\cdot,\cdot,0,\cdot), v)_{L^2(\Omega\times(0,T)}. \tag{3.2}$$

This leads to the following definition.

DEFINITION 3.1. *A solution of* (3.1) *is a function*

$$\psi(\cdot,\cdot,0,\cdot) \in C(\mathbb{R}; H^1(\Omega)) \cap C^1(\mathbb{R}; L^2(\Omega)),$$

satisfying (3.2) *for all* $\varphi_0 \in L^2(\Omega)$ *and* $\varphi_1 \in (H^1(\Omega))'$.

The definition is justified by the following claim.

PROPOSITION 3.2. *Given* $\psi_0 \in H^1(\Omega)$, $\psi_1 \in L^2(\Omega)$, *and* $v \in L^2_{loc}(\mathbb{R}; L^2\Omega))$ *arbitrarily, problem* (3.1) *has a unique solution.*

Proof. For any fixed T, the right-hand side of (3.2) defines a bounded linear form of

$$(\varphi_0,\varphi_1) \in L^2(\Omega)\times(H^1(\Omega))'.$$

Since the linear map

$$(\varphi_0,\varphi_1) \mapsto \left(\varphi, \frac{\partial\varphi}{\partial t}\right)(\cdot,\cdot,0,T)$$

is an automorphism of $L^2(\Omega)\times(H^1(\Omega))'$ onto itself, the existence of a unique couple

$$\left(\psi, \frac{\partial\psi}{\partial t}\right)(\cdot,\cdot,0,T) \in L^2(\Omega)\times(H^1(\Omega))'$$

satisfying (3.2) follows.

Since both sides of (3.2) change continuously with T, the solution also depends continuously of T, thereby completing the proof of Proposition 3.2. ∎

4. Boundary controllability of liquid containers. A crucial idea in the HUM of Lions (see [4] and [5]) was the construction of suitable controls by solving a corresponding homogeneous dual problem. Using this duality approach, we establish in this section the following controllability result for problem (3.1).

THEOREM 4.1. *Given a positive number T and initial data $\psi_0 \in H^1(\Omega)$, $\psi_1 \in L^2(\Omega)$, there exists a function $v \in L^2(0,T;L^2(\Omega))$ such that the solution of* (3.1) *satisfies*

$$\left(\psi, \frac{\partial \psi}{\partial t}\right)(\cdot,\cdot,0,T) = 0.$$

Proof. Given $\varphi_0 \in L^2(\Omega)$ and $\varphi_1 \in (H^1(\Omega))'$ arbitrarily, solve the homogeneous problem (2.1) and then solve the following nonhomogeneous problem:

$$\begin{cases} \Delta\psi = 0 \quad \text{in} \quad \Omega \times (-h,0) \times \mathbb{R}, \\ \partial_\nu \psi = 0 \quad \text{on} \quad \Gamma \times (-h,0) \times \mathbb{R}, \\ \partial_\nu \psi = 0 \quad \text{on} \quad \Omega \times \{-h\} \times \mathbb{R}, \\ \frac{\partial^2 \psi}{\partial t^2} + g\frac{\partial \psi}{\partial z} = \varphi \quad \text{on} \quad \Omega \times \{0\} \times \mathbb{R}, \\ \psi(x,y,0,T) = 0, \quad (x,y) \in \Omega, \\ \frac{\partial \psi}{\partial t}(x,y,0,T) = 0, \quad (x,y) \in \Omega. \end{cases} \tag{4.1}$$

(The well-posedness of this problem follows from Proposition 3.2 because the time 0 does not play any special role in this problem.) It is sufficient to prove that for a suitable choice of the initial data φ_0 and φ_1 we have

$$\psi(x,y,0,0) = \psi_0(x,y) \quad \text{and} \quad \frac{\partial \psi}{\partial t}(x,y,0,0) = \psi_1(x,y), \quad (x,y) \in \Omega.$$

Using the identity (3.2) of the preceding section, now we have

$$\langle (-\psi_1, \psi_0), (\varphi_0, \varphi_1) \rangle_{L^2(\Omega)\times H^1(\Omega), L^2(\Omega)\times (H^1(\Omega))'} = \int_0^T \int_\Omega |\varphi(x,y,0,t)|^2 \, dx \, dy \, dt. \tag{4.2}$$

Thanks to Proposition 2.2, the right-hand side of this identity is a positive definite quadratic form of $(\varphi_0, \varphi_1) \in L^2(\Omega) \times (H^1(\Omega))'$. Applying the Lax–Milgram theorem (or simply the Riesz–Fréchet theorem), we conclude that the linear map

$$(\varphi_0, \varphi_1) \mapsto (-\psi_1, \psi_0)$$

maps $L^2(\Omega) \times (H^1(\Omega))'$ onto $L^2(\Omega) \times H^1(\Omega)$, and the theorem follows. □

5. Spectral boundary controllability of liquid containers. The controls applied in the preceding section are difficult to realize in practice. Here we present a more realistic variant: we apply simpler controls but we only look for partial controllability by eliminating a finite number of modes from the solution.

Choose a positive integer N and N functions $f_1(x,y), \dots, f_N(x,y)$ on Ω. Using controls of the form

$$v(x,y,t) := \sum_{j=1}^{N} (a_j t + b_j) f_j(x,y),$$

where we can act by choosing suitable constants a_j and b_j for $j = 1, \dots, N$, we try to drive the system to a final state satisfying the orthogonality conditions

$$\int_\Omega \psi(x,y,0,T)e_n(x,y)\,dx\,dy = \int_\Omega \frac{\partial\psi}{\partial t}(x,y,0,T)e_n(x,y)\,dx\,dy = 0$$

for $n = 1, \dots, N$.

Expanding the given functions $f_j(x,y)$ into Fourier series

$$f_j(x,y) = \sum_{n=1}^{\infty} f_{jn}e_n(x,y), \quad j = 1, \dots, N,$$

the control takes the form

$$v(x,y,t) := \sum_{n=1}^{\infty} \left(\sum_{j=1}^{N} f_{jn}(a_j t + b_j) \right) e_n(x,y).$$

Therefore the solution of (3.1) is given by the series

$$\psi(x,y,z,t) = \sum_{n=1}^{\infty} e_n(x,y)\cosh[\lambda_n(z+h)]g_n(t),$$

where the functions $g_n(t)$ are solutions to the initial-value problems

$$\begin{aligned} &g_n''(t)\cosh(\lambda_n h) + g_n(t)g\lambda_n \sinh(\lambda_n h) = \sum_{j=1}^{N} f_{jn}(a_j t + b_j), \\ &g_n(0)\cosh(\lambda_n h) = c_n, \\ &g_n'(0)\cosh(\lambda_n h) = d_n, \end{aligned}$$

where the constants c_n and d_n depend on the initial data.

We have to prove that for any given c_n and d_n there exist constants a_n and b_n such that

$$g_n(T) = g_n'(T) = 0, \quad n = 1, \dots, N. \tag{5.1}$$

Equivalently, we have to prove that if $c_n = d_n = 0$ for $n = 1, \dots, N$, then we only have the trivial solution $a_n = b_n = 0$ for $n = 1, \dots, N$.

In this homogeneous case we explicitly compute the solution: setting

$$\omega_n = \sqrt{g\lambda_n \tanh(\lambda_n h)}$$

as before, we have

$$g_n(t) = \sum_{j=1}^{N} f_{jn} \left(\frac{\omega_n t - \sin(\omega_n t)}{\omega_n^3} a_j + \frac{1 - \cos(\omega_n t)}{\omega_n^2} b_j \right)$$

for $n = 1, \dots, N$. It follows that the conditions (5.1) are equivalent to the following system of linear equations:

$$\begin{aligned} &\sum_{j=1}^{N} f_{jn}\big(\omega_n T - \sin(\omega_n T)\big)a_j + \sum_{j=1}^{N} f_{jn}\omega_n\big(1 - \cos(\omega_n T)\big)b_j = 0, \qquad n = 1, \dots, N, \\ &\sum_{j=1}^{N} f_{jn}\big(1 - \cos(\omega_n T)\big)a_j + \sum_{j=1}^{N} f_{jn}\omega_n \sin(\omega_n T)b_j = 0, \qquad n = 1, \dots, N. \end{aligned}$$

We arrive in this way at the following result.

THEOREM 5.1. *Assume that the determinant of the above system is different from zero for some $T > 0$. Then for any given initial data $\psi_0 \in H^1(\Omega)$, $\psi_1 \in L^2(\Omega)$, there exist constants $a_1, \dots, a_N$ and $b_1, \dots, b_N$ such that, applying the control*

$$v(x, y, t) := \sum_{j=1}^{N} (a_j t + b_j) f_j(x, y),$$

the solution of (3.1) *satisfies*

$$\int_\Omega \psi(x, y, 0, T) e_n(x, y)\ dx\ dy = \int_\Omega \frac{\partial \psi}{\partial t}(x, y, 0, T) e_n(x, y)\ dx\ dy = 0$$

for $n = 1, \dots, N$.

Example 3. Let us consider the simplest case where $N = 1$ and $f_1(x, y) = e_1(x, y)$. Then, setting $\alpha := T\omega_1$ for brevity, the determinant is equal to

$$\begin{vmatrix} \alpha - \sin\alpha & \omega_1(1 - \cos\alpha) \\ 1 - \cos\alpha & \omega_1 \sin\alpha \end{vmatrix} = 2\omega_1 \sin\alpha \left(\frac{\alpha}{2} - \tan\frac{\alpha}{2} \right).$$

It follows that the determinant vanishes if and only if $\alpha = T\omega_1$ is an integer multiple of π, i.e., if and only if

$$T = \frac{k\pi}{\sqrt{g\lambda_1 \tanh(\lambda_1 h)}}$$

for some positive integer k. For all other values of $T > 0$ the problem is controllable in the above sense.

Remark. From the point of view of applications, it is interesting to find control constants a_j and b_j of a moderate size: in some sense the size of these constants may be viewed as a measure of the cost of the control. (Furthermore, for large values of the controls our model may not be any more realistic.) This is equivalent to choosing functions $f_1, \dots f_N$ such that the inverse of the matrix of the above linear system has a sufficiently large norm. On the other hand, it is also useful to find relatively simple functions f_j, concentrated on some points of the domain Ω. These two, in a sense contradictory, requirements lead to an interesting optimization problem of a geometric nature, which can be studied in special cases if, for example, Ω is a rectangle or a disk.

REFERENCES

[1] H. N. ABRAMSON, *The Dynamic Behavior of Liquids in Moving Containers*, NASA-SP-106, 1966.

[2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics* I, John Wiley & Sons, New York, 1989.

[3] D. JACKSON, *Fourier Series and Orthogonal Polynomials*, Carus Math. Monogr. 6, The Mathematical Association of America, Oberlin, OH, 1941.

[4] J.-L. LIONS, *Exact controllability, stabilization, and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[5] J.-L. LIONS, *Contrôlabilité exacte et stabilisation de systèmes distribués*, Vol. 1, Masson, Paris, 1988.

[6] S. SILVERMAN AND H. N. ABRAMSON, *Lateral sloshing in moving containers*, in The Dynamic Behavior of Liquids in Moving Containers, H. N. Abramson, Technical Report NASA-SP-106, 1966, pp. 13–78.

[7] R. TEMAM, *Navier-Stokes Equations and Nonlinear Functional Analysis*, SIAM, Philadelphia, 1983.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 697–707

CHARACTERIZATION OF THE STATE CONSTRAINED MINIMAL TIME FUNCTION*

R. J. STERN†

Abstract. A standard class of finite dimensional control systems is considered, along with a state constraint set S and a target set $\Sigma \subset S$. Under certain geometric assumptions on S and a required S-constrained small time controllability property, a proximal Hamilton–Jacobi characterization of the S-constrained minimal time function to target Σ is obtained.

Key words. state constrained minimal time function, proximal Hamilton–Jacobi inequalities, small time controllability

AMS subject classifications. 49L99, 49N99, 49J52

DOI. 10.1137/S0363012903426033

1. Introduction. We shall consider a standard finite dimenisonal control system $\dot{x}(t) = f(x(t), u(t))$, $u(\cdot) \in U$, along with a state constraint $x(t) \in S$ and a closed target set $\Sigma \subset S$. Our goal is to provide, apparently for the first time, Hamilton–Jacobi (HJ) characterizations of the S-constrained minimal time function with target Σ. The unconstrained (that is, $S = \mathbb{R}^n$) form of this problem was considered by several researchers. In Bardi [1] such a characterization was obtained under a continuity hypothesis encapsulated as a small time controllability condition. In Bardi and Staicu [2] the discontinuous case was settled for targets which are the closure of their interior, while the general case was resolved in Soravia [23]. The nonsmooth characterizations obtained in the aforementioned references are provided in viscosity terms. A proximal characterization equivalent to that of Soravia is among the results obtained by Wolenski and Zhuang [25]. There it was shown that under mild hypotheses, the unconstrained minimal time function is the unique function satisfying, among other properties, a proximal HJ equation on the complement of Σ, along with a boundary condition which takes the form of a certain proximal HJ inequality holding on Σ itself. Useful references on the minimal time function and viscosity approaches are Bardi and Capuzzo-Dolcetta [3] and Cannarsa and Sinestrari [5]. Of related interest to this discussion are recent results of Clarke and Nour [11], where the proximal HJ equation was reconsidered, but in the absence of the aforementioned boundary condition. The solutions obtained to the proximal HJ equation in this new framework were studied, and shown to be associated with geodesic trajectories.

As in [25] and [11], our methods here are based upon nonsmooth proximal monotonicity-invariance considerations as developed in Clarke, Ledyaev, Stern, and Wolenski [8], [10]. But as will be seen, the imposition of a state constraint necessitates additional techniques. Of particular interest is the role played by a required S-constrained small time controllability hypothesis. Geometric sufficient conditions for the latter property will be provided as well, in the concluding comments.

State constrained control problems have received attention in recent years. Differ-

*Received by the editors April 8, 2003; accepted for publication (in revised form) September 23, 2003; published electronically August 4, 2004. This research was supported by the Natural Sciences Engineering Research Council of Canada and Le Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec.

http://www.siam.org/journals/sicon/43-2/42603.html

†Department of Mathematics and Statistics, Concordia University, 1400 De Maisonneuve Blvd., Montreal, Quebec H3G 1M8, Canada (stern@vax2.concordia.ca).

ential games with state constraints were studied in Bardi, Koike, and Soravia [4] and Cardaliaguet, Quincampoix, and Saint-Pierre [6]. In Clarke, Rifford, and Stern [12], a state constrained Mayer problem was studied, and in Clarke and Stern [13], the problem of state constrained stabilization was considered. In the two latter references, the emphasis was on the construction of feedback controls. In Clarke and Stern [14], the value function of the state constrained problem studied in [12] was characterized in proximal HJ terms; in this regard, see also Frankowska and Vinter [18] and Vinter [24]. As is the case in [14], the method here will rely upon state constrained trajectory tracking properties developed in [12], and the proximal characterizations can be reframed, as they were in [14], in the "constrained viscosity solution" framework introduced by Soner [22]. This too will be remarked upon in the concluding comments.

We shall consider a control system of the form

$$\dot{x}(t) = f(x(t), u(t)) \quad \text{a.e.} \tag{1.1}$$

The state trajectory $x(\cdot)$ evolves in $\mathbb{R}^n$ and control functions $u(\cdot)$ are Lebesgue measurable functions $u : \mathbb{R} \to U$, where $U \subset \mathbb{R}^m$ is a compact control restraint set.

Hypotheses on the dynamics $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ are as follows, and will be assumed to hold throughout. (The Euclidean norm is denoted by $\|\cdot\|$.)

(F1) The function f is continuous and locally Lipschitz in the state variable x, uniformly for $u \in U$; that is, for each bounded set $\Gamma \subset \mathbb{R}^n$, there exists $K_\Gamma > 0$ such that

$$\|f(x,u) - f(y,u)\| \le K_\Gamma \|x - y\|$$

whenever (x,u) and (y,u) are in $\Gamma \times U$.

(F2) The function f satisfies a linear growth condition; that is, there exist positive numbers c_1, c_2 such that

$$\|f(x,u)\| \le c_1\|x\| + c_2 \quad \forall\, (x,u) \in \mathbb{R}^n \times U.$$

(F3) The velocity set $f(x,U)$ is convex for every $x \in \mathbb{R}^n$.

Under (F1)–(F2), for every initial state α and every control function $u(\cdot)$, there exists a unique trajectory $x(t) = x(t;\alpha,u(\cdot))$ defined for all $t \ge 0$ and satisfying $x(0) = \alpha$. Actually, as is explained in the concluding comments, condition (F2) can be dropped in the results to follow, but we include it now for ease of exposition. The imposition of (F3) is needed in order to have available the familiar property of sequential compactness of trajectories on compact time intervals, as is explained in [10].

The next section provides a required result on S-constrained trajectory tracking. The main result is presented in section 3, while section 4 consists of the concluding comments. Prior to proceeding, we refer the reader to [10] for all the basic definitions and facts from nonsmooth analysis that will be required.

2. *S*-constrained trajectory tracking. Geometric hypotheses on S are now posited and will be assumed to hold throughout.

(S1) S is compact and *wedged* at each $x \in \mathrm{bdry}(S)$, meaning that at each boundary point x one has pointedness of $N_S^C(x)$; that is, $N_S^C(x) \cap \{-N_S^C(x)\} = \{0\}$. Here $N_S^C(x)$ denotes the Clarke normal cone to S at x. (This is equivalent to $T_S^C(x)$ having nonempty interior for each $x \in \mathrm{bdry}(S)$, where $T_S^C(x)$ denotes the Clarke tangent cone to S at x.)

(S2) The following "strict inwardness" condition holds:

$$\min_{u\in U}\langle \zeta, f(x,u)\rangle < 0 \quad \forall\, 0 \neq \zeta \in N_S^C(x), \quad \forall\, x \in \mathrm{bdry}(S). \tag{2.1}$$

The set S being wedged at $x \in \mathrm{bdry}(S)$ is also referred to in the literature as *epi-Lipschitzness* of S at x, since it is equivalent to S being locally linearly homeomorphic to the epigraph of a Lipschitz function; see Rockafellar [20] and Clarke [7].

Remark 2.1.

(a) We will require the fact that when S is wedged at each of its boundary points, then S is the closure of its interior.

(b) (S1)–(S2) are sufficient but not necessary for S to be *weakly invariant*; that is, for any initial state $\alpha \in S$, there exists a control $u(\cdot)$ producing a trajectory $x(t) = x(t;\alpha,u(\cdot))$ with $x(0) = \alpha$ and $x(t) \in S$ for all $t \geq 0$.

Given $r \geq 0$, an *inner approximation* of S is given by

$$S_r := \{x \in S : d_{\hat{S}}(x) \geq r\}.$$

Here $\hat{S}$ denotes the closure of the complement of S, and $d_\Gamma(\cdot)$ denotes the Euclidean distance function to a set Γ. (Note that $S_0 = S$.) The geometry of inner approximations was studied in Clarke, Ledyaev, and Stern [9], as well as in [12].

We will require the following S-constrained tracking result from the proofs of Proposition 3.13 and Theorem 3.10 in [12]. This result generalizes one due to Soner [22]; see also Forcellini and Rampazzo [16] and Frankowska and Rampazzo [17].

LEMMA 2.2.

(a) *Given $T > 0$, there exists a constant $M = M(T) > 0$ such that the following holds: Let $\alpha_0 \in S$, $\alpha_1 \in S$, and let $u_0(\cdot)$ be a control function on $[0,T]$ producing a trajectory which satisfies*

$$x(t;\alpha_0,u_0(\cdot)) \in S \quad \forall\, t \in [0,T]. \tag{2.2}$$

Then there exists a control function $u_1(\cdot)$ on $[0,T]$ which produces a trajectory that satisfies

$$\|x(t;\alpha_1,u_1(\cdot)) - x(t;\alpha_0,u_0(\cdot))\| \leq M\|\alpha_1 - \alpha_0\| \quad \forall\, t \in [0,T] \tag{2.3}$$

and

$$x(t;\alpha_1,u_1(\cdot)) \in S \quad \forall\, t \in [0,T]. \tag{2.4}$$

(b) *There exists a constant $r_0 > 0$ so that the following holds: Given $T > 0$, there exists $W = W(T) > 0$ such that for any initial state $\alpha \in \mathrm{int}(S)$, if $r \in [0,r_0]$ is such that $r \leq d_{\hat{S}}(\alpha)$ and $u(\cdot)$ is a control function on $[0,T]$ such that*

$$x(t;\alpha,u(\cdot)) \in S \quad \forall\, t \in [0,T], \tag{2.5}$$

then there exists a control function $\bar{u}(\cdot)$ on $[0,T]$ such that

$$\|x(t;\alpha,\bar{u}(\cdot)) - x(t;\alpha,u(\cdot))\| \leq rW \quad \forall\, t \in [0,T] \tag{2.6}$$

and

$$x(t;\alpha,\bar{u}(\cdot)) \in S_r \quad \forall\, t \in [0,T]. \tag{2.7}$$

3. The *S*-constrained minimal time function. Let Σ be a closed subset of S, and denote by Γ_S the set of $\alpha \in S$ such that for some control $u(\cdot)$ and some $\hat{t} \geq 0$ one has

$$x(\hat{t}; \alpha, u(\cdot)) \in \Sigma$$

and

$$x(t; \alpha, u(\cdot)) \in S \quad \forall\, t \in [0, \hat{t}].$$

In other words, Γ_S consists of those initial states in S which can be controlled in finite time to the target Σ along an S-constrained trajectory. (Trivially, we have $\Sigma \subset \Gamma_S$.) Denote by $\tau(\alpha)$ the infimal time from an initial state $\alpha \in \Gamma_S$ to the target Σ along such a trajectory. By a standard sequential compactness of trajectories argument, the infimum is attained as a minimum. We go on to define an extended real-valued function $T : \mathbb{R}^n \to [0, +\infty]$ as follows:

$$T(\alpha) := \begin{cases} \tau(\alpha) & \text{if } \alpha \in \Gamma_S, \\ +\infty & \text{otherwise.} \end{cases}$$

$T(\cdot)$ is called the *S-constrained minimal time function* with respect to the target Σ. One can show that $T(x) \geq 0$ for all $x \in \mathbb{R}^n$ and $T(\alpha) = 0$ if and only if $\alpha \in \Sigma$. Observe that there are *two* ways for $T(\alpha)$ to be nonfinite:

(i) $\alpha \in S$, but Σ is not reachable from α via an S-constrained trajectory; that is, $\alpha \in S \backslash \Gamma_S$.
(ii) $\alpha \notin S$.

Let us collect some basic properties of the functions $T(\cdot)$ and $g(\cdot, \cdot)$, where

$$g(t, x) := t + T(x).$$

We denote the set of strictly positive real numbers by $\mathbb{R}_+$, and the complement of Σ by Σ^c.

(LSC) $T(\cdot)$ is lower semicontinuous at every $\alpha \in \mathbb{R}^n$; that is,

$$T(\alpha) \leq \liminf_{\alpha' \to \alpha} T(\alpha').$$

(SI) $g(\cdot, \cdot)$ is *strongly increasing* on $\mathbb{R}_+ \times \operatorname{int}(S)$. This means that for every $\alpha \in \operatorname{int}(S)$, for every trajectory of (1.1) with $x(0) = \alpha$, one has $g(t, x(t)) \geq T(\alpha)$ on $[0, s]$ for every $s > 0$ such that $x(t) \in \operatorname{int}(S)$ on $[0, s]$.

(WD) $g(\cdot, \cdot)$ is *weakly decreasing* on $\mathbb{R}_+ \times \Sigma^c$. This means that for every $\alpha \in \Sigma^c$, there exists a trajectory of (1.1) with $x(0) = \alpha$, for which $g(t, x(t)) \leq T(\alpha)$ on $[0, s]$ for every $s > 0$ such that $x(t) \in \Sigma^c$ on $[0, s]$.

Note that in (SI) and (WD), infinite values of $T(\cdot)$ and $g(\cdot, \cdot)$ are possible. In particular, in (SI), note that if a trajectory in $\operatorname{int}(S)$ exits Γ_S, it cannot re-enter Γ_S while remaining in S. Property (LSC) readily follows from a sequential compactness of trajectories argument. The strong increase property (SI) is a straightforward consequence of the principle of optimality, while the weak decrease property (WD) is due to the existence of S-constrained time-optimal trajectories from any startpoint α where $T(\alpha)$ is finite; that is, $\alpha \in \Gamma_S$. (Startpoints for which $T(\alpha) = +\infty$ trivially satisfy the weak decrease condition.)

We now introduce the *lower Hamiltonian* $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by

$$h(x, p) := \min_{u \in U} \langle p, f(x, u) \rangle, \tag{3.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product.

The next lemma provides proximal characterizations of (SI) and (WD) in terms of proximal HJ inequalities. It follows in a straightforward manner from Exercise 4.6.4(a) and Proposition 4.6.5 in [10].

LEMMA 3.1.

(i) *Property* (SI) *is equivalent to*

$$h(x, \partial_P T(x)) + 1 \geq 0 \quad \forall\, x \in \text{int}(S). \tag{3.2}$$

(ii) *Property* (WD) *is equivalent to*

$$h(x, \partial_P T(x)) + 1 \leq 0 \quad \forall\, x \in \Sigma^c. \tag{3.3}$$

Here ∂_P denotes the *proximal subdifferential*; recall that for an extended real-valued lower semicontinuous function $w : \mathbb{R}^n \to (-\infty, +\infty]$ and a point x, where $w(x) < +\infty$, $\zeta \in \partial_P w(x)$ if and only if there exists $\sigma = \sigma(x, \zeta)$ such that

$$w(y) - w(x) + \sigma \|y - x\|^2 \geq \langle \zeta, y - x \rangle$$

for all y near x. Thus (3.2) says that

$$h(x, \zeta) + 1 \geq 0 \quad \forall\, \zeta \in \partial_P T(x), \quad \forall\, x \in \text{int}(S),$$

and (3.3) says that

$$h(x, \zeta) + 1 \leq 0 \quad \forall\, \zeta \in \partial_P T(x), \quad \forall\, x \in \Sigma^c.$$

These inequalities hold vacuously when the proximal subdifferentials involved are empty, as is the case when $T(x) = +\infty$, but emptiness is not precluded even when $T(x)$ is finite.

Note also that (3.2) and (3.3) together are equivalent to $T(\cdot)$ satisfying

$$h(x, \partial_P T(x)) + 1 = 0 \quad \forall\, x \in \{\text{int}(S)\} \backslash \Sigma, \tag{3.4}$$

$$h(x, \partial_P T(x)) + 1 \geq 0 \quad \forall\, x \in \{\text{int}(S)\} \cap \Sigma, \tag{3.5}$$

and

$$h(x, \partial_P T(x)) + 1 \leq 0 \quad \forall\, x \in \{\text{bdry}(S)\} \backslash \Sigma. \tag{3.6}$$

We require the following definition, where the open unit ball is denoted $B := \{z \in \mathbb{R}^n : \|z\| < 1\}$.

DEFINITION 3.2. *A function $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ is said to be* (Σ, S)-continuous *provided that there exist $\gamma_\varphi > 0$ and a function $\omega_\varphi : [0, \gamma_\varphi] \to [0, +\infty)$ such that $\lim_{s \downarrow 0} \omega_\varphi(s) = 0$ and $\varphi(x) \leq \omega_\varphi(d_\Sigma(x))$ for all $x \in S \cap \{\Sigma + \gamma_\varphi B\}$.*

In the absence of a state constraint (i.e., $S = \mathbb{R}^n$), the minimal time function being $(\Sigma, \mathbb{R}^n)$-continuous is often referred to as *small time controllability*. Accordingly, when the S-constrained minimal time function $T(\cdot)$ is (Σ, S)-continuous, then we say that *S-constrained small time controllability* holds.

Remark 3.3. Results concerning the small time controllability property in the absence of a state constraint may be found in [3] and [5], as well as [25]. "Classical" trajectory tracking based on Gronwall's lemma features in these discusssions. Specifically, one has the following facts:

(a) Small time controllability is equivalent to $T(\cdot)$ being continuous on an open neighborhood of Σ.

(b) When small time controllability holds, then Γ, the set of startpoints controllable to Σ, is open, and small time controllability implies that the minimal time function $T(\cdot)$ is continuous on Γ.

In the state constrained case being considered, one can employ the tracking result of Lemma 2.2(a) in order to obtain the following analogues of (a) and (b):

(a′) S-constrained small time controllability is equivalent to $T(\cdot)$ being continuous on $S \cap \{\Sigma + \gamma B\}$ for some $\gamma > 0$.

(b′) When S-constrained small time controllability holds, the set Γ_S is open relative to S, and small time controllability implies that the minimal time function $T(\cdot)$ is continuous on Γ_S.

The preceding remarks are not required in the following main result.

THEOREM 3.4. *Let* (F1)–(F3), (S1)–(S2) *hold, and assume that S-constrained small time controllability holds. Then there is a unique lower semicontinuous extended real-valued function $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ which is (Σ, S)-continuous, bounded below on $\mathbb{R}^n$, identically 0 on Σ, identically $+\infty$ on S^c, and satisfies*

$$h(x, \partial_P \varphi(x)) + 1 = 0 \quad \forall\, x \in \{\mathrm{int}(S)\} \backslash \Sigma, \tag{3.7}$$

$$h(x, \partial_P \varphi(x)) + 1 \geq 0 \quad \forall\, x \in \{\mathrm{int}(S)\} \cap \Sigma, \tag{3.8}$$

and

$$h(x, \partial_P \varphi(x)) + 1 \leq 0 \quad \forall\, x \in \mathrm{bdry}(S) \backslash \Sigma. \tag{3.9}$$

That function is the S-constrained minimal time function $T(\cdot)$. Furthermore, Γ_S is open relative to S and $T(\cdot)$ is continuous on Γ_S.

In order to understand the role played by S-constrained small time controllability and (Σ, S)-continuity in this result, consider the case where $\Sigma \subset \mathrm{bdry}(S)$, and note that then the indicator function of Σ, namely,

$$\varphi(x) = \begin{cases} +\infty & \text{if } x \notin \Sigma, \\ 0 & \text{if } x \in \Sigma, \end{cases}$$

will *always* satisfy the conditions set out in the theorem; in particular (3.7)–(3.9) hold vacuously. Therefore there is no hope of uniqueness holding, in general, without further conditions. In fact, the proof will require these conditions even if $\Sigma \subset \mathrm{int}(S)$.

Proof of the theorem. That $T(\cdot)$ satisfies the stated conditions has already been explained. Hence the uniqueness assertion remains to be verified. To this end, let $\varphi(\cdot)$ be as in the statement of the theorem. The proof will follow from the verification of two claims.

Claim 1. $T(\alpha) \leq \varphi(\alpha)$ for all $\alpha \in S$.

To see this, note that (3.7) and (3.9) imply that the function $\tilde{g}(t, \alpha) := t + \varphi(\alpha)$ is weakly decreasing on $\mathbb{R}_+ \times \Sigma^c$. We need only consider $\alpha \in S \backslash \Sigma$, since if $\alpha \in \Sigma$ we have $T(\alpha) = \varphi(\alpha) = 0$. Furthermore, we can assume that $\varphi(\alpha) < +\infty$, for otherwise the claim is trivially true.

Now, for $\alpha \in S \backslash \Sigma$ with $\varphi(\alpha) < +\infty$, weak decrease implies that for some trajectory of (1.1) with $x(0) = \alpha$, one has

$$t + \varphi(x(t)) \leq \varphi(\alpha) \tag{3.10}$$

as long as $x(t) \notin \Sigma$. Since $\varphi(\cdot)$ is globally bounded below and $\varphi(\alpha)$ is finite, we conclude that the trajectory eventually enters Σ. Then $T(\alpha) < +\infty$ and

$$t + \varphi(x(t)) \leq \varphi(\alpha) \quad \forall\, t \in [0, T(\alpha)). \tag{3.11}$$

Upon invoking the lower semicontinuity of $\varphi(\cdot)$, we then have

$$\begin{aligned} T(\alpha) &= T(\alpha) + \varphi(x(T(\alpha))) \\ &\leq \liminf_{t \uparrow T(\alpha)} [t + \varphi((x(t)))] \\ &\leq \varphi(\alpha), \end{aligned}$$

proving the claim.

Claim 2. $T(\alpha) \geq \varphi(\alpha)$ for all $\alpha \in S$.

Let us first consider the case where $\alpha \in \text{int}(S)$. Without loss of generality, we can assume that $\alpha \in \Gamma_S$, for otherwise $T(\alpha) = +\infty$ and the claim will be trivial. Let $x(\cdot)$ be an S-constrained time-optimal trajectory of (1.1) to target Σ, with $x(0) = \alpha$. By the tracking result of Lemma 2.2(b), there exists $W = W(T(\alpha)) > 0$ such that for each sufficiently small $r > 0$ there exists an S_r-constrained trajectory $x_r(\cdot)$ of (1.1) on the interval $[0, T(\alpha)]$ with $x_r(0) = \alpha$ and

$$\|x_r(T(\alpha)) - x(T(\alpha))\| \leq rW. \tag{3.12}$$

Then

$$x_r(T(\alpha)) \in \Sigma + rW\overline{B}, \tag{3.13}$$

where $\overline{B}$ denotes the closure of the unit ball B.

In view of (3.7)–(3.8), the function $\varphi(\cdot)$ is strongly increasing on $\text{int}(S)$. Consequently

$$T(\alpha) + \varphi(x_r(T(\alpha))) \geq \varphi(\alpha). \tag{3.14}$$

Since $\varphi(\cdot)$ is (Σ, S)-continuous, (3.13) implies that the term $\varphi(x_r(T(\alpha)))$ can be made arbitrarily small by decreasing r, and we deduce that

$$T(\alpha) \geq \varphi(\alpha). \tag{3.15}$$

It remains to consider the case where $\alpha \in \text{bdry}(S)$. Again, we can assume that $T(\alpha)$ is finite, and we consider an S-constrained time-optimal trajectory $x(\cdot)$ of (1.1) to target Σ, with $x(0) = \alpha$. In view of Remark 2.1(a), there exists a sequence $\{\alpha_i\}$ in the interior of S, such that $\alpha_i \to \alpha$. From the preceding argument, we have that

$$T(\alpha_i) \geq \varphi(\alpha_i) \tag{3.16}$$

for each $i = 1, 2, \ldots$.

By the tracking result of Lemma 2.2(a), there exists $M = M(T(\alpha)) > 0$ such that for each i there is a trajectory $x_i(\cdot)$ of (1.1) for which $x_i(0) = \alpha_i$,

$$\|x(t) - x_i(t)\| \leq M\|\alpha_i - \alpha\| \quad \forall\, t \in [0, T(\alpha)], \tag{3.17}$$

and

$$x_i(t) \in S \quad \forall\, t \in [0, T(\alpha)]. \tag{3.18}$$

Since $x(T(\alpha)) \in \Sigma$, we have $x_i(T(\alpha)) \in \Sigma + M\|\alpha_i - \alpha\|\overline{B}$, and therefore the S-constrained small time controllability assumption, inequality (3.16), and the preceding case yield

$$T(\alpha) + \omega_T(M\|\alpha_i - \alpha\|) \geq T(\alpha_i) \geq \varphi(\alpha_i) \tag{3.19}$$

for each i. Since $\omega_T(M\|\alpha_i - \alpha\|) \to 0$ as $i \to +\infty$ and $\varphi(\cdot)$ is lower semicontinuous, we obtain $T(\alpha) \geq \varphi(\alpha)$, as required. Finally, the "furthermore" part of the statement is simply Remark 3.3(b′). □

4. Concluding comments.

4.1. Sufficient conditions for *S*-constrained small time controllability. It is possible to replace the S-constrained small time controllability hypothesis in Theorem 3.4 by stronger geometric hypotheses. Specifically, we introduce the following condition, in which $\mathrm{proj}_\Sigma(x)$ denotes the set of closest points in Σ to x, and $N_S^P(x)$ denotes the proximal normal cone to S at x; see [10].

(S3) There exist $\delta_1 > 0$, $\delta_2 > 0$ such that the following holds: For each $x \in S \cap \Sigma^c \cap \{\Sigma + \delta_1 B\}$, there exists $u(x) \in U$ for which

$$\langle \eta, f(x, u(x))\rangle \leq 0 \quad \forall\, \eta \in N_S^P(x) \tag{4.1}$$

and

$$\langle x - y, f(x, u(x))\rangle \leq -\delta_2\|x - y\| \tag{4.2}$$

for some $y \in \mathrm{proj}_\Sigma(x)$.

Intuitively, condition (S3) says that at each point x in S which is near but exterior to Σ, there exists a velocity $f(x, u(x))$ which "points into S" (this is (4.1)) while *simultaneously* "pointing towards Σ" (this is (4.2)). The "proximal aiming" technique of [8], [10] and Clarke and Wolenski [15] can now be brought to bear, whereby one can show that for some $\gamma_1 > 0$, $\gamma_2 > 0$, all limiting Euler solutions $x(\cdot)$ of the initial value problem

$$\dot{x}(t) = f(x(t), u(x(t))), \quad x(0) = \alpha \in S \cap \{\Sigma + \gamma_1 B\},$$

are bona fide trajectories of the control system (1.1) which remain in S and enter Σ prior to time $t = \gamma_2 d_\Sigma(\alpha)$. Hence one has

$$T(\alpha) \leq \gamma_2 d_\Sigma(\alpha) \quad \forall\, \alpha \in S \cap \{\Sigma + \gamma_1 B\}.$$

The foregoing comments show that we can take $\omega_T(z) = \gamma_2 z$, in the notation of Definition 3.2. One can then invoke Lemma 2.2(b) in order to show that $T(\cdot)$ is Lipschitz on $S \cap \{\Sigma + \gamma_3 B\}$ for some $\gamma_3 > 0$, and similarly to Remark 3.3(b′), one can show that $T(\cdot)$ is Lipschitz on Γ_S, where, as noted earlier, this set is open relative to S. We omit these details, as they are very similar to familiar arguments used in the case of no state constraints, as in [3] and [25].

A variant of Theorem 3.4 ensues.

COROLLARY 4.1. *Let* (F1)–(F3) *and* (S1)–(S3) *hold. Then there is a unique lower semicontinuous extended real-valued function $\varphi(\cdot)$ which is (Σ, S)-continuous, bounded below on $\mathbb{R}^n$, identically 0 on Σ, identically $+\infty$ on S^c, and satisfies*

$$h(x, \partial_P\varphi(x)) + 1 = 0 \quad \forall\, x \in \{\mathrm{int}(S)\}\backslash\Sigma, \tag{4.3}$$

$$h(x, \partial_P\varphi(x)) + 1 \geq 0 \quad \forall\, x \in \{\mathrm{int}(S)\} \cap \Sigma, \tag{4.4}$$

and

$$h(x, \partial_P \varphi(x)) + 1 \leq 0 \quad \forall\, x \in \mathrm{bdry}(S) \backslash \Sigma. \tag{4.5}$$

That function is the S-constrained minimal time function $T(\cdot)$. Furthermore, Γ_S is open relative to S and $T(\cdot)$ is Lipschitz on Γ_S.

4.2. Constrained viscosity formulation. Given a lower semicontinuous extended real-valued function $w : \mathbb{R}^n \to (-\infty, +\infty]$, we denote the *Dini subderivate* at x in the direction v by

$$Dw(x; v) := \liminf_{\substack{t \downarrow 0 \\ u \to v}} \frac{w(x + tu) - w(x)}{t}.$$

The D-subdifferential (or *viscosity subdifferential*) of $w(\cdot)$ at x is the set

$$\partial_D w(x) := \{\zeta \in \mathbb{R}^n : \langle \zeta, v \rangle \leq Dw(x; v) \ \forall\, v \in \mathbb{R}^n\}.$$

Similarly to remarks made in [14], Theorem 3.4 and Corollary 4.1 hold true if ∂_P is replaced by ∂_D in the statement, because the P- and D-subdifferentials approximate one another in a suitable sense. This in turn rests upon a theorem of Subbotin; see Proposition 3.4.5 in [10]. By Proposition 3.4.12 in [10], the ∂_D-form of the theorem can in turn be put into "constrained viscosity" terms, in Soner's terminology [22]; see also [3]. Specifically, (3.7) can be replaced by the condition

$$h(x, g'(x)) + 1 = 0 \quad \forall\, x \in \{\mathrm{int}(S)\} \backslash \Sigma \tag{4.6}$$

for any $g \in C^1(\mathbb{R}^n)$ such that $\varphi - g$ has a local minimum at x. Furthermore, conditions (3.8) and (3.9) have analogous constrained viscosity reformulations.

4.3. *S*-restricted dynamics. Suppose that the dynamics are specified only for $x \in S$; that is, $f : S \times U \to \mathbb{R}^n$. Such a restricted domain can be expected in control problems where the state is S-constrained. Again, similarly to [14], it can be shown that Theorem 3.4 remains true if (F1)–(F3) are replaced by the following *two* conditions, where (F2) is not present explicitly:

(G1) The function f is continuous on $S \times U$ and is Lipschitz in the state variable x, uniformly for $u \in U$; that is, there exists K such that

$$\|f(x, u) - f(y, u)\| \leq K \|x - y\|$$

whenever (x, u) and (y, u) are in $S \times U$.

(G2) The velocity set $f(x, U)$ is convex for every $x \in S$.

We can extend f from $S \times U$ to $\mathbb{R}^n \times U$ as follows: Let f_i denote the ith component function of f, $i = 1, 2, \ldots, n$. For each fixed $u \in U$, define a function $x \to \hat{f}_i(x, u)$ on $\mathbb{R}^n$ via

$$\hat{f}_i(x, u) = \min_{y \in S} \{f_i(y, u) + K \|y - x\|\}.$$

Then $x \to \hat{f}_i(x, u)$ agrees with $f_i(x, u)$ on S, and is globally Lipschitz of rank K, facts first noted by Hiriart-Urruty [19]; see also [10] and Rockafellar and Wets [21]. We then extend f componentwise by setting $f_i(x, u) = \hat{f}_i(x, u)$ for every $(x, u) \in \mathbb{R}^n \times U$.

The resulting function $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ is continuous on $\mathbb{R}^n \times U$ and is globally Lipschitz of rank nK in the state variable x, uniformly for $u \in U$; that is,

$$\|f(x,u) - f(y,u)\| \leq nK\|x - y\|,$$

whenever (x,u) and (y,u) are in $\mathbb{R}^n \times U$. The *global* Lipschitz condition on the extended dynamics implies the linear growth condition needed for the global extendability of solutions, and possible nonconvexity of the velocity sets $f(x,U)$ for $x \notin S$ does not cause a problem in the required state constrained tracking considerations as their proofs in [12] show.

4.4. The case of unbounded S. The main results in this article (as well as [12] and [13]) have been stated for the case of compact S, but can be generalized to the case where S is merely assumed to be closed. Then the growth condition (F2) is indispensible. The generalizations follow from appropriately localized versions of the state constrained tracking properties employed in the compact case. We omit these details here.

REFERENCES

[1] M. Bardi, *A boundary value problem for the minimum-time function*, SIAM J. Control Optim., 27 (1989), pp. 776–785.

[2] M. Bardi and V. Staicu, *The Bellman equation for time-optimal control of noncontrollable nonlinear systems*, Acta. Appl. Math., 81 (1993), pp. 201–223.

[3] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[4] M. Bardi, S. Koike, and P. Soravia, *Pursuit-evasions games with state constraints: Dynamic programming and discrete-time approximations*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 361–380.

[5] P. Cannarsa and C. Sinestrari, *Semiconcave Functions, Hamilton-Jacobi Equations, and Optimal Control*, to appear.

[6] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Pursuit differential games with state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1615–1632.

[7] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983. Republished as Classics Appl. Math. 5, SIAM, Philadelphia, 1990.

[8] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Qualitative properties of trajectories of control systems: A survey*, J. Dynam. Control Systems, 1 (1995), pp. 1–48.

[9] F. H. Clarke, Yu. S. Ledyaev, and R. J. Stern. *Complements, approximations, smoothings and invariance properties*, J. Convex Anal., 4 (1997), pp. 189–219.

[10] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer-Verlag, New York, 1998.

[11] F. H. Clarke and C. Nour, *The Hamilton-Jacobi Equation of Minimal Time Control*, Preprint.

[12] F. H. Clarke, L. Rifford, and R. J. Stern, *Feedback in state constrained optimal control*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 97–133.

[13] F. H. Clarke and R. J. Stern, *State constrained feedback stabilization*, SIAM J. Control Optim., 42 (2003), pp. 422–441.

[14] F. H. Clarke and R. J. Stern, *Hamilton-Jacobi characterization of the state constrained value*, Nonlinear Anal., to appear.

[15] F. H. Clarke and P. R. Wolenski, *Control of systems to sets and their interiors*, J. Optim. Theory Appl., 88 (1996), pp. 3–23.

[16] F. Forcellini and F. Rampazzo, *On nonconvex differential inclusions whose state is constrained in the closure of an open set. Applications to dynamic programming*, Differential Integral Equations, 12 (1999), pp. 471–497.

[17] H. Frankowska and F. Rampazzo, *Filippov's and Filippov-Wazewski's theorems on closed domains*, J. Differential Equations, 161 (2000), pp. 449–478.

[18] H. Frankowska and R. B. Vinter, *Existence of neighboring feasible trajectories: Applications to dynamic programming for state-constrained optimal control problems*, J. Optim. Theory Appl., 104 (2000), pp. 21–40.

[19] J.-B. Hiriart-Urruty, *New concepts in nondifferentiable programming*, Bull. Soc. Math. France Mém., 60 (1979), pp. 57–85.
[20] R. T. Rockafellar, *Clarke's tangent cones and boundaries of closed sets in* $\mathbb{R}^n$, Nonlinear Anal., 3 (1979), pp. 145–154.
[21] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1998.
[22] H. M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.
[23] P. Soravia, *Discontinuous viscosity solutions to Dirichlet problems for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 18 (1993), pp. 1493–1514.
[24] R. B. Vinter, *Optimal Control*, Systems Control Found. Appl., Birkhäuser, Boston, 2000.
[25] P. R. Wolenski and Y. Zhuang, *Proximal analysis and the minimal time function*, SIAM J. Control Optim., 36 (1998), pp. 1048–1072.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 708–730

SINGULAR STOCHASTIC CONTROL PROBLEMS*

F. DUFOUR† AND B. MILLER‡

Abstract. In this paper, we study an optimal singular stochastic control problem. By using a time transformation, this problem is shown to be equivalent to an auxiliary control problem defined as a combination of an optimal stopping problem and a classical control problem. For this auxiliary control problem, the controller must choose a stopping time (optimal stopping), and the new control variables belong to a compact set. This equivalence is obtained by showing that the (discontinuous) state process governed by a singular control is given by a time transformation of an auxiliary state process governed by a classical bounded control. It is proved that the value functions for these two problems are equal. For a general form of the cost, the existence of an optimal singular control is established under certain technical hypotheses. Moreover, the problem of approximating singular optimal control by absolutely continuous controls is discussed in the same class of admissible controls.

Key words. nonlinear stochastic systems, optimal control, singular control, time change

AMS subject classifications. 49J30, 49N25, 93E20

DOI. 10.1137/S0363012902412719

1. Introduction.

In this paper, the existence of optimal singular controls is studied for the nonlinear stochastic system defined by the following equation:

$$x_t \doteq \zeta + \int_0^t A(s, x_s)ds + \int_0^t B(s)du_s + \int_0^t D(s, x_s)dW_s, \tag{1}$$

where the functions A, B, D are deterministic, $\{W_t\}$ is a Brownian motion, and $\{u_t\}$ is the control. All the processes are assumed to be defined on a probability space $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$. Let $K \subset \mathbb{R}^p$ be a closed convex cone and T be the finite horizon. The class of admissible controls, labeled $\mathfrak{C}^a$, is defined by the class of K-valued, continuous on the right with left-hand limits, $\{\mathcal{F}_t\}$-progressively measurable processes for which almost every sample path is of finite variation on the interval $[0, T]$:

$$v_T^u < \infty, \qquad Q\text{-a.s.}, \tag{2}$$

where $v_t^u = |u_0| + \lim_{n\to\infty} \sum_{k=1}^n |u_{tk/n} - u_{t(k-1)/n}|$ ($|z|$ is the norm of the vector z). For an admissible control u, the cost is given by

$$J[u] = E\left[\int_0^T k(t, x_t)du_t^c + \sum_{0\le t\le T} \int_0^{\Delta v_t^u} k\left(t, x_{t-} + B(t)\frac{\Delta u_t}{\Delta v_t^u}s\right)\frac{\Delta u_t}{\Delta v_t^u}ds + g(x_T, v_T^u)\right], \tag{3}$$

*Received by the editors August 7, 2002; accepted for publication (in revised form) December 5, 2003; published electronically August 4, 2004. This research was supported by a CNRS/Russian Academy of Sciences cooperation (number PECO/NEI 9570) and in part by the Nonlinear Control Network and by Russian Basic Investigation Foundation grant 02-01-00361.
http://www.siam.org/journals/sicon/43-2/41271.html

†Corresponding author. MAB, Université Bordeaux I, 351 cours de la Liberation, 33405 Talence Cedex, France (dufour@math.u-bordeaux1.fr) and GRAPE, Université Bordeaux IV, France.

‡Institute for Information Transmission Problems, 19 Bolshoy Karetny per., Moscow 127994, Russia (bmiller@iitp.ru).

where g, k are deterministic functions and u^c is the continuous part of u (Δz_t denotes $z_t - z_{t-}$ for a process $\{z_t\}$ which is continuous on the right with left-hand limits). The interested reader may consult [42] for a nice interpretation of the cost defined by (3).

Singular stochastic control problems have received considerable attention in the literature. The authors do not pretend to present here an exhaustive panorama of singular control problems. However, the interested reader may consult the work of Boetius [9], especially the sections at the end of the chapters for an interesting and complete survey on stochastic singular control problems including theoretical results and applications.

This problem was first introduced by Bather and Chernoff [6] in 1967 by considering a simplified model for the control of a spaceship. It was noted for this special model that there was a connection between the singular control problem and optimal stopping problem. This link was established through the derivative of the value function of this initial singular control problem and the value function of the corresponding optimal stopping problem.

After this seminal work, this connection and its properties were extensively studied in different contexts but mainly in the one-dimensional case or in the multidimensional linear case. Two approaches were used: one is based on the theory of partial differential equations and on variational arguments, and can be found in the works of Alvarez [1, 2], Chow, Menaldi, and Robin [13], Karatzas [27], Karatzas and Shreve [31], and Menaldi and Taksar [36]. The other approach is related to probabilistic methods; see, for example, Baldursson [4], Boetius [8, 9], Boetius and Kohlmann [10], El Karoui and Karatzas [17, 18], Karatzas [28], and Karatzas and Shreve [29, 30].

Other problems, such as the dynamic programming principle, have been studied in a general context, for example, by Boetius [9], Haussmann and Suo [24], Fleming and Soner [21], and Zhu [43], as well as the stochastic maximum principle in [11].

Singular control problems correspond to many applications in diverse areas such as mathematical finance (see, for example, Baldursson and Karatzas [5], Chiarolla and Haussmann [12], Kobila [34], and Karatzas and Wang [33]), manufacturing systems (see, for example, Shreve, Lehoczky, and Gaver [41]), and queuing systems (see, for example, Martins and Kushner [35]).

In this paper, we focus our attention on the existence problem and the connection between the singular control and optimal stopping problems.

As we have already mentioned, the connection between singular stochastic control problems and optimal stopping problems has only been studied in the one-dimensional case or in the multidimensional linear case. A generalization of this connection to the multidimensional nonlinear case has been proposed by Benth and Reikvam [7] but under a very strong hypothesis, namely, the model of the state process must have a special structure in order to ensure that the ith component of $x(t)$ depends only on the ith component of the initial state process ζ. In the work of Boetius and Kohlmann [10], the result of Karatzas and Shreve [29] was generalized to a nonlinear one-dimensional state process. A multidimensional problem (see section 5.2 in [10]) was also considered but again under a strong hypothesis, since the control process could influence the state through only one variable.

In the present paper, this link is revisited by using a completely different approach. It will be shown that a multidimensional and nonlinear singular control problem can be converted into an auxiliary control problem where the control variables are of the classical type and where the controller must choose a stopping time (optimal stopping as described by Haussmann and Lepeltier, p. 851 in [22]). Consequently this auxiliary problem combines classical control and optimal stopping. Moreover, it will

be shown that these two optimization problems have the same value function. It must be pointed out that our result differs from existing results in the literature on two points. First, our auxiliary equivalent control problem is defined as a combination of optimal stopping and classical control, but in the literature, the equivalent control problem is formulated by a *pure* optimal stopping problem (the controller cannot influence the trajectory of the state process). Second, our connection is obtained directly through the value function, but in the literature, this link was established through the gradient of the value function of the singular control problem and the value of the related optimal stopping problem.

In this auxiliary control problem, the state processes are defined on a probability space $(\Omega, \mathcal{F}, P, \{\mathcal{G}_t\})$ by the following equations:

$$\xi_t \doteq \zeta + \int_0^t A(\eta_s, \xi_s)(1-\theta_s)ds + \int_0^t \theta_s B(\eta_s)\alpha_s ds + \int_0^t D(\eta_s, \xi_s)\sqrt{(1-\theta_s)}dV_s, \tag{4}$$

$$\eta_t \doteq \int_0^t (1-\theta_s)ds, \tag{5}$$

where $\{V_t\}$ is a Brownian motion. The new control variables are $\{(\alpha_t, \theta_t)\}$ and ρ, which is a $\{\mathcal{G}_t\}$-stopping time to be chosen by the controller. A key feature of this formulation is that the processes $\{(\alpha_t, \theta_t)\}$ belong to a compact set (labeled $\mathfrak{B}$; for the definition of this set, see the notation section at the end of the introduction). The cost is defined by

$$J[\alpha, \theta, \rho] = E\left[g(\xi_\rho, \rho - \eta_\rho) + \int_0^\rho \theta_s k(\eta_s, \xi_s)\alpha_s ds + G(\eta_\rho)\right] \tag{6}$$

(the function G is equal to $+\infty$ everywhere excepted at T, where it takes the value zero).

The idea used to show this equivalence result is based on a time transformation. This method, originally developed in deterministic control theory (for a complete exposition on the subject see the recent book [37] and the references therein), has been introduced recently in the stochastic context in [15, 38] and in [3]. In the latter reference, Alavarez, Gyllenberg, and Shepp used a time change technique to show that a one-dimensional singular control problem subject to a state-dependent killing rate is equivalent to an associated one-dimensional singular control problem with constant discount rate. Note that in this work, the time change does not depend explicitly on the control but is related to the integral of the discount factor, which is dependent on the state process.

Our result provides a set of weak hypotheses (Assumptions A1–A5, defined below) to ensure the existence of an optimal singular control for a general model, but it has the drawback that it provides no information about the nature and the properties of the optimal control. The existence of singular stochastic controls has been investigated by Haussmann and Suo for a general nonlinear model [23]. Using a compactification method, the authors show an existence result under certain technical conditions. Our result can be viewed as extending the work of Haussmann and Suo [23] in several directions.

We assume that the functions A and C must satisfy a Lipschitz condition but are not necessarily bounded as in [23], where these functions are required to be bounded continuous (for example, linear control problems cannot be considered).

Moreover, an important difference is that the form of the cost presented in our paper is more general. The part of the cost depending on the singular control in [23] has the following form:

$$E\left[\int_{[0,T)} c(s)du_s\right], \tag{7}$$

where $c(.)$ is lower semicontinuous and each of its components is strictly positive.

In our work, we propose a more general form given by

$$E\left[\int_0^T k(t,x_t)du_t^c + \sum_{0\le t\le T}\int_0^{\Delta v_t^u} k\left(t, x_{t-} + B(t)\frac{\Delta u_t}{\Delta v_t^u}s\right)\frac{\Delta u_t}{\Delta v_t^u}ds\right]. \tag{8}$$

It must be pointed out that the cost defined by (8) depends explicitly on the state process $\{x_t\}$, which is not the case in (7). Singular control problems defined with such general cost functions (see (3)) have been studied by many authors (for example, Zhu [43] studied a finite horizon problem and Taksar [42] and Davis and Zervos [14] analyzed infinite horizon problems). To the best knowledge of the authors, the work presented in this paper is the first attempt to derive an existence result for a multidimensional nonlinear model with such a general form for the cost. Zhu [43] derived a dynamic programming principle for a nonlinear model where D (with our notation) is assumed to be time independent (see the remark in [43, p. 229]) and nondegenerate. Taksar [42] considered a singular control problem where the state process satisfies an equation where A and D are time independent, A is bounded, and D is nondegenerate. The author showed the equivalence between the original singular problem and a linear programming problem. In [14], Davis and Zervos proved a verification theorem for a nonlinear time independent model. Two special one-dimensional cases were explicitly solved by the authors. The hypotheses used here are weaker than the assumptions previously cited.

Another important difference is that in [23], the variation of an admissible control needs to be integrable: $E[v_t^u] < \infty$ (see the proof of Proposition 3.4, p. 34, in Suo's thesis). Here a weaker hypothesis is introduced in the sense that the variation of an admissible control needs to satisfy the following assumption: $E[g(x_T, v_T^u)] < +\infty$, where g must satisfy $\lim_{t\to+\infty}\inf_{x\in\mathbb{R}^n} g(x,t) = +\infty$.

An interesting point is that a large number of results presented in the literature concern singular control problems where the control process is left continuous, rejecting the possibility of a jump at the terminal time (see, for example, Karatzas and Shreve [29]). In contrast, we consider singular controls which are continuous on the right and have limits on the left, allowing a jump at the terminal time. Our approach is more direct than in [23], where the left continuous control process is considered for which a modification is proposed, allowing jumps at the terminal time (see Remark 2.2 and section 4 in [23]). The advantage of our approach is that one can find an optimal solution to those singular control problems which do not admit optimal solutions when the control is assumed to be left continuous. In order to illustrate this point, the well-known example of nonexistence of Karatzas and Shreve in [29] is revisited in section 6.

In section 6, we also discuss some extensions and generalizations of the model initially presented in section 2. In particular, it is shown how our results can be modified in order to study other singular control problems, such as monotone follower problems (where the process $\{u_t\}$ is assumed to be a nondecreasing function).

The problem of approximating singular optimal control by absolutely continuous controls is studied in the final section. The main difficulty is that the control may have a jump at the terminal time and, consequently, the associated state process. Therefore, it is difficult to find a sequence of absolutely continuous control process $\{v_n(t)\}$ defined on $[0,T]$ and a sequence of filtration $\{\mathcal{F}_t^n\}$ defined on the probability space $(\Omega,\mathcal{F},P)$ satisfying both of the following conditions: $\{v_n(t)\}$ is $\{\mathcal{F}_t^n\}$-progressively measurable and $\lim_{n\to+\infty} x_n(T) = x(T)$, where $\{x_n(t)\}$ is the state process controlled by $\{v_n(t)\}$. Our equivalence result provides a way of overcoming this difficulty. However, it must be pointed out that although this problem is simpler, it remains difficult to solve mainly because one needs to approximate the combination of an optimal stopping problem (where the stopping is not necessarily bounded) and a classical control problem under the strong admissibility condition of the control and the stopping time given by $E[G(\eta_\rho)] = 0$.

In [15], the authors considered a general stochastic control problem where the controls have to satisfy an integral constraint. It was shown that there exists an optimal control within the class of generalized controls leading to impulse actions. The problem studied in [15] is related to singular control problems in the sense that the optimal generalized state process is given in terms of stochastic differential equations governed by a measure. However, the main difference is that the jump of the optimal state process cannot be explicitly expressed in terms of the jump of the optimal control process contrary to the case of the singular control problem ($\Delta x_t = B(t)\Delta u_t$). Although the general idea of time change is used here and in [15], the results presented in [15] cannot be applied to solve the problem studied in this paper. Indeed, in this work, it is required that the control process be only of finite variation contrary to [15], where the control is assumed to satisfy the integral constraint $P(\{\int_0^T |u_s|ds < M\}) = 1$ for a constant M (this point is a crucial hypothesis in [15]). Finally, it must be pointed out that in the present paper, the stopping time is not necessarily bounded, which represents a difficulty which makes the technique proposed in [15] inapplicable.

The paper is organized as follows. In section 2, we formulate the singular control problem. The description of the time transformation is presented in section 3. In section 4, an auxiliary control problem is introduced that will be shown to be equivalent to the original one. On the basis of known results, the existence theorem is proved for the auxiliary problem and consequently for the original problem in section 5. In section 6, we show how our existence result can be modified and applied to other problems, and we revisit a well-known example found in the literature. In the last part of the paper, it is shown that the optimal singular control can be approximated by continuous controls in a sense that will be defined below (see Theorem 7.5).

Now, we present some notation and terminology.

Notation. $\mathbb{N}_N$ is the set of the first N integers; that is, $\mathbb{N}_N = \{1,\dots,i,\dots,N\}$.

$\mathbb{N}^* \doteq \{k \in \mathbb{N} : k > 0\}$ and $\mathbb{R}_+ \doteq \{x \in \mathbb{R} : x \geq 0\}$.

For a vector x in $\mathbb{R}^p$, the ith component of x is denoted by x^i, $|x| \doteq \sum_{i=1}^p |x^i|$ is the norm of x, and 0_p is the zero vector in $\mathbb{R}^p$.

If A is an $m \times n$ matrix, the norm of A is defined by $|A| = \max_{|x|\leq 1} |Ax|$ and $(')$ denotes the transpose operation.

The indicator function of a set A is defined as $I_A(x)$.

The function δ defined on $\mathbb{N}\times\mathbb{N}$ is such that $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.

For $x \in \mathbb{R}$, x^+ is defined by $x^+ = \frac{1}{x}$ if $x \neq 0$ and by $x^+ = 0$ if $x = 0$.

If X is a metric space, then $\mathcal{B}(X)$ denotes its associated Borel σ-field.

A process is said to be *corlol* if it is continuous on the right and have limits on the left.

On the probability space $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$, the mathematical expectation is denoted by $E_P[.]$, and for an $\mathbb{R}^p$-valued corlol process $\{u_t\}$, the total variation of $\{u_t\}$ on the interval $[0,t]$ is defined by

$$v_t^u \doteq |u_0| + \lim_{n\to\infty} \sum_{k=1}^{n} |u_{tk/n} - u_{t(k-1)/n}|, \tag{9}$$

and $\{u_t\}$ is said to be of finite variation on $[0,t]$ if $v_t^u < +\infty$, P-a.s.

Let $\{w_t\}$ be a real-valued corlol process of finite variation on $[0,T]$. $\{w_t\}$ is the distribution function of a signed measure defined on $[0,T]$; this measure is denoted by dw.

In order to define the state processes, let us introduce the following data:

- T is a fixed real number.
- K is a subset of $\mathbb{R}^p$.
- $\mathfrak{B} \doteq \{(x,y) \in K \times [0,1] : |x| \leq 1\}$.
- $A : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$.
- $B : [0,T] \to \mathbb{R}^{n\times p}$.
- $D : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$.
- $g : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}_+$.
- $k : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}_+^p$.
- ζ is a fixed vector in $\mathbb{R}^n$.
- $G : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ such that $G(T) = 0$ and $G(t) = \infty$ for $t \neq T$.

The following assumptions will be used in the paper.

Assumption A1. The functions $A(.,.)$, $B(.)$, and $D(.,.)$ are continuous, and $\forall\, t \in \mathbb{R}_+$, there is a constant L_1 such that, $\forall\, (x,y) \in \mathbb{R}^n \times \mathbb{R}^n$,

$$|A(t,x) - A(t,y)| + |D(t,x) - D(t,y)| \leq L_1|x-y|.$$

Assumption A2. The function g is lower semicontinuous and satisfies $\lim_{t\to+\infty} \inf_{x\in\mathbb{R}^n} g(x,t) = +\infty$ and $(\forall x \in \mathbb{R}^n)$, $(\forall (y_1,y_2) \in \mathbb{R}_+ \times \mathbb{R}_+)$, if $y_1 \leq y_2$, then $g(x,y_1) \leq g(x,y_2)$.

Assumption A3. Each component of k is lower semicontinuous. There exists a continuously differentiable function $L : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ such that

$$\nabla_x L(t,x) B(t) = k(t,x). \tag{10}$$

Assumption A4. K is a closed cone which is convex.

Assumption A5. For all $(t,x) \in [0,T] \times \mathbb{R}^n$, the set $K(t,x)$, defined by

$$K(t,x) \doteq \{(A(t,x)(1-\theta) + \theta B(t)\alpha, (1-\theta)D(t,x)D(t,x)', 1-\theta) : (\alpha,\theta) \in \mathfrak{B}\},$$

is convex.

Unless clearly mentioned, we shall always follow the convention that $X_{0-} = 0$, P-a.s., for any *corlol* process $\{X_t\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$.

2. Preliminaries and statement of the problem. In this section, we formulate the stochastic control problem presented in the introduction using the formulation described in [20] and in [22].

DEFINITION 2.1. *A singular control is defined by the following term:*

$$C \doteq (\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{u_t\}, \{W_t\}, \{x_t\}),$$

where

(i) $(\Omega, \mathcal{F}, P)$ *is a complete probability space with a right continuous complete filtration* $\{\mathcal{F}_t\}$*;*

(ii) $\{u_t\}$ *is an* $\mathbb{R}^p$*-valued, corlol* $\{\mathcal{F}_t\}$*-progressively measurable process such that*

$$(\forall A \in \mathcal{B}([0,T]) \otimes \mathcal{F}), \quad \int_0^T I_A du_t \in K, \tag{11}$$

$$v_T^u < +\infty; \tag{12}$$

(iii) $\{W_t\}$ *is a standard* m*-dimensional* $\{\mathcal{F}_t\}$*-Brownian motion;*

(iv) $\{x_t\}$ *is an* $\mathbb{R}^n$*-valued, corlol* $\{\mathcal{F}_t\}$*-progressively measurable process such that* $(\forall t \in [0,T])$

$$x_t \doteq \zeta + \int_0^t A(s, x_s) ds + \int_{[0,t]} B(s) du_s + \int_0^t D(s, x_s) dW_s \tag{13}$$

and $x_{0-} = \zeta$*.*

We write $\mathfrak{C}$ *for the set of controls satisfying the previous conditions.*

The cost is given by

$$J[C] \doteq E_P \left[\int_0^T k(t, x_t) du_t^c + \sum_{0 \le t \le T} \int_0^{\Delta v_t^u} k\left(t, x_{t-} + B(t) \frac{\Delta u_t}{\Delta v_t^u} s\right) \frac{\Delta u_t}{\Delta v_t^u} ds \right.$$

$$\left. + g(x_T, v_T^u) \right]. \tag{14}$$

The set $\mathfrak{C}^a$ *of admissible controls is defined by*

$$\mathfrak{C}^a \doteq \{C \in \mathfrak{C} : J[C] < \infty\}. \tag{15}$$

The singular control problem is defined by the minimization of $J[C]$ on $\mathfrak{C}^a$.

3. Time transformation. In this section, it is assumed that on a probability space $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$ satisfying the usual hypotheses (completion and right continuity), there exists a process $\{u_t\}$ satisfying item (ii) of Definition 2.1.

Let us define the process $\{\Gamma_t\}$ by

$$\Gamma_t \doteq t + v_t^u. \tag{16}$$

$\{\Gamma_t\}$ is a corlol strictly increasing, $\{\mathcal{F}_t\}$-progressively measurable process. Denote by $\{\eta_t\}$ the right inverse of $\{\Gamma_t\}$:

$$\eta_t \doteq \inf\{s \ge 0 : \Gamma_s > t\}. \tag{17}$$

Therefore, applying Proposition 1.1 of Chapter V in [40], $\{\eta_t\}$ is a time change satisfying

$$\eta_{\Gamma_t} = t. \tag{18}$$

PROPOSITION 3.1. *There exists a process* $\{(\overline{\alpha}_t, \overline{\theta}_t)\}$, $\{\mathcal{F}_t\}$*-progressively measurable process taking value in* $\mathfrak{B}$ *such that* $(\forall j \in \mathbb{N}_p)$

$$v_t^u = \int_0^t \overline{\theta}_s d\Gamma_s \tag{19}$$

$$= \int_0^{\Gamma_t} \overline{\theta}_{\eta_s} ds, \tag{20}$$

$$u_t = \int_0^t \overline{\alpha}_s dv_s^u \tag{21}$$

$$= \int_0^{\Gamma_t} \overline{\theta}_{\eta_s} \overline{\alpha}_{\eta_s} ds. \tag{22}$$

Proof. By definition, the measure dv^u is absolutely continuous with respect to the measure $d\Gamma$ for almost all $\omega \in \Omega$. Consequently, using Proposition 3.13 of Chapter I in [26], it follows that there exists an $\{\mathcal{F}_t\}$-optional process $\{\widetilde{\theta}_t\}$ such that

$$v_t^u = \int_0^t \widetilde{\theta}_s d\Gamma_s. \tag{23}$$

Using the fact that $(\forall A \in \mathcal{B}([0,T]))\ dv^u(A) \leq d\Gamma(A)$, we obtain from (23) that

$$v_t^u = \int_0^t \overline{\theta}_s d\Gamma_s, \tag{24}$$

where $\overline{\theta}_s \doteq (0 \vee \widetilde{\theta}_s) \wedge 1$. Moreover, using Proposition 4.9 in [40, p. 8] and (24), we obtain (20).

Using the same arguments, it can be shown that there exists a process $\{\overline{\alpha}_t\}$ such that $|\overline{\alpha}_t| \leq 1$ and which satisfies (21). By combining (19), (21), and [40, Proposition 4.9, p. 8], we have (22), giving the result.

Now using (11), it can be shown easily that $\{\overline{\alpha}_t\}$ is a K-valued process giving the result. □

PROPOSITION 3.2. *The process* $\{\overline{\theta}_t\}$ *satisfies the following equality:*

$$\eta_t = \int_0^{\Gamma_t} (1 - \overline{\theta}_{\eta_s}) ds. \tag{25}$$

Proof. Combining (16) and (20), we obtain that

$$\int_0^{\Gamma_t} (1 - \overline{\theta}_{\eta_s}) ds = \Gamma_t - v_t^u = t.$$

Consequently, using (18), we have

$$\int_0^{\Gamma_t} (1 - \overline{\theta}_{\eta_s}) ds = \eta_{\Gamma_t}.$$

Since $\{\int_0^t (1 - \overline{\theta}_{\eta_s}) ds\}$ and $\{\eta_t\}$ are increasing continuous processes and $\{\Gamma_t\}$ is a strictly increasing corlol process, the result follows. □

4. Auxiliary control and equivalence result. We introduce an auxiliary control problem given in terms of a classical control problem and an optimal stopping problem. It is shown that this problem is equivalent to the initial one. The new control variables are defined by $\{(\alpha_t, \theta_t)\}$ and by ρ, which is a $\{\mathcal{G}_t\}$-stopping time to be chosen by the controller (see Definition 4.1 below). A key property of the auxiliary control problem is that the new control variables $\{(\alpha_t, \theta_t)\}$ take their values in a compact set.

DEFINITION 4.1. *An auxiliary control is defined by the following term:*

$$\Psi \doteq (\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\}, \{(\alpha_t, \theta_t)\}, \{V_t\}, \{\Lambda_t\}, \rho),$$

where

(i) $(\Omega, \mathcal{G}, Q)$ *is a complete probability space with a right continuous complete filtration* $\{\mathcal{G}_t\}$*;*

(ii) $\{(\alpha_t, \theta_t)\}$ *is a* $\mathfrak{B}$*-valued,* $\{\mathcal{G}_t\}$*-progressively measurable process;*

(iii) $\{V_t\}$ *is a standard* m*-dimensional* $\{\mathcal{G}_t\}$*-Brownian motion;*

(iv) ρ *is a* $\{\mathcal{G}_t\}$*-stopping time such that*

$$\rho < +\infty, \quad Q\text{-a.s.}; \tag{26}$$

(v) $\{\Lambda_t \doteq (\xi_t', \eta_t)'\}$ *is an* $\mathbb{R}^{n+1}$*-valued,* $\{\mathcal{G}_t\}$*-progressively measurable process such that*

$$\begin{aligned} \xi_t &\doteq \zeta + \int_0^t A(\eta_s, \xi_s)(1-\theta_s)ds + \int_0^t \theta_s B(\eta_s)\alpha_s ds \\ &\quad + \int_0^t D(\eta_s, \xi_s)\sqrt{(1-\theta_s)}dV_s, \end{aligned} \tag{27}$$

$$\eta_t \doteq \int_0^t (1-\theta_s)ds \tag{28}$$

for $t \in [0, \rho]$.

We write Υ *for the set of controls satisfying the previous conditions. The cost is given by*

$$\mathcal{M}[\Psi] \doteq E_Q\left[g(\xi_\rho, \rho - \eta_\rho) + \int_0^\rho \theta_s k(\eta_s, \xi_s)\alpha_s ds + G(\eta_\rho)\right]. \tag{29}$$

The set Υ^a *of admissible auxiliary controls is defined by*

$$\Upsilon^a \doteq \{\Psi \in \Upsilon : \mathcal{M}[\Psi] < \infty\}. \tag{30}$$

The auxiliary control problem is defined by the minimization of $\mathcal{M}[\Psi]$ on Υ^a.

In this section, the equivalence between the auxiliary and the initial control problems is shown.

THEOREM 4.2. *Assume Assumption* A1. *Let* C *be an element of* $\mathfrak{C}^a$*. Then there exists an auxiliary control* Ψ *in* Υ^a *such that*

$$\mathcal{M}[\Psi] = J[C]. \tag{31}$$

Proof. Let us define C as $\big(\overline{\Omega}, \overline{\mathcal{F}}, \overline{P}, \{\overline{\mathcal{F}}_t\}, \{u_t\}, \{W_t\}, \{x_t\}\big)$. Define the process $\{\Gamma_t\}$ by (16) and its right inverse $\{\eta_t\}$ by (17). Clearly, the probability space

$(\overline{\Omega}, \overline{\mathcal{F}}, \overline{P}, \{\overline{\mathcal{F}}_{\eta_t}\})$ satisfies the usual hypotheses. Since $\{u_t\}$ satisfies item (i) of Definition 2.1, we can apply the results of the previous section to obtain the existence of an $\{\mathcal{F}_t\}$-progressively measurable process $\{(\overline{\alpha}_t, \overline{\theta}_t)\}$ satisfying (19) and (21) (see Proposition 3.1). Therefore, we have the process $\{(\alpha_t, \theta_t)\}$, where $\alpha_t \doteq \overline{\alpha}_{\eta_t}$ and $\theta_t \doteq \overline{\theta}_{\eta_t}$ is $\{\overline{\mathcal{F}}_{\eta_t}\}$-progressively measurable.

Using Proposition 1.1 in [40, Chapter V], Γ_T is an $\{\overline{\mathcal{F}}_{\eta_t}\}$-stopping time and $\Gamma_T < +\infty$ by definition since $v_T^u < +\infty$.

Let us introduce the following stochastic differential equation:

$$\begin{aligned}\overline{\xi}_t = \zeta &+ \int_0^t A(\eta_s, \overline{\xi}_s)(1-\theta_s)ds + \int_0^t \theta_s B(\eta_s)\alpha_s ds \\ &+ \int_0^t D(\eta_s, \overline{\xi}_s)\sqrt{(1-\theta_s)(1-\theta_s)^+}dW_{\eta_s}.\end{aligned} \tag{32}$$

Using Assumption A1 and [39, Theorem 6, p. 194], it follows that the solution to the previous equation exists and is unique and continuous on the probability space $\big(\overline{\Omega}, \overline{\mathcal{F}}, \overline{P}, \{\overline{\mathcal{F}}_{\eta_t}\}\big)$.

Using [40, Proposition 4.9, p. 8], Propositions 3.1 and 3.2, and (18), it follows that $\forall\, t \in [0, \Gamma_T]$

$$\int_0^{\Gamma_t} A(\eta_s, \overline{\xi}_s)(1-\theta_s)ds = \int_0^t A(s, \overline{\xi}_{\Gamma_s})ds \quad \text{and} \quad \int_0^{\Gamma_t} \theta_s B(\eta_s)\alpha_s ds = \int_{[0,t]} B(s)du_s. \tag{33}$$

Define the process $\{\overline{W}_t\}$ by $\overline{W}_t = \int_0^t \sqrt{(1-\theta_s)(1-\theta_s)^+}dW_{\eta_s}$. It is a continuous $\{\overline{\mathcal{F}}_{\eta_t}\}$-local martingale such that

$$\langle \overline{W}^i, \overline{W}^j \rangle_t = \int_0^t (1-\theta_s)(1-\theta_s)^+ d\langle W_\eta^i, W_\eta^j \rangle_s = \delta_{ij}\int_0^t (1-\theta_s)(1-\theta_s)^+ d\eta_s = \delta_{ij}\eta_t,$$

where the last equality has been obtained using Proposition 3.2. Moreover, it is easy to show that $\overline{W}_{\Gamma_t} = W_t$. Now using [32, Proposition 4.8, p. 176] and (18), it follows that

$$\begin{aligned}\int_0^{\Gamma_t} D(\eta_s, \overline{\xi}_s)\sqrt{(1-\theta_s)(1-\theta_s)^+}dW_{\eta_s} &= \int_0^{\Gamma_t} D(\eta_s, \overline{\xi}_s)d\overline{W}_s \\ &= \int_0^t D(s, \overline{\xi}_{\Gamma_s})dW_s.\end{aligned} \tag{34}$$

Finally, combining (33) and (34), we obtain that $\{\overline{\xi}_{\Gamma_t}\}$ is a *corlol* process which satisfies (13). From Theorem 6 in [39, p. 194], it follows that (13) has a unique solution and, consequently, $\overline{\xi}_{\Gamma_t} = x_t$

Let $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{P}, \{\widetilde{\mathcal{F}}_t\})$ be a filtered probability space supporting a standard m-dimensional Brownian motion $\{\widetilde{W}_t\}$, and set

$$\Omega = \overline{\Omega} \times \widetilde{\Omega}, \qquad \mathcal{F} = \overline{\mathcal{F}} \otimes \widetilde{\mathcal{F}}, \qquad P = \overline{P} \otimes \widetilde{P}, \qquad \mathcal{F}_t = \overline{\mathcal{F}}_{\eta_t} \otimes \widetilde{\mathcal{F}}_t.$$

A process $\overline{X}$ defined on $\overline{\Omega}$ may be viewed as being defined on Ω by setting $X(\overline{\omega}, \widetilde{\omega}) = \overline{X}(\overline{\omega})$. For simplicity of exposition, the same notation will be used in the rest of the paper to identify the process $\overline{X}$ and X. This approach will also be applied to the processes defined on $\widetilde{\Omega}$.

Let us introduce the process $\{V_t\}$ defined on $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$ by

$$V_t = \int_0^t \sqrt{(1-\theta_s)^+} dW_{\eta_s} + \int_0^t \sqrt{1-(1-\theta_s)(1-\theta_s)^+} d\widetilde{W}_s. \tag{35}$$

Clearly, $\{W_{\eta_t}\}$ and $\{\widetilde{W}_t\}$ are two independent continuous $\{\mathcal{F}_t\}$-martingales. Therefore, $\{V_t\}$ is a continuous $\{\mathcal{F}_t\}$-local martingale such that

$$\langle V^i, V^j \rangle_t = \delta_{ij} \left[\int_0^t (1-\theta_s)^+ d\eta_s + \int_0^t (1-(1-\theta_s)(1-\theta_s)^+) ds \right].$$

However, using Proposition 3.2, it follows that $\langle V^i, V^j \rangle_t = \delta_{ij}\, t$, which, by Levy's characterization theorem, gives that $\{V_t\}$ is a standard m-dimensional $\{\mathcal{F}_t\}$-Brownian motion.

On $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$, let us consider the equation

$$\xi_t = \zeta + \int_0^t A(\eta_s, \xi_s)(1-\theta_s) ds + \int_0^t \theta_s B(\eta_s)\alpha_s ds + \int_0^t D(\eta_s, \xi_s)\sqrt{1-\theta_s} dV_s. \tag{36}$$

Since

$$\int_0^t D(\eta_s, \xi_s)\sqrt{(1-\theta_s)} dV_s = \int_0^t D(\eta_s, \xi_s)\sqrt{(1-\theta_s)(1-\theta_s)^+} dW_{\eta_s},$$

it can be shown that $\{\bar{\xi}_t\}$, defined by (32), is a solution to (36).

Let $(\Omega, \mathcal{G}, Q)$ be the completion of the probability space $(\Omega, \mathcal{F}, P)$. Denote by $\mathcal{N}$ the σ-field, generated by all Q-null sets. Introduce $\mathcal{G}_t = \bigcap_{s>t} (\mathcal{F}_s \vee \mathcal{N})$. Then the probability space $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\})$ satisfies the usual hypotheses (item (i) in Definition 4.1). Clearly, Γ_T is a $\{\mathcal{G}_t\}$-stopping time and $\{(\alpha_t, \theta_t)\}$ is a $\{\mathcal{G}_t\}$-progressively measurable process. Moreover, using Lemmas A.1 and A.2 in [15], it follows that $\{V_t\}$ is a $\{\mathcal{G}_t\}$-Brownian motion.

Let us define the auxiliary control Ψ by

$$\left(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\}, \{(\alpha_t, \theta_t)\}, \{V_t\}, \{(\bar{\xi}'_t, \eta_t)'\}, \Gamma_T\right).$$

To complete the proof, it remains to be shown that the costs are equal: $\mathcal{M}[\Psi] = J[C]$. Since $\eta_{\Gamma_T} = T$, we have by definition of $\{\Gamma_t\}$ that $\Gamma_T - \eta_{\Gamma_T} = v_T^u$. Therefore,

$$E_P[g(x_T, v_T^u)] = E_Q[g(\xi_{\Gamma_T}, \Gamma_T - \eta_{\Gamma_T})]. \tag{37}$$

Denote by $\{\tau_n\}_{n\in\mathbb{N}^*}$ the sequence of $\{\overline{\mathcal{F}}_t\}$-stopping times which exhausts the jumps of $\{\Gamma_t\}$. Since $\{\eta_t\}$ is a time change on $\left(\overline{\Omega}, \overline{\mathcal{F}}, \overline{P}, \{\overline{\mathcal{F}}_t\}\right)$, and using [25, Lemma 10.5(a)], it follows that $\{\Gamma_{\tau_n}\}_{n\in\mathbb{N}^*}$ is a sequence of $\{\overline{\mathcal{F}}_{\eta_t}\}$-stopping times. We have that $\{\Gamma_{\tau_n -} > t\} = \{\eta_t < \tau_n\} \in \overline{\mathcal{F}}_{\eta_t}$; therefore, $\{\Gamma_{\tau_n -}\}_{n\in\mathbb{N}^*}$ is a sequence of $\{\overline{\mathcal{F}}_{\eta_t}\}$-stopping times. Remark that

$$\bigcup_{n=1}^{\infty} [\![\Gamma_{\tau_n -}, \Gamma_{\tau_n}]\!] \subset \{(t,\omega) \in \mathbb{R}_+ \times \Omega : \theta_t = 1\}.$$

Define

$$\mathcal{D} \doteq \{(t,\omega) \in \mathbb{R}_+ \times \Omega : \theta_t = 1\} - \bigcup_{n=1}^{\infty} [\![\Gamma_{\tau_n -}, \Gamma_{\tau_n}]\!].$$

Consequently,

$$(\forall t \in [0,T]), \quad u_t = \int_0^{\Gamma_t} I_{\{\theta_s<1\}}\theta_s\alpha_s ds + \int_0^{\Gamma_t} I_{\{\theta_s=1\}}\theta_s\alpha_s ds$$
$$= \int_0^{\Gamma_t} [I_{\{\theta_s<1\}} + I_{\mathcal{D}}]\theta_s\alpha_s ds + \sum_{n\in\mathbb{N}^*} \int_{\Gamma_{\tau_n-}}^{\Gamma_{\tau_n}} \alpha_s ds\, I_{[\![\tau_n,T]\!]}.$$

For $(s,\omega) \in \bigcup_{n=1}^{\infty} [\![\Gamma_{\tau_n-}, \Gamma_{\tau_n}]\!]$, we have $I_{\{\theta_s<1\}}(\omega) + I_{\mathcal{D}}(t,\omega) = 0$. So the process $\{\int_0^t [I_{\{\theta_s<1\}} + I_{\mathcal{D}}]\theta_s\alpha_s ds\}$ is $\{\Gamma_t\}$-continuous. Consequently, the decomposition of the process $\{u_t\}$ is given by

$$u_t^c = \int_0^{\Gamma_t} [I_{\{\theta_s<1\}} + I_{\mathcal{D}}]\theta_s\alpha_s ds,$$
$$u_t^d = \sum_{n\in\mathbb{N}^*} \int_{\Gamma_{\tau_n-}}^{\Gamma_{\tau_n}} \alpha_s ds\, I_{[\![\tau_n,T]\!]}.$$

Clearly, we have

$$\int_0^{\Gamma_T} \theta_s k(\eta_s, \overline{\xi}_s)\alpha_s ds = \int_0^{\Gamma_t} k(\eta_s, \overline{\xi}_s)[I_{\{\theta_s<1\}} + I_{\mathcal{D}}]\theta_s\alpha_s ds$$
$$+ \sum_{n\in\mathbb{N}^*} \int_{\Gamma_{\tau_n-}}^{\Gamma_{\tau_n}} k(\eta_s, \overline{\xi}_s)\alpha_s ds\, I_{[\![\tau_n,T]\!]}. \tag{38}$$

Moreover, from Proposition 1.4, Chapter V in [40], we obtain that

$$\int_0^{\Gamma_t} k(\eta_s, \overline{\xi}_s)[I_{\{\theta_s<1\}} + I_{\mathcal{D}}]\theta_s\alpha_s ds = \int_0^t k(s, \overline{\xi}_{\Gamma_s}) du_s^c. \tag{39}$$

Since $\{\eta_t\}$ is the right inverse of $\{\Gamma_t\}$, we have that $(\forall (s,\omega) \in [\![\Gamma_{\tau_n-}, \Gamma_{\tau_n}]\!])$, $\eta_s = \tau_n$, and with (25), it gives $(\forall (s,\omega) \in [\![\Gamma_{\tau_n-}, \Gamma_{\tau_n}]\!])$, $\theta_s = \overline{\theta}_{\tau_n} = 1$. By definition of the processes $\{\alpha_t\}$, we have

$$(\forall (s,\omega) \in [\![\Gamma_{\tau_n-}, \Gamma_{\tau_n}]\!]), \quad \alpha_s = \overline{\alpha}_{\tau_n}. \tag{40}$$

Note that $\{\tau_n\}_{n\in\mathbb{N}^*}$ exhausts the jumps of $\{u_t\}$. Using (25), we have that

$$\Delta u_{\tau_n} = \int_{\Gamma_{\tau_n-}}^{\Gamma_{\tau_n}} \overline{\theta}_{\tau_n}\overline{\alpha}_{\tau_n} ds = \Delta\Gamma_{\tau_n}\overline{\alpha}_{\tau_n}. \tag{41}$$

Using the fact that $\Delta\Gamma_{\tau_n} = \Delta v^u_{\tau_n}$ and combining (40) and (41), we obtain that

$$(\forall (s,\omega) \in [\![\Gamma_{\tau_n-}, \Gamma_{\tau_n}]\!]), \quad \alpha_s = \frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}}. \tag{42}$$

Consequently, we have that

$$(\forall (s,\omega) \in [\![\Gamma_{\tau_n-}, \Gamma_{\tau_n}]\!]), \quad \overline{\xi}_s = \overline{\xi}_{\Gamma_{\tau_n-}} + \int_{\Gamma_{\tau_n-}}^{s} \theta_t B(\eta_t)\alpha_t dt$$
$$= \overline{\xi}_{\Gamma_{\tau_n-}} + B(\tau_n)\frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}}(s - \Gamma_{\tau_n-}). \tag{43}$$

Therefore,

$$\int_{\Gamma_{\tau_n-}}^{\Gamma_{\tau_n}} k(\eta_s,\overline{\xi}_s)\alpha_s ds = \int_{\Gamma_{\tau_n-}}^{\Gamma_{\tau_n}} k\left(\tau_n, \overline{\xi}_{\Gamma_{\tau_n-}} + B(\tau_n)\frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}}(s-\Gamma_{\tau_n-})\right)\frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}} ds$$

$$= \int_0^{\Delta v^u_{\tau_n}} k\left(\tau_n, \overline{\xi}_{\Gamma_{\tau_n-}} + B(\tau_n)\frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}}s\right)\frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}} ds. \tag{44}$$

Combining (38), (39), and (44), we obtain that

$$\int_0^{\Gamma_T} \theta_s k(\eta_s,\overline{\xi}_s)\alpha_s ds = \int_0^t k(s,\overline{\xi}_{\Gamma_s}) du^c_s$$

$$+ \sum_{0\le t\le T} \int_0^{\Delta v^u_t} k\left(t, x_{t-} + B(t)\frac{\Delta u_t}{\Delta v^u_t}s\right)\frac{\Delta u_t}{\Delta v^u_t} ds. \tag{45}$$

Using (37) and (45), we can conclude that Ψ is in Υ^a and satisfies $\mathcal{M}[\Psi] = J[C]$, thus giving the result. □

PROPOSITION 4.3. *Suppose* $(\Omega,\mathcal{G},Q,\{\widetilde{\mathcal{G}}_t\},\{(\alpha_t,\theta_t)\},\{V_t\},\{(\xi'_t,\eta_t)'\},\rho)$ *is an element in* Υ^a. *Let* $\{\Gamma_t\}$ *be the right inverse of* $\{\eta_t\}$. *Then* $\{\Gamma_{\eta_t}\}$ *is a time change on* $(\Omega,\mathcal{G},Q,\{\widetilde{\mathcal{G}}_t\})$. *Moreover, the process* $\{\int_0^t \sqrt{1-\theta_s}dV_s\}$ *is a* $\{\widetilde{\mathcal{G}}_{\Gamma_{\eta_t}}\}$ *martingale.*

Proof. The first item is a straightforward consequence from Lemma 1 in [19]. Now let us denote $\int_0^t \sqrt{1-\theta_s}dV_s$ by $\widetilde{V}_t$. Using the fact that $\Gamma_{\eta_t} \ge t$, we obtain that $\{\widetilde{V}_t\}$ is adapted to $\{\widetilde{\mathcal{G}}_{\Gamma_{\eta_t}}\}$ and $\forall\, j \in \mathbb{N}_m$

$$E_Q[(\widetilde{V}^j_{\Gamma_{\eta_t}} - \widetilde{V}^j_t)^2] = E_Q[(\widetilde{V}^j_{\Gamma_{\eta_t}})^2 - (\widetilde{V}^j_t)^2]$$
$$= E_Q[\eta_{\Gamma_{\eta_t}} - \eta_t].$$

However, $\eta_{\Gamma_{\eta_t}} = \eta_t$, implying that $\{\widetilde{V}_t\}$ and $\{\widetilde{V}_{\Gamma_{\eta_t}}\}$ are indistinguishable, thus giving the result. □

THEOREM 4.4. *Assume Assumptions* A1, A2, A3, *and* A4. *Let* Ψ *be an element in* Υ^a. *Then there exists an admissible control* C *in* $\mathfrak{C}^a$ *such that*

$$J[C] \le \mathcal{M}[\Psi]. \tag{46}$$

Proof. Denote Ψ by $(\Omega,\mathcal{G},Q,\{\widetilde{\mathcal{G}}_t\},\{(\alpha_t,\theta_t)\},\{V_t\},\{(\xi'_t,\eta_t)'\},\rho)$. Let $\{\Gamma_t\}$ be the right inverse of $\{\eta_t\}$. Using Proposition 1.1, Chapter V in [40], $\{\Gamma_t\}$ is a time change on $(\Omega,\mathcal{G},Q,\{\widetilde{\mathcal{G}}_t\})$. Consequently, let us define by $\{\tau_n\}_{n\in\mathbb{N}_*}$ the sequence of $\{\widetilde{\mathcal{G}}_{\Gamma_t}\}$-stopping times which exhaust the jumps of $\{\Gamma_t\}$. Using [19, Lemma 11(a)], it follows that $\{\Gamma_{\tau_n}\}$ is a sequence of $\{\widetilde{\mathcal{G}}_t\}$-stopping times. Using Proposition 4.3, we can define the filtration $\{\mathcal{G}_t\}$ by $\mathcal{G}_t \doteq \widetilde{\mathcal{G}}_{\Gamma_{\eta_t}}$, which is right continuous and complete. Moreover, using [19, Remark (b), p. 73], it follows that $\{\Gamma_{\tau_n-}\}$ is a sequence of $\{\mathcal{G}_t\}$-stopping times. Note that the introduction of the filtration $\{\mathcal{G}_t\}$ is necessary since $\{\Gamma_{\tau_n-}\}$ may not be a stopping time with respect to $\{\widetilde{\mathcal{G}}_t\}$.

Let us denote $\int_0^t \sqrt{1-\theta_s}dV_s$ by $\widetilde{V}_t$. Clearly, the processes $\{(\alpha_t,\theta_t)\}$, $\{(\xi'_t,\eta_t)'\}$ are $\{\mathcal{G}_t\}$-progressively measurable and ρ is a $\{\mathcal{G}_t\}$ stopping time. Applying Proposition 4.3, the process $\{\xi_t\}$ satisfies the following equation:

$$\xi_t \doteq \zeta + \int_0^t A(\eta_s,\xi_s)(1-\theta_s)ds + \int_0^t \theta_s B(\eta_s)\alpha_s ds + \int_0^t D(\eta_s,\xi_s)d\widetilde{V}_s.$$

Clearly, $\{\mathcal{G}_{\Gamma_t}\}$ is a right continuous complete filtration. Since $\Psi \in \Upsilon^a$, we have that $\Gamma_T = \rho$.

Define

$$u_t \doteq \int_0^{\Gamma_t} \theta_s \alpha_s ds. \tag{47}$$

Since $\{(\alpha_t, \theta_t)\}$ is a $\mathfrak{B}$-valued process, it follows easily that the process $\{u_t\}$ satisfies (11). Moreover, it is easy to check that $\int_0^{\Gamma_t} I_{\{\eta_s \in (a,b]\}} \theta_s \alpha_s ds = \int_{(a,b]} du_s$ for any real $a < b$. By using the fact that $B(.)$ is continuous and a monotone class theorem, we obtain that

$$\int_0^{\Gamma_t} B(\eta_s)\theta_s \alpha_s ds = \int_0^t B(s) du_s. \tag{48}$$

From Proposition 3.2 and [32, Theorem 4.13, p. 178], it follows that $\{\widetilde{V}_{\Gamma_t}\}$ is a standard m-dimensional $\{\mathcal{G}_{\Gamma_t}\}$-Brownian motion. From the definition of $\{\eta_t\}$, [40, Proposition 4.9, p. 8], [32, Theorem 4.8, p. 176], and (48), it follows that on the filtered probability space $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_{\Gamma_t}\})$, the process $\{\xi_{\Gamma_t}\}$ satisfies the following equation on $[0, T]$:

$$\xi_{\Gamma_t} = \zeta + \int_0^t A(s, \xi_{\Gamma_s}) ds + \int_{[0,t]} B(s) du_s + \int_0^t D(s, \xi_{\Gamma_s}) d\widetilde{V}_{\Gamma_s}. \tag{49}$$

From the definition of $\{u_t\}$ (see (47)), we have that

$$v_T^u \leq \int_0^\rho \theta_s \sum_{i=1}^p |\alpha_s| ds,$$

and since the process $\{(\alpha_t, \theta_t)\}$ takes its value in $\mathfrak{B}$ and using equation (28), we obtain that

$$v_T^u \leq \rho - \eta_\rho. \tag{50}$$

Since $\Psi \in \Upsilon^a$ and $\lim_{t \to +\infty} \inf_{x \in \mathbb{R}^n} g(x,t) = +\infty$ (see Assumption A2), we have that $\rho - \eta_\rho < \infty$ Q-a.s., and therefore $\{u_t\}$ is of finite variation on $[0, T]$: $v_T^u < \infty$ Q-a.s.

Using (50) and Assumption A2, it follows that

$$E_Q\left[g(\xi_\rho, \rho - \eta_\rho)\right] \geq E_Q\left[g(x_T, v_T^u)\right]. \tag{51}$$

As in the proof of Theorem 4.2, we have $(\forall (s,\omega) \in [\![\Gamma_{\tau_n -}, \Gamma_{\tau_n}]\!]), \eta_s = \tau_n$, and with (25), it gives $\theta_s = 1$. However, it must be pointed out that (42) may not be valid here. Therefore, we obtain

$$(\forall (s,\omega) \in [\![\Gamma_{\tau_n -}, \Gamma_{\tau_n}]\!]), \quad \overline{\xi}_s = \overline{\xi}_{\Gamma_{\tau_n -}} + \int_{\Gamma_{\tau_n -}}^s B(\tau_n)\alpha_t dt. \tag{52}$$

Using Assumption A3 and the previous equation, we have that

$$\begin{aligned} L(\tau_n, \xi_{\Gamma_{\tau_n}}) - L(\tau_n, \xi_{\Gamma_{\tau_n -}}) &= \int_{\Gamma_{\tau_n -}}^{\Gamma_{\tau_n}} \frac{d}{ds} L\left(\tau_n, \xi_{\Gamma_{\tau_n -}} + \int_{\Gamma_{\tau_n -}}^s B(\tau_n)\alpha_t dt\right) ds \\ &= \int_{\Gamma_{\tau_n -}}^{\Gamma_{\tau_n}} k\left(\tau_n, \xi_{\Gamma_{\tau_n -}} + \int_{\Gamma_{\tau_n -}}^s B(\tau_n)\alpha_t dt\right) \alpha_s ds. \end{aligned} \tag{53}$$

However, note that $\Delta u_{\tau_n} = \int_{\Gamma_{\tau_n-}}^{\Gamma_{\tau_n}} \alpha_t dt$, and so it follows from using Assumption A3 again that

$$\begin{aligned} L(\tau_n, \xi_{\Gamma_{\tau_n}}) - L(\tau_n, \xi_{\Gamma_{\tau_n-}}) &= L(\tau_n, \xi_{\Gamma_{\tau_n-}} + B(\tau_n)\Delta u_{\tau_n}) - L(\tau_n, \xi_{\Gamma_{\tau_n-}}) \\ &= \int_0^{\Delta v^u_{\tau_n}} \frac{d}{ds} L\left(\tau_n, \xi_{\Gamma_{\tau_n-}} + B(\tau_n)\frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}} s\right) ds \\ &= \int_0^{\Delta v^u_{\tau_n}} k\left(\tau_n, \xi_{\Gamma_{\tau_n-}} + B(\tau_n)\frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}} s\right) \frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}} ds. \end{aligned} \tag{54}$$

Combining (53) and (54), we have that

$$\begin{aligned} &\int_{\Gamma_{\tau_n-}}^{\Gamma_{\tau_n}} k\left(\tau_n, \xi_{\Gamma_{\tau_n-}} + \int_{\Gamma_{\tau_n-}}^{s} B(\tau_n)\alpha_t dt\right) \alpha_s ds \\ &\qquad = \int_0^{\Delta v^u_{\tau_n}} k\left(\tau_n, \xi_{\Gamma_{\tau_n-}} + B(\tau_n)\frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}} s\right) \frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}} ds. \end{aligned} \tag{55}$$

Similarly to the proof of Theorem 4.2, we obtain that

$$\int_0^{\rho} \theta_s k(\eta_s, \xi_s)\alpha_s ds = \int_0^T k(s, \xi_{\Gamma_s}) du^c_s + \sum_{n\in\mathbb{N}^*} \int_{\Gamma_{\tau_n-}}^{\Gamma_{\tau_n}} k(\eta_s, \xi_s)\alpha_s ds \, I_{[\![\tau_n, T]\!]}. \tag{56}$$

Using (55) and (56), we obtain that

$$\begin{aligned} \int_0^{\rho} \theta_s k(\eta_s, \xi_s)\alpha_s ds &= \int_0^T k(s, \xi_{\Gamma_s}) du^c_s \\ &\quad + \sum_{n\in\mathbb{N}^*} \int_0^{\Delta v^u_{\tau_n}} k\left(\tau_n, \xi_{\Gamma_{\tau_n-}} + B(\tau_n)\frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}} s\right) \frac{\Delta u_{\tau_n}}{\Delta v^u_{\tau_n}} ds \, I_{[\![\tau_n, T]\!]} \\ &= \int_0^T k(s, \xi_{\Gamma_s}) du^c_s + \sum_{0\le t\le T} \int_0^{\Delta v^u_t} k\left(t, \xi_{\Gamma_{t-}} + B(t)\frac{\Delta u_t}{\Delta v^u_t} s\right) \frac{\Delta u_t}{\Delta v^u_t} ds. \end{aligned} \tag{57}$$

Consequently, combining (51) and (57), the control

$$C \doteq \left(\Omega, \mathcal{G}, Q, \{\mathcal{G}_{\Gamma_t}\}, \left\{\int_0^{\Gamma_t} \theta_s \alpha_s ds\right\}, \{\widetilde{V}_{\Gamma_t}\}, \{\xi_{\Gamma_t}\}\right)$$

is such that $J[C] \le \mathcal{M}[\Psi]$, and therefore C is in $\mathfrak{C}^a$, thus giving the result. □

An immediate consequence of Theorems 4.2 and 4.4 is the following result.

THEOREM 4.5. *Assume Assumptions* A1, A2, A3*, and* A4*. The singular and the auxiliary control problems are equivalent:*

$$\inf_{C\in\mathfrak{C}^a} J[C] = \inf_{\Psi\in\Upsilon^a} \mathcal{M}[\Psi]. \tag{58}$$

5. Existence of optimal singular control. The existence of an optimal control for the singular control problem is guaranteed by the following result.

THEOREM 5.1. *Assume Assumptions* A1, A2, A3, A4*, and* A5*. For the singular control problem there exists an optimal control $C^* \in \mathfrak{C}^a$:*

$$J[C^*] = \inf_{C\in\mathfrak{C}^a} J[C]. \tag{59}$$

Proof. The existence of C^* is obtained through the equivalence result obtained in Theorem 4.5. Let us introduce the set $\mathcal{H}$ of controls $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\}, \{(\alpha_t, \theta_t)\}, \{V_t\}, \{\Lambda_t\}, \rho)$, for which $(\Omega, \mathcal{G}, Q)$ is a probability space with filtration $\{\mathcal{G}_t\}$, the processes $\{(\alpha_t, \theta_t)\}$, $\{V_t\}, \{\Lambda_t\}, \rho$ satisfy items (i)–(v) in Definition 4.1, and such that

$$\mathcal{N}[\Psi] \doteq E_Q \left[g(\xi_\rho, \rho - \eta_\rho) + \int_0^\rho \theta_s k(\eta_s, \xi_s) \alpha_s ds + G(\eta_\rho) \right].$$

Define the corresponding admissible class of controls $\mathcal{H}^a \doteq \{\Psi \in \mathcal{H} : \mathcal{N}[\Psi] < \infty\}$. Using the fact that $\mathfrak{B}$ is a compact set and Remark 4.3 in [22], and applying Corollary 4.8 in [22], it follows that there exists an optimal control $\widetilde{\Psi}$ in the class of control $\mathcal{H}^a$ such that

$$\inf_{\Psi \in \mathcal{H}^a} \mathcal{N}[\Psi] = \mathcal{N}[\widetilde{\Psi}].$$

Note that Corollary 4.8 in [22] provides a way of showing the existence of an optimal control in the class $\mathcal{H}^a$, but in order to use our equivalence result given by Theorem 4.5, we need to find an optimal control in the class of controls Υ^a. It must be pointed out that $\Upsilon^a \subset \mathcal{H}^a$ (for the definition of Υ^a, see Definition 4.1), since a control in $\mathcal{H}^a$ is defined on a filtered probability space, which is not necessarily right continuous, and is completely contrary to a control in Υ^a, which must be defined on a filtered probability space satisfying the usual hypotheses (right continuity and completeness). However, by using Lemmas A.1 and A.2 in [15], it can be shown that there exists a modification Ψ^* in Υ^a of $\widetilde{\Psi}$ such that $\mathcal{M}[\Psi^*] = \mathcal{N}[\widetilde{\Psi}]$. Since $\Upsilon^a \subset \mathcal{H}$, it follows that $\inf_{\Psi \in \Upsilon^a} \mathcal{M}[\Psi] = \mathcal{M}[\Psi^*]$. Finally, Theorem 4.5 gives the result. □

6. Extensions and example. We want to point out that our work can be extended in several directions. Hard and soft constraints can be added, such as a finite fuel constraint. The model defined by (1)–(3) is used in order to simplify and clarify the exposition and the derivation of our results. Other related problems, such as the maximum principle for singular control problems, can be studied using our approach. In [16], it is shown how this method can be applied to reexamine the maximum principle studied in [11]. In the deterministic context, such extensions have been studied in detail in the recent book [37].

For example, the results we obtained in the previous sections (the equivalence results and the existence results) remain unchanged if we assume that the control is nondecreasing. We write $\overline{\mathfrak{C}}$ for the set of controls satisfying the conditions of Definition 2.1, except that item (ii) is replaced by the following:

(ii′) $\{u_t\}$ is an $\mathbb{R}^p$-valued, nondecreasing *corlol* $\{\mathcal{F}_t\}$-progressively measurable process such that

$$(\forall A \in \mathcal{B}([0,T]) \otimes \mathcal{F}), \quad \int_0^T I_A du_t \in K,$$
$$v_T^u < +\infty.$$

Let $\overline{\mathfrak{C}}^a$ be the corresponding class of admissible control:

$$\overline{\mathfrak{C}}^a \doteq \{C \in \overline{\mathfrak{C}} : J[C] < \infty\}.$$

In the same way, we write $\overline{\Upsilon}$ for the set of controls satisfying the conditions of Definition 4.1, except that item (ii) is replaced by the following item:

(ii′) $\{(\alpha_t, \theta_t)\}$ is a $\overline{\mathfrak{B}}$-valued, $\{\mathcal{G}_t\}$-progressively measurable process, where $\overline{\mathfrak{B}} \doteq \{(x, y) \in K \times [0, 1] : (\forall j \in \mathbb{N}_p),\ 0 \leq x^j, \sum_{i=1}^{p} x^j \leq 1\}$.

It is necessary to replace $\mathfrak{B}$ by $\overline{\mathfrak{B}}$, since now we need to consider nondecreasing control processes. Moreover, denote

$$\overline{\Upsilon}^a \doteq \{\Psi \in \overline{\Upsilon} : \mathcal{M}[\Psi] < \infty\}.$$

Now let us introduce the following new assumption.

Assumption A6. For all $(t, x) \in [0, T] \times \mathbb{R}^n$, the set $\overline{K}(t, x)$ defined by

$$\overline{K}(t, x) \doteq \{(A(t, x)(1 - \theta) + \theta B(t)\alpha, (1 - \theta)D(t, x)D(t, x)', 1 - \theta) : (\alpha, \theta) \in \overline{\mathfrak{B}}\}$$

is convex.

Proceeding as before, the following can be easily shown.

COROLLARY 6.1. *Assume Assumptions* A1, A2, A3, *and* A4. *The singular and the auxiliary control problems are equivalent:*

$$\inf_{C \in \overline{\mathfrak{C}}^a} J[C] = \inf_{\Psi \in \overline{\Upsilon}^a} \mathcal{M}[\Psi].$$

COROLLARY 6.2. *Assume Assumptions* A1, A2, A3, A4, *and* A6. *For the singular control problem there exists an optimal control* $C^* \in \overline{\mathfrak{C}}^a$*:*

$$J[C^*] = \inf_{C \in \overline{\mathfrak{C}}^a} J[C].$$

Classically, optimal singular controls are studied by assuming that the class of admissible controls is characterized by left continuous processes and that the cost function has a terminal cost. Sometimes it leads to nonexistence of optimal control within the class of left continuous controls (see the well-known example in [29, section 4, p. 863] by Karatzas and Shreve). Indeed, in such examples, one may interpret this nonexistence in the following manner. The control minimizing the cost function can be left continuous on the time interval $[0, T)$ and has a jump at the terminal time. This control does not belong to the class of left continuous admissible controls but can be obtained as a limit of a sequence of left continuous control processes, implying, therefore, the nonexistence of an admissible optimal control. In order to reject such a possibility, a special assumption must be made on the terminal cost, rendering the possibility of a terminal jump unfavorable to any state (x_{T-}, v^u_{T-}) that precedes the terminal state, namely,

$$g(x_{T-} + B(T)\Delta u_T, v^u_{T-} + |\Delta u_T|) \geq g(x_{T-}, v_{T-})$$

for any admissible impulse Δu_T of the singular control applied at the terminal time, and for any (x_{T-}, v^u_{T-}). Other related conditions guaranteeing the existence of the optimal control in the class of left continuous control processes can be found, for example, in [29, Condition 4.2] or in [7, Condition 3.6] and in many other papers. Note that with our approach, we can study in a straightforward manner problems where the control is assumed to be right continuous (allowing a jump at the terminal time) and where the cost function admits a terminal cost. The advantage of our approach is that one can find an optimal solution to singular control problems which do not admit optimal solutions when the control is supposed to be left continuous.

In [29, section 4, p. 863], Karatzas and Shreve present an example of nonexistence of optimal control when the control is left continuous. With our notation, the state

$\{x_t\}$ is a real-valued process satisfying (13) with $A = 0$, $B = -1$, and $D = 1$, where the control process $\{u_t\}$ is a nondecreasing function, and where the cost is defined by $g(x, y) = x^2 + y$. However, using Corollary 6.2, we obtained that there exists an optimal control in the class of right continuous controls which shows the importance of choosing such a class of admissible controls.

7. Continuous approximation of the singular optimal control. This section addresses the problem of approximating singular optimal control C^* using absolutely continuous controls. In Theorem 7.5, it is shown that there exists a sequence of admissible continuous controls $\{C^k\}$ such that $\lim_{k\to\infty} J[C^k] = J[C^*]$. We assume in this section that $k = 0$.

The following assumption will be used in this section.

Assumption A7. There exist a bounded, continuous function $g_1 : \mathbb{R}^n \to \mathbb{R}_+$ and a Lipschitz function $g_2 : \mathbb{R}^p \to \mathbb{R}_+$ such that

$$(\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^p), \quad g(x, y) \doteq g_1(x) + g_2(y).$$

Let us denote by $\Psi^* \doteq (\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\}, \{(\alpha_t, \theta_t)\}, \{V_t\}, \{(\xi_t', \eta_t, \mu_t')'\}, \rho)$ an optimal control for the auxiliary control problem, and let $\{\Gamma_t\}$ be the right inverse of $\{\eta_t\}$.

Let us introduce the following processes on the probability space $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\})$:

$$\gamma^m \doteq \rho \wedge m, \tag{60}$$

$$(\forall t \in \mathbb{R}_+), \qquad \theta_t^m \doteq \theta_t I_{[\![0,\gamma^m]\!]}(t), \tag{61}$$

$$(\forall t \in \mathbb{R}_+), \qquad \eta_t^m \doteq \int_0^t (1 - \theta_s^m) ds, \tag{62}$$

$$(\forall t \in [0, T]), \qquad \Gamma_t^m \doteq \inf\{s \in \mathbb{R}_+ : \eta_s^m > t\}, \tag{63}$$

$$\rho^m \doteq \Gamma_T^m, \tag{64}$$

and define

$$\begin{aligned} \xi_t^m \doteq \zeta &+ \int_0^t A(\eta_s^m, \xi_s^m)(1 - \theta_s^m) ds + \int_0^t \theta_s^m B(\eta_s^m) \alpha_s ds \\ &+ \int_0^t D(\eta_s^m, \xi_s^m) \sqrt{1 - \theta_s^m} dV_s \end{aligned} \tag{65}$$

for $m \in \mathbb{N}$.

Remark 7.1. *Since the processes* $\{\Gamma_t\}$ *are continuous on the right with left-hand limits, the process* $\{\widetilde{\Gamma}_t\}$ *can be defined by*

$$(\forall t \in [0, T)), \qquad \widetilde{\Gamma}_t \doteq \lim_{\substack{s \to t \\ s < t}} \Gamma_s, \tag{66}$$

and by $\widetilde{\Gamma}_T \doteq \rho$.

LEMMA 7.2. *For each* $m \in \mathbb{N}$, *write*

$$\Psi^m \doteq \left(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\}, \{(\theta_t^m, \alpha_t)\}, \{V_t\}, \{(\xi_t^{m\prime}, \eta_t^m)'\}, \rho^m\right). \tag{67}$$

Then for each $m \in \mathbb{N}$, *the control* Ψ^m *belongs to* Υ^a *and satisfies*

$$\xi_{\rho^m}^m \xrightarrow[m\to\infty]{Q} \xi_\rho, \qquad \rho^m \xrightarrow[m\to\infty]{Q} \rho, \tag{68}$$

$$\sup_{t \in \mathbb{R}_+} |\xi_{t\wedge\rho^m}^m - \xi_{t\wedge\rho}| \xrightarrow[m\to\infty]{Q} 0. \tag{69}$$

Proof. Clearly, $\{(\theta_t^m, \alpha_t)\}$ is a $\mathfrak{B}$-valued, $\{\mathcal{G}_t\}$-progressively measurable process. From Theorem 6 in [39, p. 194], it follows that a solution to (65) exists. By definition, ρ^m is a $\{\mathcal{G}_t\}$-stopping time such that $\eta^m_{\rho^m} = T$, implying that $G(\eta^m_{\rho^m}) = 0$.

Using Assumption A7, we have

$$E_Q[g(\xi^m_{\rho^m}, \rho^m - \eta^m_{\rho^m}) + G(\eta^m_{\rho^m})] \leq L_3(1 + E_Q[g_2(\rho^m - \eta^m_{\rho^m})]).$$

However, using the definitions of ρ^m and $\{\eta^m_t\}$ (see (60) and (62)), we obtain that

$$\rho^m \leq \gamma^m + T. \tag{70}$$

Consequently, using (64), we have that $\rho^m \leq T + m$, implying $\mathcal{M}[\Psi^m] < \infty$ and $\Psi^m \in \Upsilon^a$. Moreover, we have $\{\gamma^m = \rho\} \subset \{\xi^m_{\rho^m} = \xi_\rho\}$, $\{\gamma^m = \rho\} \subset \{\rho^m = \rho\}$, and $\{\gamma^m = \rho\} \subset \{\sup_{t\in\mathbb{R}_+} |\xi^m_{t\wedge\rho^m} - \xi_{t\wedge\rho}| = 0\}$. It is easy to show that $(\forall \delta \in \mathbb{R}^*_+)$, $(\exists M \in \mathbb{N})$ such that $(\forall m \geq M)$, $Q(\{\gamma^m = \rho\}) \geq 1 - \delta$, giving (68) and (69). ∎

LEMMA 7.3. *For each $m \in \mathbb{N}$, there exists a control*

$$\Psi^{m,n} \doteq \left(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\}, \{(\theta^{m,n}_t, \alpha_t)\}, \{V_t\}, \{(\xi^{m,n\prime}_t, \eta^{m,n}_t)'\}, \rho^{m,n}\right) \in \Upsilon^a$$

such that $\{\eta^{m,n}_t\}$ is a strictly increasing process, and

$$\xi^{m,n}_{\rho^{m,n}} \xrightarrow[n\to\infty]{Q} \xi^m_{\rho^m}, \qquad \rho^{m,n} \xrightarrow[n\to\infty]{Q} \rho^m, \tag{71}$$

and

$$\rho^{m,n} \leq T + \rho^m. \tag{72}$$

Moreover, we have

$$(\forall m \in \mathbb{N}), \quad \sup_{t\in\mathbb{R}_+} |\xi^{m,n}_{t\wedge\rho^{m,n}} - \xi^m_{t\wedge\rho^m}| \xrightarrow[n\to\infty]{Q} 0. \tag{73}$$

Proof. The arguments used to show these results are similar to those presented in [15, Lemma 4.7, Proposition 4.8] and are therefore omitted. ∎

PROPOSITION 7.4. *There exists a sequence of controls $\{\Phi^k\}$ in Υ^a where*

$$\Phi^k \doteq \left(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\}, \{(\overline{\theta}^k_t, \alpha_t)\}, \{V_t\}, \{(\overline{\xi}^{k\prime}_t, \overline{\eta}^k_t)'\}, \overline{\rho}^k\right) \tag{74}$$

such that $\{\overline{\eta}^k_t\}$ is a strictly increasing process and

$$\lim_{k\to\infty} \mathcal{M}[\Phi^k] = \mathcal{M}[\Psi^*], \tag{75}$$

$$\lim_{k\to\infty} \overline{\xi}^k_{\overline{\rho}^k} = \xi_\rho, \tag{76}$$

$$\lim_{k\to\infty} \sup_{t\in\mathbb{R}_+} |\overline{\xi}^k_{t\wedge\overline{\rho}^k} - \xi_{t\wedge\rho}| = 0, \tag{77}$$

$$(\forall t \in [0, T]), \quad \lim_{k\to\infty} \overline{\Gamma}^k_t = \widetilde{\Gamma}_t, \tag{78}$$

where $\{\overline{\Gamma}^k_t\}$ denotes the right inverse of the process $\{\overline{\eta}^k_t\}$.

Proof. Using (68)–(73), a subsequence denoted by $\{\Psi^{m_k,n_k}\}$ can be extracted from $\{\Psi^{m,n}\}$ to give

$$\xi^{m_k,n_k}_{\rho^{m_k,n_k}} \xrightarrow[k\to\infty]{Q} \xi_\rho, \qquad \rho^{m_k,n_k} \xrightarrow[k\to\infty]{Q} \rho, \tag{79}$$

and

$$\sup_{t\in\mathbb{R}_+} |\xi^{m_k,n_k}_{t\wedge\rho^{m_k,n_k}} - \xi_{t\wedge\rho}| \xrightarrow[k\to\infty]{Q} 0. \tag{80}$$

Again, a subsequence denoted by

$$\left\{\Phi^k \doteq \left(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\}, \{(\overline{\alpha}^k_t, \theta_t)\}, \{V_t\}, \{(\overline{\xi}^{k\prime}_t, \overline{\eta}^k_t, \overline{\mu}^{k\prime}_t)'\}, \overline{\rho}^k\right)\right\}$$

can be extracted from $\{\Psi^{m_k,n_k}\}$ to ensure that the previous limits hold almost surely. Moreover, using arguments similar to those presented in the proof of Theorem 5.1 in [15], it can be shown that $(\forall t \in [0,T])$, $\lim_{k\to\infty} \overline{\Gamma}^k_t = \widetilde{\Gamma}_t$, where $\{\overline{\Gamma}^k_t\}$ denotes the right inverse of the process $\{\overline{\eta}^k_t\}$, giving equations (76) and (77).

By definition, we have that $\forall\, k \in \mathbb{N}$, $\{\overline{\eta}^k_t\}$ is a strictly increasing process and $G(\overline{\eta}^k_{\overline{\rho}^k}) = 0$. Now, using the fact that g_1 is bounded and continuous, we have that

$$\lim_{k\to\infty} E_Q[g_1(\overline{\xi}^k_{\overline{\rho}^k})] = E_Q[g_1(\xi_\rho)].$$

Note that we have $\eta_\rho = T$ and $\overline{\eta}^k_{\overline{\rho}^k} = T$. From (60), (70), and (72), it follows easily that $\overline{\rho}^k \le \rho + 2T$. Since g_2 is monotone increasing, we have that $g_2(\overline{\rho}^k - T) \le g_2(\rho + T)$. Using the fact that $\Psi^* \in \Upsilon^a$, we obtain that $E_Q[g_2(\rho - T)] < \infty$, and by hypothesis g_2 is Lipschitz, it follows that $E_Q[g_2(\rho + T)] < \infty$. Therefore the sequence $\{g_2(\overline{\rho}^k - T)\}$ is uniformly integrable and converges to $\{g_2(\rho - T)\}$ in L_1. Finally, we obtain that

$$\lim_{k\to\infty} E_Q[g_2(\overline{\rho}^k - \overline{\eta}^k_{\overline{\rho}^k})] = E_Q[g_2(\rho - \eta_\rho)],$$

showing the last part of the result. □

THEOREM 7.5. *There exists a sequence of control $\{C^k\}$ in $\mathfrak{C}^a$,*

$$C^k \doteq \left(\Omega, \mathcal{F}, P, \{\mathcal{F}^k_t\}, \{u^k_t\}, \{W^k_t\}, \{x^k_t\}\right),$$

such that $\{u^k_t\}$ is an absolutely continuous process and

$$\lim_{k\to\infty} J[C^k] = J[C^*], \tag{81}$$

where C^* is an optimal control for the singular control problem. Moreover, the sequence of control $\{C^k\}$ satisfies

$$(\forall t \in [0,T)), \quad \lim_{k\to\infty} x^k_t = \lim_{\substack{s\to t\\ s<t}} x_s, \tag{82}$$

$$\lim_{k\to\infty} x^k_T = x_T. \tag{83}$$

Proof. From Theorems 4.4, 4.5, and 5.1, it follows that there exists a filtered probability space $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$ satisfying the usual hypotheses such that the control defined by

$$C^* \doteq \left(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \left\{\int_0^{\Gamma_t} \theta_s \alpha_s ds\right\}, \left\{\int_0^{\Gamma_t} \sqrt{(1-\alpha_s)} dV_s\right\}, \{\xi_{\Gamma_t}\}\right)$$

is an optimal control for the singular control problem. Moreover, we have that

$$J[C^*] = \mathcal{M}[\Phi^*]. \tag{84}$$

From the sequence of control $\{\Phi^k\}$ in Υ^a defined by (74) and using Theorem 4.4, we obtain that there exists a sequence of control $\{C^k\}$ in $\mathfrak{C}^a$ where

$$C^k \doteq \left(\Omega, \mathcal{F}, P, \{\mathcal{F}_t^k\}, \{u_t^k\}, \{W_t^k\}, \{x_t^k\}\right),$$

satisfying

$$J[C^k] \leq \mathcal{M}[\Phi^k], \tag{85}$$

$$x_t^k \doteq \overline{\xi}^k_{\overline{\Gamma}^k_t}, \tag{86}$$

$$u_t^k \doteq \int_0^{\overline{\Gamma}^k_t} \overline{\theta}^k_s \alpha_s ds, \tag{87}$$

with $\{\overline{\Gamma}^k_t\}$ the right inverse of $\{\overline{\eta}^k_t\}$.

Combining (75), (84), and (85), we obtain (81). Moreover, with (77) and (78), we have

$$\lim_{k\to\infty} \overline{\xi}^k_{\overline{\Gamma}^k_t \wedge \overline{\rho}^k} = \xi_{\widetilde{\Gamma}_t \wedge \rho}.$$

Then because $\overline{\Gamma}^k_t \wedge \overline{\rho}^k = \overline{\Gamma}^k_{t\wedge T}$ and $\widetilde{\Gamma}_t \wedge \rho = \widetilde{\Gamma}_{t\wedge T}$, we obtain (82) and (83).

Since $\{\overline{\eta}^k_t\}$ is a strictly increasing continuous process, we have that $\forall\, k \in \mathbb{N}$, $\overline{\Gamma}^k_{\overline{\eta}^k_t} = t$, and $\{\overline{\Gamma}^k_t\}$ is a strictly increasing process. Differentiating the previous equality and using the fact that $\overline{\eta}^k_t = \int_0^t (1-\overline{\theta}^k_s) ds$, we obtain that $\overline{\Gamma}^k_t = \int_0^t \frac{1}{1-\overline{\theta}^k_{\overline{\Gamma}^k_s}} ds$. Consequently, we have that

$$u_t^k = \int_0^t \alpha^k_{\overline{\Gamma}^k_s} \frac{\overline{\theta}_{\overline{\Gamma}^k_s}}{1-\overline{\theta}^k_{\overline{\Gamma}^k_s}} ds,$$

showing the result. ∎

Acknowledgments. The authors would like to thank the anonymous referees for their suggestions, which have greatly improved the presentation of the paper. This work was undertaken while the second author held a visiting professorship at the University Montesquieu–Bordeaux IV. The authors are grateful to the University Montesquieu–Bordeaux IV for its support.

REFERENCES

[1] L. Alvarez, *A class of solvable singular stochastic control problems*, Stochastics Stochastics Rep., 67 (1999), pp. 83–122.

[2] L. H. R. Alvarez, *Singular stochastic control, linear diffusions, and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.

[3] L. Alvarez, M. Gyllenberg, and L. Shepp, *Optimal Harvesting in the Presence of Density-Dependent Extinction Probabilities*, TUCS Technical Report, 2001, p. 25.

[4] F. Baldursson, *Singular stochastic control and optimal stopping*, Stochastics, 21 (1987), pp. 1–40.

[5] F. Baldursson and I. Karatzas, *Irreversible investment and industry equilibrium*, Finance Stoch., 1 (1997), pp. 69–89.

[6] J. Bather and H. Chernoff, *Sequential decisions in the control of a spaceship*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. III: Physical Sciences, University of California Press, Berkeley, CA, 1967, pp. 181–207.

[7] F. Benth and K. Reikvam, *A Note on the Multi-Dimensional Monotone Follower Problem and Its Connection to Optimal Stopping*, Technical report 1999-10, University of Aarhus, Aarhus, Denmark, 1999.

[8] F. Boetius, *Bounded variation singular stochastic control and associated Dynkin game*, in Mathematical Finance, Trends Math., M. Kohlmann and S. Tang, eds., Birkhäuser, Basel, 2001, pp. 111–120.

[9] F. Boetius, *Singular Stochastic Control and Its Relations to Dynkin Game and Entry-Exit Problems*, Ph.D. thesis, University of Konstanz, 2001.

[10] F. Boetius and M. Kohlmann, *Connections between optimal stopping and singular stochastic control*, Stochastic Process. Appl., 77 (1998), pp. 253–281.

[11] A. Cadenillas and U. Haussmann, *The stochastic maximum principle for a singular control problem*, Stochastics Stochastics Rep., 49 (1994), pp. 211–237.

[12] M. B. Chiarolla and U. G. Haussmann, *Optimal control of inflation: Central bank problem*, SIAM J. Control Optim., 36 (1998), pp. 1099–1132.

[13] P. Chow, J.-L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.

[14] M. Davis and M. Zervos, *A pair of explicitly solvable singular stochastic control problems*, Appl. Math. Optim., 38 (1998), pp. 327–352.

[15] F. Dufour and B. M. Miller, *Generalized solutions in nonlinear stochastic control problems*, SIAM J. Control Optim., 40 (2002), pp. 1724–1745.

[16] F. Dufour and B. Miller, *The maximum principle for singular stochastic control problems*, in preparation.

[17] N. El Karoui and I. Karatzas, *Probabilistic aspects of finite-fuel, reflected follower problems*, Acta Appl. Math., 11 (1988), pp. 223–258.

[18] N. El Karoui and I. Karatzas, *A new approach to the skorohod problem, and its applications*, Stochastics Stochastics Rep., 34 (1991), pp. 57–82.

[19] N. El Karoui and P. Meyer, *Les changements de temps en théorie générale des processus*, in Seminaire de Probabilités XI, Lectures Notes in Math. 581, Springer-Verlag, Berlin, 1975/1976, pp. 65–78.

[20] N. El Karoui, H. Nguyen, and M. Jeanblanc-Picqué, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics, 20 (1987), pp. 169–219.

[21] W. Fleming and H. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[22] U. G. Haussmann and J.-P. Lepeltier, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.

[23] U. G. Haussmann and W. Suo, *Singular optimal controls* I: *Existence*, SIAM J. Control Optim., 33 (1995), pp. 916–936.

[24] U. G. Haussmann and W. Suo, *Singular optimal controls* II: *Dynamic programming*, SIAM J. Control Optim., 33 (1995), pp. 937–959.

[25] J. Jacod, *Calcul stochastique et problèmes de martingales*, Lectures Notes in Math. 714, Springer-Verlag, Berlin, 1979.

[26] J. Jacod and A. Shiryaev, *Limit Theorems for Stochastic Porcesses*, 2nd ed., Springer-Verlag, New York, 2003.

[27] I. Karatzas, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.

[28] I. Karatzas, *Probabilistic aspects of finite-fuel stochastic control*, Proc. Natl. Acad. Sci. USA, 82 (1985), pp. 5579–5581.
[29] I. Karatzas and S. Shreve, *Connections between optimal stopping and singular stochastic control* I. *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.
[30] I. Karatzas and S. Shreve, *Connections between optimal stopping and singular stochastic control* II. *Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–451.
[31] I. Karatzas and S. Shreve, *Equivalent models for finite-fuel stochastic control*, Stochastics, 18 (1986), pp. 245–276.
[32] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York, 1991.
[33] I. Karatzas and H. Wang, *A barrier option of American type*, Appl. Math. Optim., 42 (2000), pp. 259–279.
[34] T. Kobila, *A class of solvable stochastic investment problems involving singular controls*, Stochastics Stochastics Reports, 43 (1993), pp. 29–63.
[35] L. F. Martins and H. J. Kushner, *Routing and singular control for queueing networks in heavy traffic*, SIAM J. Control Optim., 28 (1990), pp. 1209–1233.
[36] J.-L. Menaldi and M. Taksar, *Optimal correction problem of a multidimensional stochastic system*, Automatica J. IFAC, 25 (1989), pp. 223–232.
[37] B. Miller and E. Rubinovitch, *Impulsive Control in Continuous and Discrete-Continuous Systems*, Kluwer Academic/Plenum Press, New York, 2003.
[38] B. M. Miller and W. J. Runggaldier, *Optimization of observations: A stochastic control approach*, SIAM J. Control Optim., 35 (1997), pp. 1030–1050.
[39] P. Protter, *Stochastic Integration and Differential Equations*, Springer-Verlag, New York, 1990.
[40] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion,* 3rd ed., Springer-Verlag, New York, 1999.
[41] S. E. Shreve, J. P. Lehoczky, and D. P. Gaver, *Optimal consumption for general diffusions with absorbing and reflecting barriers*, SIAM J. Control Optim., 22 (1984), pp. 55–75.
[42] M. I. Taksar, *Infinite-dimensional linear programming approach to singular stochastic control*, SIAM J. Control Optim., 35 (1997), pp. 604–625.
[43] H. Zhu, *Generalized solution in singular control: The nondegenrate problem*, Appl. Math. Optim., 25 (1992), pp. 225–245.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 731–742

PROXIMAL METHODS FOR COHYPOMONOTONE OPERATORS*

PATRICK L. COMBETTES† AND TEEMU PENNANEN‡

Abstract. Conditions are given for the viability and the weak convergence of an inexact, relaxed proximal point algorithm for finding a common zero of countably many cohypomonotone operators in a Hilbert space. In turn, new convergence results are obtained for an extended version of the proximal method of multipliers in nonlinear programming.

Key words. cohypomonotone operator, common zero problem, hypomonotone operator, method of multipliers, nonlinear programming, proximal point method, weak convergence

AMS subject classifications. 47H04, 65K10, 90C26, 90C30

DOI. 10.1137/S0363012903427336

1. Introduction. Let $\mathcal{H}$ be a real Hilbert space with scalar product $\langle \cdot\,,\cdot \rangle$, norm $\|\cdot\|$, and distance d. A basic problem in applied mathematics and optimization is to find a zero of a maximal monotone operator $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$, that is, a point $x \in \mathcal{H}$ such that $0 \in Ax$ [22, 23, 29]. Assuming $0 \in \operatorname{ran} A$, since the resolvent $(\mathrm{Id}+A)^{-1}$ of A is a firmly nonexpansive operator with fixed point set $A^{-1}(0)$, a zero of A can be constructed iteratively through the recursion

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = (\mathrm{Id}+A)^{-1}x_n. \tag{1.1}$$

Indeed, since an operator T is nonexpansive if and only if its average $(T+\mathrm{Id})/2$ is firmly nonexpansive [28, Lemma 1.1], it follows from [19, Theorem 3] that, for any $x_0 \in \mathcal{H}$, the sequence $(x_n)_{n\in\mathbb{N}}$ generated by the successive approximations (1.1) converges weakly to a zero of A (see also [17] for a special case). More generally, let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in $]0,+\infty[$ such that $\inf_{n\in\mathbb{N}}\gamma_n > 0$ and let $(e_n)_{n\in\mathbb{N}}$ be an absolutely summable sequence in $\mathcal{H}$. Then, for every $x_0 \in \mathcal{H}$, the so-called proximal point iterations $x_{n+1} = (\mathrm{Id}+\gamma_n A)^{-1}x_n + e_n$ converge weakly to a zero of A [22, Theorem 1] (see also [4] for further analysis). This result was shown in [11, Theorem 3] to remain true for the relaxed proximal iterations

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = x_n + \lambda_n\big((\mathrm{Id}+\gamma_n A)^{-1}x_n + e_n - x_n\big), \tag{1.2}$$

where $(\lambda_n)_{n\in\mathbb{N}}$ lies in $[\varepsilon, 2-\varepsilon]$ for some arbitrary $\varepsilon \in]0,1[$. A further extension was proposed in [2, Corollary 6.1(i)] (for $e_n \equiv 0$) and then in [7, Theorem 6.9(i)], where weak convergence to a common zero of a countable family of maximal monotone operators $(A_i)_{i\in I}$ was established for the iterations

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = x_n + \lambda_n\big((\mathrm{Id}+\gamma_n A_{\mathrm{i}(n)})^{-1}x_n + e_n - x_n\big), \tag{1.3}$$

where $\mathrm{i}\colon \mathbb{N} \to I$ sweeps through the indices with some regularity. It will be convenient to cast this algorithm in the following more general framework.

*Received by the editors May 7, 2003; accepted for publication (in revised form) December 1, 2003; published electronically August 4, 2004.
http://www.siam.org/journals/sicon/43-2/42733.html

†Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie – Paris 6, 75005 Paris, France (plc@math.jussieu.fr).

‡Department of Management Science, Helsinki School of Economics, 00101 Helsinki, Finland (pennanen@hkkk.fi).

ALGORITHM 1.1. *Let $(A_i)_{i\in I}$ be a countable family of set-valued operators from $\mathcal{H}$ to $\mathcal{H}$, let $(\gamma_n)_{n\in\mathbb{N}}$ and $(\lambda_n)_{n\in\mathbb{N}}$ be sequences in $]0,+\infty[$, let $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ be sequences in $\mathcal{H}$, let* i *be a mapping from $\mathbb{N}$ to I, and let x_0 be a point in $\mathcal{H}$. A sequence $(x_n)_{n\in\mathbb{N}}$ is constructed according to the updating rule*

$$(\forall n\in\mathbb{N})\quad x_{n+1}=x_n+\lambda_n(x_{n+\frac{1}{2}}+u_n-x_n), \tag{1.4}$$

where $x_{n+\frac{1}{2}}$ is a solution to the inclusion

$$v_n\in x_{n+\frac{1}{2}}-x_n+\gamma_n A_{\mathrm{i}(n)}x_{n+\frac{1}{2}}. \tag{1.5}$$

In the case of maximal monotone operators, the weak convergence properties of Algorithm 1.1 are summarized in the next theorem, which is derived from a result of [7]. This theorem captures the weak convergence results of [2, 11, 17, 22] for the proximal point algorithm, as well as standard results on the weak convergence of sequential projection methods for convex feasibility problems, such as those of [5, 6, 13], when the operators are taken to be normal cones to closed convex sets.

THEOREM 1.2. *Suppose that in Algorithm* 1.1 *the following conditions are satisfied:*

(i) (a) *For every $i\in I$, A_i is maximal monotone;*
(b) $S=\bigcap_{i\in I}A_i^{-1}(0)\neq\varnothing$.
(ii) $(\forall i\in I)(\exists\, M_i\in\mathbb{N}\smallsetminus\{0\})(\forall n\in\mathbb{N})\ \ i\in\{\mathrm{i}(n),\dots,\mathrm{i}(n+M_i-1)\}$.
(iii) $\inf_{n\in\mathbb{N}}\gamma_n>0$.
(iv) $(\exists\,\varepsilon\in\,]0,1[)(\forall n\in\mathbb{N})\ \ \varepsilon\leq\lambda_n\leq 2-\varepsilon$.
(v) $\sum_{n\in\mathbb{N}}\|u_n\|<+\infty$ *and* $\sum_{n\in\mathbb{N}}\|v_n\|<+\infty$.

Then every orbit generated by Algorithm 1.1 *converges weakly to a point in S.*

Proof. For every $n\in\mathbb{N}$, set

$$e_n=u_n+(\mathrm{Id}+\gamma_n A_{\mathrm{i}(n)})^{-1}(x_n+v_n)-(\mathrm{Id}+\gamma_n A_{\mathrm{i}(n)})^{-1}x_n. \tag{1.6}$$

Then (1.4)–(1.5) coincides with (1.3), which is itself a special case of [7, Algorithm 6.7] (obtained by taking $I^{(1)}=I^{(2)}=\varnothing$ and $(I_n)_{n\in\mathbb{N}}=(\{\mathrm{i}(n)\})_{n\in\mathbb{N}}$ there). On the other hand, since the resolvents $\big((\mathrm{Id}+\gamma_n A_{\mathrm{i}(n)})^{-1}\big)_{n\in\mathbb{N}}$ are nonexpansive [1, Proposition 3.5.3], we obtain

$$(\forall n\in\mathbb{N})\quad \|e_n\|\leq\|u_n\|+\|v_n\|. \tag{1.7}$$

Hence, (v) implies that $\sum_{n\in\mathbb{N}}\|e_n\|<+\infty$ and the claim therefore follows at once from [7, Theorem 6.9(i)]. □

Remark 1.3. The sequences $(v_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ model errors at various steps of the execution of the iterations, thereby allowing for some tolerance in the numerical implementation of the algorithm. It is clear from the above proof that, in the presence of monotone operators, the errors $(v_n)_{n\in\mathbb{N}}$ can easily be absorbed in the errors $(u_n)_{n\in\mathbb{N}}$ and are, in this sense, redundant. However, since our ultimate goal is to investigate the behavior of Algorithm 1.1 with nonmonotone operators, the use of two error sequences is required to obtain a more flexible algorithmic model. An illustration of how condition (v) can be checked in practice is provided in section 4 (see Remark 4.2).

Extensions of the basic proximal iterations (1.1) have also been investigated in another direction, namely, by relaxing the monotonicity requirements on A. The motivation for this line of work stems from the fact that proximal iterations have been observed to converge to zeros of nonmonotone operators in certain numerical

experiments, e.g., [12]. Attempts to explain this behavior in the case of general variational inclusions can be traced back to [26], where a convergence proof is given which does not assume monotonicity. However, the assumptions made in that early work are rather stringent as they impose, essentially, that the inverse of the operator be differentiable at the origin with a monotone derivative.

Relaxing the monotonicity property of an operator is equivalent to relaxing the monotonicity property of its inverse. In some applications, however, it is more natural to work directly with the inverse. For instance, since multiplier methods are based on applying the proximal algorithm to a dual formulation of the original problem, it is more pertinent to impose relaxed monotonicity conditions on the inverse of the operator. This observation was the starting point of the investigation proposed in [20], where local convergence is analyzed under the condition that the mapping be cohypomonotone, i.e., that its inverse be hypomonotone (see Definition 2.2). The analysis of [20] is incomplete, however, at least in the sense that it assumes that the proximal steps can be computed exactly. This is an unrealistic assumption in most practical applications. In [14], an effort was made to remove this assumption by investigating the convergence in the case of inexact computations under a so-called relative error criterion. The analysis of [14] requires that the values of the operator outside a certain neighborhood be discarded. However, since this neighborhood is usually unknown in concrete applications, the applicability of this conceptual analysis is limited.

The goal of this paper is to unify and extend various convergence results on proximal iterations, by investigating the asymptotic behavior of Algorithm 1.1 when applied to a family of cohypomonotone operators. Such operators are discussed in section 2. Our main result is presented in section 3, where local viability and weak convergence conditions are established for Algorithm 1.1. An application to nonlinear programming is presented in section 4, where local convergence of a relaxed inexact proximal method of multipliers is proven for a nonconvex problem.

Throughout, $B(x;\eta)$ denotes the closed ball of center $x \in \mathcal{H}$ and extended radius $\eta \in\ ]0,+\infty]$; d_C the distance function to a nonempty set $C \subset \mathcal{H}$; P_C the projection operator onto a nonempty closed convex set $C \subset \mathcal{H}$; and N_C its normal cone map. $\operatorname{Fix} T$ the set of fixed points of an operator T, $\operatorname{dom} T$ its domain, $\operatorname{ran} T$ its range, and $\operatorname{gph} T$ its graph. The complement of a set C is denoted by $\complement C$.

2. Cohypomonotone operators. Our goal is to prove the local convergence of Algorithm 1.1 under a relaxed monotonicity assumption on the operators $(A_i)_{i\in I}$ that we now define.

DEFINITION 2.1. *Let U be a subset of $\mathcal{H}^2$. The U-localization of an operator $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is the operator denoted by $A|^U\colon \mathcal{H} \rightrightarrows \mathcal{H}$ whose graph is* $\operatorname{gph}(A|^U) = U \cap \operatorname{gph} A$.

DEFINITION 2.2. *Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$, $\rho \in [0,+\infty[$, and $U \subset \mathcal{H}^2$. Then A is* [maximal] ρ-hypomonotone *on U if there exists an operator $\bar{A}\colon \mathcal{H} \rightrightarrows \mathcal{H}$ such that $\bar{A} + \rho\,\mathrm{Id}$ is [maximal] monotone and $\bar{A}|^U = A|^U$. The operator A is* [maximal] ρ-cohypomonotone *on U if A^{-1} is [maximal] ρ-hypomonotone on U.*

The above definition of ρ-hypomonotonicity on a set is related to the pointwise notion of hypomonotonicity of [25, Example 12.28] and [8] as follows: If W is a neighborhood of $x \in \mathcal{H}$ and there exists $\rho \in [0,+\infty[$ such that A is ρ-hypomonotone on $W \times \mathcal{H}$, then A is hypomonotone at x in the sense of [8, 25].

Maximal hypomonotonicity has been studied extensively in the variational analysis literature. Thus, classes of functions with hypomonotone subdifferentials have

been investigated in various settings [8, 24, 25, 27]. Interesting connections between hypomonotonicity and Aubin continuity, Lipschitz continuity, and strict graphical derivatives have also been found [15, 16]. On the other hand, maximal hypomonotonicity is a less stringent requirement than imposing the existence of Lipschitz localizations. The latter has been studied in the context of variational inequality and nonlinear programming problems, e.g., [9, 10, 15, 16]. For completeness, we provide a simple proof of this important fact.

LEMMA 2.3. *Suppose that $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ has a Lipschitz localization at a point $(x, y) \in \operatorname{gph} A$; that is, there exist open sets $X \ni x$ and $Y \ni y$ such that the mapping $z \mapsto A(z) \cap Y$ is single-valued and ρ-Lipschitz continuous on X. Then A is maximal ρ-hypomonotone on $X \times Y$.*

Proof. Set $\widetilde{A} = A|^{X\times Y} + \rho\,\mathrm{Id}$ and take $(u, v) \in X^2$. Then, by Cauchy–Schwarz,

$$\Big\langle u - v\,,\, \widetilde{A}u - \widetilde{A}v \Big\rangle \geq \|u - v\| \big(\rho\|u - v\| - \big\|A|^{X\times Y}(u) - A|^{X\times Y}(v)\big\|\big) \geq 0. \tag{2.1}$$

Hence, $\widetilde{A}$ is monotone. Let A' be a maximal monotone extension of $\widetilde{A}$ and set $\bar{A} = A' - \rho\,\mathrm{Id}$. Then $\bar{A} + \rho\,\mathrm{Id}$ is maximal monotone and, to complete the proof, it suffices to show that $\bar{A}|^{X\times Y} = A|^{X\times Y}$. By construction, $\operatorname{gph}(A|^{X\times Y}) \subset \operatorname{gph}(\bar{A}|^{X\times Y})$. Conversely, take $(\bar{x}, \bar{y}) \in \operatorname{gph} \bar{A}|^{X\times Y}$ and let $z = \bar{y} - A|^{X\times Y}(\bar{x})$. Then $(\bar{x}, \bar{y} + \rho\bar{x}) \in \operatorname{gph} A'$ and, since X is open, we have $\bar{x} + \varepsilon z \in X$ for $\varepsilon > 0$ sufficiently small. Since $\operatorname{gph}(A|^{X\times Y} + \rho\,\mathrm{Id}) \subset \operatorname{gph} A'$ and A' is monotone, we have

$$\begin{aligned} 0 &\leq \big\langle \bar{x} + \varepsilon z - \bar{x}\,,\, A|^{X\times Y}(\bar{x} + \varepsilon z) + \rho(\bar{x} + \varepsilon z) - (\bar{y} + \rho\bar{x}) \big\rangle \\ &= \big\langle \varepsilon z\,,\, A|^{X\times Y}(\bar{x} + \varepsilon z) + \rho\varepsilon z - \bar{y} \big\rangle. \end{aligned} \tag{2.2}$$

Dividing by ε and letting $\varepsilon \downarrow 0^+$, the continuity of $A|^{X\times Y}$ gives $0 \leq -\big\|A|^{X\times Y}(\bar{x}) - \bar{y}\big\|^2$, whence $(\bar{x}, \bar{y}) \in \operatorname{gph} A|^{X\times Y}$. ∎

The relevance of cohypomonotonicity in proximal methods hinges on the following identity.

LEMMA 2.4. *Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ and let $(\gamma, \rho) \in \mathbb{R}^2$, where $\gamma \neq 0$. Then*

$$\mathrm{Id} + \left(1 - \frac{\rho}{\gamma}\right)\Big(\big(\mathrm{Id} + \gamma A\big)^{-1} - \mathrm{Id}\Big) = \Big(\mathrm{Id} + (\gamma - \rho)\big(A^{-1} + \rho\,\mathrm{Id}\,\big)^{-1}\Big)^{-1}. \tag{2.3}$$

Proof. If $\gamma = \rho$, the identity is clear. Otherwise, take $(x, u) \in \mathcal{H}^2$. Then

$$\begin{aligned} u \in \Big(\mathrm{Id} + (\gamma - \rho)\big(A^{-1} + \rho\,\mathrm{Id}\,\big)^{-1}\Big)^{-1} x \\ \Leftrightarrow \frac{x - u}{\gamma - \rho} \in \big(A^{-1} + \rho\,\mathrm{Id}\,\big)^{-1} u \\ \Leftrightarrow u \in A^{-1}\left(\frac{x - u}{\gamma - \rho}\right) + \frac{\rho}{\gamma - \rho}(x - u) \\ \Leftrightarrow \frac{x - u}{\gamma - \rho} \in A\left(\frac{\gamma u - \rho x}{\gamma - \rho}\right) \\ \Leftrightarrow x \in \big(\mathrm{Id} + \gamma A\big)\left(\frac{\gamma u - \rho x}{\gamma - \rho}\right) \\ \Leftrightarrow u \in \left(\mathrm{Id} + \left(1 - \frac{\rho}{\gamma}\right)\Big(\big(\mathrm{Id} + \gamma A\big)^{-1} - \mathrm{Id}\Big)\right) x. \end{aligned} \tag{2.4}$$

∎

The above lemma states that relaxing a proximal step for the original operator A amounts to computing a proximal step for the operator $(A^{-1}+\rho\,\mathrm{Id})^{-1}$. Clearly, when A is cohypomonotone, the latter behaves locally like a monotone operator. This observation will play a central role in our convergence analysis.

3. Proximal iterations with cohypomonotone operators. In this section, we establish our main convergence result for the inexact relaxed proximal point Algorithm 1.1 with cohypomonotone operators.

THEOREM 3.1. *Suppose that in Algorithm* 1.1 *the following conditions are satisfied:*

(i) *There exist a number* $\delta\in\,]0,+\infty[$, *a sequence* $(\rho_i)_{i\in I}$ *in* $[0,+\infty[$, *and open sets* V *and* $(X_i)_{i\in I}$ *in* $\mathcal{H}$ *for which the following hold:*
  (a) $0\in V$;
  (b) *for every* $i\in I$, $S_i=X_i\cap A_i^{-1}(0)$ *is closed and* $S_i+B(0;\delta)\subset X_i$;
  (c) *for every* $i\in I$, A_i *is maximal* ρ_i*-cohypomonotone on* $V\times X_i$;
  (d) $S=\bigcap_{i\in I}S_i\neq\varnothing$.

(ii) $(\forall i\in I)(\exists\, M_i\in\mathbb{N}\smallsetminus\{0\})(\forall n\in\mathbb{N})\ \ i\in\{\mathrm{i}(n),\dots,\mathrm{i}(n+M_i-1)\}$.

(iii) $\inf_{n\in\mathbb{N}}(\gamma_n-\rho_{\mathrm{i}(n)})>0$.

(iv) $(\exists\,\varepsilon\in\,]0,1[)(\forall n\in\mathbb{N})\ \ \varepsilon\leq\frac{\lambda_n}{1-\rho_{\mathrm{i}(n)}/\gamma_n}\leq 2-\varepsilon$.

Then there exists a closed ball B *of radius* $\eta\in\,]0,+\infty]$ *centered at a point in* S *such that if the following conditions hold:*

(v) x_0 *is sufficiently close to* S, *say,*

$$d_S(x_0)<\nu=\frac{4-2\varepsilon}{5-2\varepsilon}\eta; \tag{3.1}$$

(vi) $\sum_{n\in\mathbb{N}}(\|u_n\|+2\|v_n\|)<\frac{\nu-d_S(x_0)}{2-\varepsilon}$;

(vii) *for every* $n\in\mathbb{N}$, *one selects* $x_{n+\frac{1}{2}}\in B$ *in* (1.5),

then there is one and only one orbit $(x_n)_{n\in\mathbb{N}}$ *of Algorithm* 1.1 *contained in* B *and, furthermore,* $(x_n)_{n\in\mathbb{N}}$ *converges weakly to a point in* $B\cap S$.

Proof. Take $i\in I$. By (i)(c), there exists an operator $\bar{A}_i\colon\mathcal{H}\rightrightarrows\mathcal{H}$ such that

$$(X_i\times V)\cap\operatorname{gph}\bar{A}_i=(X_i\times V)\cap\operatorname{gph}A_i \tag{3.2}$$

and $\bar{A}_i^{-1}+\rho_i\,\mathrm{Id}$ is maximal monotone. Consequently, it follows from (i)(a), (i)(b), and (i)(d) that

$$X_i\cap\bar{A}_i^{-1}(0)=X_i\cap A_i^{-1}(0)\neq\varnothing \tag{3.3}$$

is closed. Therefore, by maximal monotonicity, $\bar{A}_i^{-1}(0)=(\bar{A}_i^{-1}+\rho_i\,\mathrm{Id})(0)$ is closed and convex [1, Proposition 3.5.6]. Thus, the convex set $\bar{A}_i^{-1}(0)$ is the union of the two disjoint closed sets $X_i\cap\bar{A}_i^{-1}(0)\neq\varnothing$ and $(\complement X_i)\cap\bar{A}_i^{-1}(0)$, which forces the latter to be empty. Indeed, otherwise, by convexity of $\bar{A}_i^{-1}(0)$, we could find $a\in X_i\cap\bar{A}_i^{-1}(0)$ and $b\in(\complement X_i)\cap\bar{A}_i^{-1}(0)$ such that the closed segment $[a,b]$ is the union of the two disjoint closed sets $[a,b]\cap X_i\cap\bar{A}_i^{-1}(0)$ and $[a,b]\cap(\complement X_i)\cap\bar{A}_i^{-1}(0)$, which is impossible since $[a,b]$ is connected by [3, Théorème IV.2.5.4]. To sum up,

$$S_i=X_i\cap A_i^{-1}(0)=X_i\cap\bar{A}_i^{-1}(0)=\bar{A}_i^{-1}(0) \tag{3.4}$$

is closed and convex. It therefore follows from (i)(d) that the projection P_Sx_0 of x_0 onto S is well defined. On the other hand, (i)(a) and (i)(b) yield $0\in\operatorname{int}V$ and

$P_S x_0 \in \operatorname{int} \bigcap_{i\in I} X_i$, respectively. As a result, we can find $\eta \in \left]0, +\infty\right]$ such that

$$B(P_S x_0; \eta) \subset \bigcap_{i\in I} X_i \quad \text{and} \quad B\bigg(0; 2\eta/\inf_{n\in\mathbb{N}} \gamma_n\bigg) \subset V. \tag{3.5}$$

We now set

$$B = B(P_S x_0; \eta) \quad \text{and} \quad D = B(P_S x_0; \nu) \tag{3.6}$$

and observe that (3.1) forces

$$x_0 \in \operatorname{int} D. \tag{3.7}$$

Next, take $(x, u) \in B^2$ and $n \in \mathbb{N}$. Then it follows from (3.5) that $\big(u, (x-u)/\gamma_n\big) \in X_{\mathrm{i}(n)} \times V$. Consequently, by (3.2),

$$\begin{aligned} u \in \big(\operatorname{Id} + \gamma_n A_{\mathrm{i}(n)}\big)^{-1} x &\Leftrightarrow \bigg(u, \frac{x-u}{\gamma_n}\bigg) \in \operatorname{gph} A_{\mathrm{i}(n)} \\ &\Leftrightarrow \bigg(u, \frac{x-u}{\gamma_n}\bigg) \in \operatorname{gph} \bar{A}_{\mathrm{i}(n)} \\ &\Leftrightarrow u \in \big(\operatorname{Id} + \gamma_n \bar{A}_{\mathrm{i}(n)}\big)^{-1} x. \end{aligned} \tag{3.8}$$

Thus,

$$\big(\operatorname{Id} + \gamma_n A_{\mathrm{i}(n)}\big)^{-1}\big|^{B\times B} = \big(\operatorname{Id} + \gamma_n \bar{A}_{\mathrm{i}(n)}\big)^{-1}\big|^{B\times B}. \tag{3.9}$$

On the other hand, let

$$T_n = \operatorname{Id} + \bigg(1 - \frac{\rho_{\mathrm{i}(n)}}{\gamma_n}\bigg)\Big(\big(\operatorname{Id} + \gamma_n \bar{A}_{\mathrm{i}(n)}\big)^{-1} - \operatorname{Id}\Big). \tag{3.10}$$

Alternatively, using Lemma 2.4, we can write

$$T_n = \Big(\operatorname{Id} + \tau_n\big(\bar{A}^{-1}_{\mathrm{i}(n)} + \rho_{\mathrm{i}(n)} \operatorname{Id}\big)^{-1}\Big)^{-1}, \quad \text{where} \quad \tau_n = \gamma_n - \rho_{\mathrm{i}(n)}. \tag{3.11}$$

Since $\tau_n > 0$ by (iii), T_n is therefore the resolvent of the operator $\tau_n C_{\mathrm{i}(n)}$, where

$$C_{\mathrm{i}(n)} = \big(\bar{A}^{-1}_{\mathrm{i}(n)} + \rho_{\mathrm{i}(n)} \operatorname{Id}\big)^{-1}, \tag{3.12}$$

which is maximal monotone as the inverse of such an operator. Hence, it follows from [2, Proposition 2.3] that

$$T_n\colon \operatorname{dom} T_n = \mathcal{H} \to \mathcal{H} \quad \text{and} \quad (\forall (x, z) \in \mathcal{H} \times \operatorname{Fix} T_n) \;\; \langle z - T_n x \,, x - T_n x\rangle \leq 0, \tag{3.13}$$

which, by [7, Proposition 2.3(ii)], implies

$$\begin{aligned} &(\forall \mu \in [0, 2])(\forall (x, z) \in \mathcal{H} \times \operatorname{Fix} T_n) \\ &\qquad \|x + \mu(T_n x - x) - z\|^2 \leq \|x - z\|^2 - \mu(2-\mu)\|T_n x - x\|^2. \end{aligned} \tag{3.14}$$

Now, let

$$\mu_n = \frac{\lambda_n}{1 - \rho_{\mathrm{i}(n)}/\gamma_n}. \tag{3.15}$$

Then we get from (iv) that

$$\mu_n \in [\varepsilon, 2 - \varepsilon]. \tag{3.16}$$

We also obtain from (3.5), (3.3), (3.12), and (3.11) that

$$B \cap S_{\mathrm{i}(n)} = B \cap A_{\mathrm{i}(n)}^{-1}(0) = B \cap \bar{A}_{\mathrm{i}(n)}^{-1}(0) = B \cap C_{\mathrm{i}(n)}^{-1}(0) = B \cap \operatorname{Fix} T_n. \tag{3.17}$$

Hence, $P_S x_0 \in B \cap S \subset B \cap \operatorname{Fix} T_n$, and it results from (3.14) with $\mu = 1$ and $z = P_S x_0$ that

$$T_n(B) \subset B. \tag{3.18}$$

Let us now show that Algorithm 1.1 is viable, i.e., that the recursion (1.4)–(1.5) does generate an infinite sequence. To this end, we shall show that the sequence $(x_n)_{n\in\mathbb{N}}$ is well defined and that it lies in $\operatorname{int} D$, whereas the sequence $(x_n + v_n)_{n\in\mathbb{N}}$ lies in $\operatorname{int} B$. Since (vi) yields $\|v_0\| < \eta - d_S(x_0)$, it follows from (3.7) that $\|x_0 + v_0 - P_S x_0\| \leq \|v_0\| + d_S(x_0) < \eta$, whence $x_0 + v_0 \in \operatorname{int} B$. Now assume that, for some $n \in \mathbb{N}$, the points $(x_k)_{0\leq k\leq n}$ and $(x_k + v_k)_{0\leq k\leq n}$ lie in $\operatorname{int} D$ and $\operatorname{int} B$, respectively. Then it results from (vii) and (3.9) that (1.5) can be written as

$$x_{n+\frac{1}{2}} \in (\mathrm{Id} + \gamma_n A_{\mathrm{i}(n)})^{-1}\big|^{B\times B}(x_n + v_n) = (\mathrm{Id} + \gamma_n \bar{A}_{\mathrm{i}(n)})^{-1}\big|^{B\times B}(x_n + v_n). \tag{3.19}$$

In view of (3.15), (1.4) can now be written as

$$x_{n+1} \in x_n + \mu_n \left(1 - \frac{\rho_{\mathrm{i}(n)}}{\gamma_n}\right) \left((\mathrm{Id} + \gamma_n \bar{A}_{\mathrm{i}(n)})^{-1}\big|^{B\times B}(x_n + v_n) + u_n - x_n\right), \tag{3.20}$$

which, by virtue of (3.10), yields

$$x_{n+1} \in x_n + \mu_n \left(T_n\big|^{B\times B}(x_n + v_n) + w_n - x_n\right), \tag{3.21}$$

where

$$w_n = \left(1 - \frac{\rho_{\mathrm{i}(n)}}{\gamma_n}\right) u_n - \frac{\rho_{\mathrm{i}(n)}}{\gamma_n} v_n. \tag{3.22}$$

However, since $x_n + v_n \in B$, (3.18) yields $T_n\big|^{B\times B}(x_n + v_n) = T_n(x_n + v_n)$. Hence, since T_n is single-valued and defined everywhere on $\mathcal{H}$ (see (3.13)), we deduce from (3.21) that x_{n+1} is uniquely defined by

$$x_{n+1} = x_n + \mu_n\big(T_n(x_n + v_n) + w_n - x_n\big). \tag{3.23}$$

Now put

$$e_n = w_n + T_n(x_n + v_n) - T_n x_n. \tag{3.24}$$

Then we derive from (3.23) that

$$x_{n+1} = x_n + \mu_n(T_n x_n - x_n) + \mu_n e_n, \tag{3.25}$$

and it follows from (3.16) and (3.14) that

$$\begin{aligned} \|x_{n+1} - P_S x_0\| &\leq \|x_n - P_S x_0 + \mu_n(T_n x_n - x_n)\| + \mu_n\|e_n\| \\ &\leq \|x_n - P_S x_0\| + \mu_n\|e_n\|. \end{aligned} \tag{3.26}$$

Consequently, since $(x_k + v_k)_{0\leq k\leq n}$ lies in B, we have

$$\begin{aligned} \|x_{n+1} - P_S x_0\| &\leq \|x_0 - P_S x_0\| + \sum_{k=0}^{n} \mu_k\|e_k\| \\ &\leq d_S(x_0) + (2-\varepsilon)\sum_{k\in\mathbb{N}} \|e_k\|. \end{aligned} \tag{3.27}$$

On the other hand, it follows from (3.24), the nonexpansivity of T_n [1, Proposition 3.5.3], (3.22), and (iii) that

$$\|e_n\| \leq \|w_n\| + \|T_n(x_n + v_n) - T_n x_n\| \leq \|w_n\| + \|v_n\| \leq \|u_n\| + 2\|v_n\|. \tag{3.28}$$

Therefore, we derive from (vi) that

$$d_S(x_0) + (2-\varepsilon)\sum_{k\in\mathbb{N}} \|e_k\| < \nu, \tag{3.29}$$

and deduce from (3.27) that $\|x_{n+1} - P_S x_0\| < \nu$, i.e., $x_{n+1} \in \operatorname{int} D$. In turn, (vi) and (3.1) yield

$$\|x_{n+1} + v_{n+1} - P_S x_0\| < \nu + \|v_{n+1}\| < \nu + \frac{\nu}{2(2-\varepsilon)} = \eta, \tag{3.30}$$

i.e., $x_{n+1} + v_{n+1} \in \operatorname{int} B$. We have thus shown by induction that the entire sequence $(x_n + v_n)_{n\in\mathbb{N}}$ lies in B and that $(x_n)_{n\in\mathbb{N}}$ is a well-defined sequence which lies entirely in $\operatorname{int} D \subset \operatorname{int} B$. In view of (3.23), (3.11), and (3.12), the recursion governing the sequence $(x_n)_{n\in\mathbb{N}}$ can now be rewritten as

$$(\forall n \in \mathbb{N})\ \ x_{n+1} = x_n + \mu_n(x_{n+\frac{1}{2}} + w_n - x_n), \tag{3.31}$$

where $x_{n+\frac{1}{2}}$ is the unique solution to the inclusion

$$v_n \in x_{n+\frac{1}{2}} - x_n + \tau_n C_{\mathrm{i}(n)} x_{n+\frac{1}{2}}, \tag{3.32}$$

namely, $x_{n+\frac{1}{2}} = \big(\operatorname{Id} + \tau_n C_{\mathrm{i}(n)}\big)^{-1}(x_n + v_n) = T_n(x_n + v_n) \in B$, where the last inclusion follows from $x_n + v_n \in B$ and (3.18). In summary, since the operators $(C_i)_{i\in I}$ are maximal monotone, $\inf_{n\in\mathbb{N}} \tau_n > 0$, $(\mu_n)_{n\in\mathbb{N}}$ lies in $[\varepsilon, 2-\varepsilon]$, $\sum_{n\in\mathbb{N}} \|w_n\| \leq \sum_{n\in\mathbb{N}} \|u_n\| + \sum_{n\in\mathbb{N}} \|v_n\| < +\infty$, and $\sum_{n\in\mathbb{N}} \|v_n\| < +\infty$, it follows from Theorem 1.2 that $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point x in $\bigcap_{i\in I} C_i^{-1}(0)$. On the other hand, since $(x_n)_{n\in\mathbb{N}}$ lies in the weakly closed set B, $x \in B$. As a result, (3.17) yields $x \in B \cap \bigcap_{i\in I} C_i^{-1}(0) = B \cap S$. □

Remark 3.2. The above result unifies and extends several results found in the literature.

- If $I = \{1\}$ (a single operator is considered), $\lambda_n \equiv 1$, and $u_n \equiv 0 \equiv v_n$, then Theorem 3.1 is found in [20, Theorem 9].
- If $I = \{1\}$, $\rho_1 = 0$, and $V = \mathcal{H} = X_1$, Theorem 3.1 reduces to [11, Theorem 3], and to [22, Theorem 1] if we further assume $\lambda_n \equiv 1$.
- If $\rho_i \equiv 0$ and $V = \mathcal{H} \equiv X_i$, Theorem 3.1 reduces to Theorem 1.2.
- If $\rho_i \equiv 0$, $V = \mathcal{H} \equiv X_i$, and $u_n \equiv 0 \equiv v_n$, Theorem 3.1 corresponds to [2, Corollary 6.1(i)].
- If $I = \{1, \ldots, m\}$, $(S_i)_{i\in I}$ is a family of closed convex sets in $\mathcal{H}$ with associated projection operators $(P_i)_{i\in I}$, $\mathrm{i}\colon n \mapsto n$ modulo $m+1$, for every $i \in I$, $A_i = N_{S_i}$ (hence $\rho_i \equiv 0$ and $V = \mathcal{H} \equiv X_i$), and $u_n \equiv 0 \equiv v_n$, then Algorithm 1.1 produces the method of cyclic projections

(3.33)
$$(\forall n \in \mathbb{N})\ \ x_{n+1} = x_n + \lambda_n(P_{n \text{ modulo } m+1}x_n - x_n), \quad \text{where} \quad \varepsilon \leq \lambda_n \leq 2 - \varepsilon,$$

and Theorem 3.1 reduces to [13, Theorem 1].

4. Nonlinear programming application. Using the same arguments as in [20, section 5], one can derive multiplier methods for quite general variational inclusions by combining Theorem 3.1 with an abstract duality framework for set-valued mappings. Instead of going through all the steps and applications discussed in [20], we analyze the proximal method of multipliers for nonlinear (nonconvex) programming as an example. The proximal method of multipliers was introduced and analyzed in the convex case by Rockafellar [23] and in the nonconvex case in [20, section 7] with exact, unrelaxed iterates.

Consider the nonlinear programming problem

$$\text{minimize} \quad f_0(x) \quad \text{subject to} \quad \begin{cases} f_i(x) = 0 & \text{for } 1 \leq i \leq r, \\ f_i(x) \leq 0 & \text{for } r+1 \leq i \leq m, \end{cases} \tag{4.1}$$

where $(f_i)_{0\leq i\leq m}$ are real-valued C^2-functions defined on the standard Euclidean space $\mathbb{R}^N$. Our aim is to find Karush–Kuhn–Tucker (KKT) points for (4.1). To this end, we introduce the closed convex cone $K = \{0\}^r \times \mathbb{R}_-^{m-r}$, let $F\colon x \mapsto \big(f_1(x), \ldots, f_m(x)\big)$, and set $\mathcal{H} = \mathbb{R}^N \times \mathbb{R}^m$. We shall derive from Theorem 3.1 a local convergence result for the following proximal method of multipliers.

ALGORITHM 4.1. *Let $(x_0, y_0) \in \mathcal{H}$, let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in $]0, +\infty[$, and let $(w_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^N$. A sequence $\big((x_n, y_n)\big)_{n\in\mathbb{N}}$ is constructed according to the updating rule*

$$(\forall n \in \mathbb{N})\ \ (x_{n+1}, y_{n+1}) = (x_n, y_n) + \lambda_n\big((x_{n+\frac{1}{2}}, y_{n+\frac{1}{2}}) - (x_n, y_n)\big), \tag{4.2}$$

where $x_{n+\frac{1}{2}}$ minimizes approximately the function

$$\varphi_n\colon x \mapsto f_0(x) + \frac{1}{2\gamma_n}\|x - x_n\|^2 + \frac{1}{2\gamma_n}d_K\big(y_n + \gamma_n F(x)\big)^2 \tag{4.3}$$

in the sense that $\nabla\varphi_n(x_{n+\frac{1}{2}}) = w_n$, and

$$\begin{cases} y^i_{n+\frac{1}{2}} = y^i_n + \gamma_n f_i(x_{n+\frac{1}{2}}) & \text{for } 1 \leq i \leq r, \\ y^i_{n+\frac{1}{2}} = \max\big\{y^i_n + \gamma_n f_i(x_{n+\frac{1}{2}}), 0\big\} & \text{for } r+1 \leq i \leq m. \end{cases} \tag{4.4}$$

Remark 4.2. It was shown in the proof of [20, Theorem 19] that, under condition (i) of Theorem 4.3 and for x near $\bar{x}$, the condition $\nabla\varphi_n(x) = 0$ implies that x is a local minimizer of φ_n. For this reason, the condition $\nabla\varphi_n(x_{n+\frac{1}{2}}) = w_n$ is interpreted in Algorithm 4.1 as an approximate minimization. In practice, the parameter γ_n is often chosen adaptively while the size of the vector $\|w_n\|$ can be made arbitrarily small by choosing the stopping criterion appropriately in the minimization of φ_n in (4.3).

Define a mapping $L\colon \mathcal{H} \rightrightarrows \mathcal{H}$ by

$$L\colon (x,y) \mapsto \big(\nabla f_0(x) + \langle y\,,\nabla F(x)\rangle\,, -F(x) + N_{K^*}(y)\big), \tag{4.5}$$

where $K^* = \mathbb{R}^r \times \mathbb{R}^{m-r}_+$ is the polar cone of K. Then the KKT system for (4.1) can be written as [25, Example 11.46]

$$(0,0) \in L(x,y). \tag{4.6}$$

Let $(\bar{x},\bar{y}) \in \mathcal{H}$ be a point satisfying the KKT conditions for (4.1) and define

$$I^+ = \{1,\ldots,r\} \cup \{r+1 \leq i \leq m \mid f_i(\bar{x}) = 0 \ \text{ and } \ \bar{y}_i > 0\}. \tag{4.7}$$

Now let $l\colon (x,y) \mapsto f_0(x) + \langle y\,,F(x)\rangle$. Recall that $(\bar{x},\bar{y})$ is said to satisfy the *strong second order sufficient condition* for (4.1) if [21]

$$(\forall y \in \mathbb{R}^m)\ \begin{cases} y \neq 0 \\ (\forall i \in I^+)\ \ \langle y\,,\nabla f_i(\bar{x})\rangle = 0 \end{cases} \quad\Rightarrow\quad \big\langle y\,,\nabla^2_{xx} l(\bar{x},\bar{y})y\big\rangle > 0. \tag{4.8}$$

THEOREM 4.3. *Suppose that in Algorithm* 4.1 *the following conditions are satisfied:*

(i) $(\bar{x},\bar{y}) \in \mathcal{H}$ *is a KKT point for* (4.1) *satisfying* (4.8) *and such that the gradients* $\big(\nabla f_i(\bar{x})\big)_{i\in I^+}$ *are linearly independent;*
(ii) $\inf_{n\in\mathbb{N}}\gamma_n$ *is large enough;*
(iii) $(\exists\,\varepsilon \in\,]0,1[)(\forall n \in \mathbb{N})\ \ \varepsilon \leq \lambda_n \leq 2-\varepsilon$.

Then there exists a closed ball B *centered at* $(\bar{x},\bar{y})$ *such that if the following conditions hold:*

(iv) (x_0,y_0) *is sufficiently close to* $(\bar{x},\bar{y})$;
(v) $\sum_{n\in\mathbb{N}}\gamma_n\|w_n\|$ *is small enough;*
(vi) *for every* $n \in \mathbb{N}$, $(x_{n+\frac{1}{2}},y_{n+\frac{1}{2}}) \in B$,

then there is one and only one orbit $\big((x_n,y_n)\big)_{n\in\mathbb{N}}$ *of Algorithm* 4.1 *contained in* B *and, furthermore,* $\big((x_n,y_n)\big)_{n\in\mathbb{N}}$ *converges to* $(\bar{x},\bar{y})$.

Proof. By [18, Exemple 4.b and Proposition 7.d],

$$\nabla d^2_K = 2(\mathrm{Id} - P_K) = 2P_{K^*}. \tag{4.9}$$

Thus, it follows from (4.3) that

$$\nabla\varphi_n(x) = \nabla f_0(x) + \gamma_n^{-1}(x-x_n) + \big\langle P_{K^*}\big(y_n + \gamma_n F(x)\big)\,,\nabla F(x)\big\rangle\,. \tag{4.10}$$

Note also that $y_{n+\frac{1}{2}}$ in (4.4) can be expressed as

$$y_{n+\frac{1}{2}} = P_{K^*}\big(y_n + \gamma_n F(x_{n+\frac{1}{2}})\big) = (I + N_{K^*})^{-1}\big(y_n + \gamma_n F(x_{n+\frac{1}{2}})\big). \tag{4.11}$$

The update rules for $(x_{n+\frac{1}{2}}, y_{n+\frac{1}{2}})$ are thus equivalent to the system

$$\begin{cases} w_n = \nabla f_0(x_{n+\frac{1}{2}}) + \gamma_n^{-1}(x_{n+\frac{1}{2}} - x_n) + \left\langle y_{n+\frac{1}{2}}\,, \nabla F(x_{n+\frac{1}{2}}) \right\rangle, \\ y_n + \gamma_n F(x_{n+\frac{1}{2}}) \in y_{n+\frac{1}{2}} + N_{K^*}(y_{n+\frac{1}{2}}). \end{cases} \tag{4.12}$$

Alternatively, $(x_{n+\frac{1}{2}}, y_{n+\frac{1}{2}})$ is a solution to

$$(\gamma_n w_n, 0) \in (x_{n+\frac{1}{2}}, y_{n+\frac{1}{2}}) - (x_n, y_n) + \gamma_n L(x_{n+\frac{1}{2}}, y_{n+\frac{1}{2}}). \tag{4.13}$$

Now, set $A_1 = L$, $I = \{1\}$, and $\mathrm{i}(n) = 1$ for all $n \in \mathbb{N}$. Then, in view of (4.13), the iterations described by (4.2)–(4.4) are seen to conform to the format (1.4)–(1.5), and Algorithm 4.1 therefore fits the general framework of Algorithm 1.1. Accordingly, it suffices to verify the conditions of Theorem 3.1 to establish the claims. By [21, Theorem 4.1], condition (i) implies that L^{-1} has a Lipschitz localization at $\big((0,0),(\bar{x},\bar{y})\big)$. Therefore, by Lemma 2.3, condition (i) of Theorem 3.1 holds with $S = S_1 = \{(\bar{x},\bar{y})\}$ for some $\rho \in [0, +\infty[$. Now let $\gamma = \inf_{n\in\mathbb{N}} \gamma_n$. Then condition (iii) of Theorem 3.1 reads $\gamma > \rho$, and is trivially implied by condition (ii) above. Next, let us show that condition (iii) above implies condition (iv) of Theorem 3.1, i.e.,

$$(\exists\, \zeta \in]0,1[)(\forall n \in \mathbb{N})\ \ \zeta \leq \frac{\lambda_n}{1-\rho/\gamma_n} \leq 2 - \zeta. \tag{4.14}$$

It is readily checked that for $\gamma > 2\rho/\varepsilon$ (as is allowed by (ii) above), we have

$$\zeta = \frac{\varepsilon\gamma - 2\rho}{\gamma - \rho} \in]0, \varepsilon[\,. \tag{4.15}$$

Hence, it follows from (iii) above that

$$(\forall n \in \mathbb{N})\ \ \zeta < \varepsilon \leq \lambda_n < \frac{\lambda_n}{1-\rho/\gamma_n} \leq \frac{2-\varepsilon}{1-\rho/\gamma} = 2 - \zeta, \tag{4.16}$$

which establishes (4.14). Finally, it is clear that conditions (v)–(vii) of Theorem 3.1 are implied by conditions (iv)–(vi) above. □

Remark 4.4.

(i) In most concrete problems, it is not possible to obtain the value of ρ in the above proof [21]. As a result, condition (ii) and (v) in Theorem 4.3 are stated in qualitative terms rather than with hard bounds involving ρ.

(ii) The above result extends [20, Theorem 19] by allowing for relaxations and inexact computation of the iterates, thus making the algorithm more practical and flexible.

REFERENCES

[1] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis,* Birkhäuser, Boston, MA, 1990.

[2] H. H. Bauschke and P. L. Combettes, *A weak-to-strong convergence principle for Fejér-monotone methods in Hilbert spaces*, Math. Oper. Res., 26 (2001), pp. 248–264.

[3] N. Bourbaki, *Topologie Générale* 1–4, Masson, Paris, 1990.

[4] H. Brézis and P. L. Lions, *Produits infinis de résolvantes*, Israel J. Math., 29 (1978), pp. 329–345.

[5] F. E. Browder, *Convergence theorems for sequences of nonlinear operators in Banach spaces,* Math. Z., 100 (1967), pp. 201–225.

[6] P. L. Combettes, *Hilbertian convex feasibility problem: Convergence of projection methods*, Appl. Math. Optim., 35 (1997), pp. 311–330.
[7] P. L. Combettes, *Quasi-Fejérian analysis of some optimization algorithms*, in Inherently Parallel Algorithms for Feasibility and Optimization, D. Butnariu, Y. Censor, and S. Reich, eds., Elsevier, New York, 2001, pp. 115–152.
[8] A. Daniilidis and P. Georgiev, *Cyclic hypomonotonicity, cyclic submonotonicity, and integration*, J. Optim. Theory Appl., 122 (2004), pp. 19–40.
[9] A. L. Dontchev and R. T. Rockafellar, *Characterizations of strong regularity for variational inequalities over polyhedral convex sets,* SIAM J. Optim., 6 (1996), pp. 1087–1105.
[10] A. L. Dontchev and R. T. Rockafellar, *Characterizations of Lipschitzian stability in nonlinear programming*, in Mathematical Programming with Data Perturbations, Lecture Notes in Pure and Appl. Math. 195, Dekker, New York, 1998, pp. 65–82.
[11] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators,* Math. Program., 55 (1992), pp. 293–318.
[12] J. Eckstein and M. C. Ferris, *Smooth methods of multipliers for complementarity problems,* Math. Program., 86 (1999), pp. 65–90.
[13] L. G. Gubin, B. T. Polyak, and E. V. Raik, *The method of projections for finding the common point of convex sets*, U.S.S.R. Comput. Math. and Math. Phys., 7 (1967), pp. 1–24.
[14] A. N. Iusem, T. Pennanen, and B. F. Svaiter, *Inexact variants of the proximal point algorithm without monotonicity,* SIAM J. Optim., 13 (2003), pp. 1080–1097.
[15] A. B. Levy, *Lipschitzian multifunctions and a Lipschitzian inverse mapping theorem,* Math. Oper. Res., 26 (2001), pp. 105–118.
[16] A. B. Levy and R. A. Poliquin, *Characterizing the single-valuedness of multifunctions*, Set-Valued Anal., 5 (1997), pp. 351–364.
[17] B. Martinet, *Détermination approchée d'un point fixe d'une application pseudo-contractante. Cas de l'application prox,* C. R. Acad. Sci. Paris Sér. A Math., 274 (1972), pp. 163–165.
[18] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[19] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.
[20] T. Pennanen, *Local convergence of the proximal point algorithm and multiplier methods without monotonicity,* Math. Oper. Res., 27 (2002), pp. 170–191.
[21] S. M. Robinson, *Strongly regular generalized equations,* Math. Oper. Res., 5 (1980), pp. 43–62.
[22] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[23] R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
[24] R. T. Rockafellar, *Favorable classes of Lipschitz-continuous functions in subgradient optimization,* in Progress in Nondifferentiable Optimization, E. A. Nurminski, ed., IIASA Collaborative Proc. Ser. CP-82, International Institute for Applied Systems Analysis, Laxenburg, 1982, pp. 125–143.
[25] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis,* Springer-Verlag, New York, 1998.
[26] J. E. Spingarn, *Submonotone mappings and the proximal point algorithm,* Numer. Funct. Anal. Optim., 4 (1981/82), pp. 123–150.
[27] J. P. Vial, *Strong and weak convexity of sets and functions,* Math. Oper. Res., 8 (1983), pp. 231–259.
[28] E. H. Zarantonello, *Projections on convex sets in Hilbert space and spectral theory,* in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 237–424.
[29] E. Zeidler, *Nonlinear Functional Analysis and Its Applications II/B—Nonlinear Monotone Operators,* Springer-Verlag, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 2, pp. 743–767

MODELING AND CONTROL OF THE TIMOSHENKO BEAM. THE DISTRIBUTED PORT HAMILTONIAN APPROACH*

ALESSANDRO MACCHELLI† AND CLAUDIO MELCHIORRI†

Abstract. The purpose of this paper is to show how the Timoshenko beam can be fruitfully described within the framework of distributed port Hamiltonian (dpH) systems so that rather simple and elegant considerations can be drawn regarding both the modeling and control of this mechanical system. After the dpH model of the beam is introduced, the control problem is discussed. In particular, it is shown how control approaches already presented in the literature can be unified, and a new control methodology is presented and discussed. This control methodology relies on the generalization to infinite dimensions of the concept of structural invariant (Casimir function) and on the extension to distributed systems of the so-called control by interconnection methodology. In this way, finite dimensional passive controllers can stabilize distributed parameter systems by shaping their total energy, i.e., by assigning a new minimum in the desired equilibrium configuration that can be reached if a dissipative effect is introduced.

Key words. modeling and control of flexible structures, Stokes–Dirac structures, infinite dimensional port Hamiltonian systems, control by damping injection, Casimir functions, control by interconnection

AMS subject classifications. 35Q72, 37K99, 93C20

DOI. 10.1137/S0363012903429530

1. Introduction.

The port Hamiltonian approach has been introduced as a systematic framework for geometric modeling and control of lumped-parameter physical systems [15, 26]. The port Hamiltonian model of a finite dimensional system takes its inspiration from circuit analysis: the behavior of a physical system is the result of a network of atomic multiport elements, each of them characterized by a particular energy property. The key point is the identification of the interconnection structure, mathematically described by a Dirac structure [2, 26], a generalization of the well-known Kirchhoff laws [16]. In this way, the variation of a system's total energy is related to the power exchanged with the environment, and the dynamics is the result of internal power flows among different parts of the whole system. It has been shown that this approach can be fruitfully applied for modeling a wide class of physical (mechanical, electrical, hydraulic, and chemical) systems, and several control techniques, based on energy considerations, have been developed in order to solve the regulation problem [20, 21, 26].

In some sense, it seems natural to extend the finite dimensional Hamiltonian formulation in order to deal with distributed parameter systems. Many results on integrability, existence of solutions or stability, and several applications have been proposed in the last decades; see, for example, [24] for an application to fluid dynamics and [19] for a nice introduction and historical remarks. On the other hand, it is interesting to note that some problems regarding the treatment of *boundary conditions* are still open. In fact, most of the research activity has been focused on the study of

*Received by the editors June 9, 2003; accepted for publication (in revised form) December 18, 2003; published electronically August 27, 2004. This research activity has been performed in the context of the European project GeoPlex, reference code IST-2001-34166. Further information is available at http://www.geoplex.cc.

http://www.siam.org/journals/sicon/43-2/42953.html

†Department of Electronics, Computer Science and Systems (DEIS), University of Bologna, viale Risorgimento 2, 40136 Bologna, Italy (amacchelli@deis.unibo.it, cmelchiorri@deis.unibo.it).

infinite dimensional systems characterized by an infinite spatial domain, for which the state variables tend to zero when the spatial variable tends to infinity (with respect to some norm), or on the analysis of infinite dimensional systems with zero boundary conditions (on the finite spatial domain).

These are autonomous systems: no interaction, i.e., power exchange, with the environment is taken into account. This is a strong limitation since it is not possible to study the effect of nonzero boundary conditions (e.g., voltages and currents at both ends of a transmission line) on the dynamics of the system. In this way, it is difficult to deal with control application for infinite dimensional systems in Hamiltonian form. The controller, in fact, can act on the system only by properly modifying the boundary variables or, equivalently, by exchanging power with the (infinite dimensional) system.

From a mathematical point of view, it is not evident how a nonzero energy flow through the boundary can be incorporated into the *classical* distributed Hamiltonian framework. The key point is the notion of Dirac structure in infinite dimensions that will be defined, in this case, on a space of differential forms on the spatial domain of the system and its boundary. Since the relation between variation of internal energy and power flow through the boundary relies on the Stokes theorem [17, 18], these structures are called Stokes–Dirac structures.

Once the Stokes–Dirac structure of a particular infinite dimensional system is deduced, the distributed port Hamiltonian (dpH) model follows *automatically* [17, 18] and the control problem can be approached. When dealing with the control of distributed parameter systems, the main problem concerns the intrinsic difficulties related to the proof of stability of an equilibrium configuration. It is important to emphasize that this *limitation* does not depend on the particular approach adopted. Even if a distributed parameter system is described within the port Hamiltonian framework, the stability proof of a certain control scheme will always be a difficult task. On the other hand, the main advantages in adopting the dpH framework can be the following:

- The development of control schemes for infinite dimensional systems is usually based on energy considerations or, equivalently, the stability proof often relies on the properties of an energy-like functional, a generalization of the Lyapunov function to the distributed parameters case. The Hamiltonian description of a distributed parameter system is given in terms of time evolution of energy variables depending on the variation of the total energy of the system. In this way, the energy of the system, which is generally a *good* Lyapunov function, appears explicitly in the mathematical model of the system itself and, consequently, both the design of the control law and the proof of its stability can be deduced and presented in a more intuitive (in some sense *physical*) and elegant way.
- The port Hamiltonian formulation of distributed parameter systems deeply relies on the notion of Dirac structure, as in finite dimensions. This fact is important and allows us to go further; in particular, it is of great interest to understand whether the control schemes developed for finite dimensional port Hamiltonian systems could also be generalized in order to deal with the distributed parameter case. For example, suppose that the total energy (Hamiltonian) of the system is characterized by a minimum at the desired equilibrium configuration. This happens, for example, in the case of flexible beams, for which the zero-energy configuration corresponds to the undeformed beam. In this situation, the controller can be developed in order to

behave as a dissipative element to be connected to the system at the boundary or along the distributed port. The amount of dissipated power can be increased in order to quickly reach the configuration with minimum energy. As in the finite dimensional case, it can happen that the minimum of the energy does not correspond to a desired configuration. Then it is necessary to shape the energy function so that a new minimum is introduced. This can be achieved by generalizing the control by interconnection and energy-shaping methodology to deal with distributed parameter systems, as presented in [22], where the infinite dimensional system is a lossless transmission line.

In this paper, it is shown how the modeling and control problems of an infinite dimensional system, the Timoshenko beam, can be solved within the framework of dpH systems. Flexible beams are generally modeled according to the classical Euler–Bernoulli theory: this formulation provides a good description of the dynamical behavior of the system if the beam's cross sectional dimension is small in comparison to its length. In this case, the effects of the rotary inertia of the beam are not considered. A more accurate beam model is provided by the Timoshenko theory, according to which the rotary inertia and also the deformation due to shear are considered. The resulting Timoshenko model of the beam is generally more accurate in predicting the beam's response than the Euler–Bernoulli one but, on the other hand, it is more difficult to utilize for control purposes because of its complexity.

As already pointed out, the dpH formulation of the Timoshenko model of the beam [6] (but refer also to [7, 11, 12]) does not reduce the complexity of the model itself, but it is useful both for modeling considerations and control purposes. From the modeling point of view, the internal and external interconnections of the system are revealed: it is clear how the kinetic and potential elastic energy domains interact and how the system can exchange power with the environment through its border and/or a distributed port. Furthermore, the dpH representation of the system makes it possible to extend well-established passive control strategies that were originally developed for finite dimensional port Hamiltonian systems and to elegantly unify control approaches already presented in the literature [8, 25].

The paper is organized as follows. In section 2, a brief background on Dirac structures and on the classical formulation of the Timoshenko model of the beam is provided, and then the Stokes–Dirac structure of the Timoshenko beam is presented and the dpH model introduced. The control problem is approached in sections 3 and 4. In section 3, the *control by damping injection* methodology is extended to infinite dimensions in order to stabilize the beam in its undeformed configuration, as already presented in [8, 25]. In section 4, the *control by interconnection and energy shaping* methodology [20, 26, 21] is extended to distributed parameter systems in order to control a mechanical system made of a flexible (Timoshenko) beam with a rigid body connected at one of its extremities. The finite dimensional controller, acting on the system through the other extremity, is developed by properly extending the concept of Casimir functions to the infinite dimensional case and by generalizing the results presented in [22] (see also [12, 11]). Finally, conclusions and suggestions for future work are illustrated in section 5.

2. Timoshenko beam in dpH form.

2.1. Timoshenko beam—the *classical* formulation. According to the Timoshenko theory, the motion of a beam can be described by the following system of

PDEs:

$$\begin{aligned} \rho\frac{\partial^2 w}{\partial t^2} - K\frac{\partial^2 w}{\partial x^2} + K\frac{\partial \phi}{\partial x} &= 0, \\ I_\rho\frac{\partial^2 \phi}{\partial t^2} - EI\frac{\partial^2 \phi}{\partial x^2} + K\left(\phi - \frac{\partial w}{\partial x}\right) &= 0, \end{aligned} \tag{2.1}$$

where t is the time and $x \in [0, L]$ is the spatial coordinate along the beam in its equilibrium position, and $w(x,t)$ is the deflection of the beam from the equilibrium configuration and $\phi(x,t)$ is the rotation of the beam's cross section due to bending; the motion takes place in the wx-plane. Denote by $\mathcal{D} := [0, L]$ the spatial domain and by $\partial\mathcal{D} = \{0, L\}$ its boundary.

The coefficients ρ, I_ρ, E, and I, assumed to be constant, are the mass per unit length, the mass moment of inertia of the cross section, Young's modulus, and the moment of inertia of the cross section, respectively. The coefficient K is equal to kGA, where G is the modulus of elasticity in shear, A is the cross sectional area, and k is a constant depending on the shape of the cross section.

The mechanical energy is given by the following relation [8]:

$$\begin{aligned} \mathcal{H}(t) := \overbrace{\frac{1}{2}\int_0^L \rho\left(\frac{\partial w}{\partial t}\right)^2 + I_\rho\left(\frac{\partial \phi}{\partial t}\right)^2 dx}^{\textit{kinetic energy}} \\ \underbrace{+\frac{1}{2}\int_0^L K\left(\phi - \frac{\partial w}{\partial x}\right)^2 + EI\left(\frac{\partial \phi}{\partial x}\right)^2 dx}_{\textit{potential elastic energy}}. \end{aligned} \tag{2.2}$$

Note the presence of two interactive energy domains, the *kinetic* and the *potential elastic*.

2.2. Dirac structures. The starting point in the definition of a port Hamiltonian system (both finite and infinite dimensional) is the identification of a suitable space of power variables, strictly related to the geometry of the system, and the definition of a Dirac structure on this space of power variables, in order to describe the internal and external interconnection of the system. The Dirac structures were introduced in [2], while in [4, 26] it is pointed out that they are the geometric *tool* that allows us to formalize and generalize the notion of power-conserving interconnection.

Before stating the general definition of Dirac structure, it is necessary to introduce the space of *power variables*. Consider a linear space $\mathcal{F}$, possibly infinite dimensional (space of generalized velocities or *flows*), and denote by $\mathcal{E} = \mathcal{F}^*$ its dual (space of generalized forces or *efforts*). The space of power variables is $\mathcal{F} \times \mathcal{E}$. Then, from [18], we take the following fundamental definition.

DEFINITION 2.1 (Dirac structure). *Denote by $\mathcal{F} \times \mathcal{E}$ a space of power variables (possibly infinite dimensional). There exists on $\mathcal{F} \times \mathcal{E}$ the canonically defined symmetric bilinear form (+pairing operator)*

$$\ll (f_1, e_1), (f_2, e_2) \gg := \langle e_1, f_2\rangle + \langle e_2, f_1\rangle, \tag{2.3}$$

where $f_i \in \mathcal{F}$, $e_i \in \mathcal{E}$, $i = 1, 2$, and $\langle\cdot,\cdot\rangle$ denotes the duality product between $\mathcal{F}$ and its dual space $\mathcal{E}$. A constant Dirac structure on $\mathcal{F}$ is a linear subspace $\mathbb{D} \subset \mathcal{F} \times \mathcal{E}$ such that

$$\mathbb{D} = \mathbb{D}^\perp, \tag{2.4}$$

where $\perp$ denotes the orthogonal complement with respect to the bilinear form $\ll \cdot, \cdot \gg$.

An immediate consequence of the previous definition is that, if $(f, e) \in \mathbb{D}$, then

$$0 = \ll (f, e), (f, e) \gg = 2 \langle e, f \rangle . \tag{2.5}$$

Consequently, $\langle e, f \rangle = 0$ for every $(f, e) \in \mathbb{D}$. In other words, if $(f, e) \in \mathcal{F} \times \mathcal{E}$ is a couple of power conjugated variables, the fact that they belong to the Dirac structure $\mathbb{D}$ implies power conservation, i.e., the dual product is equal to 0. The Dirac structure is the geometrical tool by means of which it is possible to deal with power-conserving interconnection in physical systems. As will be pointed out in section 2.4 for the Timoshenko beam, once a proper interconnection structure is defined, the port Hamiltonian model of a physical system follows automatically.

2.3. Timoshenko beam—The Stokes–Dirac structure. Consider the mechanical energy (2.2). The potential elastic energy is a function of the *shear* and of the *bending*, given by the following 1-forms:

$$\epsilon_t(t, x) = \left[\frac{\partial w}{\partial x}(t, x) - \phi(t, x) \right] dx, \qquad \epsilon_r(t, x) = \frac{\partial \phi}{\partial x}(t, x) dx. \tag{2.6}$$

The associated coenergy variables are the 0-forms (functions) *shear force* and the *bending momentum*, given by $\sigma_t(t, x) = K * \epsilon_t(t, x)$ and $\sigma_r(t, x) = EI * \epsilon_r(t, x)$, where $*$ is the Hodge star operator defined, for example, in [14]. Besides, the kinetic energy is a function of the *translational* and *rotational momenta*, i.e., of the following 1-form:

$$p_t(t, x) = \rho \frac{\partial w}{\partial t}(t, x) dx, \qquad p_r(t, x) = I_\rho \frac{\partial \phi}{\partial t}(t, x) dx, \tag{2.7}$$

and the associated coenergy variables are the 0-forms *translational* and *rotational momenta*, given by $v_t(t, x) = \frac{1}{\rho} * p_t(t, x)$ and $v_r(t, x) = \frac{1}{I_\rho} * p_r(t, x)$.

Consider an n-dimensional (Riemannian) manifold $\mathcal{N}$ and denote by $\Omega^k(\mathcal{N})$ the space of k-forms on $\mathcal{N}$, i.e., the space of k-linear alternating functions. So, we have that $p_t, p_r, \epsilon_t, \epsilon_r \in \Omega^1(\mathcal{D})$ and that $w, \phi \in \Omega^0(\mathcal{D})$. If $\mathrm{d} : \Omega^k(\mathcal{N}) \to \Omega^{k+1}(\mathcal{N})$ is the exterior derivative on the space of forms, it is possible to rewrite (2.6) and (2.7) as

$$p_t = \rho * \frac{\partial w}{\partial t}, \qquad \epsilon_t = \mathrm{d}w - * \phi, \qquad p_r = I_\rho * \frac{\partial \phi}{\partial t}, \qquad \epsilon_r = \mathrm{d}\phi,$$

and the total energy (2.2) becomes the following (quadratic) functional:

$$\begin{aligned} \mathcal{H}(p_t, p_r, \epsilon_t, \epsilon_r) &= \int_{\mathcal{D}} H(p_t, p_r, \epsilon_t, \epsilon_r) \\ &= \frac{1}{2} \int_{\mathcal{D}} \left(\frac{1}{\rho} * p_t \wedge p_t + \frac{1}{I_\rho} * p_r \wedge p_r + K * \epsilon_t \wedge \epsilon_t + EI * \epsilon_r \wedge \epsilon_r \right) \end{aligned} \tag{2.8}$$

with $H : \Omega^1(\mathcal{D}) \times \cdots \times \Omega^1(\mathcal{D}) \times \mathcal{D} \to \Omega^1(\mathcal{D})$ the energy density. Consider a time function

$$(p_t(t), p_r(t), \epsilon_t(t), \epsilon_r(t)) \in \Omega^1(\mathcal{D}) \times \cdots \times \Omega^1(\mathcal{D})$$

with $t \in \mathbb{R}$, and evaluate the energy $\mathcal{H}$ along this trajectory. At any time t, the variation of internal energy, that is, the power exchanged with the environment, is

given by

$$\begin{aligned}\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} &= \int_{\mathcal{D}} \left(\delta_{p_t}\mathcal{H} \wedge \frac{\partial p_t}{\partial t} + \delta_{p_r}\mathcal{H} \wedge \frac{\partial p_r}{\partial t} + \delta_{\epsilon_t}\mathcal{H} \wedge \frac{\partial \epsilon_t}{\partial t} + \delta_{\epsilon_r}\mathcal{H} \wedge \frac{\partial \epsilon_r}{\partial t} \right) \\ &= \int_{\mathcal{D}} \left[\left(\frac{1}{\rho} * p_t \right) \wedge \frac{\partial p_t}{\partial t} + \left(\frac{1}{I_\rho} * p_r \right) \wedge \frac{\partial p_r}{\partial t} + (K * \epsilon_t) \wedge \frac{\partial \epsilon_t}{\partial t} + (EI * \epsilon_r) \wedge \frac{\partial \epsilon_r}{\partial t} \right].\end{aligned} \tag{2.9}$$

The differential forms $\frac{\partial p_t}{\partial t}$, $\frac{\partial p_r}{\partial t}$, $\frac{\partial \epsilon_t}{\partial t}$, and $\frac{\partial \epsilon_r}{\partial t}$ are the time derivatives of the energy variables p_t, p_r, ϵ_t, and ϵ_r and represent the *generalized velocities* (flows), while $\delta_{p_t}\mathcal{H}$, $\delta_{p_r}\mathcal{H}$, $\delta_{\epsilon_t}\mathcal{H}$, and $\delta_{\epsilon_r}\mathcal{H}$ are the variational derivatives of the total energy (2.8). They are related to the rate of change of the stored energy and represent the *generalized forces* (efforts).

The dpH formulation of the Timoshenko beam can be obtained either by expressing (2.1) in terms of p_t, p_r, ϵ_r, and ϵ_t as introduced in (2.6) and (2.7), or, in a more rigorous way, by revealing the underlying Dirac structure of the model. For this purpose, it is necessary to define the space of power variables. The space of flows is given by

$$\mathcal{F} := \underbrace{\Omega^1(\mathcal{D}) \times \Omega^1(\mathcal{D}) \times \Omega^1(\mathcal{D}) \times \Omega^1(\mathcal{D})}_{\textit{generalized velocities}} \times \underbrace{\Omega^0(\partial\mathcal{D}) \times \Omega^0(\partial\mathcal{D})}_{\text{border } \textit{flow}}, \tag{2.10}$$

and it is well known that the space of effort $\mathcal{E}$ is the *dual* of $\mathcal{F}$. The concept of duality over the space of forms can be given by the following proposition [17].

PROPOSITION 2.2. *Consider an n-dimensional Riemannian manifold $\mathcal{N}$. Then the dual space $(\Omega^k(\mathcal{N}))^*$ of $\Omega^k(\mathcal{N})$ can be identified with $\Omega^{n-k}(\mathcal{N})$ and the duality product between $\Omega^k(\mathcal{N})$ and $(\Omega^k(\mathcal{N}))^*$ by*

$$\langle \beta, \alpha \rangle := \int_{\mathcal{N}} \alpha \wedge \beta \tag{2.11}$$

with $\alpha \in \Omega^k(\mathcal{N})$ and $\beta \in \Omega^{n-k}(\mathcal{N})$. The same result holds for $\Omega^k(\partial\mathcal{N})$.

An immediate consequence of Proposition 2.2 is that the dual space $\mathcal{E}$ of $\mathcal{F}$, the space of efforts, can be easily identified with

$$\mathcal{E} := \underbrace{\Omega^0(\mathcal{D}) \times \Omega^0(\mathcal{D}) \times \Omega^0(\mathcal{D}) \times \Omega^0(\mathcal{D})}_{\textit{generalized forces}} \times \underbrace{\Omega^0(\partial\mathcal{D}) \times \Omega^0(\partial\mathcal{D})}_{\text{border } \textit{effort}}. \tag{2.12}$$

Thus, the duality product (2.11) and the +pairing operator (2.3) can be easily specialized in order to deal with the space of power variables $\mathcal{F} \times \mathcal{E}$ defined by (2.10) and (2.12). Suppose that

$$\left(f_{p_t}, f_{p_r}, f_{\epsilon_t}, f_{\epsilon_r}, f_b^t, f_b^r, e_{p_t}, e_{p_r}, e_{\epsilon_t}, e_{\epsilon_r}, e_b^t, e_b^r\right), \left(f_{p_t}^i, \ldots, f_b^{r,i}, e_{p_t}^i, \ldots, e_b^{r,i}\right) \in \mathcal{F} \times \mathcal{E}$$

with $i = 1, 2$. Then

$$\begin{aligned}
&\langle (e_{p_t}, e_{p_r}, e_{\epsilon_t}, e_{\epsilon_r}, e_b^t, e_b^r), (f_{p_t}, f_{p_r}, f_{\epsilon_t}, f_{\epsilon_r}, f_b^t, f_b^r) \rangle \\
&\quad := \int_{\mathcal{D}} (f_{p_t} \wedge e_{p_t} + f_{p_r} \wedge e_{p_r} + f_{\epsilon_t} \wedge e_{\epsilon_t} + f_{\epsilon_r} \wedge e_{\epsilon_r}) + \int_{\partial\mathcal{D}} \left(f_b^t \wedge e_b^t + f_b^r \wedge e_b^r \right), \\
&\ll \left(f_{p_t}^1, \ldots, f_b^{r,1}, e_{p_t}^1, \ldots, e_b^{r,1}, \right) \left(f_{p_t}^2, \ldots, f_b^{r,2}, e_{p_t}^2, \ldots, e_b^{r,2} \right) \gg \\
&\qquad := \int_{\mathcal{D}} \left(f_{p_t}^1 \wedge e_{p_t}^2 + f_{p_t}^2 \wedge e_{p_t}^1 + f_{p_r}^1 \wedge e_{p_r}^2 + f_{p_r}^2 \wedge e_{p_r}^1 \right) \\
&\qquad\quad + \int_{\mathcal{D}} \left(f_{\epsilon_t}^1 \wedge e_{\epsilon_t}^2 + f_{\epsilon_t}^2 \wedge e_{\epsilon_t}^1 + f_{\epsilon_r}^1 \wedge e_{\epsilon_r}^2 + f_{\epsilon_r}^2 \wedge e_{\epsilon_r}^1 \right) \\
&\qquad\quad + \int_{\partial\mathcal{D}} \left(f_b^{t,1} \wedge e_b^{t,2} + f_b^{t,2} \wedge e_b^{t,1} + f_b^{r,1} \wedge e_b^{r,2} + f_b^{r,2} \wedge e_b^{r,1} \right).
\end{aligned} \tag{2.13}$$

With the following proposition, the main result of this section is presented.

PROPOSITION 2.3 (the Timoshenko beam Dirac structure). *Consider the space of power variables $\mathcal{F} \times \mathcal{E}$ with $\mathcal{F}$ and $\mathcal{E}$ defined in* (2.10) *and* (2.12) *and the bilinear form (+pairing operator) $\ll \cdot, \cdot \gg$ given by* (2.13). *Define the following linear subspace $\mathbb{D}$ of $\mathcal{F} \times \mathcal{E}$:*

$$\begin{aligned}
\mathbb{D} \;=\; \Bigg\{ & (f_{p_t}, f_{p_r}, f_{\epsilon_t}, f_{\epsilon_r}, f_b^t, f_b^r, e_{p_t}, e_{p_r}, e_{\epsilon_t}, e_{\epsilon_r}, e_b^t, e_b^r) \in \mathcal{F} \times \mathcal{E} \mid \\
& \begin{bmatrix} f_{p_t} \\ f_{p_r} \\ f_{\epsilon_t} \\ f_{\epsilon_r} \end{bmatrix} = - \begin{bmatrix} 0 & 0 & \mathrm{d} & 0 \\ 0 & 0 & * & \mathrm{d} \\ \mathrm{d} & -* & 0 & 0 \\ 0 & \mathrm{d} & 0 & 0 \end{bmatrix} \begin{bmatrix} e_{p_t} \\ e_{p_r} \\ e_{\epsilon_t} \\ e_{\epsilon_r} \end{bmatrix}, \quad \begin{bmatrix} f_b^t \\ f_b^r \\ e_b^t \\ e_b^r \end{bmatrix} = \begin{bmatrix} e_{p_t} \mid_{\partial\mathcal{D}} \\ e_{p_r} \mid_{\partial\mathcal{D}} \\ e_{\epsilon_t} \mid_{\partial\mathcal{D}} \\ e_{\epsilon_r} \mid_{\partial\mathcal{D}} \end{bmatrix} \Bigg\},
\end{aligned} \tag{2.14}$$

where $\mid_{\partial\mathcal{D}}$ denotes the restriction on the border of the (spatial) domain $\mathcal{D}$. Then $\mathbb{D} = \mathbb{D}^\perp$; that is, $\mathbb{D}$ is a Dirac structure.

Proof. The proof can be divided into two steps. In the first one, it is verified that $\mathbb{D} \subseteq \mathbb{D}^\perp$ while in the second one, that $\mathbb{D}^\perp \subseteq \mathbb{D}$. Suppose that

$$\omega^i = \left(f_{p_t}^i, f_{p_r}^i, f_{\epsilon_t}^i, f_{\epsilon_r}^i, f_b^{t,i}, f_b^{r,i}, e_{p_t}^i, e_{p_r}^i, e_{\epsilon_t}^i, e_{\epsilon_r}^i, e_b^{t,i}, e_b^{r,i} \right) \in \mathcal{F} \times \mathcal{E}$$

with $i = 1, 2$; clearly, $\mathbb{D} \subseteq \mathbb{D}^\perp$ if $\forall\, \omega^1, \omega^2 \in \mathbb{D}$, it happens that $\ll \omega_1, \omega_2 \gg = 0$. From (2.13) and from the definition (2.14) of the Dirac structure $\mathbb{D}$, we have

$$\begin{aligned}
\ll \omega^1, \omega^2 \gg = & \int_{\mathcal{D}} \left[-\mathrm{d}e_{\epsilon_t}^1 \wedge e_{p_t}^2 - \mathrm{d}e_{\epsilon_t}^2 \wedge e_{p_t}^1 + (-*e_{\epsilon_t}^1 - \mathrm{d}e_{\epsilon_r}^1) \wedge e_{p_r}^2 \right] \\
& + \int_{\mathcal{D}} \left[(-*e_{\epsilon_t}^2 - \mathrm{d}e_{\epsilon_r}^2) \wedge e_{p_r}^1 + (-\mathrm{d}e_{p_t}^1 + *e_{p_r}^1) \wedge e_{\epsilon_t}^2 \right] \\
& + \int_{\mathcal{D}} \left[(-\mathrm{d}e_{p_t}^2 + *e_{p_r}^2) \wedge e_{\epsilon_t}^1 - \mathrm{d}e_{p_r}^1 \wedge e_{\epsilon_r}^2 - \mathrm{d}e_{p_r}^2 \wedge e_{\epsilon_r}^1 \right] \\
& + \int_{\partial\mathcal{D}} \left[f_b^{t,1} \wedge e_b^{t,2} + f_b^{t,2} \wedge e_b^{t,1} + f_b^{r,1} \wedge e_b^{r,2} + f_b^{r,2} \wedge e_b^{r,1} \right]
\end{aligned}$$

$$
\begin{aligned}
&= -\int_{\mathcal{D}} \left[(\mathrm{d}e^2_{p_t} \wedge e^1_{\epsilon_t} + e^2_{p_t} \wedge \mathrm{d}e^1_{\epsilon_t}) + (\mathrm{d}e^1_{p_t} \wedge e^2_{\epsilon_t} + e^1_{p_t} \wedge \mathrm{d}e^2_{\epsilon_t}) \right] \\
&\quad - \int_{\mathcal{D}} \left[(\mathrm{d}e^2_{p_r} \wedge e^1_{\epsilon_r} + e^2_{p_r} \wedge \mathrm{d}e^1_{\epsilon_r}) + (\mathrm{d}e^1_{p_r} \wedge e^2_{\epsilon_r} + e^1_{p_r} \wedge \mathrm{d}e^2_{\epsilon_r}) \right] \\
&\quad + \int_{\partial\mathcal{D}} \left[f^{t,1}_b \wedge e^{t,2}_b + f^{t,2}_b \wedge e^{t,1}_b + f^{r,1}_b \wedge e^{r,2}_b + f^{r,2}_b \wedge e^{r,1}_b \right].
\end{aligned}
$$

Since $\mathrm{d}(e^1 \wedge e^2) = \mathrm{d}e^1 \wedge e^2 + e^1 \wedge \mathrm{d}e^2$ and, for the Stokes theorem, if $\alpha \in \Omega^0(\mathcal{D})$, then $\int_{\mathcal{D}} \mathrm{d}\alpha = \int_{\partial\mathcal{D}} \alpha$, we deduce that $\ll \omega^1, \omega^2 \gg = 0$ and $\mathbb{D} \subseteq \mathbb{D}^\perp$.

In order to prove that $\mathbb{D}^\perp \subseteq \mathbb{D}$, consider $\omega^2 \in \mathbb{D}^\perp$. From the definition of Dirac structure we have that $\forall\, \omega^1 \in \mathbb{D}$, $\ll \omega^1, \omega^2 \gg = 0$. Since $\omega^1 \in \mathbb{D}$, from (2.14) we have

$$
\begin{aligned}
0 &= \ll \omega^1, \omega^2 \gg \\
&= \int_{\mathcal{D}} \left[-\mathrm{d}e^1_{\epsilon_t} \wedge e^2_{p_t} + f^2_{p_t} \wedge e^1_{p_t} + (-{*}e^1_{\epsilon_t} - \mathrm{d}e^1_{\epsilon_r}) \wedge e^2_{p_r} + f^2_{p_r} \wedge e^1_{p_r} \right] \\
&\quad + \int_{\mathcal{D}} \left[(-\mathrm{d}e^1_{p_t} + {*}e^1_{p_r}) \wedge e^2_{\epsilon_t} + f^2_{\epsilon_t} \wedge e^1_{\epsilon_t} - \mathrm{d}e^1_{p_r} \wedge e^2_{\epsilon_r} + f^2_{\epsilon_r} \wedge e^1_{\epsilon_r} \right] \\
&\quad + \int_{\partial\mathcal{D}} \left[e^1_{p_t}\,|_{\partial\mathcal{D}} \wedge e^{t,2}_b + f^{t,2}_b \wedge e^1_{\epsilon_t}\,|_{\partial\mathcal{D}} + e^1_{p_r}\,|_{\partial\mathcal{D}} \wedge e^{r,2}_b + f^{r,2}_b \wedge e^1_{\epsilon_r}\,|_{\partial\mathcal{D}} \right].
\end{aligned}
$$

From the Stokes theorem and the properties of the exterior derivative, it is possible to obtain that

$$
\begin{aligned}
0 &= \int_{\mathcal{D}} \left[e^1_{\epsilon_t} \wedge (\mathrm{d}e^2_{p_t} - {*}e^2_{p_r} + f^2_{\epsilon_t}) + e^1_{p_t} \wedge (f^2_{p_t} + \mathrm{d}e^2_{\epsilon_t}) \right] \\
&\quad + \int_{\mathcal{D}} \left[e^1_{\epsilon_r} \wedge (\mathrm{d}e^2_{p_r} + f^2_{\epsilon_r}) + e^1_{p_r} \wedge (f^2_{p_r} + {*}e^2_{\epsilon_t} + \mathrm{d}e^2_{\epsilon_r}) \right] \\
&\quad + \int_{\partial\mathcal{D}} \left[e^1_{\epsilon_t}\,|_{\partial\mathcal{D}} \wedge (-e^2_{p_t}\,|_{\partial\mathcal{D}} + f^{t,2}_b) + e^1_{\epsilon_r}\,|_{\partial\mathcal{D}} \wedge (-e^2_{p_r}\,|_{\partial\mathcal{D}} + f^{r,2}_b) \right] \\
&\quad + \int_{\partial\mathcal{D}} \left[e^1_{p_t}\,|_{\partial\mathcal{D}} \wedge (-e^2_{\epsilon_t}\,|_{\partial\mathcal{D}} + e^{t,2}_b) + e^1_{p_r}\,|_{\partial\mathcal{D}} \wedge (-e^2_{\epsilon_r}\,|_{\partial\mathcal{D}} + e^{r,2}_b) \right]
\end{aligned}
$$

for every $\omega^1 \in \mathbb{D}$. We deduce that the previous relation holds if and only if $\omega^2 \in \mathbb{D}$. So, $\mathbb{D}^\perp \subseteq \mathbb{D}$, and this completes the proof. ∎

2.4. dpH formulation of the Timoshenko beam. Consider the total energy (2.8) as the Hamiltonian of the system, i.e., a (quadratic) functional of the energy variables p_t, p_r, ϵ_t, and ϵ_r bounded from below. The rate of change of these energy variables (generalized velocities) can be connected to the Dirac structure (2.14) by setting

$$
f_{p_t} = -\frac{\partial p_t}{\partial t}, \quad f_{\epsilon_t} = -\frac{\partial \epsilon_t}{\partial t}, \quad f_{p_r} = -\frac{\partial p_r}{\partial t}, \quad f_{\epsilon_r} = -\frac{\partial \epsilon_r}{\partial t}, \tag{2.15}
$$

where the minus sign is necessary in order to have a consistent energy flow description. Moreover, the rate of change of the Hamiltonian with respect to the energy variables,

that is, its variational derivatives, can be related to the Dirac structure by setting

$$e_{p_t} = \delta_{p_t}\mathcal{H}, \quad e_{\epsilon_t} = \delta_{\epsilon_t}\mathcal{H}, \quad e_{p_r} = \delta_{p_r}\mathcal{H}, \quad e_{\epsilon_r} = \delta_{\epsilon_r}\mathcal{H}. \tag{2.16}$$

From (2.15) and (2.16), it is possible to obtain the distributed Hamiltonian formulation with boundary energy flow of the Timoshenko beam. We give the following.

DEFINITION 2.4 (dpH model of Timoshenko beam). *The dpH model of the Timoshenko beam with Dirac structure* $\mathbb{D}$ (2.14) *and Hamiltonian* $\mathcal{H}$ (2.8) *is given by*

$$\begin{bmatrix} \partial_t p_t \\ \partial_t p_r \\ \partial_t \epsilon_t \\ \partial_t \epsilon_r \end{bmatrix} = \begin{bmatrix} 0 & 0 & \mathrm{d} & 0 \\ 0 & 0 & * & \mathrm{d} \\ \mathrm{d} & -* & 0 & 0 \\ 0 & \mathrm{d} & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_{p_t}\mathcal{H} \\ \delta_{p_r}\mathcal{H} \\ \delta_{\epsilon_t}\mathcal{H} \\ \delta_{\epsilon_r}\mathcal{H} \end{bmatrix}, \quad \begin{bmatrix} f_b^t \\ f_b^r \\ e_b^t \\ e_b^r \end{bmatrix} = \begin{bmatrix} \delta_{p_t}\mathcal{H}\,|_{\partial\mathcal{D}} \\ \delta_{p_r}\mathcal{H}\,|_{\partial\mathcal{D}} \\ \delta_{\epsilon_t}\mathcal{H}\,|_{\partial\mathcal{D}} \\ \delta_{\epsilon_r}\mathcal{H}\,|_{\partial\mathcal{D}} \end{bmatrix}. \tag{2.17}$$

Since the elements of every Dirac structure satisfy the power conserving property, we have, given

$$(f_{p_t}, \ldots, f_{\epsilon_r}, e_{p_t}, \ldots, e_{\epsilon_r}, f_b^t, \ldots, e_b^r) \in \mathbb{D},$$

that

$$\int_{\mathcal{D}} (f_{p_t} \wedge e_{p_t} + f_{p_r} \wedge e_{p_r} + f_{\epsilon_t} \wedge e_{\epsilon_t} + f_{\epsilon_r} \wedge e_{\epsilon_r}) + \int_{\partial\mathcal{D}} \left(f_b^t \wedge e_b^t + f_b^r \wedge e_b^r\right) = 0$$

and, consequently, from (2.9), (2.15), and (2.16), the following proposition can be proved.

PROPOSITION 2.5 (energy balance). *Consider the dpH model of the Timoshenko beam* (2.17). *Then*

$$\begin{aligned} \frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t}(t) &= \int_{\partial\mathcal{D}} \left(e_b^t \wedge f_b^t + e_b^r \wedge f_b^r\right) \\ &= \left[e_b^t(t,L) f_b^t(t,L) + e_b^r(t,L) f_b^r(t,L)\right] - \left[e_b^t(t,0) f_b^t(t,0) + e_b^r(t,0) f_b^r(t,0)\right] \end{aligned} \tag{2.18}$$

or, in other words, the increase of total energy of the beam is equal to the power supplied through the border.

2.5. Introducing the distributed port. Power exchange through the boundaries is not the only means by which the system can interact with the environment. The "distributed control" is a well-known control technique that can be fruitfully applied to flexible structures. The actuators are connected along the flexible structure and can act on the system by applying forces/couples that are functions of the configuration of the beam. The final result is that vibrations can be damped in a more efficient way than by acting only on the border of the beam.

In order to introduce a distributed port, the space of power variables $\mathcal{F} \times \mathcal{E}$ defined in (2.10) and (2.12) and the Dirac structure $\mathbb{D}$ defined in (2.14) have to be modified. The space of power variables becomes $\mathcal{F}_d \times \mathcal{E}_d$, where

$$\mathcal{F}_d := \mathcal{F} \times \underbrace{\Omega^1(\mathcal{D}) \times \Omega^1(\mathcal{D})}_{\textit{distrib. flow}}, \qquad \mathcal{E}_d := \mathcal{E} \times \underbrace{\Omega^1(\mathcal{D}) \times \Omega^1(\mathcal{D})}_{\textit{distrib. effort}}. \tag{2.19}$$

The modified Dirac structure that incorporates the distributed port is given by the following.

PROPOSITION 2.6. *Consider the space of power variables* $\mathcal{F}_d \times \mathcal{E}_d$ *defined in* (2.19) *and the bilinear form (+pairing operator)* $\ll \cdot, \cdot \gg$ *given by* (2.13). *Define the following linear subspace* $\mathbb{D}_d$ *of* $\mathcal{F}_d \times \mathcal{E}_d$*:*

$$
\begin{aligned}
\mathbb{D}_d \;=\; \Bigg\{ & (f_{p_t}, f_{p_r}, f_{\epsilon_t}, f_{\epsilon_r}, f_b^t, f_b^r, f_d^t, f_d^r, e_{p_t}, e_{p_r}, e_{\epsilon_t}, e_{\epsilon_r}, e_b^t, e_b^r, e_d^t, e_d^r) \in \mathcal{F}_d \times \mathcal{E}_d \mid \\
& \begin{bmatrix} f_{p_t} \\ f_{p_r} \\ f_{\epsilon_t} \\ f_{\epsilon_r} \end{bmatrix} = - \begin{bmatrix} 0 & 0 & \mathrm{d} & 0 \\ 0 & 0 & * & \mathrm{d} \\ \mathrm{d} & -* & 0 & 0 \\ 0 & \mathrm{d} & 0 & 0 \end{bmatrix} \begin{bmatrix} e_{p_t} \\ e_{p_r} \\ e_{\epsilon_t} \\ e_{\epsilon_r} \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} f_d^t \\ f_d^r \end{bmatrix}, \\
& \begin{bmatrix} e_d^t \\ e_d^r \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} e_{p_t} \\ e_{p_r} \\ e_{\epsilon_t} \\ e_{\epsilon_r} \end{bmatrix}, \quad \begin{bmatrix} f_b^t \\ f_b^r \\ e_b^t \\ e_b^r \end{bmatrix} = \begin{bmatrix} e_{p_t} \mid_{\partial \mathcal{D}} \\ e_{p_r} \mid_{\partial \mathcal{D}} \\ e_{\epsilon_t} \mid_{\partial \mathcal{D}} \\ e_{\epsilon_r} \mid_{\partial \mathcal{D}} \end{bmatrix} \Bigg\},
\end{aligned}
\tag{2.20}
$$

where $|_{\partial \mathcal{D}}$ *denotes the restriction on the border of the (spatial) domain* $\mathcal{D}$. *Then* $\mathbb{D}_d = \mathbb{D}_d{}^{\perp}$*; that is,* $\mathbb{D}_d$ *is a Dirac structure.*

Proof. The proof is very similar to the one given for Proposition 2.3. □

The dpH formulation of the Timoshenko beam with boundary and distributed energy flow can be obtained simply by combining the Dirac structure $\mathbb{D}_d$ (2.20) with (2.15) and (2.16). The resulting model is given in the following.

DEFINITION 2.7. *The dpH model of the Timoshenko beam with Dirac structure* $\mathbb{D}_d$ (2.20) *and Hamiltonian* $\mathcal{H}$ (2.8) *is given by*

$$
\begin{aligned}
& \begin{bmatrix} \partial_t p_t \\ \partial_t p_r \\ \partial_t \epsilon_t \\ \partial_t \epsilon_r \end{bmatrix} = \begin{bmatrix} 0 & 0 & \mathrm{d} & 0 \\ 0 & 0 & * & \mathrm{d} \\ \mathrm{d} & -* & 0 & 0 \\ 0 & \mathrm{d} & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_{p_t} \mathcal{H} \\ \delta_{p_r} \mathcal{H} \\ \delta_{\epsilon_t} \mathcal{H} \\ \delta_{\epsilon_r} \mathcal{H} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} f_d^t \\ f_d^r \end{bmatrix}, \\
& \begin{bmatrix} e_d^t \\ e_d^r \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_{p_t} \mathcal{H} \\ \delta_{p_r} \mathcal{H} \\ \delta_{\epsilon_t} \mathcal{H} \\ \delta_{\epsilon_r} \mathcal{H} \end{bmatrix}, \quad \begin{bmatrix} f_b^t \\ f_b^r \\ e_b^t \\ e_b^r \end{bmatrix} = \begin{bmatrix} \delta_{p_t} \mathcal{H} \mid_{\partial \mathcal{D}} \\ \delta_{p_r} \mathcal{H} \mid_{\partial \mathcal{D}} \\ \delta_{\epsilon_t} \mathcal{H} \mid_{\partial \mathcal{D}} \\ \delta_{\epsilon_r} \mathcal{H} \mid_{\partial \mathcal{D}} \end{bmatrix}.
\end{aligned}
\tag{2.21}
$$

The energy balance equation (2.18) becomes

$$
\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} = \int_{\partial \mathcal{D}} \left(f_b^t \wedge e_b^t + f_b^r \wedge e_b^r \right) + \int_{\mathcal{D}} \left(f_d^t \wedge e_d^t + f_d^r \wedge e_d^r \right),
\tag{2.22}
$$

which expresses the fact that the variation of internal stored energy equals the power supplied to the system through the border and the distributed port. From a *bond graph* point of view, the Timoshenko beam can be described as in Figure 2.1, where the power flows through the border ($f_b^{t,r} \mid_{x=\cdot}, e_b^{t,r} \mid_{x=\cdot}$) and the distributed port (f_d, e_d) are shown.

3. Control by damping injection.

3.1. Introduction. In this section, some considerations about control by damping injection applied to the Timoshenko beam are presented. In order to be as general as possible, consider the dpH formulation of the Timoshenko beam with distributed

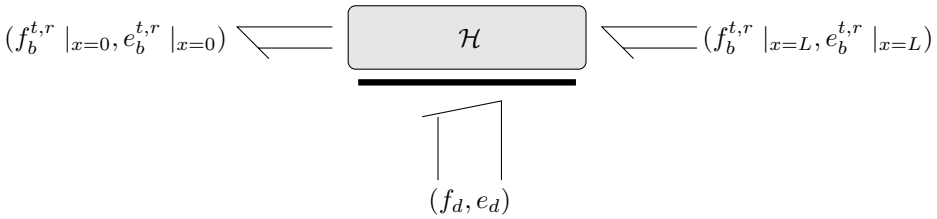


FIG. 2.1. *Bond graph representation of the Timoshenko beam.*

port (2.21). The energy functional (2.8) assumes its minimum in the *zero* configuration, i.e., when

$$p_t = 0, \qquad p_r = 0, \qquad \epsilon_t = 0, \quad \text{and} \quad \epsilon_r = 0 \tag{3.1}$$

or, equivalently, when

$$w(t,x) = \alpha^* x + d^*, \qquad \phi(t,x) = \alpha^*, \tag{3.2}$$

where the constants α^* and d^* are determined by the boundary conditions on w and ϕ. In (3.2), α^* represents the rotation angle of the beam around the point $x = 0$, while d^* is the vertical displacement in $x = 0$.

If some dissipation effect is introduced by means of a controller, it is possible to drive the state of the beam to the configuration where the (open loop) energy functional (2.8) assumes its minimum. If the controller is interconnected on the boundary of the spatial domain, we can speak about *boundary control* of the distributed parameter system (more precisely, about damping injection through the boundary). If the controller is interconnected along the distributed port, we can speak about *distributed control* of the infinite dimensional system (distributed damping injection). Energy dissipation can be introduced by *terminating* these ports with a *dissipative element*, i.e., by a generalized *impedance*, simulated by the control algorithm.

In order to simplify some stability proofs that are presented in the remaining part of this section, it is important to characterize the behavior of the Timoshenko beam equation when the energy function becomes constant and when the boundary conditions are equal to zero. We give this important remark (see [9, Assumption 5.12]).

Remark 3.1. Consider the dpH model of the Timoshenko beam (2.21). The only invariant solution compatible with $\dot{\mathcal{H}} = 0$ and with the boundary conditions

$$\begin{cases} f_b^t(0) = f_b^r(0) = 0, \\ e_b^t(L) = e_b^r(L) = 0 \end{cases} \quad \text{or} \quad \begin{cases} f_b^t(L) = f_b^r(L) = 0, \\ e_b^t(0) = e_b^r(0) = 0 \end{cases}$$

is the zero solution (3.1).

Note 3.1. More precisely, Remark 3.1 should be extended in order to also contain information about the *observability* of the infinite dimensional system (the Timoshenko beam, in this case), as discussed in [3, from page 154 onward]. These conditions can be interpreted as the generalization of the definition of detectability and observability (see [1]) to the infinite dimensional case (see also [3, pp. 227ff.]).

3.2. Boundary control. Suppose that a finite dimensional controller can be interconnected to the beam in $x = L$ and that the beam can interact with the environment in $x = 0$. Moreover, suppose that no interaction can take place through the

distributed port. The last hypothesis means that, in (2.21), it can be assumed that

$$f_d^t(t,x) = 0, \qquad f_d^r(t,x) = 0.$$

The controller is designed in order to act as if a dissipative element is connected to the power port of the beam in $x = L$, whose causality is represented in Figure 2.1. Dissipation can be introduced if it is possible to impose the following relation between flow and effort in $x = L$:

$$\left\{ \begin{array}{l} f_b^t(t,L) = -b^t(t) * e_b^t(t,L), \\ \\ f_b^r(t,L) = -b^r(t) * e_b^r(t,L) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \dfrac{1}{\rho} * p_t(L) = -b^t(\cdot) * K * \epsilon_t(L), \\ \\ \dfrac{1}{I_\rho} * p_r(L) = -b^r(\cdot) * EI * \epsilon_r(L) \end{array} \right. \tag{3.3}$$

with $b^t, b^r > 0$ functions of time t.

In this way, the energy balance equation (2.22) becomes

$$\begin{aligned} \frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t}(t) &= -\int_{x=L} \left[b^t(t)\, e_b^t \wedge *e_b^t + b^r(t)\, e_b^r \wedge *e_b^r \right] + \int_{x=0} \left(e_b^t \wedge f_b^t + e_b^r \wedge f_b^r \right) \\ &= -b^t(t) \left[K * \epsilon_t \mid_{x=L} \right]^2 - b^r(t) \left[EI * \epsilon_r \mid_{x=L} \right]^2 \\ &\quad + \left[e_b^t(t,0) f_b^t(t,0) + e_b^r(t,0) f_b^r(t,0) \right]. \end{aligned} \tag{3.4}$$

If, for example, the boundary conditions in $x = 0$ are

$$w(t,0) = 0, \qquad \phi(t,0) = 0, \tag{3.5a}$$

and, consequently,

$$f_b^t(t,0) = f_b^r(t,0) = 0, \tag{3.5b}$$

then (3.4) becomes

$$\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t}(t) = -b^t(t) \left[K * \epsilon_t \mid_{x=L} \right]^2 - b^r(t) \left[EI * \epsilon_r \mid_{x=L} \right]^2 \leq 0.$$

So it is possible to state the following proposition [8].

PROPOSITION 3.1. *Consider the dpH model of the Timoshenko beam* (2.21)*, and suppose that the boundary conditions in* $x = 0$ *are given by* (3.5) *and that the controller* (3.3) *is interconnected to the beam in* $x = L$*. Then the final configuration is* (3.2)*, with* $\alpha^* = 0$ *and* $d^* = 0$*; that is,*

$$w(t,x) = 0, \qquad \phi(t,x) = 0.$$

Proof. The proof follows from Remark 3.1 and the LaSalle theorem generalized to infinite dimensions (see [9, section 3.7]). Furthermore, it is necessary that $\alpha^* = 0$ and $d^* = 0$ in (3.2), in order to be compatible with the boundary conditions (3.5a). ∎

Note 3.2. These results were already presented in [8] using a different approach. The proposed control law was written in the following form:

$$\begin{aligned} \frac{\partial w}{\partial t}(t,L) &= -b^t(t) \cdot K \left[\frac{\partial w}{\partial x}(t,L) - \phi(t,L) \right], \\ \frac{\partial \phi}{\partial t}(t,L) &= -b^r(t) \cdot EI \frac{\partial \phi}{\partial x}(t,L), \end{aligned}$$

which is clearly equivalent to (3.3). The main advantage in approaching the problem within the framework of dpH systems is that both the way the control law is deduced and the proof of its stability can be presented in a more intuitive (in some sense *physical*) and elegant way. The same considerations hold for the distributed control of the beam by damping injection presented in the next subsection: in this case, the same results were already presented in [5], but with a different approach.

3.3. Distributed control. Following the same ideas presented in the previous section, it is possible to extend the control by damping injection to the case in which the interaction between system and controller takes place through a distributed port. In this case, the (distributed) power port has to be terminated by a desired impedance implemented by a *distributed* controller. In other words, in this section it is shown how to stabilize the Timoshenko beam with a locally distributed control based on an extension to the infinite dimensional case of the damping injection control technique.

Assume that $b_d^t(t,x)$ and $b_d^r(t,x)$ are smooth functions on $\mathcal{D}$, and suppose that it is possible to find $\bar{\mathcal{D}} \subset \mathcal{D}$ and $b_0 > 0$ such that $b_d^t(\cdot,x), b_d^r(\cdot,x) \geq b_0 > 0$ if $x \in \bar{\mathcal{D}} \subset \mathcal{D}$. By taking into account the causality of the distributed port illustrated in Figure 2.1, the dissipation effects can be introduced through the distributed port if the *controller* can impose the following relation between flows and efforts on $\mathcal{D}$:

$$\text{(3.6a)} \qquad \begin{cases} f_d^t = -b_d^t * e_d^t, \\ f_d^r = -b_d^r * e_d^r. \end{cases}$$

This relation can be equivalently written as

$$\text{(3.6b)} \qquad \begin{cases} f_d^t = -\dfrac{b_d^t}{\rho}\, p_t, \\ f_d^r = -\dfrac{b_d^r}{I_\rho}\, p_r, \end{cases}$$

and, clearly, the closed-loop system is described by the following set of PDEs:

$$\begin{aligned} \rho \frac{\partial^2 w}{\partial t^2} - K\left(\frac{\partial^2 w}{\partial x^2} - \frac{\partial \phi}{\partial x}\right) + b_d^t \frac{\partial w}{\partial t} &= 0, \\ I_\rho \frac{\partial^2 \phi}{\partial t^2} - EI \frac{\partial^2 \phi}{\partial x^2} + K\left(\frac{\partial w}{\partial x} - \phi\right) + b_d^r \frac{\partial \phi}{\partial t} &= 0, \end{aligned}$$

in which the boundary conditions still have to be specified. Moreover, the energy balance (2.22) becomes

$$\text{(3.7)} \qquad \begin{aligned} \frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} &= \int_{\partial\mathcal{D}} \left(e_b^t \wedge f_b^t + e_b^r \wedge f_b^r\right) + \int_{\bar{\mathcal{D}}} \left(e_d^t \wedge f_d^t + e_d^r \wedge f_d^r\right) \\ &= \int_{\partial\mathcal{D}} \left(e_b^t \wedge f_b^t + e_b^r \wedge f_b^r\right) - \int_{\bar{\mathcal{D}}} \left(b_d^t\, e_d^t \wedge *e_d^t + b_d^r\, e_d^r \wedge *e_d^r\right). \end{aligned}$$

Assume, for simplicity, that the beam is clamped in $x = 0$, that is,

$$\text{(3.8a)} \qquad w(t,0) = 0 \qquad \text{and} \qquad \phi(t,0) = 0,$$

and that there is no force/torque acting on $x = L$. Moreover, the boundary conditions, i.e., the values assumed by the power variables on the (power) ports on $\partial\mathcal{D}$, are given

by

$$\text{(3.8b)} \qquad \begin{cases} f_b^t(t,0) = 0, \\ f_b^r(t,0) = 0, \end{cases} \qquad \begin{cases} e_b^t(t,L) = 0, \\ e_b^r(t,L) = 0. \end{cases}$$

From (3.7), the energy balance relation (2.22) becomes

$$\text{(3.9)} \qquad \begin{aligned} \frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} &= -\int_{\bar{\mathcal{D}}} \left(b_d^t \, e_d^t \wedge *e_d^t + b_d^r \, e_d^r \wedge *e_d^r \right) \\ &= -\int_{\bar{\mathcal{D}}} \left[\frac{1}{b_d^t} \left(\frac{\partial w}{\partial t} \right)^2 + \frac{1}{b_d^r} \left(\frac{\partial \phi}{\partial t} \right)^2 \right] \mathrm{d}x \leq 0. \end{aligned}$$

So it is possible to state the following proposition.

PROPOSITION 3.2. *Consider the dpH system of the Timoshenko beam with distributed port* (2.21) *and suppose that the boundary conditions are given by* (3.5). *Then the distributed control action* (3.3) *asymptotically stabilizes the system in*

$$w(t,x) = 0 \qquad \textit{and} \qquad \phi(t,x) = 0.$$

Proof. From (3.9), we have that $\dot{\mathcal{H}} = 0$ if $\epsilon_t = \epsilon_r = 0$ and $p_t = p_r = 0$ on $\bar{\mathcal{D}}$. Consequently, from Proposition 3.1 and from the boundary conditions (3.8), we deduce that also on $\mathcal{D} \setminus \bar{\mathcal{D}}$ we have $\epsilon_t = \epsilon_r = 0$ and $p_t = p_r = 0$. The only configuration compatible with this *energy* configuration and the boundary conditions (3.8a) is clearly $w(t,x) = 0$ and $\phi(t,x) = 0$. ∎

Note 3.3. It is important to underscore that the most difficult points in the analysis of the stability of the proposed control schemes are the proof of Remark 3.1, which characterizes the invariant solutions of the Timoshenko beam equations for zero boundary conditions, and the verification of the applicability of LaSalle theorem. More details on these problems and the rigorous way to solve them can be found in [3, Chapter 5] and [9, Chapters 3 and 5].

4. Control by interconnection and energy shaping.

4.1. Model of the plant. Consider the mechanical system of Figure 4.1, in which a flexible beam, modeled according to the Timoshenko theory and whose dpH model is given by (2.17), is connected to a rigid body with mass m and inertia momentum J in $x = L$ and to a controller in $x = 0$. The controller acts on the system with a force f_c and a torque τ_c. Since the Timoshenko model of the beam is valid only for small deformations, it is possible to assume that the motion of the rigid body is the combination of a rotational and a translational motion along $x = L$. The port Hamiltonian model of the rigid body is given by

$$\text{(4.1)} \qquad \begin{aligned} \begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} &= \left(\begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & D \end{bmatrix} \right) \begin{bmatrix} \partial_q H \\ \partial_p H \end{bmatrix} + \begin{bmatrix} 0 \\ I \end{bmatrix} \mathbf{f}, \\ \mathbf{e} &= \partial_p H, \end{aligned}$$

where $q = [q_1,\ q_2]^T \in Q \subset \mathbb{R}^2$ are the generalized coordinates, with q_1 the distance from the equilibrium configuration and q_2 the rotation angle, p are the generalized momenta, $\mathbf{f}, \mathbf{e} \in \mathbb{R}^2$ are the port variables,

$$\text{(4.2)} \qquad H(q,p) := \frac{1}{2} \left(\frac{p_1^2}{m} + \frac{p_2^2}{J} \right) + V(q)$$

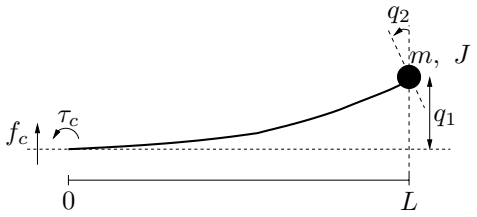


FIG. 4.1. *Flexible link with mass in $x = L$.*

FIG. 4.2. *Bond graph representation of the closed-loop system.*

is the total energy (Hamiltonian) function, with V the potential, and $D = D^{\mathrm{T}} \geq 0$ is a matrix taking into account energy dissipation.

As regards the controller, we assume that it can be modeled by means of the following finite dimensional port Hamiltonian systems:

$$\begin{aligned} \begin{bmatrix} \dot{q}_c \\ \dot{p}_c \end{bmatrix} &= \left(\begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & D_c \end{bmatrix} \right) \begin{bmatrix} \partial_{q_c} H_c \\ \partial_{p_c} H_c \end{bmatrix} + \begin{bmatrix} 0 \\ G_c \end{bmatrix} \mathbf{f}_c, \\ \mathbf{e}_c &= G_c^T \partial_{p_c} H_c, \end{aligned} \tag{4.3}$$

where $q_c \in Q_c \subset \mathbb{R}^2$ are the generalized coordinates, with $\dim(Q_c) = 2$, p_c are the generalized momenta, $\mathbf{f}_c, \mathbf{e}_c \in \mathbb{R}^2$ are the power conjugated port variables, and $D_c = D_c^{\mathrm{T}} \geq 0$ is a matrix taking into account energy dissipation. Moreover, $H_c(q_c, p_c)$ is the Hamiltonian, and it will be specified in the remaining part of this section in order to drive the whole system in a desired equilibrium configuration. Note that $\mathbf{f}_c = [f_c,\ \tau_c]^{\mathrm{T}}$.

The port causality of both the mass and the controller is assumed to be with flows as inputs and efforts as outputs. As pointed out in [23], it is possible to interconnect two port Hamiltonian systems only if a port dualization is applied on one of them. In this way, a system can have an effort as input and a flow as output. Since the port causality and orientation of the beam are given in Figure 2.1, the bond graph representation of the closed-loop system made of the Timoshenko beam, the mass in $x = L$, and the finite dimensional port Hamiltonian controller acting in $x = 0$ are given in Figure 4.2. Then, the interconnection constraints between the port variables of the subsystems are given by the following power-preserving relations:

$$\begin{cases} \begin{bmatrix} f_b^t(L) & f_f^r(L) \end{bmatrix}^{\mathrm{T}} = -\mathbf{e}, \\ \begin{bmatrix} e_b^t(L) & e_b^r(L) \end{bmatrix}^{\mathrm{T}} = \mathbf{f}, \end{cases} \qquad \begin{cases} \begin{bmatrix} f_b^t(0) & f_f^r(0) \end{bmatrix}^{T} = \mathbf{e}_c, \\ \begin{bmatrix} e_b^t(0) & e_b^r(0) \end{bmatrix}^{T} = \mathbf{f}_c. \end{cases} \tag{4.4}$$

From (2.17), (4.1), (4.3), and (4.4), it is possible to obtain the mixed finite and infinite dimensional port Hamiltonian (m-pH) representation of the closed-loop system. The total energy $\mathcal{H}_{cl}$ is defined in the extended space

$$\mathcal{X}_{cl} := \underbrace{Q \times T^*Q \times Q_c \times T^*Q_c}_{\mathcal{X}} \times \underbrace{\Omega^1(\mathcal{D}) \times \Omega^1(\mathcal{D}) \times \Omega^1(\mathcal{D}) \times \Omega^1(\mathcal{D})}_{\mathcal{X}_\infty} \tag{4.5}$$

and is given by the sum of the energy functions of the subsystems, that is,

$$\mathcal{H}_{cl} := H + H_c + \mathcal{H}. \tag{4.6}$$

Moreover, it is easy to verify that the energy rate is equal to

$$\frac{\mathrm{d}\mathcal{H}_{cl}}{\mathrm{d}t} = -\left(\frac{\partial^{\mathrm{T}}\mathcal{H}_{cl}}{\partial p} D \frac{\partial \mathcal{H}_{cl}}{\partial p} + \frac{\partial^{\mathrm{T}}\mathcal{H}_{cl}}{\partial p_c} D_c \frac{\partial \mathcal{H}_{cl}}{\partial p_c}\right),$$

where D_c and H_c have to be designed in order to drive the system in the desired equilibrium position, which is still to be specified. Following the same procedure presented in [20, 26] for finite dimensional port Hamiltonian systems, the idea is to *shape* the total energy $\mathcal{H}_{cl}$ by properly choosing the controller Hamiltonian H_c in order to have a new minimum of energy in the desired configuration that can be reached if some dissipative effect is introduced. The first step is to find the Casimir functionals of the closed-loop system.

4.2. Casimir functionals for the closed-loop system. The applicability of the control by interconnection and energy shaping relies on the possibility of relating the controller state variables to the state variables of the plant by means of Casimir functions [13]. Equivalently, we can say that the controller structure is chosen in order to constrain the closed-loop trajectory to evolve on a particular submanifold of the whole state space. The key point is, then, to find necessary and sufficient conditions on the existence of Casimir functions for a given dynamical system. Since a Casimir function is a structural invariant, that is, a scalar function defined on the state space of a dynamical system which is constant along its trajectories independently from the Hamiltonian function (see [20] and [26, p. 87]), a possible generalization can be given by means of the following definition [10].

DEFINITION 4.1 (Casimir functionals). *Consider a scalar function* $\mathcal{C} : \mathcal{X}_{cl} \to \mathbb{R}$ *defined on the extended state space* (4.5). *Then,* $\mathcal{C}$ *is a Casimir functional for the m-pH of Figure* 2.1 *if and only if*

$$\frac{\mathrm{d}\mathcal{C}}{\mathrm{d}t} = 0 \qquad \forall \mathcal{H}_{cl} : \mathcal{X}_{cl} \to \mathbb{R},$$

where $\mathcal{H}_{cl}$ *has the structure given in* (4.6).

The definition is quite general. In the case under study, we have that

$$\begin{aligned}\frac{\mathrm{d}\mathcal{C}}{\mathrm{d}t} &= \frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial q}\dot{q} + \frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial p}\dot{p} + \frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial q_c}\dot{q}_c + \frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial p_c}\dot{p}_c \\ &\quad + \int_{\mathcal{D}} \left(\frac{\partial p_t}{\partial t} \wedge \delta_{p_t}\mathcal{C} + \frac{\partial p_r}{\partial t} \wedge \delta_{p_r}\mathcal{C} + \frac{\partial \epsilon_t}{\partial t} \wedge \delta_{\epsilon_t}\mathcal{C} + \frac{\partial \epsilon_r}{\partial t} \wedge \delta_{\epsilon_r}\mathcal{C}\right)\end{aligned}$$

and, from (2.17), (4.1), (4.3), and the interconnection constraints (4.4), we obtain

$$\begin{aligned}\frac{\mathrm{d}\mathcal{C}}{\mathrm{d}t} &= \frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial q}\frac{\partial H}{\partial p} + \frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial p}\left\{-\frac{\partial H}{\partial q} - D\frac{\partial H}{\partial p} + [\delta_{\epsilon_t}\mathcal{H}\mid_{x=L} \quad \delta_{\epsilon_r}\mathcal{H}\mid_{x=L}]^{\mathrm{T}}\right\} \\ &\quad + \frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial q_c}\frac{\partial H_c}{\partial p_c} + \frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial p_c}\left\{-\frac{\partial H_c}{\partial q_c} - D_c\frac{\partial H_c}{\partial p_c} + G_c[\delta_{\epsilon_t}\mathcal{H}\mid_{x=0} \quad \delta_{\epsilon_r}\mathcal{H}\mid_{x=0}]^{\mathrm{T}}\right\} \\ &\quad + \int_{\mathcal{D}} [\mathrm{d}\delta_{\epsilon_t}\mathcal{H} \wedge \delta_{p_t}\mathcal{C} + (\mathrm{d}\delta_{\epsilon_r}\mathcal{H} + *\delta_{\epsilon_t}\mathcal{H}) \wedge \delta_{p_r}\mathcal{C}] \\ &\quad + \int_{\mathcal{D}} [(\mathrm{d}\delta_{p_t}\mathcal{H} - *\delta_{p_r}\mathcal{H}) \wedge \delta_{\epsilon_r}\mathcal{C} + \mathrm{d}\delta_{p_r}\mathcal{H} \wedge \delta_{\epsilon_t}\mathcal{C}].\end{aligned} \tag{4.7}$$

The integral term in (4.7) is equal to

$$\begin{aligned}
&\int_{\mathcal{D}} \left[\mathrm{d}\left(\delta_{\epsilon_t}\mathcal{H}\wedge\delta_{p_t}\mathcal{C}\right)+\mathrm{d}\left(\delta_{\epsilon_r}\mathcal{H}\wedge\delta_{p_r}\mathcal{C}\right)+\mathrm{d}\left(\delta_{p_t}\mathcal{H}\wedge\delta_{\epsilon_t}\mathcal{C}\right)+\mathrm{d}\left(\delta_{p_r}\mathcal{H}\wedge\delta_{\epsilon_r}\mathcal{C}\right)\right]\\
&\qquad -\int_{\mathcal{D}}\left[\delta_{p_t}\mathcal{H}\wedge\mathrm{d}\delta_{\epsilon_t}\mathcal{C}+\delta_{p_r}\mathcal{H}\wedge\left(\mathrm{d}\delta_{\epsilon_r}\mathcal{C}+*\delta_{\epsilon_t}\mathcal{C}\right)\right]\\
&\qquad\qquad -\int_{\mathcal{D}}\left[\delta_{\epsilon_t}\mathcal{H}\wedge\left(\mathrm{d}\delta_{p_t}\mathcal{C}-*\delta_{p_r}\mathcal{C}\right)+\delta_{\epsilon_r}\mathcal{H}\wedge\mathrm{d}\delta_{p_r}\mathcal{C}\right],
\end{aligned}\tag{4.8}$$

where, from the Stokes theorem, the first term can be written as

$$\begin{aligned}
&\int_{\partial\mathcal{D}}\left[\delta_{\epsilon_t}\mathcal{H}\mid_{\partial\mathcal{D}}\wedge\delta_{p_t}\mathcal{C}\mid_{\partial\mathcal{D}}+\cdots+\delta_{p_r}\mathcal{H}\mid_{\partial\mathcal{D}}\wedge\delta_{\epsilon_r}\mathcal{C}\mid_{\partial\mathcal{D}}\right]\\
&\qquad=\int_{\partial\mathcal{D}}\begin{bmatrix}\delta_{p_t}\mathcal{C}\mid_{\partial\mathcal{D}} & \delta_{p_r}\mathcal{C}\mid_{\partial\mathcal{D}}\end{bmatrix}\begin{bmatrix}\delta_{\epsilon_t}\mathcal{H}\mid_{\partial\mathcal{D}} & \delta_{\epsilon_r}\mathcal{H}\mid_{\partial\mathcal{D}}\end{bmatrix}^{\mathrm{T}}\\
&\qquad\quad+\int_{\partial\mathcal{D}}\begin{bmatrix}\delta_{\epsilon_t}\mathcal{C}\mid_{\partial\mathcal{D}} & \delta_{\epsilon_r}\mathcal{C}\mid_{\partial\mathcal{D}}\end{bmatrix}\begin{bmatrix}\delta_{p_t}\mathcal{H}\mid_{\partial\mathcal{D}} & \delta_{p_r}\mathcal{H}\mid_{\partial\mathcal{D}}\end{bmatrix}^{\mathrm{T}}.
\end{aligned}\tag{4.9}$$

From (4.1) and (4.3) and the interconnection constraints (4.4), we have that

$$\begin{bmatrix}\delta_{p_t}\mathcal{H}\mid_{x=L}\\ \delta_{p_r}\mathcal{H}\mid_{x=L}\end{bmatrix}=-e=-\frac{\partial H}{\partial p}\quad\text{and}\quad\begin{bmatrix}\delta_{p_t}\mathcal{H}\mid_{x=0}\\ \delta_{p_r}\mathcal{H}\mid_{x=0}\end{bmatrix}=e_c=G_c^{\mathrm{T}}\frac{\partial H_c}{\partial p_c}.$$

Then, combining (4.7) with (4.8) and (4.9), we obtain that

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{C}}{\mathrm{d}t} \;=\; & -\frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial p}\frac{\partial H}{\partial q}-\frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial p_c}\frac{\partial H_c}{\partial q_c}\\
&-\left\{-\frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial q}+\frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial p}D+\begin{bmatrix}\delta_{\epsilon_t}\mathcal{C}\mid_{x=L} & \delta_{\epsilon_r}\mathcal{C}\mid_{x=L}\end{bmatrix}\right\}\frac{\partial H}{\partial p}\\
&-\left\{-\frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial q_c}+\frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial p_c}D_c+\begin{bmatrix}\delta_{\epsilon_t}\mathcal{C}\mid_{x=0} & \delta_{\epsilon_r}\mathcal{C}\mid_{x=0}\end{bmatrix}G_c^{\mathrm{T}}\right\}\frac{\partial H_c}{\partial p_c}\\
&+\left\{\frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial p}+\begin{bmatrix}\delta_{p_t}\mathcal{C}\mid_{x=L} & \delta_{p_r}\mathcal{C}\mid_{x=L}\end{bmatrix}\right\}\begin{bmatrix}\delta_{\epsilon_t}\mathcal{H}\mid_{x=L} & \delta_{\epsilon_r}\mathcal{H}\mid_{x=L}\end{bmatrix}^{\mathrm{T}}\\
&+\left\{\frac{\partial^{\mathrm{T}}\mathcal{C}}{\partial p_c}-\begin{bmatrix}\delta_{p_t}\mathcal{C}\mid_{x=0} & \delta_{p_r}\mathcal{C}\mid_{x=0}\end{bmatrix}\right\}\begin{bmatrix}\delta_{\epsilon_t}\mathcal{H}\mid_{x=0} & \delta_{\epsilon_r}\mathcal{H}\mid_{x=0}\end{bmatrix}^{\mathrm{T}}\\
&-\int_{\mathcal{D}}\left[\delta_{p_t}\mathcal{H}\wedge\mathrm{d}\delta_{\epsilon_t}\mathcal{C}+\delta_{p_r}\mathcal{H}\wedge\left(\mathrm{d}\delta_{\epsilon_r}\mathcal{C}+*\delta_{\epsilon_t}\mathcal{C}\right)\right]\\
&-\int_{\mathcal{D}}\left[\delta_{\epsilon_t}\mathcal{H}\wedge\left(\mathrm{d}\delta_{p_t}\mathcal{C}-*\delta_{p_r}\mathcal{C}\right)+\delta_{\epsilon_r}\mathcal{H}\wedge\mathrm{d}\delta_{p_r}\mathcal{C}\right]
\end{aligned}$$

has to be equal to zero for every Hamiltonian H, H_c, and $\mathcal{H}$ (see Definition 4.1). This is true if and only if

$$\begin{aligned}
&\mathrm{d}\delta_{\epsilon_t}\mathcal{C}=0, && \mathrm{d}\delta_{p_t}\mathcal{C}-*\delta_{p_r}\mathcal{C}=0,\\
&\mathrm{d}\delta_{\epsilon_r}\mathcal{C}+*\delta_{\epsilon_t}\mathcal{C}=0, && \mathrm{d}\delta_{p_r}\mathcal{C}=0,\\
&\frac{\partial\mathcal{C}}{\partial p}=0,\quad \frac{\partial\mathcal{C}}{\partial p_c}=0,\quad \begin{bmatrix}\delta_{p_t}\mathcal{C}\mid_{x=L}\\ \delta_{p_r}\mathcal{C}\mid_{x=L}\end{bmatrix}=0,\quad \begin{bmatrix}\delta_{p_t}\mathcal{C}\mid_{x=0}\\ \delta_{p_r}\mathcal{C}\mid_{x=0}\end{bmatrix}=0,\\
&\frac{\partial\mathcal{C}}{\partial q}=\begin{bmatrix}\delta_{\epsilon_t}\mathcal{C}\mid_{x=L}\\ \delta_{\epsilon_r}\mathcal{C}\mid_{x=L}\end{bmatrix}, && \frac{\partial\mathcal{C}}{\partial q_c}=G_c\begin{bmatrix}\delta_{\epsilon_t}\mathcal{C}\mid_{x=0}\\ \delta_{\epsilon_r}\mathcal{C}\mid_{x=0}\end{bmatrix}.
\end{aligned}\tag{4.10}$$

In other words, the following proposition has been proved [11, 12].

PROPOSITION 4.2. *Consider the m-pH system of Figure* 4.2*, that is, the result of the power conserving interconnection* (4.4) *of the subsystems* (2.17), (4.1)*, and* (4.3)*. If* $\mathcal{X} \times \mathcal{X}_\infty$ *is the extended state space of the system, introduced in* (4.5)*, then a functional* $\mathcal{C} : \mathcal{X} \times \mathcal{X}_\infty \to \mathbb{R}$ *is a Casimir for the closed-loop system if and only if conditions* (4.10) *hold.*

Since the necessary and sufficient conditions for the existence of Casimir functions have been deduced, the control problem can be approached.

4.3. Control by energy shaping of the Timoshenko beam. In order to control the flexible beam with the finite dimensional controller (4.3), the first step is to find Casimir functionals for the closed-loop system that can relate the state variables of the controller q to the state *variables* that describe the configuration of the flexible beam and the mass connected to its extremity. In particular, we are looking for some functionals $\tilde{\mathcal{C}}_i$, $i = 1, 2$, such that

$$\mathcal{C}_i(q, p, q_c, p_c, p_t, p_r, \epsilon_t, \epsilon_r) := q_{c,i} - \tilde{\mathcal{C}}_i(q, p, p_c, p_t, p_r, \epsilon_t, \epsilon_r) \quad \text{with} \quad i = 1, 2$$

are Casimir functionals for the closed-loop system, i.e., satisfying the conditions of Proposition 4.2.

First of all, from (4.10), we note immediately that every Casimir functional cannot depend on p and p_c. Moreover, since it is necessary that $\mathrm{d}\delta_{\epsilon_t}\mathcal{C}_i = 0$ and $\mathrm{d}\delta_{p_r}\mathcal{C}_i = 0$, we deduce that $\delta_{\epsilon_t}\mathcal{C}_i$ and $\delta_{p_r}\mathcal{C}_i$ have to be constant as a function on x on $\mathcal{D}$ and their value will be determined by the boundary conditions on $\mathcal{C}_i$. Since, from (4.3), $\delta_{p_r}\mathcal{C}_i \,|_{\partial\mathcal{D}} = 0$, we deduce that $\delta_{p_r}\mathcal{C}_i = 0$ on $\mathcal{D}$. Since $\mathrm{d}\delta_{p_t}\mathcal{C}_i = *\delta_{p_r}\mathcal{C}_i = 0$, then, from the boundary conditions, we deduce that also $\delta_{p_t}\mathcal{C}_i = 0$ on $\mathcal{D}$. As a consequence, all the admissible Casimir functionals are also independent from p_t and p_r. In other words, we are interested in finding Casimir functionals in the following form:

$$\mathcal{C}_i(q, q_c, \epsilon_t, \epsilon_r) := q_{c,i} - \tilde{\mathcal{C}}_i(q, \epsilon_t, \epsilon_r), \qquad i = 1, 2.$$

Assuming $G_c = I$, we have that

$$\frac{\partial \mathcal{C}_1}{\partial q_c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \delta_{\epsilon_t}\mathcal{C}_1 \,|_{x=0} \\ \delta_{\epsilon_r}\mathcal{C}_1 \,|_{x=0} \end{bmatrix} \tag{4.11}$$

and, consequently, $\delta_{\epsilon_t}\mathcal{C}_1 = 1$ on $\mathcal{D}$. From (4.10), we have that $\mathrm{d}\delta_{\epsilon_r}\mathcal{C}_1 = -*\delta_{\epsilon_t}\mathcal{C}_1 = -*1 = -\mathrm{d}x$; then, $\delta_{\epsilon_r}\mathcal{C}_1 = -x + c_1$, where c_1 is determined by the boundary conditions. Since, from (4.11), $\delta_{\epsilon_r}\mathcal{C}_1 \,|_{x=0} = 0$, then $c_1 = 0$; moreover, we deduce that $\delta_{\epsilon_r}\mathcal{C} \,|_{x=L} = -L$, i.e., a new boundary condition in $x = L$. A consequence is that

$$\frac{\partial \mathcal{C}_1}{\partial q} = \begin{bmatrix} \delta_{\epsilon_t}\mathcal{C}_1 \,|_{x=L} \\ \delta_{\epsilon_r}\mathcal{C}_1 \,|_{x=L} \end{bmatrix} = \begin{bmatrix} 1 \\ -L \end{bmatrix}.$$

The first conclusion is that

$$\mathcal{C}_1(q, q_c, \epsilon_t, \epsilon_r) = q_{c,1} - (Lq_2 - q_1) - \int_{\mathcal{D}} (x\epsilon_r - \epsilon_t) \tag{4.12}$$

is a Casimir for the closed-loop system. Following the same procedure, it is possible to calculate $\mathcal{C}_2$. From (4.10), we have that

$$\frac{\partial \mathcal{C}_2}{\partial q_c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \delta_{\epsilon_t}\mathcal{C}_2 \,|_{x=0} \\ \delta_{\epsilon_r}\mathcal{C}_2 \,|_{x=0} \end{bmatrix},$$

and then $\delta_{\epsilon_t}\mathcal{C}_2 = 0$ on $\mathcal{D}$; moreover, $\mathrm{d}\delta_{\epsilon_r}\mathcal{C}_2 = 0$ and, consequently, $\delta_{\epsilon_r}\mathcal{C}_2 = 1$ on $\mathcal{D}$ since (4.6) holds. Again from (4.10), we deduce that

$$\frac{\partial \mathcal{C}_2}{\partial q} = \begin{bmatrix} \delta_{\epsilon_t}\mathcal{C}_2 \mid_{x=L} \\ \delta_{\epsilon_r}\mathcal{C}_2 \mid_{x=L} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

So we can state that

$$\mathcal{C}_2(q, q_c, \epsilon_t, \epsilon_r) = q_{c,2} + q_2 + \int_{\mathcal{D}} \epsilon_r \tag{4.13}$$

is another Casimir functional for the closed-loop system. In conclusion, the following proposition has been proved [11, 12].

PROPOSITION 4.3. *Consider the m-pH system of Figure* 4.2*, that is, the result of the power conserving interconnection* (4.4) *of the subsystems* (2.17), (4.1)*, and* (4.3)*. Then* (4.12) *and* (4.13) *are Casimir functionals for this system.*

Note 4.1. Since $\mathcal{C}_i$, $i = 1, 2$, are Casimir functionals, they are invariant for the system of Figure 4.2. Then, for every energy function H_c of the controller, we have that

$$q_{c,1} = (Lq_2 - q_1) + \int_{\mathcal{D}} (x\epsilon_r - \epsilon_t) + C_1, \qquad q_{c,2} = -q_2 - \int_{\mathcal{D}} \epsilon_r + C_2, \tag{4.14}$$

where C_1 and C_2 depend on the initial conditions. If the initial configuration of the system is known, then it is possible to assume these constants are equal to zero. Since H_c is an arbitrary function of q_c, it is possible to *shape* the total energy function of the closed-loop system in order to have a minimum of energy in a desired configuration: if some dissipation effect is present, the new equilibrium configuration will be reached.

Suppose that the potential energy V in (4.2) is equal to

$$V(q_1, q_2) := \frac{1}{2}\left(k_1 q_1^2 + k_2 q_2^2\right) \tag{4.15}$$

with $k_1, k_2 > 0$. In other words, suppose that a translational and rotational spring is acting on the rigid body in $x = L$. Furthermore, suppose that $(q^*, 0)$, with $q^* = [q_1^* \; q_2^*]^{\mathrm{T}}$, is the desired equilibrium configuration of the mass (4.1). Then, the corresponding equilibrium configuration of the beam can be calculated as the solution of (2.17) with

$$\frac{\partial p_t}{\partial t} = \frac{\partial p_r}{\partial t} = \frac{\partial \epsilon_t}{\partial t} = \frac{\partial \epsilon_r}{\partial t} = 0 \quad \text{on} \quad \mathcal{D}$$

and with boundary conditions (in $x = L$) given by

$$\begin{bmatrix} f_b^t(L) \\ f_b^r(L) \end{bmatrix} = \frac{\partial H}{\partial p}(q^*, 0) = 0, \qquad \begin{bmatrix} e_b^t(L) \\ e_b^r(L) \end{bmatrix} = \frac{\partial H}{\partial q}(q^*, 0) = \begin{bmatrix} k_1 q_1^* \\ k_2 q_2^* \end{bmatrix}. \tag{4.16}$$

From (2.17), we have that the equilibrium configuration has to satisfy the following system of PDEs:

$$\begin{cases} \mathrm{d}\delta_{\epsilon_t}\mathcal{H} = 0, \\ *\delta_{\epsilon_t}\mathcal{H} + \mathrm{d}\delta_{\epsilon_r}\mathcal{H} = 0, \end{cases}$$

whose solution, compatible with the boundary conditions (4.16), is equal to

$$\begin{cases} {\epsilon_t}^*(x,t) = \dfrac{k_1}{K} q_1^*, \\ {\epsilon_r}^*(x,t) = \dfrac{k_1 q_1^*}{EI}(L-x) + \dfrac{k_2 q_2^*}{EI}. \end{cases} \tag{4.17}$$

Furthermore, at the equilibrium, it is easy to compute that $p_t = {p_t}^* = 0$ and that $p_r = {p_r}^* = 0$. From (4.14) and (4.17), define

$$\begin{aligned} q_{c,1}^* &= q_{c,1}(q_1^*, q_2^*, {\epsilon_t}^*, {\epsilon_r}^*) = L q_2^* - q_1^* + \int_0^L (x {\epsilon_r}^* - {\epsilon_t}^*)\,\mathrm{d}x \\ &= L q_2^* - q_1^* + \int_0^L \left[\frac{k_1 q_1^*}{EI}(L-x)x + \frac{k_2 q_2^*}{EI} x - \frac{k_1}{K} q_1^* \right] \mathrm{d}x \\ &= \left(\frac{k_1}{EI}\frac{L^3}{6} - \frac{k_1}{K} L - 1 \right) q_1^* + \left(\frac{k_2}{EI}\frac{L^2}{2} + L \right) q_2^*, \\ q_{c,2}^* &= q_{c,2}(q_2^*, {\epsilon_r}^*) = -q_2^* - \int_0^L \epsilon_r \mathrm{d}x \\ &= -q_2^* - \int_0^L \left[\frac{k_1 q_1^*}{EI}(L-x) + \frac{k_2 q_2^*}{EI} \right] \mathrm{d}x = -\frac{k_1}{EI}\frac{L^2}{2} q_1^* - \left(\frac{k_2}{EI} L + 1 \right) q_2^*. \end{aligned}$$

Note that, at the equilibrium, $p_c = p_c^* = 0$. The energy function H_c of the controller (4.3) will be developed in order to regulate the closed-loop system in the configuration

$$\chi^* = (q^*, p^*, p_c^*, {p_t}^*, {p_r}^*, {\epsilon_t}^*, {\epsilon_r}^*).$$

In the remaining part of this section it will be proved that, by choosing the controller energy as

$$\begin{aligned} H_c(p_c, q_c) = {} & \frac{1}{2} p_c^{\mathrm{T}} M_c^{-1} p_c + \frac{1}{2} K_{c,1}(q_{c,1} - q_{c,1}^*)^2 + \frac{1}{2} K_{c,2}(q_{c,2} - q_{c,2}^*)^2 \\ & + \Psi_1(q_{c,1}) + \Psi(q_{c,2}) \end{aligned} \tag{4.18}$$

with $M_c = M_c^{\mathrm{T}} > 0$, $K_{c,1}, K_{c,1} > 0$, and Ψ_1, Ψ_2 functions still to be specified, the configuration χ^* is stable.

Remark 4.1. It is important to point out that the proposed control methodology is *solution free*; that is, a stabilizing controller is provided by (4.3) and (4.18) but the problem of the existence of a solution of the PDE modeling the Timoshenko beam is not approached. As discussed in [9, Example 5.6], the Timoshenko beam equation generates a contraction semigroup; thus the equation has a unique classical solution. Furthermore, when closing the loop, what we obtain is a *hybrid system*, that is, a system consisting of a coupled PDE with an ODE, and also in this case it should be necessary to check under which conditions a classical solution exists. The problem can be solved by extending the approach proposed in [9, section 4.6.1], for the boundary control of an Euler–Bernoulli beam by means of a PI + strain feedback controller, to the case discussed in this paper.

As in the case of a finite dimensional Hamiltonian system, the stability of an m-pH system can be proved if it can be shown that the equilibrium is a strict extremum of the total energy of the closed-loop system. The only difference is that, in order

to prove the stability for the infinite dimensional part, it is necessary to fix a norm: it is important to note that the stability with respect to this norm, in general, will not assure the stability with respect to a different one (see, e.g., [24, p. 114]). The stability definition in the sense of Lyapunov for mixed finite and infinite dimensional systems can be given as follows [24, Definition 4.18].

DEFINITION 4.4 (Lyapunov stability for *mixed* systems). *The equilibrium configuration χ^* for a mixed finite and infinite dimensional system is said to be stable in the sense of Lyapunov with respect to the norm $\|\cdot\|$ if, for every $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that*

$$\|\chi(0) - \chi^*\| < \delta_\epsilon \quad \Rightarrow \quad \|\chi(t) - \chi*\| < \epsilon$$

$\forall t > 0$, where $\chi(0)$ is the initial configuration of the system.

As proposed in [24, pp. 116–117] and in [22], in order to verify the stability of χ^*, it is necessary to show that it is an extremum of the closed-loop energy function $\mathcal{H}_{cl}$ introduced in (4.6), with H_c given by (4.18); that is, the condition

$$\nabla \mathcal{H}_{cl}(\chi^*) = 0 \tag{4.19}$$

must hold. Moreover, if $\Delta\chi$ is the displacement from the equilibrium configuration χ^*, introduce the nonlinear functional

$$\mathcal{N}(\Delta\chi) := \mathcal{H}_{cl}(\chi^* + \Delta\chi) - \mathcal{H}_{cl}(\chi^*) \tag{4.20}$$

that is proportional to the second variation of $\mathcal{H}_{cl}$. Then the configuration χ^* is stable if it is possible to find $\gamma_1, \gamma_2, \alpha > 0$ such that [24, Theorem 4.20]

$$\gamma_1 \|\Delta\chi\|^2 \leq \mathcal{N}(\Delta\chi) \leq \gamma_2 \|\Delta\chi\|^\alpha . \tag{4.21}$$

Denote by χ the state variable of the closed-loop system. From (2.8), (4.2), (4.15), and (4.18), the total energy function is given by

$$\begin{aligned}
\mathcal{H}_{cl}(\chi) = {} & \frac{1}{2}\left(\frac{p_1^2}{m} + \frac{p_2^2}{J}\right) + \frac{1}{2}\left(k_1 q_1^2 + k_2 q_2^2\right) \\
& + \frac{1}{2}\int_{\mathcal{D}} \left(\frac{1}{\rho} p_t \wedge *p_t + \frac{1}{I_\rho} p_r \wedge *p_r + K\epsilon_t \wedge *\epsilon_t + EI\epsilon_r \wedge *\epsilon_r\right) \\
& + \frac{1}{2} p_c^{\mathrm{T}} M_c^{-1} p_c + \frac{1}{2} K_{c,1}(q_{c,1} - q_{c,1}^*)^2 + \frac{1}{2} K_{c,2}(q_{c,2} - q_{c,2}^*)^2 \\
& + \Psi_1(q_{c,1}) + \Psi(q_{c,2}).
\end{aligned}$$

The first step in the stability proof is to find under which conditions, that is, for what particular choice of the functions Ψ_1 and Ψ_2, relation (4.19) is satisfied. We have that

$$\nabla \mathcal{H}_{cl}(\chi) = \begin{bmatrix} \partial_p \mathcal{H}_{cl} \\ \partial_q \mathcal{H}_{cl} \\ \delta_{p_t} \mathcal{H}_{cl} \\ \delta_{p_r} \mathcal{H}_{cl} \\ \delta_{\epsilon_t} \mathcal{H}_{cl} \\ \delta_{\epsilon_r} \mathcal{H}_{cl} \end{bmatrix} = \begin{bmatrix} \partial_p H \\ \partial_q (H + H_c) \\ \delta_{p_t} \mathcal{H} \\ \delta_{p_r} \mathcal{H} \\ \delta_{\epsilon_t} (\mathcal{H} + H_c) \\ \delta_{\epsilon_r} (\mathcal{H} + H_c) \end{bmatrix}.$$

Clearly,

$$\frac{\partial H}{\partial p}(\chi^*) = 0, \quad \frac{\partial H_c}{\partial p_c}(\chi^*) = 0, \quad \delta_{p_t}\mathcal{H}(\chi*) = 0, \quad \text{and} \quad \delta_{p_r}\mathcal{H}(\chi^*) = 0.$$

Furthermore,

$$\begin{aligned}
\frac{\partial \mathcal{H}_{cl}}{\partial q_1} &= k_1 q_1 - K_{c,1}\left(q_{c,1} - q^*_{c,1}\right) - \frac{\partial \Psi_1}{\partial q_{c,1}}, \\
\frac{\partial \mathcal{H}_{cl}}{\partial q_2} &= k_2 q_2 + K_{c,1}\left(q_{c,1} - q^*_{c,1}\right) L - K_{c,2}\left(q_{c,2} - q^*_{c,2}\right) + \frac{\partial \Psi_1}{\partial q_{c,1}} L - \frac{\partial \Psi_2}{\partial q_{c,2}},
\end{aligned}$$

and

$$\begin{aligned}
\delta_{\epsilon_t}\mathcal{H}_{cl} &= K*\epsilon_t - K_{c,1}\left(q_{c,1} - q^*_{c,1}\right) - \frac{\partial \Psi_1}{\partial q_{c,1}}, \\
\delta_{\epsilon_r}\mathcal{H}_{cl} &= EI*\epsilon_r + K_{c,1}\left(q_{c,1} - q^*_{c,1}\right) x - K_{c,2}\left(q_{c,2} - q^*_{c,2}\right) + \frac{\partial \Psi_1}{\partial q_{c,1}} x - \frac{\partial \Psi_2}{\partial q_{c,2}}.
\end{aligned}$$

Then $\nabla \mathcal{H}_{cl}(\chi^*) = 0$ if

$$\begin{cases}
\Psi_1(q_{c,1}) = k_1 q_1^* q_{c,1} + \psi_{c,1}, \\
\Psi_2(q_{c,2}) = \left(k_2 q_2^* + k_1 q_1^* L\right) q_{c,2} + \psi_{c,2},
\end{cases}$$

with $\psi_{c,1}$ and $\psi_{c,2}$ arbitrary constants. Once the equilibrium is assigned in χ^*, it is necessary to verify the convexity condition (4.21) in χ^* on the nonlinear functional $\mathcal{N}$. After simple calculations, it can be obtained (see [10]) that

$$\begin{aligned}
\|\Delta\chi\| = {} & \frac{1}{2}\Delta p^{\mathrm{T}} M^{-1} \Delta p + \frac{1}{2}\Delta p_c^{\mathrm{T}} M_c^{-1} \Delta p_c + \frac{1}{2} k_1 \Delta q_1^2 + \frac{1}{2} k_2 \Delta q_2^2 \\
& + \frac{1}{2}\int_0^L \left(\frac{1}{\rho}\Delta p_t \wedge *p_t + \frac{1}{I_\rho}\Delta p_r \wedge *p_r + K \Delta\epsilon_t \wedge *\Delta\epsilon_t + EI \Delta\epsilon_r \wedge *\Delta\epsilon_r \right) \\
& + \frac{1}{2} K_{c,1} \left[L\Delta q_2 - \Delta q_1 + \int_0^L \left(x\Delta\epsilon_r - \Delta\epsilon_t\right) \right]^2 + \frac{1}{2} K_{c,2} \left[\Delta q_2 + \int_0^L \Delta\epsilon_r \right]^2 .
\end{aligned}$$

The convexity condition (4.21) requires a norm in order to be verified: a possible choice can be

$$\begin{aligned}
\|\chi\|^2 = {} & \frac{1}{2}\Delta p^{\mathrm{T}} \Delta p + \frac{1}{2}\Delta p_c^{\mathrm{T}} \Delta p_c + \frac{1}{2}\Delta q_1^2 + \frac{1}{2}\Delta q_2^2 \\
& + \frac{1}{2}\int_0^L \left(\Delta p_t \wedge *p_t + \Delta p_r \wedge *p_r + \Delta\epsilon_t \wedge *\Delta\epsilon_t + \Delta\epsilon_r \wedge *\Delta\epsilon_r \right).
\end{aligned}$$

Then, in (4.21), assume that

$$\gamma_1 = \frac{1}{2}\min\left\{ |M^{-1}|,\ |M_c^{-1}|,\ k_1,\ k_2,\ \frac{1}{\rho},\ \frac{1}{I_\rho},\ K,\ EI \right\}.$$

Moreover, if

$$\tilde{\gamma}_2 = \frac{1}{2}\max\left\{ |M^{-1}|,\ |M_c^{-1}|,\ k_1,\ k_2,\ \frac{1}{\rho},\ \frac{1}{I_\rho},\ K,\ EI,\ K_{c,1},\ K_{c,2} \right\},$$

we have that

$$\mathcal{N}(\Delta\chi) \leq \tilde{\gamma}_2 \|\Delta\chi\| + \tilde{\gamma}_2 \left[L\Delta q_2 - \Delta q_1 + \int_0^L (x\Delta\epsilon_r - \Delta\epsilon_t) \right]^2 + \tilde{\gamma}_2 \left[\Delta q_2 + \int_0^L \Delta\epsilon_r \right]^2.$$

Since

$$\begin{aligned}
\left[L\Delta q_2 - \Delta q_1 + \int_0^L (x\Delta\epsilon_r - \Delta\epsilon_t) \right]^2 &\leq 2\left(L\Delta q_2 - \Delta q_1\right)^2 + 2\left[\int_0^L (x\Delta\epsilon_r - \Delta\epsilon_t) \right]^2 \\
&\leq 4\left(|\Delta q_1|^2 + L^2|\Delta q_2|^2\right) + 4\left(\int_0^L x\Delta\epsilon_r \right)^2 + 4\left(\int_0^L \Delta\epsilon_t \right)^2 \\
&\leq 4\left(|\Delta q_1|^2 + L^2|\Delta q_2|^2\right) + 4L\left(\int_0^L \Delta\epsilon_t \wedge *\Delta\epsilon_t + L\int_0^L \Delta\epsilon_r \wedge *\Delta\epsilon_r \right), \\
\left[\Delta q_2 + \int_0^L \Delta\epsilon_r \right]^2 &\leq 2|\Delta q_2|^2 + 2\left(\int_0^L \Delta\epsilon_r \right)^2 \leq 2|\Delta q_2|^2 + 2L\int_0^L \Delta\epsilon_r \wedge *\Delta\epsilon_r,
\end{aligned}$$

it is possible to satisfy (4.21) by choosing $\alpha = 2$ and

$$\gamma_2 = \tilde{\gamma}_2 \cdot \max\left\{4,\ 4L^2 + 2,\ 4L,\ 4L^2 + 2L\right\},$$

which completes the stability proof. In other words, the following proposition has been proved.

PROPOSITION 4.5. *Consider the m-pH system of Figure* 4.2*, that is, the result of the power conserving interconnection* (4.4) *of the subsystems* (2.17), (4.1)*, and* (4.3)*. If in* (4.3) *it is assumed that* $G_c = I$ *and* H_c *is chosen according to* (4.18)*, then the configuration* χ^* *is stable in the sense of Lyapunov, i.e., in the sense of Definition* 4.4*.*

5. Conclusions. Once the Timoshenko model of the beam has been reformulated within the framework of dpH systems, some considerations about control strategies of the flexible beam have been presented. In particular, the well-known control by damping injection is extended to distributed parameter systems in order to stabilize the beam acting through its boundary and/or its distributed port. Some well-known results already presented in the literature are obtained in this new framework.

Moreover, it has been shown that it is possible to extend the energy shaping by interconnection control technique to treat mixed finite and infinite dimensional systems following the same ideas presented in [22], of which this work is a continuation. In particular, the control of a mechanical system made of a flexible beam with a rigid body connected at one of its extremities has been presented. The finite dimensional controller, acting on the system through the other extremity, is developed by properly extending the concept of Casimir functions to infinite dimensions. The main advantage is that the controller is suitable for a clear physical interpretation, with the drawback that the whole approach is *solution free*, in the sense of Remark 4.1.

Future work will deal with the extension of these concepts to the modeling and control of simple kinematic chains with flexible links.

REFERENCES

[1] C. I. Byrnes, A. Isidori, and J. C. Willems, *Passivity, feedback equivalence, and the global stabilization of minimum phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1228–1240.

[2] T. J. Courant, *Dirac manifolds*, Trans. Amer. Math. Soc., 319 (1990), pp. 631–661.

[3] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite Dimensional Linear Systems Theory*, Texts in Appl. Math. 21, Springer-Verlag, New York, 1995.

[4] M. Dalsmo and A. J. van der Schaft, *On representations and integrability of mathematical structures in energy-conserving physical systems*, SIAM J. Control Optim., 37 (1999), pp. 54–91.

[5] S. Dong-Hua and F. De-Xing, *Exponential stabilization of the Timoshenko beam with locally distributed feedback*, in Proceedings of the 14th IFAC World Congress, Beijing, People's Republic of China, H. F. Chen, D.-Z. Cheng, and J.-F. Zhang, eds., Elsevier, New York, 1999.

[6] G. Golo, V. Talasila, and A. J. van der Schaft, *A Hamiltonian formulation of the Timoshenko beam model*, in Proceedings of Mechatronics 2002, University of Twente, The Netherlands, 2002, pp. 544–553.

[7] G. Golo, A. J. van der Schaft, and S. Stramigioli, *Hamiltonian formulation of planar beams*, in Proceedings of the 2nd IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control, A. Astolfi, A. J. van der Schaft, and F. Gordillo, eds., Sevilla, Spain, Elsevier, New York, 2003, pp. 169–174.

[8] J. U. Kim and Y. Renardy, *Boundary control of the Timoshenko beam*, SIAM J. Control. Optim., 25 (1987), pp. 1417–1429.

[9] Z. H. Luo, B. Z. Guo, and O. Morgul, *Stability and Stabilization of Infinite Dimensional Systems with Applications*, Communications and Control Engineering Series, Springer-Verlag, London, 1999.

[10] A. Macchelli, *Port Hamiltonian Systems. A Unified Approach for Modeling and Control Finite and Infinite Dimensional Physical Systems*, Ph.D. thesis, University of Bologna – DEIS, Bologna, Italy, 2003. Available at http://www-lar.deis.unibo.it/woda/data/deis-lar-publications/e499.Document.pdf

[11] A. Macchelli and C. Melchiorri, *Control by interconnection of the Timoshenko beam*, in Proceedings of the 2nd IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control, 2003, pp. 175–182.

[12] A. Macchelli and C. Melchiorri, *Distributed port Hamiltonian formulation of the Timoshenko beam: Modeling and control*, in Proceedings of the 4th MATHMOD, Vienna, Austria, 2003.

[13] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Springer-Verlag, New York, 1994.

[14] J. E. Marsden and T. S. Ratiu, *Geometry of Nonlinear Systems*, 2001. This book is freely available at http://www.cds.caltech.edu/∼marsden.

[15] B. M. Maschke and A. J. van der Schaft, *Port controlled Hamiltonian systems: Modeling origins and system theoretic properties*, in Proceedings of the Third Conference on Nonlinear Control Systems (NOLCOS), Bordeaux, France, 1992.

[16] B. M. Maschke and A. J. van der Schaft, *Interconnection of systems: The network paradigm*, in Proceedings of the 35th IEEE Conference on Decision and Control, New York, 1996, pp. 207–212.

[17] B. M. Maschke and A. J. van der Schaft, *Port controlled Hamiltonian representation of distributed parameter systems*, in Workshop on Modeling and Control of Lagrangian and Hamiltonian Systems, Princeton, NJ, 2000.

[18] B. M. Maschke and A. J. van der Schaft, *Fluid dynamical systems as Hamiltonian boundary control systems*, in Proceedings of the 40th IEEE Conference on Decision and Control, Vol. 5, Orlando, FL, 2001, pp. 4497–4502.

[19] P. J. Olver, *Application of Lie Groups to Differential Equations*, Springer-Verlag, New York, 1993.

[20] R. Ortega, A. J. van der Schaft, B. M. Maschke, and G. Escobar, *Energy-shaping of port-controlled Hamiltonian systems by interconnection*, in Proceedings of the IEEE Conference on Decision and Control, Vol. 2, Phoenix, AZ, 1999, pp. 1646–1651.

[21] R. Ortega, A. J. van der Schaft, B. M. Maschke, and G. Escobar, *Interconnection and damping assignment passivity-based control of port-controlled Hamiltonian systems*, Automatica, 38 (2000), pp. 585–596.

[22] H. Rodriguez, A. J. van der Schaft, and R. Ortega, *On stabilization of nonlinear distributed parameter port-controlled Hamiltonian systems via energy shaping*, in Proceedings of the 40th IEEE Conference on Decision and Control, Vol. 1, Orlando, FL, 2001, pp. 131–136.

[23] S. Stramigioli, *Modeling and IPC Control of Interactive Mechanical Systems: A Coordinate-Free Approach*, Springer-Verlag, London, 2001.

[24] G. E. Swaters, *Introduction to Hamiltonian Fluid Dynamics and Stability Theory*, Chapman and Hall/CRC, London, 2000.

[25] S. W. Taylor, *Boundary Control of the Timoshenko Beam with Variable Physical Characteristics*, Technical report, University of Auckland, New Zealand, 1997.

[26] A. J. van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*, in Communication and Control Engineering, Springer-Verlag, London, 2000.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 769–791

LINEAR HAMILTONIAN BEHAVIORS AND BILINEAR DIFFERENTIAL FORMS*

P. RAPISARDA† AND H. L. TRENTELMAN‡

Abstract. We study linear Hamiltonian systems using bilinear and quadratic differential forms. Such a representation-free approach allows us to use the same concepts and techniques to deal with systems isolated from their environment and with systems subject to external influences and allows us to study systems described by higher-order differential equations, thus dispensing with the usual point of view in classical mechanics of considering first- and second-order differential equations only.

Key words. linear Hamiltonian systems, two-variable polynomial matrices, bilinear and quadratic differential forms, behavioral system theory

AMS subject classifications. 93A10, 93A30, 93C05, 37J99, 70H50

DOI. 10.1137/S0363012902414664

1. Introduction.

This paper aims to give a unified treatment of linear Hamiltonian systems using the formalism of bilinear and quadratic differential forms introduced in [24]. We consider systems with and without external influences, and we deal with both cases using the same techniques and the same concepts. Moreover, we formulate concepts and study the properties of Hamiltonian systems in a representation-free way, thus dispensing with the usual point of view in mechanics and in physics (see, for example, [1]) of concentrating on first-order representations in the (generalized) coordinates and the (generalized) momenta. Instead of *postulating* the existence of a function (the Lagrangian, or the Hamiltonian) on the basis of physical considerations (conservation of energy, etc.) and deducing from such a function the equations of motion, we proceed by assuming that a set of linear differential equations with constant coefficients describing the system is given, and we *deduce* the Hamiltonian nature of the system from such equations, by proving the existence of certain bilinear functionals of the variables of the system and of their derivatives satisfying some additional property. Our approach is of a system-theoretic nature rather than derived from the study of mechanics: as happens in optimal control theory for linear systems, we consider the interplay of (quadratic and bilinear) functionals of the system variables and of the equations of motion as the central object of study when dealing with Hamiltonian systems.

In this paper we also reconcile our point of view with that of classical mechanics by showing how to construct a "generalized Lagrangian" on the basis of the equations of the system, in the sense that the trajectories of the system are stationary with respect to such a quadratic functional of the variables of the system and their derivatives. In this context, the concept of internal force also arises naturally from the equations describing the system: in this paper we show that generalized internal forces can be defined which depend on higher-order derivatives of the external variables and not only first-order ones at most, as happens in classical mechanics.

*Received by the editors September 16, 2002; accepted for publication (in revised form) October 17, 2003; published electronically September 18, 2004.
http://www.siam.org/journals/sicon/43-3/41466.html

†Department of Mathematics, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands (P.Rapisarda@math.unimaas.nl).

‡Research Institute for Mathematics and Computer Science, P.O. Box 800, 9700 AV Groningen, The Netherlands (H.L.Trentelman@math.rug.nl).

The works more closely connected with the approach proposed in this paper are [8] and [7, 19]. The approach of [8] is devoted to examining the consequences of the Hamiltonian symmetry of the transfer function on the possibility of realizing a linear system in some special form; the treatment is carried out in the framework of polynomial models. In the present paper we obtain some of the results of [8] when treating Hamiltonian systems with external influences. The work of [7, 19] (see also [6, 12, 18, 20]) deals with nonlinear systems and consequently has a larger application area than the one illustrated in the present paper. In some cases, most notably in the study of external characterizations of (nonlinear) Hamiltonian systems, such an approach provides results which indeed have a more general nature than some of those presented in this article. However, we believe that the approach presented in this paper, although applicable in its present form only to finite-dimensional linear systems, is relevant for the following reasons. First, the simpler structure of linear systems and the array of algebraic techniques available in the behavioral framework to deal with them allow us to devise constructive methods based on polynomial algebra in order to solve many of the problems arising when considering linear Hamiltonian systems, for example, the computation of "conserved quantities," of special representations, etc. (an example of the technical difficulties involved in solving similar problems in the nonlinear case is given in section III of [6]). Second, the representation-free approach that we pursue allows us to describe systems of a different nature using the same formalism, independent of the domain of application. Such a feature of our approach is especially relevant in view of the potential application of our techniques in the description of possibly infinite-dimensional nonmechanical systems, for example, those arising in the theory of fields. Moreover, proceeding directly from the equations of motion allows us to study Hamiltonianity also for complex systems (for example, those resulting from the interconnection of many simple subsystems), for which the identification of functionals representing the "conserved quantities" is not immediate; this is of particular interest when considering the application of the results presented in this paper to computer-assisted modeling and simulation.

The paper is organized as follows: in section 2 we review some notions regarding linear differential systems and bilinear differential forms, which form the setting in which we study linear Hamiltonian systems. In section 3 we define Hamiltonianity for autonomous systems. In classical mechanics, a Hamiltonian system consists of the trajectories which are stationary with respect to a Lagrangian function; in section 4 we show how such a point of view fits with our definition of Hamiltonianity, and we define the notion of generalized Lagrangian. In section 5 we consider the notion of internal forces, which we propose to see as latent variables arising naturally from the equations describing an autonomous Hamiltonian system. The relationship between internal forces and external variables in an autonomous Hamiltonian system forms the basis for our definition of a controllable Hamiltonian system. In section 6 we discuss our results and outline some directions for future research.

We give a few words on notation. The space of $\mathtt{n}$-dimensional real, respectively, complex, vectors is denoted by $\mathbb{R}^{\mathtt{n}}$, respectively, $\mathbb{C}^{\mathtt{n}}$, and the space of $\mathtt{m} \times \mathtt{n}$ real, respectively, complex, matrices, by $\mathbb{R}^{\mathtt{m}\times\mathtt{n}}$, respectively, $\mathbb{C}^{\mathtt{m}\times\mathtt{n}}$. Whenever one of the two dimensions is not specified, a bullet $\bullet$ is used so that, for example, $\mathbb{C}^{\bullet\times\mathtt{n}}$ denotes the set of complex matrices with $\mathtt{n}$ columns and an unspecified number of rows. In order to enhance readability, when dealing with a vector space $\mathbb{R}^{\bullet}$ whose elements are commonly denoted with w, we use the notation $\mathbb{R}^{\mathtt{w}}$ (note the typewriter font type); similar considerations hold for matrices representing linear operators on such spaces.

Given two column vectors x and y, we denote with $\mathrm{col}(x, y)$ the vector obtained by stacking x over y; a similar convention holds for the stacking of matrices with the same number of columns. If $A \in \mathbb{R}^{\mathtt{m}\times\mathtt{n}}$, then $A^T \in \mathbb{R}^{\mathtt{n}\times\mathtt{m}}$ denotes its transpose. If $A_i \in \mathbb{R}^{\mathtt{k}_i\times\mathtt{k}_i}$, $i = 1, \ldots, \mathtt{m}$, then block $\mathrm{diag}(A_i)_{i=1,\ldots,\mathtt{m}}$ denotes the $(\sum_{i=1}^{\mathtt{m}} \mathtt{k}_i) \times (\sum_{i=1}^{\mathtt{m}} \mathtt{k}_i)$ matrix having the A_i's on the main diagonal.

The ring of polynomials with real coefficients in the indeterminate ξ is denoted by $\mathbb{R}[\xi]$; the ring of two-variable polynomials with real coefficients in the indeterminates ζ and η is denoted by $\mathbb{R}[\zeta, \eta]$. A polynomial p in the indeterminate ξ is called *even* if $p(\xi) = p(-\xi)$ and *odd* if $p(-\xi) = -p(\xi)$. The space of all $\mathtt{n} \times \mathtt{m}$ polynomial matrices in the indeterminate ξ is denoted by $\mathbb{R}^{\mathtt{n}\times\mathtt{m}}[\xi]$, and that consisting of all $\mathtt{n} \times \mathtt{m}$ polynomial matrices in the indeterminates ζ and η is denoted by $\mathbb{R}^{\mathtt{n}\times\mathtt{m}}[\zeta, \eta]$. Given a matrix $R \in \mathbb{R}^{\mathtt{n}\times\mathtt{m}}[\xi]$, we define $R^\sim(\xi) := R^T(-\xi) \in \mathbb{R}^{\mathtt{m}\times\mathtt{n}}[\xi]$. If $R(\xi)$ has complex coefficients, then $R^\sim(\xi)$ denotes the matrix obtained from R by substituting $-\xi$ in place of ξ, transposing, and conjugating.

We denote with $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{q}})$ the set of infinitely often differentiable functions from $\mathbb{R}$ to $\mathbb{R}^{\mathtt{q}}$, and we denote with $\mathfrak{D}(\mathbb{R}, \mathbb{R}^{\mathtt{q}})$ the subset of $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{q}})$ consisting of compact support functions.

Finally, if K is an $\mathtt{n} \times \mathtt{n}$ matrix, the bilinear form on $\mathbb{R}^{\mathtt{n}}$ defined by $(x_1, x_2) \mapsto x_1^T K x_2$ is denoted by $\langle x_1, x_2\rangle_K$. If $K = K^T$, then it also induces a quadratic form $x \to x^T K x$, which we denote with $|x|_K^2$.

2. Basics. In order to make the paper as self-contained as possible, we now illustrate some basic notions regarding linear differential behaviors and bilinear and quadratic differential forms; detailed expositions of such concepts can be found, respectively, in [15] and in [24]. We conclude the section with a brief introduction to Hamiltonian constant matrices, which are relevant in the discussion of state-space representations of Hamiltonian behaviors.

2.1. Linear differential behaviors. A *linear differential behavior* is a linear subspace $\mathfrak{B}$ of $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}})$ consisting of all solutions w of a given system of linear constant-coefficient differential equations. Such a set is represented as

$$R\left(\frac{d}{dt}\right) w = 0, \tag{2.1}$$

where $R \in \mathbb{R}^{\bullet\times\mathtt{w}}[\xi]$; (2.1) is called a *kernel representation* of the behavior $\mathfrak{B} := \{w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}}) \mid w \text{ satisfies (2.1) }\}$, and w is called the *manifest* or *external variable* of $\mathfrak{B}$. The class of all such behaviors is denoted with $\mathfrak{L}^{\mathtt{w}}$.

When modeling physical systems from first principles, we often introduce a number of *latent* (or *auxiliary*) variables ℓ besides the manifest ones: thus *latent variable representations*

$$R\left(\frac{d}{dt}\right) w = M\left(\frac{d}{dt}\right) \ell \tag{2.2}$$

are obtained. Equation (2.2) describes the *full behavior*

$$\mathfrak{B}_{\mathrm{f}} := \{(w, \ell) \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}+\mathtt{l}}) \mid (2.2) \text{ holds}\},$$

and we call the projection of $\mathfrak{B}_{\mathrm{f}}$ on the w variable, i.e.,

$$\mathfrak{B} := \{w \mid \exists \ell \text{ such that (2.2) holds}\},$$

the *manifest behavior* associated with (2.2). It can be shown that $\mathfrak{B}$ can also be described in kernel form, i.e., $\mathfrak{B} = \ker R'(\frac{d}{dt})$ for a suitable $R' \in \mathbb{R}^{\bullet \times \mathtt{w}}[\xi]$. The computation of such an R' from R and M is called the *elimination of the latent variable* ℓ.

When the matrix R in (2.2) is the $\mathtt{w}$-dimensional identity, we call

$$w = M\left(\frac{d}{dt}\right)\ell \tag{2.3}$$

an *image representation* of $\mathfrak{B}$. A behavior can be represented by (2.3) if and only if each of its kernel representations is associated with a polynomial matrix $R \in \mathbb{R}^{\bullet \times \mathtt{w}}[\xi]$ such that $\operatorname{rank}(R(\lambda))$ is constant for all $\lambda \in \mathbb{C}$, or equivalently, $\mathfrak{B}$ is controllable in the behavioral sense (see Chapter 5 of [15]). The latent variable ℓ in (2.3) is called *observable* from w if $[w = M(\frac{d}{dt})\ell = 0] \Longrightarrow [\ell = 0]$. It can be shown that this is the case if and only if the matrix $M(\lambda)$ has full column rank for all $\lambda \in \mathbb{C}$.

An important class of behaviors is that of *autonomous behaviors*, which admit kernel representations (2.1) in which the matrix R is $\mathtt{w} \times \mathtt{w}$ and nonsingular. Given an autonomous behavior $\mathfrak{B} \in \mathcal{L}^{\mathtt{w}}$, the $\mathtt{w} \times \mathtt{w}$ matrices associated with any two kernel representations of $\mathfrak{B}$ have the same Smith form (see, for example, section 6.3.3 of [10]). The diagonal elements in such a Smith form are nonzero polynomials called the *invariant polynomials of* $\mathfrak{B}$; the product of such polynomials of $\mathfrak{B}$ is denoted by $\chi_{\mathfrak{B}}$ and is called the *characteristic polynomial of* $\mathfrak{B}$.

By permuting the components of w with a permutation matrix $\Pi \in \mathbb{R}^{\mathtt{w} \times \mathtt{w}}$ if necessary, we can write $\Pi w = \operatorname{col}(u, y)$ with y having $\operatorname{rank}(R)$ components and u having $\mathtt{w} - \operatorname{rank}(R)$ components, so that $\mathfrak{B}$ admits the representation $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$, with P square and nonsingular. We call such a partition of the external variables of $\mathfrak{B}$ an *input/output* (*i/o*) partition, u the *input variable*, y the *output variable*, and the rational matrix $P^{-1}Q$ the *transfer function* associated with the given i/o partition. Observe that in general many choices are possible for the permutation matrix Π above: the i/o partition is not unique. Observe also that $P^{-1}Q$ is not necessarily a matrix of proper rational functions; however, among all i/o partitions for $\mathfrak{B}$, there exists at least one whose corresponding transfer function is proper.

The number of input variables is an invariant denoted with $\mathtt{m}(\mathfrak{B})$; evidently, the number $\mathtt{p}(\mathfrak{B}) := \mathtt{w} - \mathtt{m}(\mathfrak{B})$ of output variables is also an invariant. If $\mathfrak{B}$ is autonomous, it has no input variables; in other words, $\mathtt{m}(\mathfrak{B}) = 0$, or equivalently $\mathtt{p}(\mathfrak{B}) = \mathtt{w}$. If $\mathfrak{B}$ is controllable, then it admits an observable image representation (2.3) with $M \in \mathbb{R}^{\mathtt{w} \times \mathtt{1}}[\xi]$; an i/o partition then corresponds to a partition of M as $M = \operatorname{col}(U, Y)$ with $U \in \mathbb{R}^{\mathtt{1} \times \mathtt{1}}[\xi]$ nonsingular; note that $\mathtt{m}(\mathfrak{B}) = \mathtt{1}$, the dimension of ℓ. In such a case the transfer function from u to y is the matrix of rational functions $G = YU^{-1}$.

In this paper we also use the concept of state and of state representation (see [16] for a thorough discussion). A latent variable ℓ is a *state variable* for $\mathfrak{B}$ if and only if $\mathfrak{B}$ admits a representation (2.2) of first order in ℓ and zeroth order in w: $E\frac{d\ell}{dt} + F\ell + Gw = 0$. Such a representation is called a *state representation* of $\mathfrak{B}$. The minimal number of state variables that can be used in order to represent $\mathfrak{B}$ in state-space form is an invariant called the *McMillan degree* of $\mathfrak{B}$ and is denoted $\mathtt{n}(\mathfrak{B})$. By combining the notion of state with that of inputs and outputs we arrive at the *input/state/output representation* (*i/s/o*) $\frac{d}{dt}x = Ax + Bu$, $y = Cx + Du$, $w = \operatorname{col}(u, y)$.

2.2. Bilinear and quadratic differential forms. In modeling and control problems it is often necessary to study certain functionals of the system variables and their derivatives; when considering linear systems, such functionals are quadratic.

In [24] the parametrization of such functionals using two-variable polynomial matrices has been studied in detail, resulting in the definition of bilinear and quadratic differential form and in the development of a calculus with applications in stability theory, optimal and H_∞-control, and dissipativity theory. Two-variable polynomials and their algebraic properties have been used before in systems theory, for example, by Kalman [11] and Willems and Fuhrmann [22] in the context of stability analysis. We also refer to the pioneering work of Brockett [2] on path independence of integrals of quadratic functionals in the system variables and their derivatives, which prefigures some of the results obtained in [24]. In this section we review those definitions and results of the framework developed in [24] which are used in the rest of this paper.

First, some words about bilinear forms on abstract vector spaces. A *bilinear form* $\mathcal{L}$ on a vector space $\mathbb{V}$ over $\mathbb{R}$ is a mapping $\mathcal{L} : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ that is linear in each of its arguments separately. We sometimes denote a bilinear form as $\mathcal{L}|_{\mathbb{V}}$ in order to emphasize its domain. The *rank* of a bilinear form $\mathcal{L}|_{\mathbb{V}}$ equals the number of independent linear functionals $\mathcal{L}(\cdot, v)$, where v ranges over $\mathbb{V}$. A bilinear form $\mathcal{L}|_{\mathbb{V}}$ is called *nondegenerate* if for all $v \in \mathbb{V}$ we have that $\mathcal{L}(\cdot, v) = 0$ is equivalent with $v = 0$, i.e., $[\mathcal{L}(\mathbb{V}, v) = 0] \Leftrightarrow [v = 0]$. The bilinear form $\mathcal{L}$ on $\mathbb{V}$ is called *skew-symmetric* if for all $v_1, v_2 \in \mathbb{V}$ we have $\mathcal{L}(v_1, v_2) = -\mathcal{L}(v_2, v_1)$. A *symplectic space* is a pair $(\mathbb{V}, \mathcal{L})$, where $\mathbb{V}$ is a vector space over $\mathbb{R}$ and $\mathcal{L}$ is a nondegenerate, skew-symmetric, bilinear form on $\mathbb{V}$; in such a case $\mathcal{L}$ is called a *symplectic form* on $\mathbb{V}$. If $\mathbb{V}$ is a finite-dimensional space on $\mathbb{R}$, then nondegenerate symplectic forms are in one-one correspondence with nonsingular skew-symmetric matrices in the sense that for every symplectic form $\mathcal{L}$ there exists a nonsingular $K \in \mathbb{R}^{\mathtt{n}\times\mathtt{n}}$ with $K^T = -K$ such that $\mathcal{L}(x, y) = \langle x, y\rangle_K$, and conversely, every such K defines a symplectic form on $\mathbb{V}$. Obviously, such K exists only if $\mathtt{n}$ is even.

Next, we examine bilinear differential forms. Let $\Phi \in \mathbb{R}^{\mathtt{w}_1\times\mathtt{w}_2}[\zeta, \eta]$; then $\Phi(\zeta, \eta) = \sum_{h,k=0}^{N} \Phi_{h,k}\zeta^h\eta^k$, where $\Phi_{h,k} \in \mathbb{R}^{\mathtt{w}_1\times\mathtt{w}_2}$ and N is a nonnegative integer. The two-variable polynomial matrix Φ induces the bilinear functional acting on $\mathtt{w}_1$-, respectively, $\mathtt{w}_2$-dimensional infinitely differentiable trajectories, defined as $L_\Phi(w_1, w_2) = \sum_{h,k=0}^{N} (\frac{d^h w_1}{dt^h})^T \Phi_{h,k} \frac{d^k w_2}{dt^k}$. Such a functional is called a *bilinear differential form* (BDF). L_Φ is *skew-symmetric*, meaning $L_\Phi(w_1, w_2) = -L_\Phi(w_2, w_1)$ for all w_1, w_2, if and only if Φ is a *skew-symmetric* two-variable polynomial matrix, i.e., if $\mathtt{w}_1 = \mathtt{w}_2$ and $\Phi(\zeta, \eta) = -\Phi^T(\eta, \zeta)$.

A two-variable polynomial matrix $\Phi(\zeta, \eta)$ is called *symmetric* if $\mathtt{w}_1 = \mathtt{w}_2 = \mathtt{w}$ and $\Phi(\zeta, \eta) = \Phi^T(\eta, \zeta)$. In such a case, Φ induces also a quadratic functional acting on $\mathtt{w}$-dimensional infinitely smooth trajectories as $Q_\Phi(w) := L_\Phi(w, w)$. We will call Q_Φ the *quadratic differential form* (QDF) associated with Φ.

With every $\Phi \in \mathbb{R}^{\mathtt{w}_1\times\mathtt{w}_2}[\zeta, \eta]$ we associate its *coefficient matrix* $\tilde{\Phi}$, which is defined as the infinite matrix $\tilde{\Phi} := (\Phi_{i,j})_{i,j=0,\ldots}$. Observe that although $\tilde{\Phi}$ is infinite, only a finite number of its entries are nonzero. Note that Φ is skew-symmetric if and only if $\tilde{\Phi}^T = -\tilde{\Phi}$; also, Φ is symmetric if and only if $\tilde{\Phi}^T = \tilde{\Phi}$.

The association of two-variable polynomial matrices with BDFs and QDFs allows us to develop a calculus that has applications in stability theory, optimal control, and H_∞-control. We restrict our attention only to those concepts that are used in this paper. One of them is the map $\partial : \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta] \longrightarrow \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$ defined by $\partial\Phi(\xi) := \Phi(-\xi, \xi)$. Observe that if $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta]$ is symmetric, then $\partial\Phi$ is *para-Hermitian*, i.e., $\partial\Phi = (\partial\Phi)^\sim$, and if Φ is skew-symmetric, then $\partial\Phi$ is *skew para-Hermitian*, i.e., $(\partial\Phi)^\sim = -\partial\Phi$. Given a BDF L_Ψ we define its *derivative* as the BDF L_Φ defined by $L_\Phi(w_1, w_2) := \frac{d}{dt}(L_\Psi(w_1, w_2))$ for all w_1, w_2. In terms of the two-variable polynomial

matrices associated with the BDFs, the relationship between a BDF and its derivative is expressed as $\Phi(\zeta,\eta) = (\zeta+\eta)\Psi(\zeta,\eta)$. The notion of a derivative of a QDF is analogous and algebraically characterized in the same way; we will not repeat its definition here.

We now discuss the notions of *rank* and of *nondegeneracy of BDFs*. Let $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ and $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$. Then the BDF L_Φ induces a bilinear form on the real vector space $\mathfrak{B}$ by assigning to $(v,w) \in \mathfrak{B}\times\mathfrak{B}$ the real number $L_\Phi(v,w)(0)$. We denote this bilinear form by $L_\Phi|_{\mathfrak{B}}$. We can hence speak about the rank and the nondegeneracy of this induced bilinear form. In particular, $L_\Phi|_{\mathfrak{B}}$ is nondegenerate if for all $w \in \mathfrak{B}$ we have $[L_\Phi(\mathfrak{B},w)(0)=0] \Leftrightarrow [w=0]$. If $\mathfrak{B}$ is autonomous, then the following result, whose proof is easy and is left to the reader, holds.

PROPOSITION 2.1. *Let $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$, and let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ be autonomous. Let $\frac{d}{dt}x = Ax$, $w = Cx$ be a state representation of $\mathfrak{B}$, with full behavior $\mathfrak{B}_{\mathrm{f}} = \{(x,w)|\frac{d}{dt}x = Ax, w = Cx\}$. Define $N_\infty := \mathrm{col}(CA^i)_{i=0,\dots}$; note that it has an infinite number of rows. Then $L_\Phi(w_1,w_2) = x_1^T N_\infty^T \tilde{\Phi} N_\infty x_2$ $\forall (x_1,w_1)$, $(x_2,w_2) \in \mathfrak{B}_{\mathrm{f}}$. Consequently,* rank $(L_\Phi|_{\mathfrak{B}}) =$ rank $(N_\infty^T\tilde{\Phi}N_\infty)$.

We can now characterize nondegeneracy in terms of rank.

PROPOSITION 2.2. *Let $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$, and let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ be autonomous, with McMillan degree $\mathtt{n}(\mathfrak{B})$. Then $L_\Phi|_{\mathfrak{B}}$ is nondegenerate if and only if* $\mathrm{rank}(L_\Phi|_{\mathfrak{B}}) \geq \mathtt{n}(\mathfrak{B})$.

Proof. Let $\frac{d}{dt}x = Ax$, $w = Cx$ be a minimal state representation of $\mathfrak{B}$. Then $N_\infty^T\tilde{\Phi}N_\infty$ has size $\mathtt{n}(\mathfrak{B}) \times \mathtt{n}(\mathfrak{B})$. We now prove that $L_\Phi|_{\mathfrak{B}}$ is nondegenerate if and only if $N_\infty^T\tilde{\Phi}N_\infty$ is nonsingular. This will prove the claim.

(Only if) Assume $N_\infty^T\tilde{\Phi}N_\infty x_0 = 0$. Define w by $w(t) = Ce^{At}x_0$. For an arbitrary $w' = Ce^{At}x_0'$, it holds that $L_\Phi(w',w)(0) = x_0'^T N_\infty^T\tilde{\Phi}N_\infty x_0 = 0$; by the nondegeneracy of L_Φ, we conclude that $w=0$. Minimality of $\frac{d}{dt}x = Ax$, $w = Cx$ implies $x_0 = 0$.

(If) Let $w = Ce^{At}x_0 \in \mathfrak{B}$, and assume $L_\Phi(w',w)(0) = 0$ for all $w' = Ce^{At}x_0' \in \mathfrak{B}$. Then clearly $x_0'^T N_\infty^T\tilde{\Phi}N_\infty x_0 = 0$ for all x_0', so $N_\infty^T\tilde{\Phi}N_\infty x_0 = 0$. Since $N_\infty^T\tilde{\Phi}N_\infty$ is nonsingular, this implies $x_0 = 0$ and consequently $w = 0$. □

2.3. Hamiltonian matrices. Given a symplectic form $\langle x,y\rangle_K$ on $\mathbb{R}^{\mathtt{n}}$, a linear map $A : \mathbb{R}^{\mathtt{n}} \to \mathbb{R}^{\mathtt{n}}$ (or matrix $A \in \mathbb{R}^{\mathtt{n}\times\mathtt{n}}$) is called *Hamiltonian* if $\langle Ax,y\rangle_K + \langle x,Ay\rangle_K = 0$ for all $x,y \in \mathbb{R}^{\mathtt{n}}$; equivalently $A^TK + KA = 0$. For the purposes of this paper, we are especially interested in the invariant polynomials of $\xi I - A$, where A is Hamiltonian; we call them the *invariant polynomials of* A. The relevant result is the following.

PROPOSITION 2.3. *Let $A \in \mathbb{R}^{\mathtt{n}\times\mathtt{n}}$ be Hamiltonian. Then its invariant polynomials are either even or odd, and the odd ones can be divided into pairs, so that the multiplicity of zero as a root is the same for the polynomials of each pair.*

Proof. In order to prove our statement we use the results of [4]. Such results make use of the concept of *elementary divisors* of A, i.e., the irreducible factors of the invariant polynomials of $\xi I - A$, which are in one-one correspondence with the diagonal blocks appearing in the Jordan form of A (see, for example, section VII.7 of [9]).

In Theorem 2.2 of [4] it is proved that if A is Hamiltonian, then its elementary divisors $q_i \in \mathbb{C}[\xi]$ are either even polynomials: $q_i(\xi) = (\xi^2 + a^2)^{\mathtt{k}}$, with $a \in \mathbb{R}$; or, if they are not even, then they occur in pairs: as well as $q_i(\xi) = (\xi-\lambda)^{\mathtt{k}}$, $\lambda \in \mathbb{C}$, also $q_i^{\sim}(\xi) = (-\xi-\lambda)^{\mathtt{k}}$ appears. Observe that if $\lambda = 0$, then such paired elementary divisors are necessarily $\pm\xi^{2\mathtt{k}+1}$, and if $\lambda \neq 0$, then they are coprime with each other.

Conclude from these remarks that $\xi I - A$ is Smith equivalent to a diagonal form, where on the diagonal appear either even polynomials, or pairs of polynomials of the

form $\xi^{2\mathtt{k}+1}$. We now use the argument of Theorem 2 of [5] in order to prove the claim. Let $1 \leq \mathtt{k} \leq \mathtt{n}$, and consider the set $M_{\mathtt{k}}$ consisting of all $\mathtt{k} \times \mathtt{k}$ minors of $\xi I - A$; observe that $M_{\mathtt{k}}$ contains only even or odd polynomials. Now consider the greatest common divisor $\Delta_{\mathtt{k}}$ of the polynomials in $M_{\mathtt{k}}$, and observe that for every $m \in M_{\mathtt{k}}$ there exists $m' \in \mathbb{R}[\xi]$ such that $m = \Delta_{\mathtt{k}} m'$; moreover, since m is either even or odd, it holds that $m^{\sim} = \pm m = \Delta_{\mathtt{k}}^{\sim} m'^{\sim}$. It follows that $\Delta_{\mathtt{k}}^{\sim}$ divides every polynomial in $M_{\mathtt{k}}$. Consequently $\Delta_{\mathtt{k}}^{\sim}$ divides $\Delta_{\mathtt{k}}$, and by symmetry also the converse holds. It follows that $\Delta_{\mathtt{k}} = \pm \Delta_{\mathtt{k}}^{\sim}$. Since the invariant polynomials are obtained dividing $\Delta_{\mathtt{k}}$ by $\Delta_{\mathtt{k}-1}$ (with $\Delta_0 := 1$), it follows that the invariant polynomials of $\xi I - A$ are either even or odd.

In order to prove the claim regarding the paired odd polynomials, assume by contradiction that there exists a pair of odd invariant polynomials of A for which zero is a root with multiplicity $2\mathtt{k}_1 + 1$ and $2\mathtt{k}_2 + 1$, respectively, $\mathtt{k}_1 \neq \mathtt{k}_2$, and which cannot be paired otherwise. Observe that $\xi^{2\mathtt{k}_1+1}$ and $\xi^{2\mathtt{k}_2+1}$ are elementary divisors of A. Conclude that in the Jordan form of A there are two blocks associated with zero, of dimension $2\mathtt{k}_1 + 1$ and $2\mathtt{k}_2 + 1$, respectively, which cannot be paired otherwise. This, however, is in contradiction with the results of [4] on the elementary divisors. This concludes the proof. □

3. Autonomous Hamiltonian systems. The definition of an autonomous Hamiltonian system is as follows.

DEFINITION 3.1. *Let $\mathfrak{B} \in \mathcal{L}^{\mathtt{w}}$ be autonomous. $\mathfrak{B}$ is called Hamiltonian if there exists a bilinear differential form L_Ψ, such that*

(i) $\frac{d}{dt} L_\Psi(w_1, w_2) = 0$ *for all* w_1, $w_2 \in \mathfrak{B}$;
(ii) L_Ψ *is skew-symmetric;*
(iii) $L_\Psi|_{\mathfrak{B}}$ *is nondegenerate.*

In Definition 3.1 no assumption on the number $\mathtt{w}$ of external variables of $\mathfrak{B}$ is made. This point of view is in contrast with the usual definition of an autonomous Hamiltonian system, in which a symplectic structure on the space of the external variables (and consequently, an even number of such variables) is assumed. We believe that in order to investigate linear, finite-dimensional Hamiltonian systems, Definition 3.1 is a natural starting point, more so than the classical one in mechanics, as argued in the following examples.

Example 3.2. Consider a spring-mass system without friction, with behavior $\mathfrak{B}$ represented by the equation $m \frac{d^2}{dt^2} w$ where w is the displacement from the equilibrium position. The BDF L_Ψ induced by $\Psi(\zeta, \eta) = m(\zeta - \eta)$ is skew-symmetric. In fact, $L_\Psi(w_1, w_2) = m(\frac{d}{dt} w_1) w_2 - m(\frac{d}{dt} w_2) w_1$. It is easily seen that $\frac{d}{dt} L_\Psi(w_1, w_2) = 0$ for all $w_1, w_2 \in \mathfrak{B}$. Also

$$L_\Psi(w_1, w_2)(0) = \begin{pmatrix} w_1(0) \\ \frac{d}{dt} w_1(0) \end{pmatrix}^T \begin{pmatrix} 0 & -m \\ m & 0 \end{pmatrix} \begin{pmatrix} w_2(0) \\ \frac{d}{dt} w_2(0) \end{pmatrix},$$

which clearly defines a nondegenerate bilinear form on $\mathfrak{B}$. It follows that this spring-mass system with only one external variable is Hamiltonian according to Definition 3.1. It is difficult to understand why, in order to study the Hamiltonianity of such a system from the point of view of classical mechanics, one should first transform the natural second-order differential equation description into a first-order representation in which the position *and* the momentum of the mass are the external variables and then study the symplectic structure of the resulting state-space system.

The previous example illustrates but one situation in which a representation-free definition of Hamiltonianity appears to be more natural than the classical one. The

argument for a definition of Hamiltonianity independent of the particular representation at hand becomes even stronger if one realizes that very often a dynamical system is described by a set of higher-order differential equations, obtained, for example, after elimination of auxiliary variables. The following example illustrates this point.

Example 3.3. Consider two masses m_1 and m_2 attached to springs with constants k_1 and k_2. The first mass is connected to the second one via the first spring, and the second mass is connected to a "wall" with the second spring. Denote by w_1 and w_2 the positions of the masses. Then we can write down the equation of the system as $m_1\frac{d^2w_1}{dt^2} + k_1w_1 - k_1w_2 = 0$, $-k_1w_1 + m_2\frac{d^2w_2}{dt^2} + (k_1+k_2)w_2 = 0$. Eliminate w_2 from the equations and take the position w_1 of the first mass as our external variable w. The behavior $\mathfrak{B}$ of w is represented by

$$m_1m_2\frac{d^4}{dt^4}w + (k_1m_1 + k_2m_1 + k_1m_2)\frac{d^2}{dt^2}w + k_1k_2w = 0.$$

In order to simplify the notation, define $r_0 := k_1k_2$, $r_2 := k_1m_1 + k_2m_1 + k_1m_2$, and $r_4 := m_1m_2$, so that the equation describing w can be rewritten as $r(\frac{d}{dt})w = 0$, where $r(\xi) := r_0 + r_2\xi^2 + r_4\xi^4$. Define the skew-symmetric polynomial $\Psi(\zeta,\eta)$ by $\Psi(\zeta,\eta) = r_2(\zeta-\eta) + r_4(\zeta^3-\eta^3) + r_4(\zeta\eta^2-\zeta^2\eta)$. Observe that $(\zeta+\eta)\Psi(\zeta,\eta) = r(\zeta) - r(\eta)$, and consequently $\frac{d}{dt}L_\Psi(v,w) = (r(\frac{d}{dt})v)^Tw - v^T(r(\frac{d}{dt})w) = 0$ for all $v, w \in \mathfrak{B}$. Moreover, $\Psi(\eta,\zeta) = -\Psi(\zeta,\eta)$, implying that the BDF $L_\Psi(v,w)$ is skew-symmetric. Finally, since the coefficient matrix $\tilde{\Psi}$ of $\Psi(\zeta,\eta)$ is nonsingular, $L_\Psi(v,w)(0)$ clearly defines a nondegenerate bilinear form on $\mathfrak{B}$. Hence the behavior $\mathfrak{B}$ is a Hamiltonian system in the sense of Definition 3.1.

The following theorem gives conditions under which a given autonomous linear differential behavior is Hamiltonian.

THEOREM 3.4. *Let $\mathfrak{B} \in \mathcal{L}^{\mathbb{w}}$ be autonomous. The following conditions are equivalent:*

(1) *$\mathfrak{B}$ is Hamiltonian;*
(2) *every invariant polynomial of $\mathfrak{B}$ is either even or odd, and the odd invariant polynomials can be divided into pairs so that the multiplicity of zero as a root is the same for the polynomials of each pair;*
(3) *there exists a minimal state representation $\frac{d}{dt}x = Ax$, $w = Cx$ of $\mathfrak{B}$, and a symplectic form $\langle\cdot,\cdot\rangle_K$ on the state space $\mathbb{R}^{\mathbb{n}(\mathfrak{B})}$ such that A is a Hamiltonian matrix;*
(4) *for any minimal state representation $\frac{d}{dt}x = Ax$, $w = Cx$ of $\mathfrak{B}$ there exists a symplectic form $\langle\cdot,\cdot\rangle_K$ on the state space $\mathbb{R}^{\mathbb{n}(\mathfrak{B})}$ such that A is a Hamiltonian matrix.*

Proof. We prove (1) $\Rightarrow$ (4) $\Rightarrow$ (3) $\Rightarrow$ (2) $\Rightarrow$ (1).

((1) $\Rightarrow$ (4)) Let L_Ψ satisfy Definition 3.1. Let $\frac{d}{dt}x = Ax$, $w = Cx$ be a minimal state representation of $\mathfrak{B}$. From Proposition 2.1 we have that for $K_\Psi := N_\infty^T\tilde{\Psi}N_\infty \in \mathbb{R}^{\mathbb{n}(\mathfrak{B})\times\mathbb{n}(\mathfrak{B})}$ we have $L_\Psi(w_1,w_2) = x_1^TK_\Psi x_2$ for all (w_i,x_i) $(i=1,2)$ satisfying the system equations and, moreover, $\operatorname{rank}(L_\Psi|_{\mathfrak{B}}) = \operatorname{rank}(K_\Psi)$. Since $L_\Psi|_{\mathfrak{B}}$ is nondegenerate, it follows from Proposition 2.2 that $\operatorname{rank}(K_\Psi) = \mathbb{n}(\mathfrak{B})$. Consequently K_Ψ is nonsingular. Conclude from $\frac{d}{dt}L_\Psi(w_1,w_2) = 0$ for all $w_1, w_2 \in \mathfrak{B}$ that $x_1^T(A^TK_\Psi + K_\Psi A)x_2 = 0$ for all x_i such that $\frac{d}{dt}x_i = Ax_i$, $i = 1,2$. This implies $A^TK_\Psi + K_\Psi A = 0$.

((4) $\Rightarrow$ (3)) is trivial.

((3) $\Rightarrow$ (2)) Conclude from Proposition 2.3 that the invariant polynomials of A are either even or odd and that the odd ones come in pairs and have zero as a root with the same multiplicity. Now let $R \in \mathbb{R}^{\bullet\times\mathbb{w}}[\xi]$ be such that $\mathfrak{B} = \ker R(\frac{d}{dt})$. Since $\mathfrak{B}$ is

autonomous, R has full column rank. Let λ_i, $i = 1, \dots, \mathtt{w}$, be the invariant polynomials of R. Let U and V be the unimodular matrices such that $R = U\mathrm{col}(\Lambda, 0)V$, with $\Lambda := \mathrm{diag}(\lambda_i)_{i=1,\dots,\mathtt{w}}$, the Smith form of R. Then $\mathfrak{B}$ is alternatively represented by $w = V(\frac{d}{dt})^{-1}w'$, $0 = \Lambda(\frac{d}{dt})w'$. We now construct a minimal state representation of $\mathfrak{B}$. For $i = 1, \dots, \mathtt{w}$, define $A_{c,i}$ to be the companion matrix associated with the ith invariant polynomial; also, let C_i be the first vector of the canonical basis of $\mathbb{R}^{\deg(\lambda_i)}$. Define $\hat{A} :=$ block $\mathrm{diag}(A_{c,i})_{i=1,\dots,\mathtt{w}}$, $\hat{C} :=$ block $\mathrm{diag}(C_i)_{i=1,\dots,\mathtt{w}}$. Then $\frac{d}{dt}x = \hat{A}x$, $w' = \hat{C}x$ is a minimal state representation of $\ker \Lambda(\frac{d}{dt})$. In order to come up with a minimal state representation of $\mathfrak{B}$, let $V^{-1}(\xi) = V_0 + V_1\xi + \cdots + V_N\xi^N$; then it is easy to verify that $\frac{d}{dt}x = \hat{A}x$, $w = (V_0\hat{C} + V_1\hat{C}\hat{A} + \cdots + V_N\hat{C}\hat{A}^{N-1})x$ is such a representation. Observe also that $\xi I - \hat{A}$ has the same invariant polynomials of $\xi I - A$, since $\frac{d}{dt}x = Ax$, $w = Cx$ is another minimal state-space representation of $\mathfrak{B}$ (see Lemma 6.3-19 of [10]). This concludes the proof of ((3) $\Rightarrow$ (2)).

((2) $\Rightarrow$ (1)) We first derive a special representation of $\mathfrak{B}$. Let $\mathfrak{B} = \ker R(\frac{d}{dt})$ be a minimal kernel representation of $\mathfrak{B}$, and let $R = U\Delta V$ be a Smith decomposition of R, with Δ the diagonal matrix of the invariant polynomials. Denote the even invariant polynomials of $\mathfrak{B}$ with λ_i, where $i = 1, \dots, \mathtt{e}$. We denote the odd, paired, invariant polynomials with μ_i, $i = 1, \dots, \mathtt{w} - \mathtt{e}$ (observe that $\mathtt{w} - \mathtt{e}$ is even). Now reorder if necessary the invariant polynomials so that the first $\mathtt{e}$ diagonal entries of Δ are the λ_i and the last $\mathtt{w}-\mathtt{e}$ the μ_i. From the division property of the invariant polynomials and from the pairing property of the odd invariant polynomials, it follows that we can write $\mu_{2i+1}(\xi) = \xi\pi_i(\xi)$, $\mu_{2i+2}(\xi) = \xi\pi_i(\xi)g_i(\xi)g_i(-\xi)$ with π_i even and g_i and $g_i^\sim$ coprime; in other words, $g(0) \neq 0$. (Observe that g_i in general has complex coefficients, and consequently in the following it may be necessary to work with polynomial matrices with complex coefficients.)

From these considerations it follows that each 2×2 submatrix $\mathrm{diag}(\mu_{2i+1}, \mu_{2i+2})$, $i = 0, \dots, \frac{\mathtt{w}-\mathtt{e}}{2} - 1$, is Smith-equivalent to

$$\Delta_i'(\xi) = \begin{pmatrix} 0 & -\xi\pi_i(\xi)g_i(-\xi) \\ \xi\pi_i(\xi)g_i(\xi) & 0 \end{pmatrix};$$

in other words, there exist unimodular matrices T_i and $S_i \in \mathbb{C}^{2\times 2}[\xi]$ such that $\Delta_i' = T_i\mathrm{diag}(\mu_{2i+1}, \mu_{2i+2})S_i$. Now define

$$T = \mathrm{diag}(I_r, T_1, \dots, T_{\frac{\mathtt{w}-\mathtt{e}}{2}})$$

and

$$S = \mathrm{diag}(I_r, S_1, \dots, S_{\frac{\mathtt{w}-\mathtt{e}}{2}}),$$

and observe that $T\Delta S = \mathrm{diag}(\Lambda, \Delta_1', \dots, \Delta_{\frac{\mathtt{w}-\mathtt{e}}{2}}') =: \hat{\Delta}$, where $\Lambda := \mathrm{diag}(\lambda_1, \dots, \lambda_{\mathtt{e}})$. Conclude that $(S^{-1}V)^\sim TU^{-1}R = (S^{-1}V)^\sim \hat{\Delta}S^{-1}V =: R' \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$ is another kernel representation of $\mathfrak{B}$ satisfying $R' = R'^\sim$. It is such a representation that we use in order to come up with a BDF as in Definition 3.1.

Consider the unimodular transformation of the external variables represented by $w' := (S^{-1}V)(\frac{d}{dt})w$, and observe that the ith component of w', $i = 1, \dots, \mathtt{e}$, satisfies $\lambda_i(\frac{d}{dt})w_i' = 0$, while the remaining components satisfy $\Delta_i'(\frac{d}{dt})\mathrm{col}(w_i', w_{i+1}') = 0$, $i = 0, \dots, \frac{\mathtt{w}-\mathtt{e}}{2} - 1$. We proceed to construct a skew-symmetric BDF L_{Ψ_i} acting on the ith component w_i' and satisfying Definition 3.1. From such BDFs we will construct a BDF for $\mathfrak{B}$ with the right properties.

We begin by computing such a BDF for the case of even invariant polynomials. Since $\lambda_i - \lambda_i^\sim = 0$, it follows from Theorem 3.1 of [24] that there exists $\Psi_i \in \mathbb{R}[\zeta, \eta]$ such that $(\zeta + \eta)\Psi_i(\zeta, \eta) = \lambda_i(\zeta) - \lambda_i(\eta)$. Observe that $\Psi_i(\zeta, \eta)$ is skew-symmetric and moreover, $\frac{d}{dt} L_{\Psi_i}(w_1, w_2) = 0$ for all $w_1,\ w_2 \in \ \ker \lambda_i(\frac{d}{dt})$. We now prove that L_{Ψ_i} is nondegenerate. Write $\lambda_i(\xi) = \lambda_{i0} + \lambda_{i2}\xi^2 + \cdots + \lambda_{i,2n_i}\xi^{2n_i}$. It is easy to verify that the coefficient matrix $\tilde{\Psi}_i$ of $\Psi(\zeta, \eta)$ is a $2n_i \times 2n_i$ left-upper-triangular matrix with $\pm\lambda_{i,2n_i}$ on the main antidiagonal; as a consequence, the bilinear differential form $L_{\Psi_i}|_{\ker(\lambda_i(\frac{d}{dt}))}$ is nondegenerate. This settles the case of even invariant polynomials.

We now examine the case of paired odd invariant polynomials. Observe first that since π_i is even, the two-variable polynomial $\zeta\pi_i(\zeta)g_i(\zeta) + \eta\pi_i(\eta)g_i(-\eta)$ is divisible by $\zeta + \eta$. Now define the skew-symmetric two-variable polynomial matrix $\Psi_i(\zeta, \eta) \in \mathbb{C}^{2\times 2}[\zeta, \eta]$ as

$$\Psi_i(\zeta, \eta) := \begin{pmatrix} 0 & \frac{\zeta\pi_i(\zeta)g_i(\zeta)+\eta\pi_i(\eta)g_i(-\eta)}{\zeta+\eta} \\ -\frac{\zeta\pi_i(\zeta)g_i(-\zeta)+\eta\pi_i(\eta)g_i(\eta)}{\zeta+\eta} & 0 \end{pmatrix}.$$

We observe that $\frac{d}{dt} L_{\Psi_i}(\mathrm{col}(w_i', w_{i+1}'), \mathrm{col}(\bar{w}_i', \bar{w}_{i+1}')) = 0$ for each pair $\mathrm{col}(w_i', w_{i+1}')$, $\mathrm{col}(\bar{w}_i', \bar{w}_{i+1}')$ of trajectories in $\ker \Delta_i'(\frac{d}{dt})$. We now prove that $L_{\Psi_i}|_{\ker\Delta_i'(\frac{d}{dt})}$ is nondegenerate.

Let $\deg(\pi_i) = 2K_i$, $\deg(g_i) = L_i$. It is a matter of straightforward verification to prove that the coefficients of the terms $\zeta^k\eta^{L_i+2K_i-k}$, $k = 0, \ldots, L_i + 2K_i$, in $\frac{\zeta\pi_i(\zeta)g_i(\zeta)+\eta\pi_i(\eta)g_i(-\eta)}{\zeta+\eta}$ are equal to $g_{i,L_i}\pi_{i,2K_i} \neq 0$. It follows that the coefficient matrix $\tilde{\Psi}_i$ is a block-left-upper-triangular matrix with on the diagonal 2×2 blocks of the form

$$(-1)^k \begin{pmatrix} 0 & g_{i,L_i}\pi_{i,2K_i} \\ -g_{i,L_i}\pi_{i,2K_i} & \end{pmatrix}.$$

We conclude from this that $\tilde{\Psi}_i$ is nonsingular, and consequently the bilinear form $L_{\Psi_i}|_{\ker\Delta_i'(\frac{d}{dt})}$ is nondegenerate. This settles the case of paired odd invariant polynomials.

In order to complete the proof of the claim (2) $\Rightarrow$ (1), assume $\Psi_i(\zeta, \eta)$ has been constructed as described above; now define

$$\Psi(\zeta, \eta) := (S^{-1}V)(\zeta)^T \mathrm{diag}(\Psi_i(\zeta, \eta))(S^{-1}V)(\eta) \in \mathbb{R}^{\mathbf{q}\times\mathbf{q}}[\zeta, \eta],$$

where S, V are the unimodular matrix involved in obtaining the special decomposition of R'. (Observe that if some $\Psi_i(\zeta, \eta)$ has complex coefficients, then transposition *and* complex conjugation are required.) $\Psi(\zeta, \eta)$ induces a skew-symmetric BDF L_Ψ whose derivative is zero along $\mathfrak{B}$. The nondegeneracy of $L_\Psi|_{\mathfrak{B}}$ follows immediately from the nondegeneracy of the forms $L_{\Psi_i}|_{\ker\lambda_i(\frac{d}{dt})}$ and $L_{\Psi_i}|_{\ker\Delta_i'(\frac{d}{dt})}$. This concludes the proof. □

4. Hamiltonian systems and the Euler–Lagrange equations. In section 3 we introduced the notion of Hamiltonianity on the basis of the interplay of a skew-symmetric BDF with the dynamics of the system, without reference to the notion of Lagrangian as occurs in classical mechanics. In this section we reconcile the classical point of view with our standpoint.

We begin by introducing the notion of stationarity of a trajectory with respect to a QDF. Let $\Phi \in \mathbb{R}^{\mathbf{w}\times\mathbf{w}}[\zeta, \eta]$ be symmetric and consider the corresponding QDF

$Q_\Phi(w)$ on $\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathbb{w}})$. For a given w we define the *cost degradation* of adding the compact-support function $\delta \in \mathfrak{D}(\mathbb{R},\mathbb{R}^{\mathbb{w}})$ to w as

$$J_w(\delta) := \int_{-\infty}^{+\infty} (Q_\Phi(w+\delta) - Q_\Phi(w))dt.$$

The cost degradation equals $J_w(\delta) = \int_{-\infty}^{+\infty} Q_\Phi(\delta)dt + 2\int_{-\infty}^{+\infty} L_\Phi(w,\delta)dt$, and we call the second integral on the right of the equality sign the *variation associated with* w. It defines a linear functional which associates with every $\delta \in \mathfrak{D}(\mathbb{R},\mathbb{R}^{\mathbb{w}})$ a real number $2\int_{-\infty}^{+\infty} L_\Phi(w,\delta)dt$. We call w a *stationary trajectory* of Q_Φ if the variation associated with w is the zero functional. The following proposition establishes a representation of all stationary trajectories of given QDF Q_Φ. Recall that, for a given two-variable polynomial matrix $\Phi(\zeta,\eta)$, $\partial\Phi(\xi)$ is defined as the one-variable polynomial matrix $\Phi(-\xi,\xi)$.

PROPOSITION 4.1. *Let $\Phi(\zeta,\eta) \in \mathbb{R}^{\mathbb{w}\times\mathbb{w}}[\zeta,\eta]$ be symmetric. Then $w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathbb{w}})$ is a stationary trajectory of the QDF Q_Φ if and only if w satisfies the differential equation*

$$\partial\Phi\left(\frac{d}{dt}\right) w = 0. \tag{4.1}$$

Proof. Factor $\Phi(\zeta,\eta) = M^T(\zeta)\Sigma M(\eta)$, with Σ a nonsingular signature matrix, and $M(\xi) = M_0 + M_1\xi + M_2\xi^2 + \cdots + M_L\xi^L$ (see section 3 of [24]). Integrating by parts on $\delta \in \mathfrak{D}(\mathbb{R},\mathbb{R}^{\mathbb{w}})$, the variation $\int_{-\infty}^{+\infty}(M(\frac{d}{dt})w)^T\Sigma M(\frac{d}{dt})\delta dt$ is seen to be equal to

$$\sum_{k=1}^{L}\sum_{j=k}^{L}(-1)^{k-1}\delta^{(j-k)}M_j^T\Sigma\left(M\left(\frac{d}{dt}\right)w\right)^{(k-1)}\Bigg|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty}\delta^T\left(M\left(-\frac{d}{dt}\right)^T\Sigma M\left(\frac{d}{dt}\right)w\right)dt.$$

Such a quantity is zero if and only if $M(-\frac{d}{dt})^T\Sigma M(\frac{d}{dt})w = 0$; equivalently, $\partial\Phi(\frac{d}{dt})w = 0$. ∎

From the classical theory of calculus of variations it is well known that the stationary trajectories for a given functional can be characterized in terms of the so-called higher-order Euler equations, often called the Euler–Poisson equations. If the functional is given by a QDF Q_Φ, then (4.1) can indeed be interpreted as a classical Euler–Poisson equation. In order to verify this, let $\tilde{\Phi}_{\text{eff}} := (\Phi_{k,\ell})_{k,\ell=0,\ldots,L}$. Now observe that $Q_\Phi(w)$ can be written as $F(w,w^{(1)},w^{(2)},\ldots,w^{(L)})$, with the functional $F : \mathbb{R}^{\mathbb{w}}\times\mathbb{R}^{\mathbb{w}}\times\cdots\times\mathbb{R}^{\mathbb{w}} \to \mathbb{R}$ defined by

$$F(w_0,w_1,w_2,\ldots,w_L) := \operatorname{col}(w_0,w_1,w_2,\ldots,w_L)^T\ \tilde{\Phi}_{\text{eff}}\ \operatorname{col}(w_0,w_1,w_2,\ldots,w_L).$$

In terms of this functional F, the stationary trajectories w are the solutions of the Euler–Poisson equation

$$\left(\frac{\partial F}{\partial w_0} - \frac{d}{dt}\frac{\partial F}{\partial w_1} + \frac{d^2}{dt^2}\frac{\partial F}{\partial w_2} - \cdots + (-1)^L\frac{d^L}{dt^L}\frac{\partial F}{\partial w_L}\right)(w^{(0)},\ldots,w^{(L)}) = 0. \tag{4.2}$$

It is a matter of straightforward computation to see that the equations given by (4.1) and (4.2) indeed coincide. Henceforth we will, for a given QDF Q_Φ, refer to the differential equation (4.1) as the *Euler–Poisson equation associated with* Q_Φ.

According to the *principle of least action*, the motions that are possible in a mechanical system can be obtained as the stationary trajectories of the Lagrangian, the difference between the kinetic and potential energy of the system, which is in general represented as a function of displacement and velocity. Accordingly, the corresponding Euler–Poisson equation is a system of second-order differential equations, called the *Euler–Lagrange equations* associated with the mechanical system. Thus, the possible motions in a mechanical system form a behavior represented by the Euler–Lagrange equations. We now study the converse problem (called the "inverse problem of the calculus of variations"; see [17]): *under which conditions does a linear differential behavior* $\mathfrak{B}$ *(typically described by a system of higher-order linear differential equations) consist of the stationary trajectories with respect to some functional interpretable as a Lagrangian (i.e., a functional that represents the difference between kinetic and potential energy in a suitable sense), and how does one construct such a functional on the basis of the equations describing* $\mathfrak{B}$*?* It turns out that under mild assumptions, this is the case if and only if $\mathfrak{B}$ is a Hamiltonian system. This leads us to define the notion of generalized position and generalized Lagrangian and to address the issue of the existence of second-order latent variable representations of Hamiltonian behaviors. See also [7] and [6], where the inverse problem is considered for nonlinear i/o systems, and a characterization of Hamiltonian systems is given in terms of the properties of the i/o differential equations describing them.

THEOREM 4.2. *Let* $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ *be autonomous, and assume that* $\chi_{\mathfrak{B}}$*, the characteristic polynomial of* $\mathfrak{B}$*, has no root in zero,* $\chi_{\mathfrak{B}}(0) \neq 0$*. Then the following statements are equivalent:*

(1) $\mathfrak{B}$ *is Hamiltonian.*

(2) $\mathtt{n}(\mathfrak{B})$ *is even, and there exists a full column rank matrix* $P(\xi) \in \mathbb{R}^{\mathtt{q}\times\mathtt{w}}[\xi]$*, and nonsingular matrices* $M = M^T, K = K^T \in \mathbb{R}^{\mathtt{q}\times\mathtt{q}}$*, with* $\mathtt{q} := \mathtt{n}(\mathfrak{B})/2$*, such that* $\mathfrak{B}$ *is equal to the space of all stationary trajectories with respect to the QDF*

$$Q_L(w) = \left|\frac{d}{dt}P\left(\frac{d}{dt}\right)w\right|^2_M - \left|P\left(\frac{d}{dt}\right)w\right|^2_K;$$

equivalently, $\mathfrak{B} = \{w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}}) \mid \partial L(\frac{d}{dt})w = 0\}$*, with* $L(\zeta, \eta)$ *defined by*

$$L(\zeta, \eta) := P(\zeta)^T(\zeta\eta M - K)P(\eta). \tag{4.3}$$

Furthermore, if any of these conditions holds, then P*,* M*, and* K *satisfying the conditions in* (2)*. can be chosen in such a way that in addition*

$$\frac{d}{dt}Q_H(w) = 0 \text{ for all } w \in \mathfrak{B}, \tag{4.4}$$

where $Q_H(w) := |\frac{d}{dt}P(\frac{d}{dt})w|^2_M + |P(\frac{d}{dt})w|^2_K$.

Proof. ((1) $\Rightarrow$ (2)) Since $\mathfrak{B}$ is Hamiltonian and $\chi_{\mathfrak{B}}(0) \neq 0$, $\mathfrak{B}$ has only even invariant polynomials. We reduce to the scalar case by use of the Smith form. Consider a minimal representation of the behavior $\mathfrak{B}$ as $\mathfrak{B} = \ker R(\frac{d}{dt})$, and let $R = U\Delta V$ be the Smith decomposition of R, with Δ being a diagonal matrix. Define the behavior $\mathfrak{B}' := V(\frac{d}{dt})\mathfrak{B}$ with manifest variable w', and observe that $\mathfrak{B}' = \ker \Delta(\frac{d}{dt})$. We now examine each of the behaviors $\mathfrak{B}'_i := \ker \lambda_i(\frac{d}{dt})$, $i = 1, \ldots, \mathtt{w}$, one at a time.

Let $\lambda_i(\xi) := \lambda_{i0} + \lambda_{i2}\xi^2 + \cdots + \lambda_{i,2L_i}\xi^{2L_i}$, $\lambda_{i,2L_i} \neq 0$. Consider the following two matrices:

$$M_i := \begin{pmatrix} \lambda_{i,2} & \lambda_{i,4} & \lambda_{i,6} & \cdots & \lambda_{i,2L_i} \\ \lambda_{i,4} & \lambda_{i,6} & \lambda_{i,8} & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \lambda_{i,2L_i-2} & \lambda_{i,2L_i} & 0 & \cdots & 0 \\ \lambda_{i,2L_i} & 0 & 0 & \cdots & 0 \end{pmatrix} \tag{4.5}$$

and

$$K_i := \begin{pmatrix} \lambda_{i,0} & 0 & 0 & \cdots & 0 \\ 0 & -\lambda_{i,4} & -\lambda_{i,6} & \dots & -\lambda_{i,2L_i} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & -\lambda_{i,2L_i-2} & -\lambda_{i,2L_i} & \cdots & 0 \\ 0 & -\lambda_{i,2L_i} & 0 & \cdots & 0 \end{pmatrix}. \tag{4.6}$$

It is immediate to see that M_i is nonsingular; the nonsingularity of K_i follows from $\lambda_{i,2L_i} \neq 0$ and the fact that $\lambda_{i,0} \neq 0$, since $\chi_{\mathfrak{B}}(0) \neq 0$. Let $E_i(\xi) := \mathrm{col}(1, \xi^2, \dots, \xi^{2L_i-2})$ and $e_i := \mathrm{col}(1, 0, \dots, 0) \in \mathbb{R}^{L_i \times 1}$. Then the following equation holds:

$$(M_i\xi^2 + K_i)E_i(\xi) = \lambda_i(\xi)e_i. \tag{4.7}$$

This implies $E_i^T(-\xi)(M_i\xi^2 + K_i)E_i(\xi) = \lambda_i(\xi)$. Define $E(\xi) :=$ block diag$(E_i(\xi))$, $M =$ block diag(M_i), and $K =$ block diag(K_i); then $\Delta(\xi) = E^T(-\xi)((M\xi^2+K)E(\xi)$. Finally, let $P(\xi) := E(\xi)V(\xi)$. Since $\mathfrak{B}' = \ker \Delta(\frac{d}{dt})$, also $\mathfrak{B} = \ker P^T(-\frac{d}{dt})(M\frac{d^2}{dt^2} + K)P(\frac{d}{dt})$. By defining $L(\zeta, \eta) := P^T(\zeta)M\zeta\eta - K)P(\eta)$ we then obtain $\mathfrak{B} = \ker \partial L(\frac{d}{dt})$; equivalently, $\mathfrak{B}$ is the space of stationary trajectories of the QDF $Q_L(w)$.

((2) $\Rightarrow$ (1)) Since $\mathfrak{B} = \ker \partial\Phi(\frac{d}{dt})$, the claim is proved if we show that the invariant polynomials of ∂L are all even. Observe first that ∂L is para-Hermitian. Now let $1 \leq \mathtt{k} \leq \mathtt{w}$, and consider the set $M_{\mathtt{k}}$ consisting of all $\mathtt{k} \times \mathtt{k}$ minors of $\partial L(\xi)$. Observe that since $\partial L = (\partial L)^{\sim}$, if $m \in M_{\mathtt{k}}$, then also $m^{\sim} \in M_{\mathtt{k}}$. Now use an argument analogous to that used in the proof of Proposition 2.3 in order to conclude that the invariant polynomials of ∂L are either even or odd. Conclude from $\chi_{\mathfrak{B}}(0) \neq 0$ that there are no odd invariant polynomials; then it follows from statement (2) of Theorem 3.4 that $\mathfrak{B}$ is Hamiltonian.

The rest of the claim of the theorem follows easily from the definition of L. This concludes the proof. ∎

If, in statement (2) of Theorem 4.2, we interpret $q = P(\frac{d}{dt})w$ as generalized position, then $\frac{d}{dt}q = \frac{d}{dt}P(\frac{d}{dt})w$ is generalized velocity, and consequently $|\frac{d}{dt}P(\frac{d}{dt})w|^2_M = |\frac{d}{dt}q|^2_M$ and $|P(\frac{d}{dt})w|^2_K = |q|^2_K$ can be interpreted, respectively, as kinetic and potential energy. From this point of view, the QDF $Q_L(w)$ can be interpreted as a Lagrangian, and the QDF $Q_H(w)$ can be interpreted as a Hamiltonian of the system.

The equation $\partial L(\frac{d}{dt})w = 0$ is the Euler–Poisson equation associated with the QDF Q_L. Motivated by the fact that $Q_L(w)$ can be interpreted as a Lagrangian of the system $\mathfrak{B}$, we also call it an *Euler–Lagrange equation* associated with the system $\mathfrak{B}$.

The next theorem relates Hamiltonianity with the existence of a latent variable representation of second order in the latent variable $q = P(\frac{d}{dt})w$, with P as in Theorem 4.2.

THEOREM 4.3. *Let $\mathfrak{B} \in \mathcal{L}^{\mathtt{w}}$ be autonomous. Assume that $\chi_{\mathfrak{B}}(0) \neq 0$, i.e., its characteristic polynomial has no root in 0. Then the following statements are equivalent:*

(1) *$\mathfrak{B}$ is Hamiltonian.*

(2) *$\mathtt{n}(\mathfrak{B})$ is even, and there exist $M = M^T, K = K^T \in \mathbb{R}^{\mathtt{q}\times\mathtt{q}}$ nonsingular, $C_1 \in \mathbb{R}^{\mathtt{w}\times\mathtt{q}}$ and $C_2 \in \mathbb{R}^{\mathtt{w}\times\mathtt{q}}$, with $\mathtt{q} := \mathtt{n}(\mathfrak{B})/2$, such that*

$$M\frac{d^2}{dt^2}q + Kq = 0, \qquad C_1 q + C_2\frac{d}{dt}q = w \tag{4.8}$$

is an observable latent variable representation of $\mathfrak{B}$ with latent variable q. Furthermore, if $M\frac{d^2}{dt^2}q + Kq = 0$, $w = C_1q + C_2\frac{d}{dt}q$ is an observable latent variable representation of $\mathfrak{B}$ with latent variable q, then the equations $\bar{M}\frac{d^2}{dt^2}\bar{q} + \bar{K}\bar{q} = 0$, $w = \bar{C}_1\bar{q} + \bar{C}_2\frac{d}{dt}\bar{q}$ form an observable latent variable representation of $\mathfrak{B}$ with latent variable $\bar{q}$ if and only if there exist $S, T \in \mathbb{R}^{\mathtt{q}\times\mathtt{q}}$ such that the matrix

$$\begin{pmatrix} S & T \\ -TM^{-1}K & S \end{pmatrix} \tag{4.9}$$

is nonsingular, and the relations $SM^{-1}K = \bar{M}^{-1}\bar{K}S$, $TM^{-1}K = \bar{M}^{-1}\bar{K}T$, $C_1 = \bar{C}_1S - \bar{C}_2TM^{-1}K$, and $C_2 = \bar{C}_1T + \bar{C}_2S$ hold.

Proof. ((1) $\Rightarrow$ (2)) Define M_i and K_i by (4.5) and (4.6), respectively. From (4.7) it follows that $w_i' \in \mathfrak{B}_i' = \ker \lambda_i(\frac{d}{dt})$ if and only if $(M_i\frac{d^2}{dt^2} + K_i)E_i(\frac{d}{dt})w_i' = 0$. Using this, it is easily seen that the equations $M_i\frac{d^2q}{dt^2} + K_iq = 0$, $q_i = E_i(\frac{d}{dt})w_i'$ form an observable latent variable representation of $\mathfrak{B}_i'$ with latent variable q_i. Next consider the equations $M_i\frac{d^2q_i}{dt^2} + K_iq_i = 0$, $e_i^Tq_i = w_i'$, where as before $e_i := \mathrm{col}(1, 0, \ldots, 0) \in \mathbb{R}^{L_i\times 1}$. Using the special structure of M_i and K_i, it can be verified that also these equations form an observable latent variable representation of $\mathfrak{B}_i'$. Now define $C :=$ block diag(e_i^T). Then clearly $M\frac{d^2q}{dt^2} + Kq = 0$, $Cq = w'$ is an observable latent variable representation of $\mathfrak{B}'$. An observable latent variable representation of the original system $\mathfrak{B}$ is then obtained by replacing the equation $w' = Cq$ by $w = V^{-1}(\frac{d}{dt})Cq$. Note that $V^{-1}(\xi)C$ is a polynomial matrix and that any derivative $q^{(i)}$ with $i \geq 2$ can be expressed in terms of q or $q^{(1)}$ using the first equation in (4.8). From this we conclude that matrices C_1 and C_2 exist such that $M\frac{d^2q}{dt^2} + Kq = 0$, $w = C_1q + C_2\frac{d}{dt}q$ is an observable latent variable representation of $\mathfrak{B}$.

((2) $\Rightarrow$ (1)) By defining $x_1 = q$, $x_2 = \frac{dq}{dt}$, and $C = (C_1 \quad C_2)$, the state-space representation of $\mathfrak{B}$ given by $\frac{d}{dt}x_1 = x_2$, $\frac{d}{dt}x_2 = M^{-1}Kx_1$, $w = Cx$ is obtained. Clearly, (x_1, x_2) is observable from w, so this state-space representation is minimal. Observe also that

$$\begin{pmatrix} 0 & I \\ -M^{-1}K & 0 \end{pmatrix}^T \begin{pmatrix} 0 & M \\ -M & 0 \end{pmatrix} + \begin{pmatrix} 0 & M \\ -M & 0 \end{pmatrix}\begin{pmatrix} 0 & I \\ -M^{-1}K & 0 \end{pmatrix} = 0.$$

Since M is nonsingular, we conclude from statement (3) of Theorem 3.4 that $\mathfrak{B}$ is Hamiltonian.

Assume now that $M\frac{d^2}{dt^2}q + Kq = 0$, $w = C_1q + C_2\frac{d}{dt}q$ is an observable latent variable representation of $\mathfrak{B}$ with latent variable q. Suppose the equations $\bar{M}\frac{d^2}{dt^2}\bar{q} +$

$\bar{K}\bar{q} = 0$, $w = \bar{C}_1\bar{q} + \bar{C}_2\frac{d}{dt}\bar{q}$ form an observable latent variable representation of $\mathfrak{B}$ with latent variable $\bar{q}$. Then clearly $\frac{d}{dt}x = Ax$, $w = Cx$ and $\frac{d}{dt}\bar{x} = \bar{A}\bar{x}$, $w = \bar{C}\bar{x}$ with

$$A := \begin{pmatrix} 0 & I \\ -M^{-1}K & 0 \end{pmatrix}, \qquad \bar{A} := \begin{pmatrix} 0 & I \\ \bar{M}^{-1}\bar{K} & 0 \end{pmatrix},$$

$$C := \begin{pmatrix} C_1 & C_2 \end{pmatrix}, \ \bar{C} := \begin{pmatrix} \bar{C}_1 & \bar{C}_2 \end{pmatrix}$$

are two minimal state representations of $\mathfrak{B}$. Consequently there exists a nonsingular $F \in \mathbb{R}^{2\mathtt{q}\times 2\mathtt{q}}$ such that $\bar{A}F = FA$ and $\bar{C}F = C$. Using the special structure of the state-space representations it is easily seen that F must be of the form (4.9) and that the relations in the claim must hold. Conversely, by defining F by (4.9), we see that $\bar{A}F = FA$ and $\bar{C}F = C$, so that the corresponding state realizations have the same manifest behavior $\mathfrak{B}$. □

Remark 4.4. Note that in the proof of implication (2) ⇒ (1) the assumption that $\chi_{\mathfrak{B}}(\xi)$ has no root in $\xi = 0$ is not used. Our proof of implication (1) ⇒ (2) does use this assumption. At present we do not have a proof of or a counterexample to this implication without the assumption $\chi_{\mathfrak{B}}(0) \neq 0$.

In classical mechanics, the variable q consists of the position of the masses, and (4.8) is obtained by writing down Newton's second law. Theorem 4.3 shows that a Hamiltonian behavior can *always* be interpreted in some sense as a *"mechanical* system," with q "generalized position," $\frac{d}{dt}q$ "generalized velocity," M a "mass matrix," and K a matrix of "elastic constants." Further, the theorem characterizes all such representations. Of course $\frac{d}{dt}q^T M \frac{d}{dt}q$ can be interpreted as kinetic energy, $q^T Kq$ as potential energy, $(\frac{d}{dt}q)^T M(\frac{d}{dt}q) - q^T Kq$ as a Lagrangian, and $(\frac{d}{dt}q)^T M(\frac{d}{dt}q) + q^T Kq$ (which is constant along solutions q of the first equation in (4.8)) as total energy.

The similarity between (4.8) and the second-order representation typical of conservative mechanical systems should not, however, be pushed too far. Indeed, the reader can verify that the matrix M defined in the proof of Theorem 4.3 is in general not positive-definite, as a bona fide mass matrix should be. Nor would the "kinetic" and "potential" terms $\mid P(\frac{d}{dt})w \mid^2_M$ and $\mid P(\frac{d}{dt})w \mid^2_K$ have in general the physical dimensions of energies. For further discussion on these issues see the following example.

Example 4.5. Consider the configuration of Example 3.3. As shown in that example, the behavior $\mathfrak{B}$ of the position w of the first mass is represented by

$$r\left(\frac{d}{dt}\right)w = m_1m_2\frac{d^4}{dt^4}w + (k_1m_1 + k_2m_1 + k_1m_2)\frac{d^2}{dt^2}w + k_1k_2w = 0.$$

In order to obtain an observable second-order latent variable representation (4.8) of $\mathfrak{B}$ we proceed as in the proof of Theorem 4.3 and define the latent variable $q = \mathrm{col}(w, \frac{d^2}{dt^2}w)$. Then

$$M := \begin{pmatrix} r_2 & r_4 \\ r_4 & 0 \end{pmatrix}, \quad K := \begin{pmatrix} r_0 & 0 \\ 0 & -r_4 \end{pmatrix}, \quad C_1 = 1, \quad C_2 = 0.$$

The matrix M is not positive-definite, as can be verified choosing $k_1 = k_2 = 1\ \frac{\mathrm{N}}{\mathrm{m}}$ and $m_1 = m_2 = 1$ kg. Observe also that with such a choice of M, a physical interpretation of $\frac{d}{dt}q^T M \frac{d}{dt}q$ is impossible, since the physical dimensions of such a quantity are not those of an energy. However, by choosing

$$\bar{K} := \begin{pmatrix} k_2 & \frac{k_2m_1}{k_1} \\ \frac{k_2m_1}{k_1} & \frac{m_1^2(k_1+k_2)}{k_1^2} \end{pmatrix}, \quad \bar{M} := \begin{pmatrix} m_1+m_2 & \frac{m_1m_2}{k_1} \\ \frac{m_1m_2}{k_1} & \frac{m_1^2m_2}{k_1^2} \end{pmatrix}, \quad \bar{C}_1 := 1, \quad \bar{C}_2 := 0,$$

we obtain an alternative second-order latent variable representation of $\mathfrak{B}$, with the same latent variable as before. Note that $\bar{M}^{-1}\bar{K} = M^{-1}K$ so that this alternative representation is obtained from the original one by taking in Theorem 4.3 $S = I$, $T = 0$. For this second choice of representation, the generalized kinetic energy $(\frac{d}{dt}q)^T M(\frac{d}{dt}q)$ and generalized potential energy $q^T Kq$ coincide with the *physical* kinetic energy and potential energy of the system with the two masses, that is, $E_{\text{kin}}(w_1, w_2) = \frac{1}{2}(m_1(\frac{d}{dt}w_1)^2 + m_2(\frac{d}{dt}w_2)^2)$ and $E_{\text{pot}}(w_1, w_2) = \frac{1}{2}(k_1 w_1^2 - 2k_1 w_1 w_2 + (k_1 + k_2)w_2^2)$. This can be verified observing that the generalized position $q = (w, \frac{d^2}{dt^2}w)$ is related to the actual position (w_1, w_2) as $w_1 = w$, $w_2 = w + \frac{m_1}{k_1}\frac{d^2}{dt^2}w$.

It is a matter for further investigation to see whether and how a physically consistent choice of the matrices M and K can always be performed. Such an issue is particularly pressing when considering the use of the procedures presented in this paper for computer-assisted modeling and simulation.

5. Internal forces and controllable Hamiltonian systems. In this section we define internal forces as auxiliary variables and we show how they can be obtained from a higher-order Lagrangian such as that introduced in Theorem 4.2. The notion of internal force obtained in this way brings us in a natural way to the definition of a controllable Hamiltonian system; in this section we also give various characterizations of such systems in terms of their kernel, image, or state-space representations.

According to Theorem 4.2, an autonomous system $\mathfrak{B} \in \mathcal{L}^{\mathtt{w}}$ with $\chi_{\mathfrak{B}}(0) \neq 0$ is Hamiltonian if and only if there exists a polynomial matrix $P \in \mathbb{R}^{\mathtt{q}\times\mathtt{w}}[\xi]$ with full column rank, $M = M^T, K = K^T \in \mathbb{R}^{\mathtt{q}\times\mathtt{q}}$ nonsingular, with $\mathtt{q} := \mathfrak{n}(\mathfrak{B})/2$, such that $\mathfrak{B}$ has a kernel representation $P^T(-\frac{d}{dt})(M\frac{d^2}{dt^2} + K)P(\frac{d}{dt})w = 0$. Obviously, a latent variable representation with latent variable f of $\mathfrak{B}$ is then given by the equations

$$\begin{aligned} P^T\left(-\frac{d}{dt}\right) M\frac{d^2}{dt^2}P\left(\frac{d}{dt}\right) w &= f, \\ f &= -P^T\left(-\frac{d}{dt}\right) KP\left(\frac{d}{dt}\right) w. \end{aligned} \tag{5.1}$$

For a given $w \in \mathfrak{B}$, the associated $f = -P^T(-\frac{d}{dt})KP(\frac{d}{dt})w$ is called the *internal force* associated with w. Observe that there are as many internal forces as there are external variables, and that in the case of systems described by differential equations of order higher than two, the internal force depends on higher-order derivatives of the external variable (see Chapter 2, section 31 of [21], where, in the context of the dynamics of a moving charge in an electromagnetic field, an internal force is considered which depends on a potential function which depends on position *and* velocity). Such a definition harmonizes with the classical mechanics point of view of seeing the internal force as an auxiliary variable of the same dimension as the external variables and coming from some potential function depending on the configuration (external) variables. Indeed, when applied to the prototypical mechanical system $M\frac{d^2}{dt^2}q + Kq = 0$, $w = q$, equations (5.1) result in the internal force being defined as $f = -Kq$ and coming from the potential $V(q) = q^T Kq$.

The notion of internal force sheds light on the structure of autonomous Hamiltonian systems. Define $\mathfrak{B}_1, \mathfrak{B}_2 \in \mathcal{L}^{\mathtt{w}+\mathtt{f}}$ by

$$\mathfrak{B}_1 := \left\{ \begin{pmatrix} w \\ f \end{pmatrix} \mid \begin{pmatrix} P^T(-\frac{d}{dt})M\frac{d^2}{dt^2}P(\frac{d}{dt}) & -I \end{pmatrix} \begin{pmatrix} w \\ f \end{pmatrix} = 0 \right\}, \tag{5.2}$$

$$\mathfrak{B}_2 := \left\{ \begin{pmatrix} w \\ f \end{pmatrix} \mid \begin{pmatrix} P^T(-\frac{d}{dt})KP(\frac{d}{dt}) & I \end{pmatrix} \begin{pmatrix} w \\ f \end{pmatrix} = 0 \right\}. \tag{5.3}$$

Note that both $\mathfrak{B}_1$ and $\mathfrak{B}_2$ are controllable linear differential systems. The set of trajectories (w, f) compatible with the laws of both $\mathfrak{B}_1$ and $\mathfrak{B}_2$ is the behavior $\mathfrak{B}_1 \cap \mathfrak{B}_2$; we call it the *full interconnection* of $\mathfrak{B}_1$ and $\mathfrak{B}_2$. Now assume that P, M, and K in (5.1) have been computed as in the proof of ((1) $\Rightarrow$ (2)) of Theorem 4.2. It is a matter of straightforward verification to see that in such a case $P^T(-\xi)MP(\xi)$ is nonsingular; this implies that in $\mathfrak{B}_1$, f is input and w is output; it can also be verified that in $\mathfrak{B}_2$, w is input and f is output. In such a case the interconnection of $\mathfrak{B}_1$ and $\mathfrak{B}_2$ is called a *feedback interconnection* (see [23]).

We conclude that *any autonomous Hamiltonian behavior* $\mathfrak{B}$ *is the feedback interconnection of two systems, the first one* ($\mathfrak{B}_1$) *having a free* f *variable and the second one* ($\mathfrak{B}_2$) *imposing on such a variable the additional constraint represented by the second equation in* (5.1). From this standpoint, in $\mathfrak{B}_1$, f is an *external force* which can be chosen freely, while $\mathfrak{B}_2$ constrains it to be a function of the external variables w. This point of view has much in common with the notion of "Hamiltonian interconnection" introduced in [18], where the concept of an open Hamiltonian system is introduced from a system-theoretic point of view.

Example 5.1. We consider again the system described in Examples 3.3 and 4.5. Following the procedure illustrated above, $\mathfrak{B}_1$ and $\mathfrak{B}_2$ defined in (5.2) and (5.3) are described, respectively, by

$$\begin{aligned} &(m_1 + m_2)w + 2\frac{m_1 m_2}{k_1}\frac{d^2}{dt^2}w + \frac{m_1^2 m_2}{k_1^2}\frac{d^4}{dt^4}w = f, \\ &f = k_2 w + \frac{2k_2 m_1}{k_1}\frac{d^2}{dt^2}w + \left(\frac{m_1^2}{k_1} + \frac{k_2 m_1^2}{k_1^2}\right)\frac{d^4}{dt^4}w. \end{aligned} \tag{5.4}$$

Using the fact that w satisfies the fourth-order differential equation

$$m_1 m_2 \frac{d^4}{dt^4}w + (k_1 m_1 + k_2 m_1 + k_1 m_2)\frac{d^2}{dt^2}w + k_1 k_2 w = 0,$$

the expression for f obtained from the second equation in (5.4) can be rewritten in terms of the generalized position $q = \mathrm{col}(w, \frac{d^2}{dt^2}w)$ as

$$f = \left(k_2 - \frac{k_2 m_1}{m_2} - \frac{k_2^2 m_1}{k_1 m_2}\right)w + \left(-m_1 + \frac{k_2 m_1}{k_1} - \frac{m_1^2}{m_2} - 2\frac{k_2 m_1^2}{k_1 m_2} - \frac{k_2^2 m_1^2}{k_1^2 m_2}\right)\frac{d^2}{dt^2}w.$$

Such an expression can also be given in terms of the positions of the two masses described in Example 3.3 as

$$f = \left(k_2 + \frac{k_1 m_1}{m_2} + \frac{k_2 m_1}{m_2}\right)w_1 + \left(k_2 - k_1 - \frac{k_1 m_1}{m_2} - \frac{2k_2 m_1}{m_2} - \frac{k_2^2 m_1}{k_1 m_2}\right)w_2.$$

The physical interpretation of such a quantity is not easy, though it should be remarked that the *physical dimensions* of such a latent variable are indeed those of a force.

When modeling physical phenomena, closed (i.e., autonomous) systems are the exception rather than the rule: the environment in which the system is embedded almost always interacts with it, exerting some influence. Sometimes it is reasonable to assume that the way in which the environment interacts with the system—in other

words, the generating mechanism of the external influences—depends only on the attributes of the system itself, the paramount example of such a situation being the motion of a point mass or charge in a force field depending on its position. By modeling the external influence as a function of the system "configuration" we obtain an autonomous system, whose evolution depends only on its laws of motion and the "initial state."

If we take such a point of view when considering the description of $\mathfrak{B}$ as the interconnection of $\mathfrak{B}_1$ and $\mathfrak{B}_2$ defined in (5.2) and (5.3), it is natural to consider $\mathfrak{B}_1$ as a model of an open system with external influences modeled by f and $\mathfrak{B}_2$ as a description of the way in which f depends on the variables w. In principle different constraints could be imposed by $\mathfrak{B}_2$ on f, and therefore it is natural to consider the open system $\mathfrak{B}_1$ as the starting point for a study of controllable Hamiltonian systems and to investigate the consequences of the Hamiltonianity of $\mathfrak{B}$ on $\mathfrak{B}_1$. This leads us to the definition of *controllable Hamiltonian behavior*, which we presently give.

Observe first that $\mathfrak{B}_1$ defined by (5.2) has as many outputs (external variables w) as inputs (the auxiliary variables f), in accordance with the point of view adopted in classical mechanics of considering the configuration variables as manifest ones and the external forces as inputs, each acting on a configuration variable (as in the case of collocated sensors and actuators; see section 12.1 of [12]). Now consider two compact-support trajectories $(w_i, f_i) \in \mathfrak{B}_1$, $i = 1, 2$, and compute the integral $\int_{-\infty}^{+\infty} w_1 f_2 - w_2 f_1 \, \mathrm{d}t$. Integrating by parts using (5.2) and the fact that the trajectories (w_i, f_i) are compact support, it is not difficult to verify that such an integral is zero. Such an observation brings us to the notion of a *controllable Hamiltonian system.*

DEFINITION 5.2. *Let $\mathfrak{B} \in \mathcal{L}^{\mathtt{w}}$ be controllable, with $\mathtt{w}$ even. Denote*

$$J_{\mathtt{w}} := \begin{pmatrix} 0 & I_{\frac{\mathtt{w}}{2}} \\ -I_{\frac{\mathtt{w}}{2}} & 0 \end{pmatrix}. \tag{5.5}$$

$\mathfrak{B}$ is called Hamiltonian if for all trajectories w_1, $w_2 \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}, \mathbb{R}^{\mathtt{w}})$ we have

$$\int_{-\infty}^{+\infty} L_{J_{\mathtt{w}}}(w_1, w_2) dt = 0.$$

We discuss the relationship of Definition 5.2 with other notions of Hamiltonianity in Remarks 5.5 and 5.6 below. We proceed by illustrating Definition 5.2 with an example and then give a number of characterizations of Hamiltonianity for controllable systems in Theorem 5.4, the main result of this section.

Example 5.3. Take the same system considered in Example 4.5, but with an external force applied to the first mass. Choose as external variables the position q of the first mass and the external force f; then it is easy to see that the behavior of the system is represented by the equation

$$m_1 m_2 \frac{d^4 q}{dt^4} + (m_1 k_1 + m_1 k_2 + m_2 k_1) \frac{d^2 q}{dt^2} + k_1 k_2 q = m_2 \frac{d^2 f}{dt^2} + (k_1 + k_2) f.$$

In order for this system to be controllable, the polynomials $d(\xi) := m_1 m_2 \xi^4 + (m_1 k_1 + m_1 k_2 + m_2 k_1)\xi^2 + k_1 k_2$ and $n(\xi) := m_2 \xi^2 + k_1 + k_2$ must be coprime. In that case the system also admits an observable image representation induced by the polynomial matrix $M(\xi) := \mathrm{col}(n(\xi), d(\xi))$. We now show that this system is Hamiltonian. Observe that for any pair of compact-support trajectories $w_i = M(\frac{d}{dt})\ell_i$, $i = 1, 2$, it holds that $L J_2(w_1, w_2) = L_\Phi(\ell_1, \ell_2)$, where $\Phi(\zeta, \eta) := M(\zeta)^T J_2 M(\eta)$. In order

to prove that $\int_{-\infty}^{+\infty} LJ_2(w_1, w_2)dt = 0$, observe that $\Phi(-\xi, \xi) = 0$. Conclude from Theorem 3.1 of [24] that there exists $\Psi \in \mathbb{R}[\zeta, \eta]$ such that $\Phi(\zeta, \eta) = (\zeta + \eta)\Psi(\zeta, \eta)$, equivalently, $\frac{d}{dt}L_\Psi = L_\Phi$. Now using the fact that the latent variable trajectories ℓ_i also have compact support, we can infer that $\int_{-\infty}^{+\infty} L_\Phi(\ell_1, \ell_2)dt = L_\Psi(\ell_1, \ell_2)\big|_{-\infty}^{+\infty} = 0$.

In order to state the main result of this section, consisting of several alternative characterizations of Hamiltonianity for controllable behaviors, we need to introduce the notion of an orthogonal of a controllable behavior. Given a controllable linear differential behavior $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$, we define its *orthogonal complement* $\mathfrak{B}^\perp$ as

$$\mathfrak{B}^\perp := \left\{ w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}}) \,\middle|\, \int_{-\infty}^{+\infty} w^T w' \, \mathrm{d}t = 0 \text{ for all } w' \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}, \mathbb{R}^{\mathtt{w}}) \right\}.$$

The orthogonal $\mathfrak{B}^\perp$ is again an element of $\mathfrak{L}^{\mathtt{w}}$, and it is controllable (see section 10 of [24]).

THEOREM 5.4. *Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ be controllable, with $\mathtt{w}$ even. Let $J_{\mathtt{w}}$ be given by* (5.5)*. Then the following statements are equivalent:*

(1) $\mathfrak{B}$ *is Hamiltonian.*
(2) $\mathfrak{B} = (J_{\mathtt{w}}\mathfrak{B})^\perp$.
(3) $M^\sim J_{\mathtt{w}} M = 0$ *for each* M *such that* $w = M(\frac{d}{dt})\ell$ *is an image representation of* $\mathfrak{B}$.
(4) $RJ_{\mathtt{w}}R^\sim = 0$ *for each* R *such that* $R(\frac{d}{dt})w = 0$ *is a kernel representation of* $\mathfrak{B}$.
(5) *For every i/o partition* $\mathrm{col}(u, y) = \Pi w$ *of* $\mathfrak{B}$ *the transfer function* G *from* u *to* y *satisfies* $G^\sim\Sigma = \Sigma G$, *with* Σ *the* $\frac{\mathtt{w}}{2} \times \frac{\mathtt{w}}{2}$ *signature matrix determined by*

$$\Pi J_{\mathtt{w}} \Pi^T = \begin{pmatrix} 0 & \Sigma \\ -\Sigma & 0 \end{pmatrix}. \tag{5.6}$$

(6) $\mathtt{n} := \mathtt{n}(\mathfrak{B})$ *is even, and there exists a minimal i/s/o representation*

$$\frac{d}{dt}x = Ax + Bu, \quad y = Cx + Du, \quad \mathrm{col}(u, y) = \Pi w$$

of $\mathfrak{B}$, *such that* $J_{\mathtt{n}}A + A^T J_{\mathtt{n}} = 0$, $\Sigma D = D^T\Sigma$, *and* $B^T J_{\mathtt{n}} = -\Sigma C$, *with* Σ *the* $\frac{\mathtt{w}}{2} \times \frac{\mathtt{w}}{2}$ *signature matrix determined by* (5.6).
(7) $\mathtt{n} := \mathtt{n}(\mathfrak{B})$ *is even, and for every minimal i/s/o representation*

$$\frac{d}{dt}x = Ax + Bu, \quad y = Cx + Du, \quad \mathrm{col}(u, y) = \Pi w$$

of $\mathfrak{B}$, *there exists a nonsingular skew-symmetric matrix* $K \in \mathbb{R}^{\mathtt{n}\times\mathtt{n}}$ *such that* $KA + A^T K = 0$, $\Sigma D = D^T\Sigma$, *and* $B^T K = -\Sigma C$, *with* Σ *the* $\frac{\mathtt{w}}{2} \times \frac{\mathtt{w}}{2}$ *signature matrix determined by* (5.6).

Proof. ((1) $\Leftrightarrow$ (3)) Let $w = M(\frac{d}{dt})\ell$ be an image representation of $\mathfrak{B}$. Observe that $\int_{-\infty}^{\infty} L_{J_{\mathtt{w}}}(w_1, w_2)dt = 0$ for all $w_1, w_2 \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}, \mathbb{R}^{\mathtt{w}})$ if and only if $\int_{-\infty}^{\infty} L_\Psi(\ell_1, \ell_2)dt = 0$ for all ℓ_1, ℓ_2 of compact support, where $\Psi(\zeta, \eta) := M(\zeta)^T J_{\mathtt{w}} M(\eta)$. By Theorem 3.1 of [24] this holds if and only if $M^\sim J_{\mathtt{w}} M = 0$.

((3) $\Leftrightarrow$ (5)) Let Π be a $\mathtt{w} \times \mathtt{w}$ permutation matrix such that for $\mathrm{col}(u, y) \in \Pi\mathfrak{B}$, u is input and y is output. Let $w = M(\frac{d}{dt})\ell$ be any image representation of $\mathfrak{B}$ with M full column rank. Then correspondingly $\Pi M = \mathrm{col}(U, Y)$, with $\det(U) \neq 0$. The transfer matrix from u to y is equal to the matrix of rational functions $G = YU^{-1}$. We have

$$M^\sim J_{\mathtt{w}} M = M^\sim \Pi^T \Pi J_{\mathtt{w}} \Pi^T \Pi M = M^\sim \Pi^T \begin{pmatrix} 0 & \Sigma \\ -\Sigma & 0 \end{pmatrix} \Pi M = U^\sim \Sigma Y - Y^\sim \Sigma U,$$

with Σ a nonsingular $\frac{\mathtt{w}}{2} \times \frac{\mathtt{w}}{2}$ signature matrix. Since $M^\sim J_{\mathtt{w}} M = 0$ we obtain $U^\sim \Sigma Y - Y^\sim \Sigma U = 0$, equivalently, $G^\sim \Sigma = \Sigma G$. Conversely, if $G^\sim \Sigma = \Sigma G$ then take a coprime factorization $G = YU^{-1}$. Then $M := \Pi^T \mathrm{col}(U, Y)$ yields an observable image representation $w = M(\frac{d}{dt})\ell$ of $\mathfrak{B}$ which clearly satisfies $M^\sim J_{\mathtt{w}} M = 0$. It then follows easily that $M^\sim J_{\mathtt{w}} M = 0$ for any M such that $w = M(\frac{d}{dt})\ell$ is an image representation of $\mathfrak{B}$.

((3) $\Rightarrow$ (2)) Observe that $M^T(-\frac{d}{dt})w' = 0$ is a kernel representation of $\mathfrak{B}^\perp$ and $M^T(-\frac{d}{dt})J_{\mathtt{w}} w'' = 0$ is a kernel representation of $(J_{\mathtt{w}}\mathfrak{B})^\perp$. From $M^\sim J_{\mathtt{w}} M = 0$ it thus follows that $\mathfrak{B} \subseteq (J_{\mathtt{w}}\mathfrak{B})^\perp$. The equivalence ((3) $\Leftrightarrow$ (5)) shows that every transfer matrix of $\mathfrak{B}$ is square, and consequently $\mathtt{m}(\mathfrak{B}) = \mathtt{p}(\mathfrak{B})$. Hence we have $\mathtt{m}((\mathtt{J_w}\mathfrak{B})^\perp) = \mathtt{p}(\mathtt{J_w}\mathfrak{B}) = \mathtt{p}(\mathfrak{B}) = \mathtt{m}(\mathfrak{B})$. Using the calculus of behavioral equations (see [15]) it is not difficult to prove that two controllable behaviors $\mathfrak{B}_1$ and $\mathfrak{B}_2$, with the same number of inputs such that $\mathfrak{B}_1 \subseteq \mathfrak{B}_2$, must be equal. This implies that, in fact, the equality $\mathfrak{B} = (J_{\mathtt{w}}\mathfrak{B})^\perp$ holds.

((2) $\Rightarrow$ (1)) $\mathfrak{B} \subseteq (J_{\mathtt{w}}\mathfrak{B})^\perp$ by definition implies that $\int_{-\infty}^{\infty} L_{J_{\mathtt{w}}}(w_1, w_2)dt = 0$ for all $w_1, w_2 \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}, \mathbb{R}^{\mathtt{w}})$.

((2) $\Leftrightarrow$ (4)) Let $R(\frac{d}{dt})w = 0$ be a kernel representation of $\mathfrak{B}$. Then $w' = R^T(-\frac{d}{dt})\ell$ is an image representation of $\mathfrak{B}^\perp$ and $w'' = J_{\mathtt{w}} R^T(-\frac{d}{dt})\ell$ is an image representation of $(J_{\mathtt{w}}\mathfrak{B})^\perp$ (see section 10 of [24]). Clearly $(J_{\mathtt{w}}\mathfrak{B})^\perp \subseteq \mathfrak{B}$ implies $RJ_{\mathtt{w}}R^\sim = 0$. Conversely, if $RJ_{\mathtt{w}}R^\sim = 0$ then $(J_{\mathtt{w}}\mathfrak{B})^\perp \subseteq \mathfrak{B}$. Also, it is easily seen that $RJ_{\mathtt{w}}R^\sim = 0$ implies condition (5), so that $\mathtt{p}(\mathfrak{B}) = \mathtt{m}(\mathfrak{B})$. By the same argument used in the proof of the implication (3) $\Rightarrow$ (2) this yields that $\mathfrak{B} = (J_{\mathtt{w}}\mathfrak{B})^\perp$.

((5) $\Rightarrow$ (7)) Let (A, B, C, D) be the quadruple of matrices associated with a minimal i/s/o representation of $\mathfrak{B}$. This yields an i/o partition with transfer matrix $G(\xi) = D + C(\xi I - A)^{-1}B$. By minimality of the i/s/o representation it follows that the pair (C, A) is observable; moreover, by controllability of $\mathfrak{B}$, it follows that the pair (A, B) is controllable. Recall that there exists a nonsingular signature matrix Σ such that $\Sigma G = G^\sim \Sigma$. This implies that $(A, B, \Sigma C, \Sigma D)$ and $(-A^T, C^T\Sigma, -B^T, D^T\Sigma)$ are minimal realizations of the same transfer matrix. Consequently there exists a unique nonsingular matrix K such that $-A^T = KAK^{-1}$, $C^T\Sigma = KB$, $-B^T = \Sigma C K^{-1}$, $\Sigma D = D^T\Sigma$. It is easily verified that also $-A^T = (-K^T)A(-K^{-T})$, $C^T\Sigma = (-K^T)B$, $-B^T = \Sigma C(-K^{-T})$. Due to the uniqueness of K, it follows that $K^T = -K$, i.e., K is skew-symmetric.

((7) $\Rightarrow$ (6)) Let (A, B, C, D) be the quadruple of matrices associated with a minimal i/s/o representation of $\mathfrak{B}$. Let K be nonsingular and skew-symmetric and let Σ be a nonsingular signature matrix such that $KA + A^TK = 0$, $\Sigma D = D^T\Sigma$, and $B^TK = -\Sigma C$. There exists a nonsingular matrix S such that $S^T J_{\mathtt{n}} S = K$. Define $\hat{A} := SAS^{-1}$, $\hat{B} = SB$, $\hat{C} = CS^{-1}$, and $\hat{D} = D$. This quadruple also defines a minimal i/s/o representation of $\mathfrak{B}$. Moreover, it is easily verified that $J_{\mathtt{n}}\hat{A} + \hat{A}^T J_{\mathtt{n}} = 0$, $\Sigma\hat{D} = \hat{D}^T\Sigma$, and $\hat{B}^T J_{\mathtt{n}} = -\Sigma\hat{C}$.

((6) $\Rightarrow$ (3)) If an i/s/o representation of $\mathfrak{B}$ exists satisfying the conditions in (6), then it follows that the transfer matrix from u to y satisfies $G^\sim\Sigma = \Sigma G$. Take a coprime factorization $G = YU^{-1}$. Then $M := \Pi^T \mathrm{col}(U, Y)$ yields an observable image representation $w = M(\frac{d}{dt})\ell$ of $\mathfrak{B}$ which clearly satisfies $M^\sim J_{\mathtt{w}} M = 0$. It then follows easily that $M^\sim J_{\mathtt{w}} M = 0$ for any M such that $w = M(\frac{d}{dt})\ell$ is an image representation of $\mathfrak{B}$. □

Remark 5.5. The definition of a nonlinear Hamiltonian i/o system put forward in [7, 6] is based on the self-adjointness of the i/o map of the system. We now discuss how such a point of view relates with that given in Definition 5.2 and elaborated on

in Theorem 5.4. Assume that the variables w of the controllable system described in kernel form by $R(\frac{d}{dt})w = 0$, R of full row rank, are partitioned as $w = \mathrm{col}(u, y)$, with u consisting of $\frac{\mathtt{w}}{2}$ input variables, and y consisting of $\frac{\mathtt{w}}{2}$ output variables. Then the matrix R can be partitioned as $R = \begin{pmatrix} P & -Q \end{pmatrix}$, with P square and invertible.

Under these assumptions, the condition $RJ_{\mathtt{w}}R^{\sim} = 0$ appearing in statement (4) of Theorem 5.4 reads $QP^{\sim} - PQ^{\sim} = 0$, which is equivalent with formula (51) of [6], once it is recalled that an image representation of the adjoint system (equivalently, of the orthogonal behavior) to $\ker R(\frac{d}{dt})$ is given by $\mathrm{Im}\ R^T(-\frac{d}{dt})$.

Remark 5.6. The characterization of Hamiltonian transfer functions given in statement (5) of Theorem 5.4 is the same given in [3] in the context of i/s/o systems and in [8] in the polynomial model context. See also [20], where an external characterization of the adjoint of a linear system is given, and several of the techniques used (see, in particular, section III loc. cit.) foreshadow those based on the calculus of Q/BDFs used in the present paper.

Remark 5.7. It follows from Theorem 5.4 (for example, by applying condition (4)) that the systems $\mathfrak{B}_1$ and $\mathfrak{B}_2$ represented by, respectively, (5.2) and (5.3) are controllable Hamiltonian systems. In other words, if $\mathfrak{B}$ is an autonomous Hamiltonian system with $\chi_{\mathfrak{B}}(0) \neq 0$, then there exist two controllable Hamiltonian behaviors $\mathfrak{B}_i$, $i = 1, 2$, such that $\mathfrak{B}$ is the feedback interconnection of $\mathfrak{B}_1$ and $\mathfrak{B}_2$.

We conclude this section with two examples of controllable Hamiltonian systems.

Example 5.8. Newton's second law defines a controllable Hamiltonian system $\mathfrak{B} = \{(F, q) \mid F = m\frac{d^2}{dt^2}q\}$, as it is easy to verify using, for example, statement (4) of Theorem 5.4.

Example 5.9. Consider a parallel interconnection of a capacitor C with an inductance L subject to an external current I_e. Assume that we choose as external variables for such a system the external current and the magnetic flux ϕ_L in the inductance; it is easy to verify that in such a case the system equation is $(\frac{d^2}{dt^2} + \frac{1}{CL})\phi_L - \frac{1}{C}I_e = 0$. We show that this system is Hamiltonian. An observable image representation of the system is induced by the matrix $M(\xi) = \mathrm{col}(1, C\xi^2 + \frac{1}{L})$. Consider that $M(\zeta)^T J_2 M(\eta) = (\zeta+\eta)\Psi(\zeta, \eta)$ with $\Psi(\zeta, \eta) := C(\zeta-\eta)$. Consequently such a BDF satisfies $\frac{d}{dt}L_\Psi = L_{J_2}$ on $\mathfrak{B}$, and therefore $\int_{-\infty}^{+\infty} L_{J_2}(w_1, w_2) = 0$ for all w_1, $w_2 \in \mathfrak{B}$.

6. Conclusions. In this paper we have used the formalism of bilinear and quadratic differential forms in order to study Hamiltonian systems. The approach followed in this paper is representation-free, i.e., independent of the existence of a special representation of the system, such as a transfer function or a state-space representation. However, we have also given a characterization of Hamiltonianity for various system representations such as kernel, state-space, and transfer function, in the case of autonomous systems (Theorem 3.4) and of controllable ones (Theorem 5.4). We have also proposed a definition of generalized total energy and generalized Lagrangian (see section 4), and we introduced the notion of generalized internal forces for systems described by higher-order differential equations (see section 5).

The major limitation of the present work is its treatment of Hamiltonianity for the controllable and autonomous case only, leaving out the general case of a system comprising a nonzero controllable part and a nonzero "autonomous part" (for the difficulty in defining uniquely such a subbehavior, see [15, p. 192]). The development of a general theory of linear Hamiltonian behaviors that includes completely controllable and completely autonomous behaviors as special cases is a pressing issue in our research.

In view of the encouraging results of the application of quadratic differential forms in the context of infinite-dimensional systems (see [13, 14]), it can be hoped that some of the results presented in this paper can be generalized also to systems described by linear constant-coefficient partial differential equations; such an area of research is presently under investigation. Another direction in which the research presented in this paper is being extended is that of the connections between Hamiltonian systems and optimal control.

Acknowledgments. The authors thank Prof. Dr. Ir. Jan C. Willems and Prof. Dr. A. J. van der Schaft, on whom several preliminary versions of this paper were inflicted; they stimulated several discussions conducive to the elaboration of very valuable ideas and pointed out important references in the literature. Many thanks also to Dr. H. K. Pillai for his help in understanding the more technical points of [5].

REFERENCES

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, Berlin, 1978.
[2] R. W. Brockett, *Path integrals, Lyapunov functions, and quadratic minimization*, in Proceedings of the 4th Allerton Conference on Circuit and System Theory, University of Illinois, Monticello, IL, 1966, pp. 685–698.
[3] R. W. Brockett and A. Rahimi, *Lie algebras and linear differential equations*, in Ordinary Differential Equations, L. Weiss, ed., Academic Press, New York, 1972, pp. 379–386.
[4] A. Ciampi, *Classification of Hamiltonian linear systems*, Indiana Univ. Math. J., 23 (1973), pp. 513–526.
[5] T. Cotroneo and H. K. Pillai, *Linear variational behaviors*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999.
[6] P. E. Crouch, F. Lamnabhi-Lagarrigue, and A. J. v.d. Schaft, *Adjoint and Hamiltonian input-output differential equations*, IEEE Trans. Automat. Control, 40 (1995), pp. 603–615.
[7] P. E. Crouch and A. J. van der Schaft, *Variational and Hamiltonian Control Systems*, Lecture Notes in Control and Inform. Sci. 101, Springer-Verlag, New York, 1987.
[8] P. A. Fuhrmann, *On Hamiltonian rational transfer functions*, Linear Algebra Appl., 63 (1984), pp. 1–93.
[9] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.
[10] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[11] R. E. Kalman, *Algebraic characterization of polynomials whose zeros lie in certain algebraic domains*, Proc. Natl. Acad. Sci. USA, 64 (1969), pp. 818–823.
[12] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[13] H. K. Pillai and S. Shankar, *A behavioral approach to control of distributed systems*, SIAM J. Control Optim., 37 (1998), pp. 388–408.
[14] H. K. Pillai and J. C. Willems, *Lossless and dissipative distributed systems*, SIAM J. Control Optim., 40 (2002), pp. 1406–1430.
[15] J. W. Polderman and J. C. Willems, *Introduction to Mathematical System Theory: A Behavioral Approach*, Springer-Verlag, Berlin, 1997.
[16] P. Rapisarda and J. C. Willems, *State maps for linear systems*, SIAM J. Control Optim., 35 (1997), pp. 1053–1091.
[17] R. M. Santilli, *Foundations of Theoretical Mechanics* I, Springer-Verlag, New York, 1978.
[18] A. J. van der Schaft, *Hamiltonian dynamics with external forces and observations*, Math. Systems Theory, 15 (1982), pp. 145–168.
[19] A. J. van der Schaft, *System Theoretic Description of Physical Systems*, CWI Tract 3, CWI, Amsterdam, 1984.
[20] A. J. van der Schaft, *Duality for linear systems: External and state space characterization of the adjoint system*, in Analysis of Controlled Dynamical Systems, B. Bonnard, B. Bride, J.- P. Gauthier, and I. Kupka, eds., Birkhäuser, Boston, 1991, pp. 393–403.
[21] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies, with an Introduction to the Problem of Three Bodies*, Cambridge University Press, Cambridge, UK, 1988.

[22] J. C. Willems and J. C. Fuhrmann, *Stability theory for high-order systems*, Linear Algebra Appl., 167 (1992), pp. 131–149.
[23] J. C. Willems, *On interconnections, control, and feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 326–339.
[24] J. C. Willems and H. L. Trentelman, *On quadratic differential forms*, SIAM J. Control Optim., 36 (1998), pp. 1703–1749.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 792–812

COMPUTING THE EFFECTIVE HAMILTONIAN USING A VARIATIONAL APPROACH*

DIOGO A. GOMES† AND ADAM M. OBERMAN‡

Abstract. A numerical method for homogenization of Hamilton–Jacobi equations is presented and implemented as an L^∞ calculus of variations problem. Solutions are found by solving a nonlinear convex optimization problem. The numerical method is shown to be convergent, and error estimates are provided. One and two dimensional examples are worked in detail, comparing known results with the numerical ones and computing new examples. The cases of nonstrictly convex Hamiltonians and Hamiltonians for which the cell problem has no solution are treated.

Key words. Hamilton–Jacobi, homogenization, numerics, calculus of variations

AMS subject classifications. 37, 49, 65, 35

DOI. 10.1137/S0363012902417620

1. Introduction.

Given the Hamiltonian $H(p, x)$, which is smooth, convex in p, and periodic in the second variable x, we are interested in finding for a given $P \in \mathbb{R}^n$ a periodic solution of the Hamilton–Jacobi equation

$$H(P + D_x u, x) = \bar{H}(P). \tag{HB}$$

For each fixed P the problem (HB) can be regarded as a nonlinear eigenvalue problem for the function $u(x)$ and the number $\bar{H}(P)$. We regard $\bar{H}(P)$, the *effective Hamiltonian*, as a function of the parameter P. It encodes information about $H(x, p)$, as we shall describe below.

Problem (HB) requires that we determine for a given $P \in \mathbb{R}^n$ the pair $(u, \bar{H}(P))$. Classical solutions do not exist for all P, so viscosity solutions [CIL92, BCD97, FS93] are used.

Solving (HB) directly involves finding the viscosity solution to a degenerate elliptic partial differential equation coupled to an unknown constant, $\bar{H}(P)$. While this may be done, it is not an easy task. We choose instead to reduce the problem of finding the (approximate) effective Hamiltonian to a finite dimensional convex optimization problem, which may be solved numerically using optimization routines.

Numerical computations of effective Hamiltonians have been done by [EMS95, KBM01], with applications to front propagation and combustion. At the time of preparation of this manuscript, the authors also discovered work by [Qia01]. The numerical approach taken by these authors was to find the effective Hamiltonian by partial differential equations methods.

In this work we circumvent the difficulties of solving (HB) by computing $\bar{H}(P)$ without finding the solution u. Our methods are based on the representation formula

$$\bar{H}(P) = \inf_{\phi \in C^1_{per}} \sup_x H(P + D_x \phi, x) \tag{1}$$

*Received by the editors November 11, 2002; accepted for publication (in revised form) January 5, 2004; published electronically September 18, 2004.
http://www.siam.org/journals/sicon/43-3/41762.html

†Instituto Superior Téchnico, Av. Rovisco Parents, 1049-001 Lisbon, Portugal (dgomes@math.ist.utl.pt). The research of this author was partially supported by FCT/POCTI/FEDER.

‡Department of Mathematics, University of Texas, Austin, TX 78712 (oberman@math.utexas.edu). Current address: Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

due, for strictly convex Hamiltonians, to [CIPP98], in which the infimum is taken over all periodic C^1 functions, $C^1(\mathbb{T}^n)$. This formula is a problem in the calculus of variations problem in L^∞. Such problems were studied by Aronsson in the 1960s [Aro66, Aro65]. Recently there has been a renewed interest in calculus of variations in L^∞ [BJW01a, BJW01b, Bar94]. Related methods in the calculus of variations in L^∞ were applied to the effective Hamiltonian problem in [Eva03].

In this paper we always assume that H is convex but not necessarily strictly convex. This assumption has implications for the existence and smoothness of solutions of (HB). For instance, if H is strictly convex, then there are viscosity solutions of (HB) which are Lipschitz continuous. However, if strict convexity fails, solutions may (see section 5.2) or may not (see section 5.3) exist, and the degree of smoothness will depend on the Hamiltonian in question.

1.1. Applications. Computing the effective Hamiltonian is relevant to several classes of applications: homogenization problems, the long time behavior of Hamilton–Jacobi equations, classical mechanics, Aubry–Mather theory, ergodic control, and front propagation.

In *homogenization problems* [LPV88, Con95], if w^ϵ solves

$$-w_t^\epsilon + H\left(D_x w^\epsilon, \frac{x}{\epsilon}\right) = 0,$$

then, as ϵ goes to 0, the solution w^ϵ converges to w^0, which is a solution of the limiting problem

$$-w_t^0 + \bar{H}(D_x w^0) = 0.$$

The effective Hamiltonian also appears in the study of *long time limits of viscosity solutions of Hamilton–Jacobi equations*:

$$-w_t + H(P + D_x w, x) = 0.$$

It turns out that $w(x,t) - \bar{H}(P)\,t$ converges as $t \to -\infty$ to a stationary solution of (HB) [Fat98b, BS00]; see also [AI01, CDI01].

In *classical mechanics* [AKN97], smooth solutions u of (HB) yield a canonical change of coordinates $X(p,x)$ and $P(p,x)$ defined by the equations

$$p = P + D_x u, \qquad X = x + D_P u. \tag{2}$$

This would simplify the Hamiltonian dynamics

$$\dot{x} = -D_p H(p,x), \qquad \dot{p} = D_x H(p,x) \tag{3}$$

into the trivial dynamics

$$\dot{P} = 0, \qquad \dot{X} = -D_P \bar{H}(P).$$

In other words, for each P there is an invariant torus in which the dynamics is simply a rotation. However, (HB) does not admit smooth solutions in general (see section 4), and one must deal with viscosity solutions [CIL92, BCD97, FS93, Eva98].

In *Aubry–Mather theory* [Mat89a, Mat89b, Mat91, Mn92, Mn96], instead of looking for invariant tori, one looks for probability measures μ on $\mathbb{T}^n \times \mathbb{R}^n$ that minimize the average action

$$\int L(x,v) + P \cdot v \, d\mu \tag{4}$$

and satisfy a holonomy condition

$$\int v D_x \phi \, d\mu = 0$$

for all $\phi(x) \in C^1(\mathbb{T}^n)$.

Here $L(x, v)$ is the Legendre transform of $H(p, x)$, defined by

$$L(x, v) = \sup_p -v \cdot p - H(p, x).$$

The supports of these measures are called the Aubry–Mather sets and are the natural generalizations of invariant tori. Recent results [E99, Fat97a, Fat97b, Fat98a, Fat98b, CIPP98], some by one of the authors [EG01, EG02, Gom01b], show that viscosity solutions encode the Aubry–Mather sets. In particular, we have

$$\int L(x, v) + P \cdot v d\mu = -\bar{H}(P),$$

and the support of the Mather measure is a subset of the graph

$$(x, -D_p H(P + D_x u, x))$$

for any viscosity solution of (HB).

In the Mather set the asymptotics of the Hamiltonian dynamics are controlled by viscosity solutions. Indeed let (x, p) be any point in $\mathbb{T}^n \times \mathbb{R}^n$. Consider its flow by the Hamilton equations (3). If (x, p) belongs to any Mather set, then

$$\frac{x(T)}{T} \to Q$$

as $T \to \infty$ for some vector $Q \in \mathbb{R}^n$, with $Q = D_P \bar{H}(P)$ for some P if $\bar{H}$ is differentiable.

Equation (HB) and related stationary first and second order Hamilton–Jacobi equations are also important to the *ergodic control problem* [Ari98, Ari97]. While this article was being reviewed, the authors became aware of [FSar]. Aubry–Mather theory can also be generalized to second order equations [Gom02], and many of the techniques that we develop can be generalized appropriately.

The computation of effective Hamiltonians has applications to the *propagation of flame fronts in combustion*: Hamilton–Jacobi equations which are homogeneous of order one, for example $u_t = c|Du|$, can represent the evolution of a propagating front moving in the normal direction with speed c. If the front is propagating in a periodic media, an equation of the form $u_t = c(x)|Du|$ holds, where $c(x)$ is positive and periodic in x. In this case, solving a homogenization problem gives the effective or averaged front speed. As mentioned earlier, numerical computations for this problem have been performed by [EMS95, KBM01].

2. Solvability and approximation of the homogenization problem.

2.1. Solvability of the homogenization problem. We start this section by reviewing some results concerning the function $\bar{H}(P)$. In particular, we recall the uniqueness result of $\bar{H}$ from [LPV88], and we generalize a representation formula for $\bar{H}$ due to [CIPP98].

PROPOSITION 2.1 (from Lions, Papanicolao, Varadhan [LPV88]). *There is at most one value $\bar{H}$ for which* (HB) *has a periodic viscosity solution.*

Proof. Suppose, for contradiction, that (HB) admits viscosity solutions u_1 and u_2 for $\bar{H} = \bar{H}_1, \bar{H}_2$, respectively, with $\bar{H}_1 > \bar{H}_2$. We may assume $v_1 \equiv u_1 + C > u_2$ for a sufficiently large positive constant C. For ϵ sufficiently small

$$\epsilon v_1 + H(D_x v_1, x) \geq \epsilon u_2 + H(D_x u_2, x)$$

in the viscosity sense. The comparison principle [BCD97] then implies $v_1 \leq u_2$, which is a contradiction. □

Next we prove a representation formula for $\bar{H}$. Our result extends [CIPP98] to nonstrictly convex Hamiltonians H, for which the solution may fail to be Lipschitz.

PROPOSITION 2.2 (from Contreras, Iturriaga, Paternain, Paternain [CIPP98]). *Suppose that H is periodic in x and convex in p (strict convexity is not required). Suppose further that there exists a viscosity solution u of* (HB). *Then*

$$\bar{H} = \inf_{\psi \in C^1(\mathbb{T}^n)} \sup_{x \in \mathbb{T}^n} H(D_x \psi, x), \tag{5}$$

in which the infimum is taken over the space $C^1(\mathbb{T}^n)$ of periodic functions.

First we recall some facts concerning the sup convolution. The proof may be found in [FS93].

LEMMA 2.3. *Suppose u is a viscosity of* (HB). *Define*

$$u_\epsilon(x) = \sup_y \left[u(y) - \frac{|x-y|^2}{\epsilon} \right]. \tag{6}$$

Then

1. *$u_\epsilon \to u$ uniformly as $\epsilon \to 0$,*
2. *u_ϵ is semiconvex,*
3. *u_ϵ satisfies*

$$H(D_x u_\epsilon, x) \leq \bar{H} + o(1),$$

in the viscosity sense and almost everywhere.

Proof of Proposition 2.2. Let

$$\bar{H}^* = \inf_{\psi \in C^1(\mathbb{T}^n)} \sup_{x \in \mathbb{T}^n} H(D_x \phi, x).$$

At some point x_0, $u - \psi$ has a local minimum. By the viscosity property

$$H(D_x \psi(x_0), x_0) \geq \bar{H},$$

which implies $\bar{H}^* \geq \bar{H}$.

Let η_ϵ be a smoothing kernel. Set $v_\epsilon = u_\epsilon * \eta_\epsilon$. Then, using convexity,

$$H(D_x v_\epsilon(x), x) \leq \int H(D_x u_\epsilon(y), y) \eta_\epsilon(x-y) dy + o(1) \leq \bar{H} + o(1),$$

and thus $\bar{H}^* \leq \bar{H}$. □

Before proceeding with the discretization of this problem we will prove an elementary bound.

PROPOSITION 2.4. *We have*

$$\inf_{\psi \in C^1(\mathbb{T}^n)} \sup_{x \in \mathbb{T}^n} H(D_x \psi, x) \geq \inf_{x \in \mathbb{T}^n} H(0, x).$$

Proof. For any function $\psi \in C^1(\mathbb{T}^n)$ there is a point x_0 for which $D_x \phi(x_0) = 0$. Therefore $\sup_{x \in \mathbb{T}^n} H(D_x \psi, x) \geq H(0, x_0) \geq \inf_{x \in \mathbb{T}^n} H(0, x)$. □

2.2. Approximation. The next issue we study is the approximation of the problem (1).

To this effect, consider a triangulation of $\mathbb{T}^n$ with cells of diameter smaller than h. Let $C(T_h)$ be the collection of continuous piecewise linear grid functions which interpolate given nodal values. We avoid the use of the term finite elements to emphasize the pointwise nature of the approximation.

The following proposition is an approximation result: we do not assume the existence of a viscosity solution of (HB).

PROPOSITION 2.5. *Suppose $H(p,x)$ is convex in p. Then*

$$\inf_{\psi\in C^1(\mathbb{T}^n)} \sup_x H(D_x\psi, x) = \lim_{h\to 0} \inf_{\phi\in C(T_h)} \operatorname*{esssup}_x H(D_x\phi, x).$$

Proof. Fix $\epsilon > 0$. Let ψ be a C^1 function for which

$$\sup_{x\in\mathbb{T}^n} H(D_x\psi, x) \le \inf_{\psi\in C^1(\mathbb{T}^n)} \sup_{x\in\mathbb{T}^n} H(D_x\psi, x) + \epsilon.$$

Because ψ is C^1, $D_x\psi$ is uniformly continuous. Thus, for h sufficiently small, there is $\phi \in C(T_h)$ such that

$$\operatorname*{esssup}_{x\in\mathbb{T}^n} |D_x\phi - D_x\psi| \le \epsilon. \tag{7}$$

In fact, in each triangle along an edge e_i with length $|e_i|$ pointing in the direction ν_i we have

$$D_x\psi\cdot\nu_i = \frac{1}{|e_i|}\int_{e_i} D_x\phi\cdot dz = D_x\phi(\bar{x})\cdot\nu_i + o(1),$$

in which $\bar{x}$ is, for instance, the center of the triangle. Since the shape factor is bounded, there are at least n edges ν_i linearly independent such that

$$|\det[\nu]| \ge \theta \tag{8}$$

for some θ for all triangles. Therefore

$$D_x\psi = D_x\phi(\bar{x}) + o(1),$$

as required.

This therefore implies

$$\operatorname*{esssup}_{x\in\mathbb{T}^n} H(D_x\phi, x) \le \sup_{x\in\mathbb{T}^n} H(D_x\psi, x) + O(\epsilon),$$

by the Lipschitz continuity of H in p. Thus, taking first $\lim_{h\to 0}\inf_{\phi\in C(T_h)}$ and then $\inf_{\psi\in C^1(\mathbb{T}^n)}$, we obtain

$$\lim_{h\to 0}\inf_{\phi\in C(T_h)} \operatorname*{esssup}_{x\in\mathbb{T}^n} H(D_x\phi, x) \le \inf_{\psi\in C^1(\mathbb{T}^n)} \sup_{x\in\mathbb{T}^n} H(D_x\psi, x) + O(\epsilon).$$

Send $\epsilon\to 0$.

To prove the converse inequality observe that if $\phi\in C(T_h)$, η_ϵ is a smooth mollifier, and $\psi = \eta_\epsilon * \phi$, then convexity yields

$$H(D_x\psi(x), x) \le \int H(D_x\phi(y), y)\eta_\epsilon(x-y)dy + O(\epsilon),$$

for every x, and so

$$H(D_x\psi(x),x) \le \operatorname*{esssup}_{x\in\mathbb{T}^n} H(D_x\phi(x),x) + O(\epsilon).$$

Thus, taking first $\inf_{\psi\in C^1}$ and then $\lim_{h\to 0}\inf_{\phi\in C(T_h)}$,

$$\inf_{\psi\in C^1}\sup_x H(D_x\psi,x) \le \lim_{h\to 0}\inf_{\phi\in C(T_h)}\operatorname*{esssup}_{x\in\mathbb{T}^n} H(D_x\phi,x) + O(\epsilon).$$

Since ϵ is arbitrary, we have the claim. □

Before stating and proving an improved version of the previous proposition for the case in which (HB) has a viscosity solution, we record some important properties of the L^∞ calculus of variations problem. First observe that

$$\mathcal{H}(\phi) = \sup_{x\in\mathbb{T}^n} H(D_x\phi,x)$$

is a convex, but not strictly convex, functional. Therefore for any ϕ_1 and ϕ_2 we have

$$\mathcal{H}(\lambda\phi_1 + (1-\lambda)\phi_2) \le \lambda\mathcal{H}(\phi_1) + (1-\lambda)\mathcal{H}(\phi_2).$$

This in particular implies that any local minimum is a global minimum.

However, in general the minimizers are not unique, and it is not true that a minimizing sequence will converge to a viscosity solution of

$$H(D_xu,x) = \bar{H}.$$

For example, $H = p^2/2 + \cos x$ has $\bar{H} = 1$, and $u \equiv 0$ is a minimizer. However, $H(D_xu,x) \ne \bar{H}$.

A similar argument applied to the discretized problem yields that

$$\mathcal{H}_h(\phi) = \operatorname*{esssup}_{x\in\mathbb{T}^n} H(D_x\phi,x)$$

is convex for $\phi \in C(T_h)$, and so a local minimum is a global minimum.

PROPOSITION 2.6. *The approximate Hamiltonian*

$$\bar{H}_h(P) = \inf_{\phi\in C(T_h)}\operatorname*{esssup}_{x\in\mathbb{T}^n} H(P + D_x\phi,x)$$

is convex in P.

Proof. Let $P_1, P_2 \in \mathbb{R}^n$, and let $\phi_1, \phi_2 \in C(T_h)$ be the corresponding minimizers. Let $0 \le \lambda \le 1$, and set $P = \lambda P_1 + (1-\lambda)P_2$ and $\phi = \lambda\phi_1 + (1-\lambda)\phi_2$. Then, for any x we have

$$H(P + D_x\phi,x) \le \lambda H(P_1 + D_x\phi_1,x) + (1-\lambda)H(P_2 + D_x\phi_2,x).$$

Thus

$$\operatorname*{esssup}_{x\in\mathbb{T}^n} H(P + D_x\phi,x) \le \lambda\bar{H}_h(P_1) + (1-\lambda)\bar{H}_h(P_2),$$

and so

$$\bar{H}_h(P) = \inf_{\phi\in C(T_h)}\operatorname*{esssup}_{x\in\mathbb{T}^n} H(P + D_x\phi,x) \le \lambda\bar{H}_h(P_1) + (1-\lambda)\bar{H}_h(P_2)$$

if $P = \lambda P_1 + (1-\lambda)P_2$. □

THEOREM 2.7. *For any convex Hamiltonian $H(p,x)$ for which* (HB) *has a viscosity solution*

$$\bar{H} \leq \inf_{\phi \in C(T_h)} \operatorname*{esssup}_x H(D_x\phi, x).$$

If there exists a globally C^2 solution of (HB)*, then*

$$\inf_{\phi \in C(T_h)} \operatorname*{esssup}_x H(D_x\phi, x) = \bar{H} + O(h).$$

If (HB) *has a Lipschitz solution (for instance, if $H(p,x)$ is strictly convex in p), then*

$$\inf_{\phi \in C(T_h)} \operatorname*{esssup}_x H(D_x\phi, x) = \bar{H} + O(h^{1/2}).$$

If H is convex but not strictly convex and (HB) *has a viscosity solution, then*

$$\inf_{\phi \in C(T_h)} \operatorname*{esssup}_x H(D_x\phi, x) = \bar{H} + o(1).$$

Proof. Observe that

$$\bar{H} = \inf_{\psi \in C^1(\mathbb{T}^n)} \sup_x H(D_x\psi, x) \leq \inf_{\phi \in C(T_h)} \operatorname*{esssup}_x H(D_x\phi, x),$$

because by convexity we can associate to each $\phi \in C(T_h)$ a function

$$\psi = \phi * \eta_\epsilon \in C^1(\mathbb{T}^n)$$

such that

$$\sup_x H(D_x\psi, x) \leq \operatorname*{esssup}_x H(D_x\phi, x) + O(\epsilon),$$

for arbitrary $\epsilon > 0$, as seen in the previous proposition.

To prove the second assertion suppose that u is a C^2 viscosity solution of (HB). Fix h and construct a function $\phi_u \in C(T_h)$ by interpolating linearly the values of u at the nodal points. In each triangle T^i, the oscillation of the derivative of u is $O(h)$, since u is C^2. Thus, proceeding as in the proof of (7), we obtain

$$D_x\phi_u(x) = D_x u(x) + O(h)$$

for any x. Since $H(D_x u, x) = \bar{H}$, at every point $x \in T^i$ we have

$$H(D_x\phi_u, x) \leq \bar{H} + O(h).$$

This implies

$$\inf_{\phi \in C(T_h)} \operatorname*{esssup}_{x \in \mathbb{T}^n} H(D_x\phi, x) \leq \bar{H} + O(h).$$

If u is a Lipschitz viscosity solution, let $\tilde{u} = \eta_{h^{1/2}} * u$. Observe that

$$|D^2_{xx}\tilde{u}| \leq \frac{C}{h^{1/2}}$$

and

$$H(D_x\tilde{u}, x) \leq \bar{H} + O(h^{1/2}).$$

Construct a function $\phi_u \in C(T_h)$ by interpolating linearly the values of $\tilde{u}$ at the nodal points. In each triangle T^i, the oscillation of the derivative of $\tilde{u}$ is $O(h^{1/2})$. Thus

$$D_x\phi_u(x) = D_x\tilde{u}(x) + O(h^{1/2})$$

for any x. Since $H(D_x\tilde{u}, x) \leq \bar{H} + O(h^{1/2})$, for every point $x \in T^i$,

$$H(D_x\phi_u, x) \leq \bar{H} + O(h^{1/2}).$$

This implies

$$\inf_{\phi \in C(T_h)} \operatorname{esssup}_{x \in \mathbb{T}^n} H(D_x\phi, x) \leq \bar{H} + O(h^{1/2}).$$

In the final case of nonstrictly convex Hamiltonians, the sup convolution (6) with $\epsilon = h^{1/3}$ yields a function $u_{h^{1/3}}$ that satisfies

$$H(D_x u_{h^{1/3}}, x) \leq \bar{H} + o(1)$$

almost everywhere and has Lipschitz constant bounded by $Ch^{-1/3}$. Define $\tilde{u} = \eta_{h^{1/3}} * u_{h^{1/3}}$, which satisfies

$$H(D_x\tilde{u}, x) \leq \bar{H} + o(1)$$

and

$$|D^2_{xx}\tilde{u}| \leq \frac{C}{h^{2/3}}.$$

Since in each triangle the oscillation of the derivative is $O(h^{1/3})$, we obtain

$$H(D_x\phi_u, x) \leq \bar{H} + o(1),$$

thereby proving the last statement of the theorem. □

A corollary to the previous theorem is the following.

COROLLARY 2.8. *Suppose $\xi_h \in \mathbb{R}^n$ is a supporting plane for $\bar{H}_h(P)$ that converges as $h \to 0$ to ξ. Then ξ is a supporting hyperplane for $\bar{H}(P)$. As a consequence, if $\bar{H}(P)$ is differentiable at P, then ξ_h converges to the unique supporting hyperplane of $\bar{H}(P)$ at P.*

Proof. The previous theorem asserts that $\bar{H}(P)$ converges uniformly to $\bar{H}(P)$. In Proposition 2.6 we proved that $\bar{H}_h(P)$ is convex. Therefore the corollary follows from a standard convex analysis argument. □

3. Numerical implementation. In this section we discuss the numerical implementation of the fully discretized minimax problem (4). There are two parts to the discussion: (i) implementing the discrete version of the problem and (ii) solving the resulting optimization problem. If the discretization is performed properly, the resulting minimax problem is convex, and standard routines can be used to find the solution.

3.1. Discretization. In the last section we discussed the approximation of the infinite dimensional problem

$$\bar{H}(P) = \inf_{\phi \in C^1_{per}} \sup_x H(P + D_x\phi, x)$$

by the finite dimensional problem

$$\inf_{\phi \in C(T_h)} \operatorname*{esssup}_x H(D_x\phi, x),$$

for ϕ in the space of continuous piecewise linear grid functions.

To fully discretize the problem, we make a further approximation: we discretize the spatial variable by computing the supremum only at the nodes x_i, which gives the minimax problem

$$\min_{\phi \in C(T_h)} \max_{x_i} H(D_x\phi, x_i) \tag{9}$$

for x_i at the nodal points of the grid function space. The spatial approximation introduces a small additional error of $O(h)$, which is proportional to the Lipschitz constant (in the x variable) of H.

The minimax problem (9) is a finite dimensional nonlinear optimization problem which can by solved using standard optimization routines.

Discretization in one dimension. We first present the discretization scheme in one dimension. Choosing n to be the number of nodes, we get a partition of $\mathbb{T}$, the unit interval with periodic boundary conditions, into n intervals of length $h = 1/n$. For any ϕ in the grid function space $C(T_h)$, we identify ϕ with the vector of values on the nodes

$$\phi \text{ is identified with } u = (u_1, \dots, u_n) = (\phi(0), \dots \phi(ih), \dots \phi((n-1)h)).$$

Then, choosing $x_i = (i + 1/2)h$ to be the midpoint of the interval gives the discretization

$$H(\phi_x, x) = H\left(\frac{u_{i+1} - u_i}{h}, x_i\right) \text{ on } T_i = [ih, (i+1)h]. \tag{10}$$

As long as $H(p, x)$ is convex in p, for each x, the right-hand side of (10) is convex in u_{i+1} and u_i. Taking the maximum over the nodes gives a convex function of n variables to be minimized.

Discretization in two dimensions. Next, in two dimensions, take an $n \times n$ grid for $\mathbb{T}^2$, the unit square with periodic boundary conditions. Create a regular tiling by triangles as follows. To each node i, j, let

$$T^{\pm}_{i,j} = \text{ the triangle with vertices } (i, j), (i \pm 1, j), (i, j \pm 1).$$

For ϕ in the grid function space $C(T_h)$, we identify ϕ with the matrix of values on the nodes:

$$\phi \text{ is identified with } u = (u_{i,j}) = (\phi(ih, jh)), \quad i, j = 1, \dots, n.$$

As a result we have $2n^2$ triangles on which ϕ is linear.

On each triangle, choosing $x^{\pm}_{i,j}, y^{\pm}_{i,j}$ to be a point in the middle of $T^{\pm}_{i,j}$, we get the discretization

$$H(\phi_x, \phi_y) = H^{\pm}\left(\frac{u_{i\pm1,j} - u_{i,j}}{h}, \frac{u_{i,j\pm1} - u_{i,j}}{h}, x^{\pm}_i, y^{\pm}_i\right) \text{ on } T^{\pm}_{i,j}. \tag{11}$$

As long as $H(p, x)$ is convex in p, for each x, the right-hand side of (11) is convex in the variables $u_{i,j}$ and $u_{i\pm1,j\pm1}$.

Taking the maximum over the triangles gives a convex function of $2n^2$ variables to be minimized. Alternately, we can take $x^{\pm}_{i,j} = x_{i,j}, y^{\pm}_{i,j} = y^{i,j}$ and take the maximum $H^{\pm}$ with these values to reduce the number of variables by a factor of two.

3.2. Numerical solution of the minimax problem. The implementation required only that a suitable discretization of the Hamiltonian be given. This discretization takes the form of a map from $\mathbb{R}^n$ to $\mathbb{R}^n$, or a map from $\mathbb{R}^{2n}$ to $\mathbb{R}^{2n}$, in the case of one and two dimensional Hamiltonians, respectively. Each component of the map is convex in each of the variables, and the map is sparse, in the sense that it depends on only a small number of other variables.

Taking the maximum of the components of the map gives a convex function which is to be minimized. In general, there are many minimizers, but the minimum is unique.

At this point, publicly available routines for convex optimization may be used to solve the problem. We carried out the implementation in MATLAB, using the Optimization Toolbox. The Optimization Toolbox contains an assortments of routines for solving multidimensional nonlinear optimization problems, some of which take advantage of sparse linear algebra. We used `fminimax`, which is specially designed for minimax problems but does not use sparse linear algebra. A possible alternative would have been to use the a more general solver with sparse linear algebra, which might have performed better on larger problems. For the problems we implemented, which were of modest size (128 variables in two dimensions), the minimization problem was solved in a few seconds on a laptop computer. Solving for a range of a few hundred values of P took between twenty minutes and a few hours to compute for each problem.

For background on convex optimization we refer to [Fle80]. Briefly, the minimax is solved by searching for the worst of the objective functions, then improving that function by solving a sequence of quadratic programs, which are in turn computed by solving a sequence of linear equations. The error in the solution of the optimization problem is insignificant compared to the discretization error.

3.3. Error estimates. There are three main issues which may lead to errors in the numerical computation of $\bar{H}(P)$: the error which arises from the discretization (which was discussed in the previous section), the error involved in computing the esssup approximately, and the error in solving the discrete problem numerically.

To compute the essential supremum we chose to evaluate the function at a single point in each node. This gives an additional contribution to the error of $O(h)$, depending on the Lipschitz constant of the Hamiltonian. This discretization error was nonnegligible, but it could be eliminated by computing the maximum at the endpoints of the nodes, instead of the middle of each node. Since a linear function on a segment of a triangle achieves its maximum at the nodes, this supremum would have been computed accurately up to $O(h^2)$, decreasing the discretization error at the expense of increasing the number of functions to be evaluated in the minimax.

Improved convergence estimates. Because of the improved convergence when smooth solutions exist, for most values of P we expected to get, and indeed we saw,

linear convergence. Nevertheless, there should be examples where the convergence is sublinear.

Global convergence of $\bar{H}_h(P)$ to $\bar{H}(P)$ may be better than expected. Since $\bar{H}_h(P)$ is convex, and is an upper bound for $\bar{H}(P)$, if for some values of P we get $\bar{H}_h(P)$ accurately (for instance, because there is a smooth solution of (HB)), that immediately implies improved bounds for other values of P. If, for instance, $\bar{H}_h(P) = \bar{H} + O(h)$ in a set of full measure, then immediately one gets $\bar{H}_h(P) = \bar{H}(P) + O(h)$ in the remaining points.

Secondly, in Theorem 2.7 we constructed the approximate minimizer from a viscosity solution. However, there may be other minimizers which may be smoother than the viscosity solution.

Finally, in the proof we used convolutions with a smoothing kernel, to get estimates of the form $H(D_x\tilde{u}, x) \leq \bar{H} + O(h^\alpha)$ for some exponent α. In practice the inequality may be strict at points in which the original viscosity solution is not smooth, which could help to improve the estimates since we are taking suprema.

4. Validation. We begin by studying a one dimensional case for which explicit analytical information is available. This analytical information is used to validate the numerical method.

Theorem 2.7 gives convergence of order $O(h)$ when there exists a smooth solution of (HB). Despite the lack of smoothness in the solution, we obtained convergence rates of $O(h)$.

4.1. Analytical results. Consider the Hamiltonian corresponding to a one dimensional pendulum with mass and length normalized to unity,

$$H(p, x) = \frac{p^2}{2} - \cos 2\pi x.$$

For this Hamiltonian one can find explicitly the solution of (HB).

PROPOSITION 4.1. *The solution* $(u, \bar{H}(P))$ *of* (HB)*, when* H *corresponds to the one dimensional pendulum, is given by*

$$u(x) = \int_0^x -P + s(y)\sqrt{2(\bar{H}(P) + \cos 2\pi y)}\, dy, \tag{12}$$

where $|s(y)| = 1$*, with* $\bar{H}(P) = 1$ *for* $|P| \leq 4\pi^{-1}$ *and*

$$P = \pm \int_0^1 \sqrt{2(\bar{H}(P) + \cos 2\pi y)}\, dy \tag{13}$$

otherwise.

Proof. For each $P \in \mathbb{R}$ and a.e. (almost every) $x \in \mathbb{R}$, the solution $u(P, x)$ satisfies

$$\frac{(P + D_x u)^2}{2} = \bar{H}(P) + \cos 2\pi x.$$

This implies $\bar{H}(P) \geq 1$ and so

$$D_x u = -P \pm \sqrt{2(\bar{H}(P) + \cos 2\pi x)}, \quad \text{a.e. } x \in \mathbb{R}.$$

Thus (12) holds for $|s(y)| = 1$.

TABLE 1
Computed values for $\bar{H}$ as a function of the number of points, comparing the sine and cosine potential, with $P = 0.5$ and $\bar{H}(0.5) = 1$.

Number of points n =	9	17	33	65
Values (using sine)	0.98480	0.99573	0.99887	0.99971
Max of sin on the grid	0.98480	0.99573	0.99887	0.99971
Values (using cosine)	1.00000	1.00000	1.00000	1.00000
Max of cosine on the grid	1.00000	1.00000	1.00000	1.00000

Because H is convex in p and u is a viscosity solution, u is semiconcave, and so the only possible discontinuities in the derivative of u are the ones that satisfy $D_x u(x^-) - D_x u(x^+) > 0$ (see [Eva98]). Therefore s can change sign from 1 to -1 at any point, but jumps from -1 to 1 can happen only when $\sqrt{2(\bar{H}(P) + \cos 2\pi x)} = 0$.

If we require 1-periodicity, then there are two cases. (i) If $\bar{H}(P) > 1$, the solution is C^1 since $\sqrt{2(\bar{H}(P) + \cos 2\pi y)}$ is never zero. These solutions correspond to invariant tori. In this case P and $\bar{H}(P)$ satisfy (13). It is easy to check that this equation has a solution $\bar{H}(P)$ whenever

$$|P| \geq \int_0^1 \sqrt{2(1 + \cos 2\pi y)}dy,$$

that is, whenever $|P| > 4\pi^{-1}$. (ii) Otherwise, when $|P| \leq 4\pi^{-1}$, $\bar{H}(P) = 1$ and $s(x)$ can have a discontinuity. Indeed, $s(x)$ jumps from -1 to 1 when $x = \frac{1}{2} + k$, with $k \in \mathbb{Z}$, and there is a point x_0, defined by the equation

$$-\int_0^1 s(y)\sqrt{2(1 + \cos 2\pi y)}dy = P,$$

in which $s(x)$ jumps from 1 to -1. In this last case the graph $(x, P + D_x u)$ is a backwards invariant set contained in the unstable manifold of the hyperbolic equilibria of the pendulum. The graph of $\bar{H}(P)$ has a flat spot near $P = 0$. □

This example also shows that (HB) does not have a unique solution. Indeed, $\cos 2\pi x$ is also 2-periodic. So if we look for 2-periodic solutions, we find out that for $|P|$ small we can have two points where the derivative is discontinuous, and we can choose one of them freely because our only constraint is periodicity. Note, however, that the value of $\bar{H}$ is uniquely determined and is the same whether we look for 1- or 2-periodic solutions.

4.2. Validation in one dimension. To test our numerical method we varied the number of nodes n, computing $\bar{H}$ for $n = 8, 16, 32, 64$ and with $P = 0.5$. Here the exact value is $\bar{H} = 1$.

For each of the above values of n, the exact answer was achieved with an error bounded by 10^{-10}. However, this is an artifact of the special nature of the example, related to the fact that we resolved the maximum of $\sin(x)$ well. To illustrate this, we use a poorer choice of n values and get a larger error, as seen from Table 1. With these choices of n we see that the discrete problem is solved to within the tolerance of 10^{-6}, but the exact problem is solved only up to the resolution error of the Hamiltonian.

We repeated the same test for $P = 2$, which puts us in the strictly convex part of $\bar{H}(P)$. The value $\bar{H}(2) = 2.0637954$ can be obtained by solving the equation

$$2 = \int_0^1 \sqrt{2(\bar{H}(2) - \cos(2\pi x))}dx$$

TABLE 2

Computed error for $\bar{H}$ as a function of the number of points, comparing the sine and cosine potential, with $P = 2$.

Number of points n =	8	16	32	64	96
Error (using sine) $\times 10^{-5}$	0.1912	0.0013	0.0009	0.0005	0.0183
Error (using cosine) $\times 10^{-5}$	0.1912	-0.0013	-0.0014	0.0074	0.0017

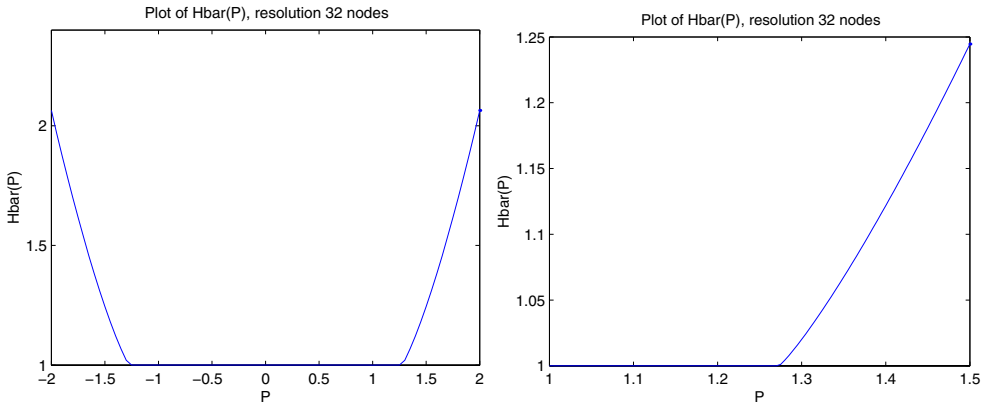


FIG. 1. *$\bar{H}(P)$ for the pendulum, over the range $[-2, 2]$ (left), and an enlargement near the lower-right corner.*

with respect to $\bar{H}(2)$. The error is plotted in Table 2; in this case except for the smallest (n = 8) discretization, the computed values fall within the tolerance of the scheme. The conclusion is that the dominant error comes from the discretization, not the optimization routines.

The values of $\bar{H}(P)$, computed with a resolution of 32 points, are shown in Figure 1. The largest error occurs near the corner. The location of the corner is close to the analytical value $P = 4\pi^{-1} \simeq 1.27324$.

4.3. Validation in two dimensions. In this section we study two uncoupled pendulums. This problem is a direct sum of the one dimensional case, so it is used to validate the method in two dimensions.

The Hamiltonian corresponding to two pendulums is

$$H(p_x, p_y, x, y) = \frac{p_x^2}{2} + \frac{p_y^2}{2} + \cos 2\pi x + \cos 2\pi y.$$

The effective Hamiltonian is thus

$$\bar{H}(P_x, P_y) = \bar{H}_0(P_x) + \bar{H}_0(P_y),$$

in which $\bar{H}_0$ is the effective Hamiltonian for a one dimensional pendulum. For instance, $\bar{H}(1.5) = 1.244638$, $\bar{H}(2.5) = 3.165327$, and thus the analytical value is $\bar{H}(1.5, 2, 5) = 4.4099660$.

As an accuracy test in the two dimensional case, we computed for $P = (1.5, 2.5)$ and $n = 8, 12, 16, 32$ the value of $\bar{H}(P)$ and the corresponding error; see Table 3.

TABLE 3
Computed values for $\bar{H}(1.5, 2.5)$ as a function of the number of points.

n	8	12	16	24	Exact
Values FE	4.6521	4.5627	4.5216	4.4836	4.4099660
Error FE	0.24	0.14	0.11	0.07	

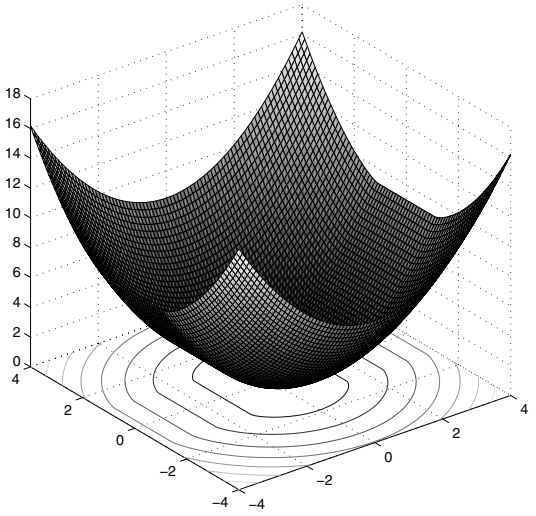


FIG. 2. *Visualization of $\bar{H}(P)$ for the potential $V(x, y) = \cos(2\pi x)$.*

5. Computational results.

5.1. Strictly convex Hamiltonians. In this section we study several strictly convex Hamiltonians. Unless stated otherwise, all the numerical examples use $n = 8$, that is, 64 nodes and 128 triangular elements.

Example 1 (pendulum and a free particle). Potential: $V(x, y) = \cos(2\pi x)$. This potential is y-independent, and therefore $\bar{H}(P_1, P_2)$ for P_2 fixed should be the same as the effective Hamiltonian for the one dimensional pendulum. With P_1 fixed, $\bar{H}$ should be a parabola in P_2. We solved for $\bar{H}(P)$ with a resolution of .1 for $P \in [-4, 4] \times [-4, 4]$. The result is presented in Figure 2.

Example 2 (potential $V(x, y) = \cos(2\pi x)\cos(2\pi y)$). We computed $\bar{H}(P)$ with a resolution in P of .1 for $P \in [-4, 4] \times [-4, 4]$. The result is presented in Figure 3.

Example 3 (potential $V(x, y) = \cos(2\pi x) + \cos(2\pi y) + \cos(2\pi(x - y))$). We computed $\bar{H}(P)$ with a resolution of .25 in P for $P \in [-3, 3] \times [-3, 3]$. The result is presented in Figure 4.

Example 4 (double pendulum). The double pendulum is a well known non-integrable example for which the effective Hamiltonian is not known. The Hamiltonian for the double pendulum is

$$H(p_x, p_y, x, y) = \frac{p_x^2 - 2p_x p_y \cos(2\pi(x - y)) + 2p_y^2}{2 - \cos^2(2\pi(x - y))} + 2\cos 2\pi x + \cos 2\pi y.$$

We computed the values of $\bar{H}(P)$ with a resolution in P of .2 for P in $[-5, 5] \times [-5, 5]$. The result is presented in Figure 5.

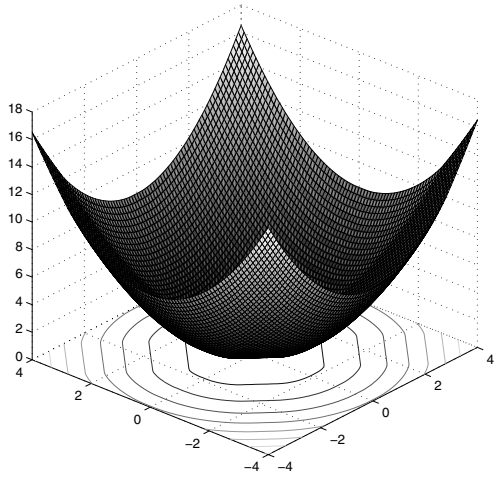


FIG. 3. *Visualization of* $\bar{H}(P)$ *for the potential* $V(x,y) = \cos(2\pi x)\cos(2\pi y)$.

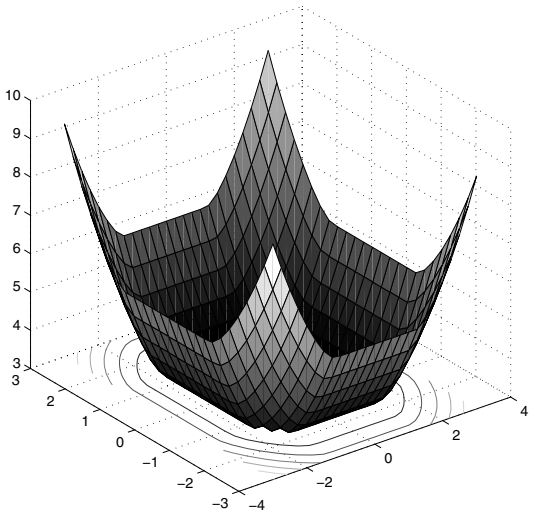


FIG. 4. *Visualization of* $\bar{H}(P)$ *for the potential* $V(x,y) = \cos(2\pi x) + \cos(2\pi y) + \cos(2\pi(x-y))$.

5.2. Nonstrictly convex problems. In this section we study several examples, in which H is convex but not strictly convex, for which there is a viscosity solution of (HB).

Example 5 (linear nonresonant). Consider the linear (nonresonant) Hamiltonian

$$H(p,x) = \omega \cdot p + V(x,y). \tag{14}$$

Suppose u is a smooth viscosity solution of (HB) for this Hamiltonian. The divergence theorem yields

$$\int_{\mathbb{T}^n} \omega \cdot D_x u = 0.$$

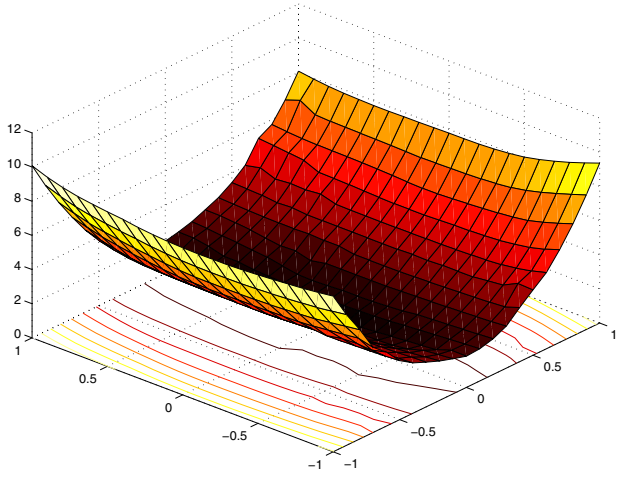


FIG. 5. *Visualization of $\bar{H}(P)$ for the double pendulum.*

Therefore

$$\bar{H}(0) = \int_{\mathbb{T}^n} V \tag{15}$$

and $\bar{H}(P) = \bar{H}(0) + \omega \cdot P$. For the example $u_x + \sqrt{2}u_y + \cos(2\pi x)$ we obtained $D_P\bar{H} = (1, \sqrt{2})$ and $\bar{H}(0,0) = 0$. Despite the fact that the vector $(1,1)$ is rationally dependent, the Hamilton–Jacobi equation $u_x + u_y + \cos(2\pi x)$ is nonresonant because of the nature of the potential. Numerically we obtained $D_P\bar{H} = (1,1)$ and $\bar{H}(0,0) = 0$. In this (linear) case the optimization routine converged very quickly.

Example 6 (time-periodic). Another example is a periodic time-dependent one-space-dimension Hamilton–Jacobi equation

$$-u_t + H(D_x u, x, t) = \bar{H}.$$

There exists a unique value $\bar{H}$ for which this problem admits space-time-periodic solutions [EG02]. Moreover, this solution is Lipschitz, and thus we have a $O(h)$ convergence.

Note also that $P = (P_t, P_x)$ but $\bar{H}(P)$ is linear in P_t, and so we may as well consider the problem

$$\inf_{\phi} \sup_{(x,t)} -\phi_t + H(P_x + D_x\phi, x, t) = \bar{H}(P_x).$$

For the forced pendulum

$$H(p, x) = \frac{p^2}{2} + \cos 2\pi x + \sin 2\pi x \sin 2\pi t.$$

We set $P_t = 0$ and plot $\bar{H}(P_x)$; see Figure 6.

Observe that the maximum of the potential $\cos 2\pi x + \sin 2\pi x \sin 2\pi t$ is $\sqrt{2}$, which coincides with the minimum of $\bar{H}(P)$.

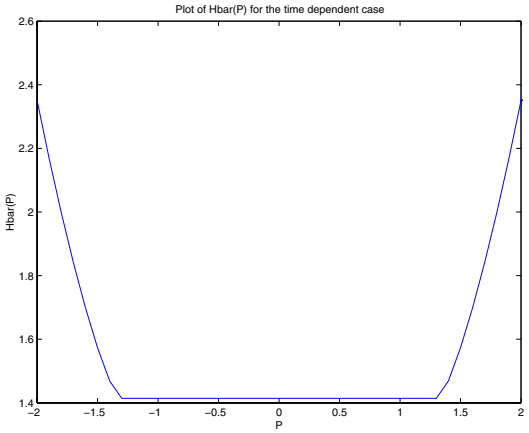


FIG. 6. *Values of $\bar{H}(P)$ for the time-periodic Hamiltonian.*

Example 7 (vakonomic). Finally, we study an example of a nonstrictly convex Hamiltonian which satisfies commutation relations related to vakonomic mechanics [AKN97],

$$H(p,x) = \frac{|f_1 \cdot Du|^2}{2} + \frac{|f_2 \cdot Du|^2}{2} + V(x,y).$$

Here the vector fields f_1, f_2 do not span $\mathbb{R}^2$ in every point, but when we consider the commutator $[f_1, f_2]$, we have that $f_1, f_2, [f_1, f_2]$ span $\mathbb{R}^2$ in every point. In this situation (HB) has Hölder continuous viscosity solutions [EJ89, Gom01a].

We choose $V = 0$, $f_1 = (0,1)$, and $f_2 = (\cos 2\pi y, \sin 2\pi y)$. If $\sin 2\pi y = 0$, $f_2 = (0, \pm 1)$, and so f_1 and f_2 are linearly dependent. However,

$$[f_1, f_2] = 2\pi(-\sin 2\pi y, \cos 2\pi y),$$

and so the vectors $f_1, f_2, [f_1, f_2]$ always span $\mathbb{R}^2$. Therefore there is a Hölder continuous viscosity solution.

In fact, this example can be reduced to a one dimensional problem. The Hamilton–Jacobi equation is

$$\frac{\cos^2 2\pi y}{2}(P_x + u_x)^2 + \frac{(1 + \sin^2 2\pi y)}{2}(P_y + u_y)^2 + \sin 2\pi y \cos 2\pi u (P_x + u_x)(P_y + u_y) = \bar{H}(P_x, P_y).$$

Since there is no explicit dependence in x, there are solutions independent of x, given by the equation

$$\frac{\cos^2 2\pi y}{2} P_x^2 + \frac{(1 + \sin^2 2\pi y)}{2}(P_y + u_y)^2 + \sin 2\pi y \cos 2\pi y P_x (P_y + u_y) = \bar{H}(P_x, P_y).$$

Since this equation is strictly convex in u_y, there is a Lipschitz solution.

To remove this degeneracy we considered the potential $V(x,y) = \cos 2\pi x + \sin 2\pi(x - y)$, for which the previous reduction procedure does not work. The result is presented in Figure 7.

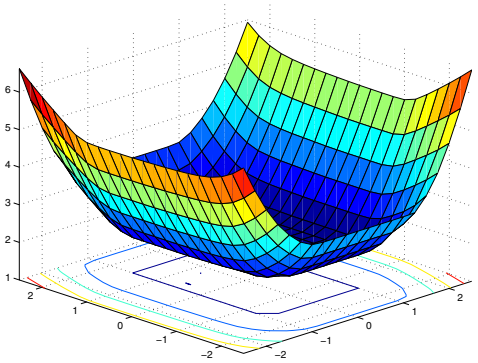


FIG. 7. *Values of $\bar{H}(P)$ for the time-periodic Hamiltonian.*

5.3. Nonexistence of viscosity solutions. There are situations where there do not exist viscosity solutions to (HB), but where $\bar{H}$ can still be defined by solving a more general problem; see [BS00, BS01] and [LS03]. In some of these situations, the solution of the minimax problem (1) may exist and give a consistent result.

We work out two interesting examples and try to explain the results obtained numerically.

The problem

$$\alpha u^\alpha + H(P + D_x u^\alpha, x) = 0, \tag{16}$$

which (when $\alpha \neq 0$) has a unique solution, is considered in [LS03]. Sending $\alpha \to 0$ gives the effective Hamiltonian

$$\bar{H}(P) \equiv \lim_{\alpha \to 0} \alpha u^\alpha. \tag{17}$$

These results are consistent with (HB) but determine a value $\bar{H}(P)$ even under weaker conditions, for example, as long as αu^α converges uniformly to a constant, which happens when

$$u^\alpha - \min_x u^\alpha \to \text{ bounded function of } x$$

uniformly.

For example, in the simpler case when H is strictly convex in p, we get that $u^\alpha - \min_x u^\alpha$ is bounded uniformly in α, since in this case the solutions u^α are Lipschitz independently of α.

The result (1) may also give a correct value for $\bar{H}$ in these more general situations.

PROPOSITION 5.1. *Let u^α be a solution of* (16)*, and suppose that αu^α converges uniformly to a constant number $\bar{H}(P)$. Then*

$$\bar{H}(P) = \lim_{\alpha \to 0} \alpha u^\alpha = \inf_{\phi} \sup_{x \in \mathbb{T}^n} H(P + D_x \phi, x).$$

Proof. 1. Define $\bar{H}_\alpha \equiv -\alpha \min_x u^\alpha$ and

$$v^\alpha \equiv u^\alpha + \frac{\bar{H}_\alpha}{\alpha},$$

so that $\min_x v^\alpha = 0$. We will demonstrate $\bar{H}_\alpha \to \bar{H}$. We have

$$\bar{H} = \lim_{\alpha\to 0} H(P + D_x u^\alpha, x) = \lim_{\alpha\to 0} -\alpha u^\alpha = \lim_{\alpha\to 0} \alpha(u^\alpha - \min_x u^\alpha) + \alpha \min_x u^\alpha = \bar{H}_\alpha.$$

2. Let v_α^ϵ denote the sup convolution of v_α, and let $\phi = \eta_\epsilon * v_\alpha^\epsilon$. Then

$$H(D_x\phi, x) \le \bar{H}_\alpha + O(\epsilon).$$

Therefore

$$\inf_\phi \sup_{x\in\mathbb{T}^n} H(D_x\phi, x) \le \bar{H}_\alpha \to \bar{H}.$$

3. Now let

$$e_\alpha = \sup_x \alpha v_\alpha,$$

which converges to 0.

Let ϕ be any function. Then $v_\alpha - \phi$ has a local minimum at a point x_0. At this point

$$\alpha v_\alpha(x_0) + H(D_x\phi(x_0), x_0) \ge \bar{H}_\alpha.$$

Thus

$$e_\alpha + H(D_x\phi(x_0), x_0) \ge \bar{H}_\alpha,$$

and so

$$\sup_{x\in\mathbb{T}^n} H(D_x\phi, x) \ge \bar{H}_\alpha - e_\alpha \to \bar{H}.$$

Therefore $\inf_\phi \sup_{x\in\mathbb{T}^n} H(D_x\phi, x) \ge \bar{H}$. □

Example 8 (quasiperiodic Hamiltonians). We consider an example from [LS03] for which there is no viscosity solution to (HB), yet where $\bar{H}(P)$ can be determined from (17). Let

$$H(p_x, p_y, x, y) = |p_x + \alpha p_y| + \sin(x) + \sin(y)$$

with α irrational. We computed $\bar{H}(P)$ numerically from (1). The results are presented in Figure 8.

Example 9 (linear resonant). Resonant linear Hamiltonians (14) may fail to have a viscosity solution. An example is

$$(0,1)\cdot Du + \sin(2\pi x) = \bar{H}.$$

The formula (15) yields $\bar{H}(0) = 0$ if there is a solution of (HB). However, we have

$$\inf_\phi \sup_x H(D_x\phi, x) = 1.$$

Let ϕ be an arbitrary periodic function. Set $x_0 = 1/4$, so that $\sin 2\pi x_0 = 1$. Then $\phi(x_0, y)$ is a periodic function of y, and so $D_y\phi(x_0, y) = 0$ at some $y = y_0$. Thus

$$\sup_x H(D_x\phi, x) \ge H(D_x\phi(x_0, y_0), x_0, y_0) = 1.$$

Numerically we obtained $D_P\bar{H} = (0,1)$ and $\bar{H}(0,0) = 1$. This is interesting, because $\bar{H}(0,0)$ should be 0, not 1, and this shows the nonexistence of solutions.

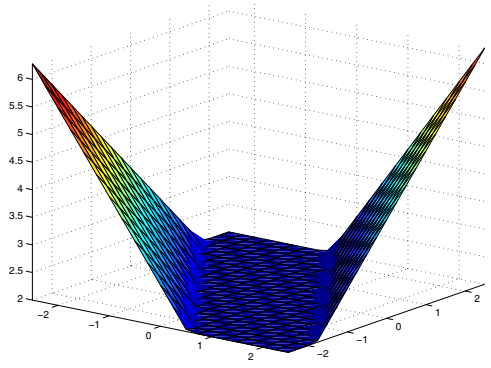


FIG. 8. *Values of $\bar{H}(P)$ for the quasi-periodic Hamiltonian.*

REFERENCES

[AI01] O. ALVAREZ AND H. ISHII, *Hamilton–Jacobi equations with partial gradient and application to homogenization*, Comm. Partial Differential Equations, 26 (2001), pp. 983–1002.

[AKN97] V. I. ARNOLD, V. V. KOZLOV, AND A. I. NEISHTADT, *Mathematical Aspects of Classical and Celestial Mechanics*, Springer-Verlag, Berlin, 1997 (translated from the 1985 Russian original by A. Iacob).

[Ari97] M. ARISAWA, *Ergodic problem for the Hamilton–Jacobi–Bellman equation.* I. *Existence of the ergodic attractor*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 415–438.

[Ari98] M. ARISAWA, *Ergodic problem for the Hamilton–Jacobi–Bellman equation.* II. Ann. Inst. H. Poincaré Anal. Non Linéaire, 15 (1998), pp. 1–24.

[Aro65] G. ARONSSON, *Minimization problems for the functional* $\sup_x F(x, f(x), f'(x))$, Ark. Mat., 6 (1965), pp. 33–53.

[Aro66] G. ARONSSON, *Minimization problems for the functional* $\sup_x F(x, f(x), f'(x))$. II. Ark. Mat., 6 (1966), pp. 409–431.

[Bar94] E. N. BARRON, *Optimal control and calculus of variations in* L^∞, in Optimal Control of Differential Equations (Athens, OH, 1993), Dekker, New York, 1994, pp. 39–47.

[BCD97] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser Boston, Boston, MA, 1997.

[BJW01a] E. N. BARRON, R. R. JENSEN, AND C. Y. WANG, *The Euler equation and absolute minimizers of* L^∞ *functionals*, Arch. Ration. Mech. Anal., 157 (2001), pp. 255–283.

[BJW01b] E. N. BARRON, R. R. JENSEN, AND C. Y. WANG, *Lower semicontinuity of* L^∞ *functionals*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 18 (2001), pp. 495–517.

[BS00] G. BARLES AND P. E. SOUGANIDIS, *On the large time behavior of solutions of Hamilton–Jacobi equations*, SIAM J. Math. Anal., 31 (2000), pp. 925–939.

[BS01] G. BARLES AND P. E. SOUGANIDIS, *Space-time periodic solutions and long-time behavior of solutions to quasi-linear parabolic equations*, SIAM J. Math. Anal., 32 (2001), pp. 1311–1323.

[CDI01] I. CAPUZZO-DOLCETTA AND H. ISHII, *On the rate of convergence in homogenization of Hamilton–Jacobi equations*, Indiana Univ. Math. J., 50 (2001), pp. 1113–1129.

[CIL92] M. G. CRANDALL, H. ISHII, AND P.-L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[CIPP98] G. CONTRERAS, R. ITURRIAGA, G. P. PATERNAIN, AND M. PATERNAIN, *Lagrangian graphs, minimizing measures and Mañé's critical values*, Geom. Funct. Anal., 8 (1998), pp. 788–809.

[Con95] M. CONCORDEL, *Periodic Homogenization of Hamilton–Jacobi Equations*, Ph.D. thesis, Department of Mathematics, University of California at Berkeley, Berkeley, CA, 1995.

[E99] Weinan E., *Aubry–Mather theory and periodic solutions of the forced Burgers equation*, Comm. Pure Appl. Math., 52 (1999), pp. 811–828.

[EG01] L. C. Evans and D. Gomes, *Effective Hamiltonians and averaging for Hamiltonian dynamics*. I., Arch. Ration. Mech. Anal., 157 (2001), pp. 1–33.

[EG02] L. C. Evans and D. Gomes, *Effective Hamiltonians and averaging for Hamiltonian dynamics*. II., Arch. Ration. Mech. Anal., 161 (2002), pp. 271–305.

[EJ89] L. C. Evans and M. R. James, *The Hamilton–Jacobi–Bellman equation for time-optimal control*, SIAM J. Control Optim., 27 (1989), pp. 1477–1489.

[EMS95] P. F. Embid, A. J. Majda, and P. E. Souganidis, *Comparison of turbulent flame speeds from complete averaging and the G-equation*, Phys. Fluids, 7 (1995), pp. 2052–2060.

[Eva98] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[Eva03] L. C. Evans, *Some new PDE methods for weak KAM theory*, Calc. Var. Partial Differential Equations, 17 (2003), pp. 159–177.

[Fat97a] A. Fathi, *Solutions KAM faibles conjuguées et barrières de Peierls*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 649–652.

[Fat97b] A. Fathi, *Théorème KAM faible et théorie de Mather sur les systèmes Lagrangiens*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 1043–1046.

[Fat98a] A. Fathi, *Orbite hétéroclines et ensemble de Peierls*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1213–1216.

[Fat98b] A. Fathi, *Sur la convergence du semi-groupe de Lax-Oleinik*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 267–270.

[Fle80] R. Fletcher, *Practical methods of optimization, Vol.* 1*, Unconstrained optimization*, John Wiley & Sons, Chichester, UK, 1980.

[FS93] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[FSar] A. Fathi and A. Siconolfi, *Existence of C^1 critical subsolutions of the Hamilton–Jacobi equation*, Invent. Math., 155 (2004), pp. 363–388.

[Gom01a] D. Gomes, *Hamilton–Jacobi Methods for Vakonomic Mechanics*, preprint, 2001; available online at http://www.math.ist.utl.pt/~dgomes/.

[Gom01b] D. Gomes, *Viscosity solutions of Hamilton–Jacobi equations, and asymptotics for Hamiltonian systems*, Cal. Var. Partial Differential Equations, 14 (2002), pp. 345–357.

[Gom02] D. A. Gomes, *A stochastic analogue of Aubry–Mather theory*, Nonlinearity, 15 (2002), pp. 581–603.

[KBM01] B. Khouider, A. Bourlioux, and A. J. Majda, *Parametrizing the burning speed enhancement by small-scale periodic flows,* I*., Unsteady shears, flame residence time and bending*, Combust. Theory Model., 5 (2001), pp. 295–318.

[LPV88] P. L. Lions, G. Papanicolau, and S. R. S. Varadhan, *Homogenization of Hamilton–Jacobi Equations*, manuscript, 1988.

[LS03] P.-L. Lions and P. E. Souganidis, *Correctors for the homogenization of Hamilton–Jacobi equations in the stationary ergodic setting*, Comm. Pure Appl. Math., 56 (2003), pp. 1501–1524.

[Mat89a] J. N. Mather, *Minimal action measures for positive-definite Lagrangian systems*, in Proceedings of the 9th International Congress on Mathematical Physics (Swansea, UK, 1988), Hilger, Bristol, UK, 1989, pp. 466–468.

[Mat89b] J. N. Mather, *Minimal measures*, Comment. Math. Helv., 64 (1989), pp. 375–394.

[Mat91] J. N. Mather, *Action minimizing invariant measures for positive definite Lagrangian systems*, Math. Z., 207 (1991), pp. 169–207.

[Mn92] R. Mañé, *On the minimizing measures of Lagrangian dynamical systems*, Nonlinearity, 5 (1992), pp. 623–638.

[Mn96] R. Mañé, *Generic properties and problems of minimizing measures of Lagrangian systems*, Nonlinearity, 9 (1996), pp. 273–310.

[Qia01] J. Qian, *A note on a numerical method for effective Hamiltonians*, J. Comput. Appl. Math., submitted.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 813–830

EXISTENCE OF OPTIMAL STOCHASTIC CONTROLS AND GLOBAL SOLUTIONS OF FORWARD-BACKWARD STOCHASTIC DIFFERENTIAL EQUATIONS *

MARCO FUHRMAN† AND GIANMARIO TESSITORE‡

Abstract. We consider an optimal stochastic control problem, assuming Lipschitz conditions and allowing degeneracy of the diffusion coefficient, under some structural constraint on the state equation. We formulate the problem in the strong form; i.e., we fix the probability space. We relate the value function and the feedback law to a forward-backward stochastic differential system. We prove existence and uniqueness of a global solution to the latter and deduce existence and, in some cases, uniqueness of an optimal control. To solve the (coupled) forward-backward system we use a priori estimates which follow from its control-theoretic interpretation.

Key words. stochastic optimal control, forward-backward stochastic differential equations

AMS subject classifications. 93E20, 60H10

DOI. 10.1137/S0363012903428664

1. Introduction.

We consider a controlled Markov process X^u in $\mathbb{R}^n$, on a time interval $[t,T] \subset [0,T]$, described by an Ito stochastic differential equation of the form

$$\text{(1.1)} \quad \begin{cases} dX^u_\tau = f(\tau, X^u_\tau)\, d\tau + g(\tau, X^u_\tau) r(\tau, X^u_\tau, u_\tau)\, d\tau + g(\tau, X^u_\tau)\, dW_\tau, \quad \tau \in [t,T], \\ X^u_t = x \in \mathbb{R}^n, \end{cases}$$

where $u_\cdot$ is the control process, W is a standard Wiener process in $\mathbb{R}^d$, and f, r, g are functions with values in $\mathbb{R}^n$, $\mathbb{R}^d$, $\mathbb{R}^{n\times d}$, respectively. We introduce the cost functional

$$J(t,x,u_\cdot) = \mathbb{E}\int_t^T l(\tau, X^u_\tau, u_\tau)\, d\tau + \mathbb{E}\,\phi(X^u_T),$$

where l, ϕ are real functions. We wish to minimize the cost over all admissible controls, i.e., processes $\{u_s, s \in [0,T]\}$ taking values in a prescribed set $\mathcal{U} \subset \mathbb{R}^m$ and predictable with respect to the (augmented) filtration generated by W. We introduce the value function

$$V(t,x) = \inf_{u_\cdot} J(t,x,u_\cdot), \qquad t \in [0,T],\ x \in \mathbb{R}^n,$$

where the infimum is taken over all admissible controls $u_\cdot$.

The particular form of the control system is essential for our results, but it covers a number of interesting cases. For instance, if $d = n$ and g takes values in the set of invertible matrices, then starting from a system of the form

$$\begin{cases} dX^u_\tau = h(\tau, X^u_\tau, u_\tau)\, d\tau + g(\tau, X^u_\tau)\, dW_\tau, \qquad s \in [t,T], \\ X^u_t = x, \end{cases}$$

*Received by the editors May 30, 2003; accepted for publication (in revised form) January 31, 2004; published electronically September 18, 2004.

http://www.siam.org/journals/sicon/43-3/42866.html

†Dipartimento di Matematica, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy (marco.fuhrman@polimi.it). This author was partially supported by the European Community's Human Potential Programme under contract HPRN-CT-2002-00279, QP-Applications.

‡Dipartimento di Matematica, Università di Parma, via d'Azeglio 85, 43100 Parma, Italy (gianmario.tessitore@unipr.it). This author was partially supported by the European Community's Human Potential Programme under contract HPRN-CT-2002-00281, Evolution Equations.

where h takes values in $\mathbb{R}^n$, we can arrive at the form (1.1) setting $r = g^{-1}h$. We also note that in the special case $r(\tau, x, u) = u$, the term $u_\tau\, d\tau + dW_\tau$ admits a natural interpretation as a *control affected by noise*.

On the functions f, g, r we impose some Lipschitz and linear growth (or boundedness) conditions, but no further regularity is assumed (see Hypothesis 2.1 below for precise statements). Moreover, we do not assume any kind of nondegeneracy on the diffusion coefficient g. We solve the optimal control problem completely; i.e., we prove existence and we characterize the value function and the optimal control law. We also show solvability of the closed-loop equation. Under some mild additional assumptions we prove uniqueness of the optimal control and of the solution of the closed-loop equation.

We stress that these results are proved for the *strong formulation* of the control problem, that is, when the probability space in which (1.1) is considered is prescribed. This is a distinctive feature of this paper.

We note that under the present assumptions, due to lack of regularity of the coefficients and possible degeneracy of g, the Hamilton–Jacobi–Bellman equation has only (nondifferentiable) viscosity solutions; see [7] and the references within. In general this is not sufficient either to prove existence of the optimal control or to find an optimal feedback law since this usually involves the gradient of a solution of the Hamilton–Jacobi–Bellman equation. Consequently, an appropriate formulation and well-posedness of the closed-loop equation seem to be a difficult task here. We also notice that even when the Hamilton–Jacobi–Bellman equation has differentiable solutions, well-posedness, in strong sense, of the closed-loop equation is not obvious (see [3]).

Our starting point is the well-known idea of replacing the approach based on the Hamilton–Jacobi–Bellman equation with the analysis of a forward-backward stochastic differential system: we recall that if we define the hamiltonian function

$$\psi(t,x,z) = \inf_{u\in\mathcal{U}}\{l(t,x,u) + zr(t,x,u)\}, \qquad t\in[0,T],\ x\in\mathbb{R}^n,\ z\in\mathbb{R}^d, \tag{1.2}$$

and we consider the uncoupled forward-backward system

$$\begin{cases} dX_\tau = f(\tau, X_\tau)\, d\tau + g(\tau, X_\tau)\, dW_\tau, & \tau\in[t,T],\\ X_t = x,\\ dY_\tau = Z_\tau\, dW_\tau - \psi(\tau, X_\tau, Z_\tau)\, d\tau, & \tau\in[t,T],\\ Y_T = \phi(X_T), \end{cases} \tag{1.3}$$

then it is not difficult to prove that $J^*(t,x) := Y_t$ is a lower bound for the value function; see [6] or Proposition 2.7 and Corollary 2.6 below. At this point, if the control problem is considered in its *weak formulation*, that is, the probability space in which (1.1) is solved is not fixed (see, e.g., [7]), then it is easily proved that the lower bound $J^*(t,x)$ is attained and the process Z in (1.3) allows us to construct an optimal control. In this case the closed-loop equation can be solved but only in the sense of weak solutions of stochastic differential equations. See [5], [12], and [8] for other results stating, in wider generality, existence of optimal control for the weak formulation of the problem. Finally, we notice that backward stochastic differential equations have also been successfully applied to the search for equilibrium points in stochastic differential games (see [9], [10], [11]). In particular, the basic idea in [9] is, in a certain sense, comparable to the one inspiring the present paper; namely, the proof of existence of open-loop Nash equilibrium points relies on an existence and uniqueness result for coupled forward-backward stochastic systems under monotonicity assumptions.

Coming back to our optimal control problem in its original strong formulation, the main idea for circumventing the difficulties outlined above is to relate the existence of an optimal control to the solvability of a different, coupled forward-backward stochastic differential system. Namely, we assume that the infimum in (1.2) is attained at some $u = \gamma(t, x, z)$ and we consider

$$\begin{cases} dX_\tau = f(\tau, X_\tau)\, d\tau + g(\tau, X_\tau) r(\tau, X_\tau, \gamma(\tau, X_\tau, Z_\tau))\, d\tau + g(\tau, X_\tau)\, dW_\tau, \quad \tau \in [t, T], \\ X_t = x, \\ dY_\tau = Z_\tau\, dW_\tau - l(\tau, X_\tau, \gamma(\tau, X_\tau, Z_\tau))\, d\tau, \quad \tau \in [t, T], \\ Y_T = \phi(X_T). \end{cases} \tag{1.4}$$

We prove that if this forward-backward system has a solution (an appropriate predictable process (X, Y, Z) in $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d$), then the control $u_\tau = \gamma(\tau, X_\tau, Z_\tau)$ is optimal, the corresponding trajectory X^u coincides with X, and the optimal cost $V(t, x)$ is equal to Y_t. We then prove, imposing only some Lipschitz and boundedness conditions on γ (see Hypothesis 2.4 and Examples 1 and 2 below), that the system (1.4) is uniquely solvable. Moreover, we show, still relying only on properties of the forward-backward system, that the optimal control u found above is in feedback form, i.e., $u_\tau = \underline{u}(\tau, X_\tau)$ for some deterministic function $\underline{u}$. Finally, if the minimum in (2.3) is unique, we obtain uniqueness of the optimal control and of the solution to the closed-loop equation corresponding to the feedback $\underline{u}$. In particular, we give results for well-posedness of the closed-loop equations under conditions that do not guarantee the classical Lipzchitz conditions on its coefficients.

It is well known that solvability of forward-backward systems of general form is particularly delicate, even if the classical conditions of Lipschitz continuity and linear growth are imposed on the coefficients. Global solutions, i.e., for arbitrary T, for the system (1.4) have been found under various assumptions. In the so-called four step scheme introduced in [16] (see also [17]), as well as in the extensions proved in [4], analytic results on partial differential equations are used (see [15]) which require either regularity of the coefficients [17] or nondegeneracy of the diffusion coefficient g [4]. Several other results give sufficient conditions for the solvability of forward-backward systems: in [20] existence and uniqueness are obtained, by purely probabilistic techniques, under lipschitzianity and monotonicity requirements, in [13] and [21] dissipativity conditions are assumed, and finally in [23] the regularity requirements in [14] are relaxed but only existence is proved.

None of these results is directly applicable to (1.4). To our knowledge, for the system (1.4) only a local solvability result is available in the literature (see [4], [1], or [17]). To obtain a global solution we argue as follows. By the above mentioned local result we know that a unique solution exists in $[T - \delta, T]$ when δ is small enough. Then we prove that if (X, Y, Z) solves the system (1.4) in an arbitrary interval $[t, T]$ with initial condition $X_t = x \in \mathbb{R}^n$ and final condition $Y_T = \phi(X_T)$, then Y_t is equal to the value function $V(t, x)$. At this point we can rely on an estimate of the form

$$|V(t, x_1) - V(t, x_2)| \le c|x_1 - x_2|, \qquad t \in [0, T],\ x_1, x_2 \in \mathbb{R}^n,$$

which is easily obtained from the definition of the value function (and is basically known); thus denoting (X^i, Y^i, Z^i), $i = 1, 2$, the solutions in $[t, T]$ corresponding to initial points x_i $(i = 1, 2)$, we obtain

$$|Y_t^1 - Y_t^2| \le c|x_1 - x_2|,$$

with c independent of t. The above can be considered as a sort of a priori estimate for the Lipschitz dependence of Y_t on x obtained by means of the control-theoretic interpretation of the forward-backward system. This allows us to conclude global existence and uniqueness of the solution to system (1.4) by an argument introduced in [4] consisting in solving the system recursively starting from T in small intervals with fixed length and suitable final conditions.

Finally, in view of the great attention that forward-backward systems have received recently, we stress the fact that our original problem has led us to focus on a special class of such systems for which existence of a unique global solution is not a direct consequence of the results already existing in the literature but rather follows from its control-theoretic interpretation.

2. Formulation of the problem and preliminary results. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, and let $\mathcal{N}$ be the family of elements of $\mathcal{F}$ with probability zero. Let $\{W_t,\ t \in [0,T]\}$ be a standard Wiener process defined in $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{R}^d$. By $(\mathcal{F}_t)_{t\in[0,T]}$ we denote the filtration generated by W augmented by the null sets of $\mathcal{F}$; namely, $\mathcal{F}_t$ denotes the σ-algebra generated by $\mathcal{N}$ and by the random variables W_s, $s \in [0,t]$. We are given a set $\mathcal{U} \subset \mathbb{R}^m$ and functions

$$
\begin{aligned}
&f : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n,\\
&g : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{n\times d},\\
&r : [0,T] \times \mathbb{R}^n \times \mathcal{U} \to \mathbb{R}^d,\\
&\phi : \mathbb{R}^n \to \mathbb{R},\\
&l : [0,T] \times \mathbb{R}^n \times \mathcal{U} \to \mathbb{R}
\end{aligned}
$$

satisfying the following assumptions.

Hypothesis 2.1. $\mathcal{U}$ is a Borel subset of $\mathbb{R}^m$, the functions f, g, r, ϕ, l are Borel measurable, the function $x \to f(t,x)$ is continuous on $\mathbb{R}^n$ for every $t \in [0,T]$, and there exists a constant C such that

$$
\begin{aligned}
&|\phi(x) - \phi(x')| \le C|x - x'|,\\
&\langle f(t,x) - f(t,x'), x - x'\rangle \le C|x - x'|^2,\\
&|g(t,x) - g(t,x')| \le C|x - x'|,\\
&|r(t,x,u) - r(t,x',u')| + |l(t,x,u) - l(t,x',u')| \le C(|x - x'| + |u - u'|),\\
&|f(t,x)| \le C(1 + |x|),\\
&|g(t,x)| + |r(t,x,u)| + |l(t,0,0)| \le C,\\
&l(t,0,u) \ge -C
\end{aligned}
$$

for every $t \in [0,T]$, $x, x' \in \mathbb{R}^n$, $u, u' \in \mathcal{U}$.

This implies in particular

$$
-C(1 + |x|) \le l(t,x,u) \le C(1 + |x| + |u|), \qquad t \in [0,T],\ x \in \mathbb{R}^n,\ u \in \mathcal{U}. \tag{2.1}
$$

We also note that if $r(t,x,u) = u$, then $\mathcal{U}$ is required to be bounded.

A stochastic process $\{u_s,\ s \in [0,T]\}$ in $\mathbb{R}^m$ will be called an *admissible control* if it is predictable with respect to $(\mathcal{F}_t)$ and it takes values in $\mathcal{U}$.

For any admissible control $u_\cdot$ and any $(t,x) \in [0,T] \times \mathbb{R}^n$ we consider the process $\{X^u_\tau,\ \tau \in [t,T]\}$, solution of the Ito stochastic equation: $\mathbb{P}$-a.s.,

$$
X^u_\tau = x + \int_t^\tau f(\sigma, X^u_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X^u_\sigma) r(\sigma, X^u_\sigma, u_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X^u_\sigma)\, dW_\sigma, \quad \tau \in [t,T].
$$

Taking into account that, in particular, the function $g(t,\cdot)r(t,\cdot,u)$ is Lipschitz, uniformly with respect to t and u, it is easy to check that a classical solvability result applies to this equation: namely, under Hypothesis 2.1, there exists a continuous, $(\mathcal{F}_\tau)$-adapted solution, unique up to indistinguishability. It also satisfies $\mathbb{E}\sup_{\tau\in[t,T]}|X^u_\tau|^2<\infty$.

We introduce the cost functional

$$J(t,x,u_\cdot)=\mathbb{E}\int_t^T l(\sigma,X^u_\sigma,u_\sigma)\,d\sigma+\mathbb{E}\phi(X^u_T).$$

As a consequence of our assumptions, in particular the first inequality in (2.1), it is easy to check that for some constant c independent of $u_\cdot$ we have

$$\mathbb{E}\int_t^T l^-(\sigma,X^u_\sigma,u_\sigma)\,d\sigma+\mathbb{E}|\phi(X^u_T)|\le c,$$

where l^- denotes the negative part of l. It follows that for all admissible controls $u_\cdot$ the cost $J(t,x,u_\cdot)$ is a well-defined element in $(-\infty,+\infty]$, and its values are bounded below by $-c$ (which may depend on t and x). We also note that the cost is not identically $+\infty$.

For given $x\in\mathbb{R}^n$ and $t\in[0,T]$, we consider the problem of minimizing $J(t,x,u_\cdot)$ over all admissible controls $u_\cdot$. Any admissible control which minimizes $J(t,x,\cdot)$, if it exists, is called optimal for the control problem starting from x at time t. The minimal value of the cost is then called the optimal cost.

We introduce the value function $V:[0,T]\times\mathbb{R}^n\to\mathbb{R}$ defined by

$$V(t,x)=\inf_{u_\cdot}J(t,x,u_\cdot),\qquad t\in[0,T],\ x\in\mathbb{R}^n,$$

where the infimum is taken over all admissible controls $u_\cdot$. By the previous discussion V is clearly well defined. The following lemma gives an a priori estimate on the Lipschitz constant of $V(t,\cdot)$, which is of basic importance for what follows.

LEMMA 2.2. *Assume Hypothesis* 2.1. *Then there exists a constant c such that*

$$|V(t,x)-V(t,x')|\le c|x-x'|$$

for every $t\in[0,T]$, $x,x'\in\mathbb{R}^n$.

Proof. We fix an arbitrary $u_\cdot$ and we denote by X (respectively, X') the corresponding solution starting from x (respectively, x') at time t. By c we denote a positive constant which may vary from line to line but does not depend on t, x, x', and $u_\cdot$. Applying the Ito formula to $\{|X_\tau-X'_\tau|^2,\tau\in[t,T]\}$ and taking expectation, we obtain

$$\begin{aligned}\mathbb{E}|X_\tau-X'_\tau|^2&=|x-x'|^2+\mathbb{E}\int_t^\tau\langle X_\sigma-X'_\sigma,f(\sigma,X_\sigma)-f(\sigma,X'_\sigma)\rangle\,d\sigma\\&\quad+\mathbb{E}\int_t^\tau\langle X_\sigma-X'_\sigma,g(\sigma,X_\sigma)r(\sigma,X_\sigma,u_\sigma)-g(\sigma,X'_\sigma)r(\sigma,X'_\sigma,u_\sigma)\rangle\,d\sigma\\&\quad+\mathbb{E}\int_t^\tau|g(\sigma,X_\sigma)-g(\sigma,X'_\sigma)|^2\,d\sigma,\qquad\tau\in[t,T].\end{aligned}$$

Recalling that the function $g(t,\cdot)r(t,\cdot,u)$ is Lipschitz, uniformly with respect to $t\in[0,T]$ and $u\in\mathcal{U}$, we easily obtain

$$\mathbb{E}|X_\tau-X'_\tau|^2\le|x-x'|^2+c\mathbb{E}\int_t^\tau|X_\sigma-X'_\sigma|^2\,d\sigma,\qquad\tau\in[t,T],$$

and therefore, by the Gronwall lemma, $\mathbb{E}|X_\tau - X'_\tau|^2 \le c|x-x'|^2$.

Moreover, our assumptions on l and ϕ imply that for all x and x' in $\mathbb{R}^n$,

$$J(t,x,u_\cdot) < +\infty \quad \Rightarrow \quad J(t,x',u_\cdot) < +\infty,$$
$$|J(t,x,u_\cdot) - J(t,x',u_\cdot)| \le c\mathbb{E}\int_t^T |X_\sigma - X'_\sigma|\, d\sigma + c\mathbb{E}|X_T - X'_T| \le |x-x'|.$$

To conclude, for fixed $\epsilon > 0$, we choose $u^\epsilon_\cdot$ such that $J(t,x,u^\epsilon_\cdot) \le V(t,x) + \epsilon$. Then

$$V(t,x') \le J(t,x',u^\epsilon_\cdot) \le J(t,x,u^\epsilon_\cdot) + c|x-x'| \le V(t,x) + c|x-x'| + \epsilon,$$

and the claim follows inverting the role of x and x'. □

We introduce the hamiltonian function $\psi : [0,T] \times \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ setting

$$\psi(t,x,z) = \inf_{u\in\mathcal{U}}\{l(t,x,u) + zr(t,x,u)\}, \qquad t \in [0,T],\ x \in \mathbb{R}^n,\ z \in \mathbb{R}^d. \tag{2.2}$$

By the boundedness of r and by (2.1), ψ is well defined as a real-valued function. For further use we prove some additional properties of ψ.

LEMMA 2.3. *Assume Hypothesis* 2.1. *Then there exists a constant c such that*

$$|\psi(t,0,0)| \le c, \qquad |\psi(t,x,z) - \psi(t,x',z')| \le c|z-z'| + c|x-x'|(1+|z|+|z'|)$$

for every $t \in [0,T]$, $x, x' \in \mathbb{R}^n$, $z, z' \in \mathbb{R}^d$.

Proof. Since $l(t,0,u) \ge -C$, therefore $\psi(t,0,0) = \inf_{u\in\mathcal{U}} l(t,0,u) \ge -C$. From (2.1) we obtain, for an arbitrary fixed element $u_0 \in \mathcal{U}$,

$$\psi(t,0,0) = \inf_{u\in\mathcal{U}} l(t,0,u) \le l(t,0,u_0) \le C(1+|u_0|),$$

and the first inequality of the lemma is proved.

For $u \in \mathcal{U}$ we have

$$\begin{aligned} l(t,x,u) + zr(t,x,u) &\le l(t,x',u) + z'r(t,x',u) + |l(t,x,u) - l(t,x',u)| \\ &\quad + |zr(t,x,u) - z'r(t,x',u)| \\ &\le l(t,x',u) + z'r(t,x',u) + |l(t,x,u) - l(t,x',u)| \\ &\quad + |(z-z')r(t,x,u)| + |z'(r(t,x,u) - r(t,x',u))| \\ &\le l(t,x',u) + z'r(t,x',u) + C|x-x'| \\ &\quad + C|z-z'| + C|z'||x-x'|. \end{aligned}$$

Taking the infimum over all $u \in \mathcal{U}$, we obtain

$$\psi(t,x,z) - \psi(t,x',z') \le C|x-x'| + C|z-z'| + C|z'||x-x'| \le c|z-z'| + c|x-x'|(1+|z|+|z'|)$$

for some c. Exchanging x, z with x', z', we get the conclusion. □

We will assume that the infimum in (2.2) is attained at some u which can be chosen to depend measurably on t, x, z, and the resulting function satisfies additional conditions. More precisely we assume the following.

Hypothesis 2.4. There exists a Borel measurable function $\gamma : [0,T] \times \mathbb{R}^n \times \mathbb{R}^d \to \mathcal{U}$ such that

$$\psi(t,x,z) = l(t,x,\gamma(t,x,z)) + zr(t,x,\gamma(t,x,z)), \qquad t \in [0,T],\ x \in \mathbb{R}^n,\ z \in \mathbb{R}^d, \tag{2.3}$$

and such that, for some constant C, we have

$$|\gamma(t,x,z) - \gamma(t,x',z')| \le C(|x-x'| + |z-z'|),$$
$$|\gamma(t,0,0)| \le C$$

for every $t \in [0,T]$, $x, x' \in \mathbb{R}^n$, $z, z' \in \mathbb{R}^d$.

In general we do not assume that, for any given $t \in [0,T]$, $x \in \mathbb{R}^n$, $z \in \mathbb{R}^d$, the infimum in (2.2) is attained at a unique point. However, this will be required for some of our results below; in this case the equality $\psi(t,x,z) = l(t,x,u) + zr(t,x,u)$ implies $u = \gamma(t,x,z)$.

Next we present two examples where our assumptions on ℓ, r, and $\mathcal{U}$, including Hypothesis 2.4, are satisfied.

Example 1. Let $d = m = 1$, $r(t,x,u) = u$, $\mathcal{U} = [-\delta, \delta]$.

Suppose that ℓ satisfies the assumptions in Hypothesis 2.1 and moreover that it can be extended to a function on $[0,T] \times \mathbb{R}^n \times \mathbb{R}$, still denoted by $\ell(t,x,z)$, such that $\ell(t,\cdot,\cdot) \in C^{1,2}(\mathbb{R}^n \times \mathbb{R}; \mathbb{R})$ and, for a suitable $c > 0$,

$$\frac{1}{c} \le \frac{\partial^2 \ell}{\partial u^2}(t,x,u) \le c, \qquad \left| \frac{\partial \ell}{\partial x}(t,x,u) \right| \le c$$

for all $t \in [0,T]$, $x \in \mathbb{R}^n$, $u \in \mathbb{R}$.

Under the previous assumptions we have

$$\lim_{u \to +\infty} \frac{\partial \ell}{\partial u}(t,x,u) = +\infty, \qquad \lim_{u \to -\infty} \frac{\partial \ell}{\partial u}(t,x,u) = -\infty.$$

Moreover, the infimum in (2.2) is attained at a unique point,

$$\gamma(t,x,z) = (-\delta) \vee (\widehat{u}(t,x,z) \wedge \delta),$$

where $\widehat{u}(t,x,z)$ is the unique solution of the equation

$$\frac{\partial \ell}{\partial u}(t,x,\widehat{u}(t,x,z)) = -z.$$

Thus the partial derivatives

$$\frac{\partial \widehat{u}}{\partial x}(t,x,z) = -\left[\frac{\partial^2 \ell}{\partial u^2}(t,x,\widehat{u}(t,x,z)) \right]^{-1} \frac{\partial \ell}{\partial x}(t,x,\widehat{u}(t,x,z)),$$
$$\frac{\partial \widehat{u}}{\partial z}(t,x,z) = -\left[\frac{\partial^2 \ell}{\partial u^2}(t,x,\widehat{u}(t,x,z)) \right]^{-1}$$

are bounded, and consequently $\gamma(t,\cdot,\cdot)$ is Lipschitz continuous uniformly in t.

Example 2. Let $\ell(t,x,u) = \ell^0(t,x) + |u|^2$, $r(t,x,u) = Bu$ with $\ell^0 : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ such that $\ell^0(\cdot,0)$ is bounded, $\ell^0(t,\cdot)$ is Lipschitz uniformly with respect to t, and $B \in L(\mathbb{R}^m, \mathbb{R}^d)$. Moreover, let $\mathcal{U} = \{u \in \mathbb{R}^m : |u| \le \delta\}$, where $\delta > 0$. Then the infimum in (2.2) is attained at a unique point:

$$\gamma(z) = \begin{cases} -(1/2)B^*z & \text{if } |B^*z| \le 2\delta, \\ -\delta |B^*z|^{-1} B^*z & \text{if } |B^*z| > 2\delta, \end{cases}$$

and clearly γ is a Lipschitz function.

In Proposition 2.5 and Corollary 2.6 we collect some results relating the optimal control problem introduced above to suitable backward stochastic differential equations. Such results are mainly known. From now on we assume that Hypotheses 2.1 and 2.4 hold.

Let $\widetilde{W}$ be a standard Wiener process in $\mathbb{R}^d$, defined in some complete probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$. For $0 \le t \le \tau \le T$, we denote by $\widetilde{\mathcal{F}}_t$ (respectively, $\widetilde{\mathcal{F}}_{[t,\tau]}$) the σ-algebra generated by $\widetilde{W}_s$, $s \in [0,t]$ (respectively, $s \in [t,\tau]$), and augmented by the null sets of $\widetilde{\mathcal{F}}$.

For fixed $t \in [0,T]$ and $x \in \mathbb{R}^n$ we consider the equation

$$\widetilde{X}_\tau = x + \int_t^\tau f(\sigma, \widetilde{X}_\sigma)\, d\sigma + \int_t^\tau g(\sigma, \widetilde{X}_\sigma)\, d\widetilde{W}_\sigma, \qquad \tau \in [t,T]. \tag{2.4}$$

The solution $\{\widetilde{X}_\tau,\ \tau \in [t,T]\}$ is a continuous process in $\mathbb{R}^n$, adapted to the filtration $(\widetilde{\mathcal{F}}_{[t,\tau]})_{\tau\in[t,T]}$. Moreover, the law of $(\widetilde{W}, \widetilde{X})$ is uniquely determined by x, f, and g. Next we consider the backward stochastic differential equation

$$\widetilde{Y}_\tau + \int_\tau^T \widetilde{Z}_\sigma\, d\widetilde{W}_\sigma = \phi(\widetilde{X}_T) + \int_\tau^T \psi(\sigma, \widetilde{X}_\sigma, \widetilde{Z}_\sigma)\, d\sigma, \qquad \tau \in [t,T]. \tag{2.5}$$

By a solution of (2.5) we mean a process $\{(\widetilde{Y}_\tau, \widetilde{Z}_\tau),\ \tau \in [t,T]\}$ with values in $\mathbb{R}\times\mathbb{R}^d$, predictable with respect to $(\widetilde{\mathcal{F}}_{[t,\tau]})_{\tau\in[t,T]}$, such that, denoting by $\widetilde{\mathbb{E}}$ the expectation with respect to $\widetilde{\mathbb{P}}$, we have

$$\widetilde{\mathbb{E}} \sup_{\tau\in[t,T]} |\widetilde{Y}_\tau|^2 + \widetilde{\mathbb{E}} \int_t^T |\widetilde{Z}_\tau|^2\, d\tau < \infty,$$

and $\widetilde{\mathbb{P}}$-a.s., $\{\widetilde{Y}_\tau,\ \tau \in [t,T]\}$ has continuous trajectories and the equality (2.5) holds.

From Lemma 2.3 it follows that

$$\widetilde{\mathbb{E}} \int_t^T |\psi(\sigma, \widetilde{X}_\sigma, 0)|^2\, d\sigma < \infty.$$

Moreover, for a suitable constant c, $\mathbb{P}$-a.s.,

$$|\psi(\sigma, \widetilde{X}_\sigma, z) - \psi(\sigma, \widetilde{X}_\sigma, z')| \le c|z - z'|, \qquad z, z' \in \mathbb{R}^d, \sigma \in [t,T].$$

Therefore, we conclude that there exists a solution $(\widetilde{Y}, \widetilde{Z})$ of (2.5) on the interval $[t,T]$, $\widetilde{Y}$ is unique up to indistinguishability, and $\widetilde{Z}$ is unique up to modification; a proof of this result can be found, for instance, in [18, Theorem 1.3] or [19, Theorem 2.2]. From this proof it also follows that the law of $(\widetilde{Y}, \widetilde{Z})$ is uniquely determined by the law of $(\widetilde{W}, \widetilde{X})$ and by ϕ and ψ. We note that $\widetilde{Y}_t$, being measurable with respect to the degenerate σ-algebra $\widetilde{\mathcal{F}}_{[t,t]}$, is deterministic; in particular, $\widetilde{Y}_t = \widetilde{\mathbb{E}}\widetilde{Y}_t$ depends only on the law of $\widetilde{Y}$, and thus it is a functional of t, x, f, g, ϕ, ψ.

To stress dependence on the parameters t and x, we will denote the solution of (2.4)–(2.5) by $\{(\widetilde{X}_\tau(t,x), \widetilde{Y}_\tau(t,x), \widetilde{Z}_\tau(t,x)),\ \tau \in [t,T]\}$. We set

$$J^*(t,x) = \widetilde{Y}_t(t,x).$$

The previous discussion shows that $J^*(t,x)$ is a number whose value is uniquely determined by t, x, f, g, ψ, and ϕ. The relevance of $J^*(t,x)$ to our control problem is explained in the following proposition.

PROPOSITION 2.5. *For every $t \in [0,T]$ and $x \in \mathbb{R}^n$ and for every admissible control $u_\cdot$, we have $J^*(t,x) \le J(t,x,u_\cdot)$.*

Proof. Let $\{X_\tau^u,\ \tau \in [t,T]\}$ be the solution to (1.1) corresponding to $u_\cdot$. We define the process

$$W_\tau^u = W_\tau + \int_{\tau\wedge t}^{\tau} r(\sigma, X_\sigma^u, u_\sigma)\, d\sigma, \qquad \tau \in [0,T], \tag{2.6}$$

and we note that X^u solves the equation

$$X_\tau^u = x + \int_t^\tau f(\sigma, X_\sigma^u)\, d\sigma + \int_t^\tau g(\sigma, X_\sigma^u)\, dW_\sigma^u, \qquad \tau \in [t,T]. \tag{2.7}$$

Since r is bounded, by the Girsanov theorem there exists a probability measure $\mathbb{P}^u$ on $(\Omega, \mathcal{F})$ such that W^u is a Wiener process under $\mathbb{P}^u$. Moreover, the density $\rho = d\mathbb{P}/d\mathbb{P}^u$ satisfies $\mathbb{E}^u\rho^p < \infty$ for every $p \in [1,\infty)$ (where $\mathbb{E}^u$ denotes expectation with respect to $\mathbb{P}^u$). Let us consider the backward equation for the unknown process $\{(Y_\tau^u, Z_\tau^u),\ \tau \in [t,T]\}$:

$$Y_\tau^u + \int_\tau^T Z_\sigma^u\, dW_\sigma^u = \phi(X_T^u) + \int_\tau^T \psi(\sigma, X_\sigma^u, Z_\sigma^u)\, d\sigma, \qquad \tau \in [t,T]. \tag{2.8}$$

By the result recalled earlier, there exists a unique solution (Y^u, Z^u) of this equation. Comparing (2.7)–(2.8) with (2.4)–(2.5), we conclude that $J^*(t,x) = Y_t^u$.

Next we write (2.8) with respect to W:

$$\begin{aligned} Y_\tau^u + \int_\tau^T Z_\sigma^u\, dW_\sigma + \int_\tau^T Z_\sigma^u r(\sigma, X_\sigma^u, u_\sigma)\, d\sigma \\ = \phi(X_T^u) + \int_\tau^T \psi(\sigma, X_\sigma^u, Z_\sigma^u)\, d\sigma, \qquad \tau \in [t,T]. \end{aligned} \tag{2.9}$$

Now we notice that

$$\begin{aligned} \mathbb{E}\left(\int_0^T |Z_\sigma^u|^2 d\sigma\right)^{1/2} &= \mathbb{E}^u\left[\rho\left(\int_0^T |Z_\sigma^u|^2 d\sigma\right)^{1/2}\right] \\ &\le (\mathbb{E}^u\,\rho^2)^{1/2}\left(\mathbb{E}^u \int_0^T |Z_\sigma^u|^2 d\sigma\right)^{1/2} < +\infty. \end{aligned}$$

Thus the stochastic integral in (2.9) has zero expectation with respect to the original probability $\mathbb{P}$. So, if we set $\tau = t$ in (2.9) and we take the expectation with respect to $\mathbb{P}$, we obtain

$$J^*(t,x) = Y_t^u = \mathbb{E}\phi(X_T^u) + \mathbb{E}\int_t^T [\psi(\sigma, X_\sigma^u, Z_\sigma^u) - Z_\sigma^u r(\sigma, X_\sigma^u, u_\sigma)]\, d\sigma.$$

Adding and subtracting $\mathbb{E}\int_t^T l(\sigma, X_\sigma^u, u_\sigma)\, d\sigma$, we arrive at 1:

$$J^*(t,x) = J(t,x,u_\cdot) + \int_t^T [\psi(\sigma, X_\sigma^u, Z_\sigma^u) - Z_\sigma^u r(\sigma, X_\sigma^u, u_\sigma) - l(\sigma, X_\sigma^u, u_\sigma)]\, d\sigma. \tag{2.10}$$

By the definition of ψ (formula (2.2)) the term in the square brackets is nonpositive, and consequently $J(t,x,u_\cdot) \ge J^*(t,x)$. ∎

We notice that relation (2.10) is a version of the *fundamental relation* in terms of backward stochastic differential equations. It immediately yields important consequences.

COROLLARY 2.6. *We fix $t \in [0,T]$, $x \in \mathbb{R}^n$.*

If, for an admissible control $u_\cdot$, we have $J(t,x,u_\cdot) = J^*(t,x)$, then $u_\cdot$ is optimal for the control problem starting from x at time t.

If an admissible control $u_\cdot$ verifies

$$u_\tau = \gamma(\tau, X^u_\tau, Z^u_\tau), \qquad \mathbb{P}\text{-a.s. for a.e. } \tau \in [t,T], \tag{2.11}$$

then $J(t,x,u_\cdot) = J^*(t,x)$ and $u_\cdot$ is optimal.

Conversely, suppose that the infimum in (2.2) is attained at a unique point and that an admissible control $u_\cdot$ satisfies $J(t,x,u_\cdot) = J^*(t,x)$. Then (2.11) holds and moreover we have, $\mathbb{P}$-a.s.,

$$\begin{cases} X^u_\tau = x + \displaystyle\int_t^\tau f(\sigma, X^u_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X^u_\sigma) r(\sigma, X^u_\sigma, \gamma(\sigma, X^u_\sigma, Z^u_\sigma))\, d\sigma + \int_t^\tau g(\sigma, X^u_\sigma)\, dW_\sigma, \\ Y^u_\tau + \displaystyle\int_\tau^T Z^u_\sigma\, dW_\sigma = \phi(X^u_T) + \int_\tau^T l(\sigma, X^u_\sigma, \gamma(\sigma, X^u_\sigma, Z^u_\sigma))\, d\sigma, \qquad \tau \in [t,T]. \end{cases} \tag{2.12}$$

Proof. The optimality of $u_\cdot$ when $J(t,x,u_\cdot) = J^*(t,x)$ follows immediately from Proposition 2.5. Moreover, if (2.11) holds, then $J(t,x,u_\cdot) = J^*(t,x)$ by (2.10).

For the converse, the equalities $J(t,x,u_\cdot) = J^*(t,x) = Y^u_t$ and (2.10) imply that

$$\int_t^T [\psi(\sigma, X^u_\sigma, Z^u_\sigma) - Z^u_\sigma r(\sigma, X^u_\sigma, u_\sigma) - l(\sigma, X^u_\sigma, u_\sigma)]\, d\sigma = 0, \qquad \mathbb{P}\text{-a.s.}$$

By our assumption on ψ (compare formula (2.2)) this is possible only if (2.11) holds.

Finally, writing (2.7)–(2.8) in terms of W and replacing u_τ by $\gamma(\tau, X^u_\tau, Z^u_\tau)$, we obtain

$$\begin{cases} X^u_\tau = x + \displaystyle\int_t^\tau f(\sigma, X^u_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X^u_\sigma) r(\sigma, X^u_\sigma, \gamma(\sigma, X^u_\sigma, Z^u_\sigma))\, d\sigma + \int_t^\tau g(\sigma, X^u_\sigma)\, dW_\sigma, \\ Y^u_\tau + \displaystyle\int_\tau^T Z^u_\sigma\, dW_\sigma + \int_\tau^T Z^u_\sigma r(\sigma, X^u_\sigma, \gamma(\sigma, X^u_\sigma, Z^u_\sigma))\, d\sigma = \phi(X^u_T) + \int_\tau^T \psi(\sigma, X^u_\sigma, Z^u_\sigma)\, d\sigma, \end{cases}$$

and recalling the identity (2.3), we get the final assertion of the corollary. □

In order to prove the existence of an optimal control, Corollary 2.6 suggests that it might be useful to investigate general properties of the following fully coupled forward-backward stochastic differential system:

$$\begin{cases} X_\tau = \xi + \displaystyle\int_t^\tau f(\sigma, X_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X_\sigma) r(\sigma, X_\sigma, \gamma(\sigma, X_\sigma, Z_\sigma))\, d\sigma + \int_t^\tau g(\sigma, X_\sigma)\, dW_\sigma, \\ Y_\tau + \displaystyle\int_\tau^T Z_\sigma\, dW_\sigma = \phi(X_T) + \int_\tau^T l(\sigma, X_\sigma, \gamma(\sigma, X_\sigma, Z_\sigma))\, d\sigma, \qquad \tau \in [t,T], \end{cases} \tag{2.13}$$

where t is fixed in $[0,T]$ and ξ is a given $\mathcal{F}_t$-measurable random variable in $\mathbb{R}^n$, satisfying $\mathbb{E}|\xi|^2 < \infty$. We denote by $\mathcal{F}^\xi_{[t,\tau]}$ the σ-algebra generated by ξ and $\{W_s - W_t : s \in [t,\tau]\}$ augmented by the class $\mathcal{N}$ of zero probability sets in $\mathcal{F}$. Notice that $\mathcal{F}_{[t,\tau]} \subset \mathcal{F}^\xi_{[t,\tau]} \subset \mathcal{F}_\tau$, $\tau \in [t,T]$. By a solution of (2.13) we mean an $(\mathcal{F}^\xi_{[t,\tau]})_{\tau\in[t,T]}$-predictable process $\{(X_\tau, Y_\tau, Z_\tau),\ \tau \in [t,T]\}$, with values in $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d$, such that

$$\mathbb{E} \sup_{\tau\in[t,T]} |X_\tau|^2 + \mathbb{E} \sup_{\tau\in[t,T]} |Y_\tau|^2 + \mathbb{E} \int_t^T |Z_\tau|^2\, d\tau < \infty$$

and, $\mathbb{P}$-a.s., X and Y have continuous trajectories and equalities (2.13) hold. With slight abuse of terminology to shorten some of the statements, we will say that (X, Y, Z) *is a solution on* $[t, T]$ *with boundary conditions* ξ *and* ϕ. In the following we say that such a solution is unique if the following holds: if (X', Y', Z') is another solution, then (X, Y) is indistinguishable from (X', Y') and Z is a modification of Z'.

When $\xi = x \in \mathbb{R}^n$ is deterministic, $\mathcal{F}^{\xi}_{[t,\tau]} = \mathcal{F}_{[t,\tau]}$. In this case Y_t is deterministic, since it is measurable with respect to the degenerate σ-algebra $\mathcal{F}_{[t,t]}$.

PROPOSITION 2.7. *Suppose that, for some* $(t, x) \in [0, T] \times \mathbb{R}^n$, *the forward-backward system* (2.13) *has a solution* (X, Y, Z) *on* $[t, T]$ *with boundary conditions* $\xi = x$ *and* ϕ. *Then setting* $u^*_\tau = \gamma(\tau, X_\tau, Z_\tau)$, $\tau \in [t, T]$, *the process* $u^*_\cdot$ *is optimal for the control problem starting from* x *at time* t *with optimal cost* $J(t, x, u^*_\cdot) = J^*(t, x) = Y_t$.

Proof. It is clear that $u^*_\cdot$ is admissible and that X is the corresponding solution (i.e., $X = X^{u^*}$). Moreover, by definition of γ the system (2.13) can be rewritten

$$\begin{cases} X_\tau = x + \displaystyle\int_t^\tau f(\sigma, X_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X_\sigma) r(\sigma, X_\sigma, u^*_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X_\sigma)\, dW_\sigma, \\ Y_\tau + \displaystyle\int_\tau^T Z_\sigma\, dW_\sigma = \phi(X_T) + \int_\tau^T [\psi(\sigma, X_\sigma, u^*_\sigma) - Z_\sigma r(\sigma, X_\sigma, u^*_\sigma)]\, d\sigma, \quad \tau \in [t, T]. \end{cases}$$

And letting W^{u^*} be defined as in (2.6), we obtain

$$\begin{cases} X_\tau = x + \displaystyle\int_t^\tau f(\sigma, X_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X_\sigma)\, dW^{u^*}_\sigma, \\ Y_\tau + \displaystyle\int_\tau^T Z_\sigma\, dW^{u^*}_\sigma = \phi(X_T) + \int_\tau^T \psi(\sigma, X_\sigma, u^*_\sigma) d\sigma, \qquad \tau \in [t, T]. \end{cases}$$

Comparing this with (2.7) and (2.8), we get $Y_\tau = Y^{u^*}_\tau$, $Z_\tau = Z^{u^*}_\tau$. Thus $Y_t = Y^{u^*}_t = J^*(t, x)$. Since by definition $u^*_\tau = \gamma(\tau, X^{u^*}_\tau, Z^{u^*}_\tau)$, the claim follows by Proposition 2.6. □

In the following section we will show that a solution of the forward-backward system actually exists, and this will lead to the desired existence of an optimal control in the strong formulation.

3. Main results. In this section we will assume that Hypotheses 2.1 and 2.4 hold.

For $t \in [0, T]$, $x \in \mathbb{R}^n$, $z \in \mathbb{R}^d$, we set for brevity

$$f_0(t, x, z) = f(t, x) + g(t, x) r(t, x, \gamma(t, x, z)), \qquad l_0(t, x, z) = l(t, x, \gamma(t, x, z)),$$

so that the forward-backward system (2.13) can be written

$$\begin{cases} X_\tau = \xi + \displaystyle\int_t^\tau f_0(\sigma, X_\sigma, Z_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X_\sigma)\, dW_\sigma, \\ Y_\tau + \displaystyle\int_\tau^T Z_\sigma\, dW_\sigma = \phi(X_T) + \int_\tau^T l_0(\sigma, X_\sigma, Z_\sigma)\, d\sigma, \qquad \tau \in [t, T], \end{cases}$$

and we note that there exists a constant c such that

$$\begin{aligned} &\langle f_0(t, x, z) - f_0(t, x', z), x - x' \rangle \le c|x - x'|^2, \\ &|f_0(t, x, z) - f_0(t, x, z')| \le c|z - z'|, \\ &|l_0(t, x, z) - l_0(t, x', z')| \le c(|x - x'| + |z - z'|), \\ &|f_0(t, x, z)| + |l_0(t, x, z)| \le c(1 + |x| + |z|). \end{aligned}$$

Further, let us denote by L a Lipschitz constant for ϕ:

$$|\phi(x) - \phi(x')| \le L|x - x'|, \qquad x, x' \in \mathbb{R}^n.$$

Concerning the solvability of the forward-backward system (2.13), we will use the following proposition; this is Theorem 1.1 in [4], which generalizes an earlier result by [1] (see also [17] and the references therein).

PROPOSITION 3.1. *There exists $\delta > 0$, depending only on f_0, l_0, g, and L, with the following property: for any $t \in [0,T]$ satisfying $t \ge T - \delta$ and any $\mathcal{F}_t$-measurable ξ satisfying $\mathbb{E}|\xi|^2 < \infty$ there exists a unique solution of* (2.13).

When $\xi = x$ we denote the solution by $\{(X_\tau(t,x), Y_\tau(t,x), Z_\tau(t,x)),\ \tau \in [t,T]\}$, to stress dependence on the parameters (t,x), and we recall that $Y_t(t,x)$ is deterministic. It can be shown that the solution can be chosen in such a way that, for fixed t, the mapping $(\omega,\tau,x) \to (X_\tau(t,x,\omega), Y_\tau(t,x,\omega), Z_\tau(t,x,\omega))$ is measurable with respect to $\mathcal{P}_{[t,T]} \times \mathcal{B}(\mathbb{R}^n)$, where $\mathcal{B}(\mathbb{R}^n)$ denotes the Borel sets of $\mathbb{R}^n$ and $\mathcal{P}_{[t,T]}$ denotes the predictable σ-algebra on $\Omega \times [t,T]$ with respect to the filtration $(\mathcal{F}_{[t,\tau]})_{\tau \in [t,T]}$. In the following we always assume that this measurability property holds.

The general solution of the system (2.13) can be represented in terms of this family of solutions; more precisely, given t and ξ as in Proposition 3.1, it is easy to verify that the solution on $[t,T]$ with boundary conditions ξ and ϕ is

$$X_\tau = X_\tau(t,\xi), \quad Y_\tau = Y_\tau(t,\xi), \quad Z_\tau = Z_\tau(t,\xi), \qquad \tau \in [t,T].$$

Remark 3.2. With the present notation, Proposition 2.7 states that if $t \ge T - \delta$, then there exists an optimal control for the control system starting from x at time t, with optimal cost $V(t,x) = Y_t(t,x)$.

Proposition 3.1 is a local existence and uniqueness result for solutions of (2.13). In order to obtain a global result we need to study extensions of solutions. This is the subject of the following lemma.

LEMMA 3.3. *Let $0 \le t < s < T$, and assume that for every $x \in \mathbb{R}^n$ there exists a solution $(X^1(x), Y^1(x), Z^1(x))$ on $[s,T]$ with boundary conditions x and ϕ such that the mapping $(\omega,\tau,x) \to (X^1_\tau(x,\omega), Y^1_\tau(x,\omega), Z^1_\tau(x,\omega))$ is measurable with respect to $\mathcal{P}_{[s,T]} \times \mathcal{B}(\mathbb{R}^n)$.*

Setting $V(x) = Y^1_s(x)$, $x \in \mathbb{R}^n$, assume that there exists a solution (X,Y,Z) on $[t,s]$ with boundary conditions ξ and V. Then defining, for $\tau \in (s,T]$,

$$X_\tau = X^1_\tau(X_s), \quad Y_\tau = Y^1_\tau(X_s), \quad Z_\tau = Z^1_\tau(X_s),$$

the process $\{(X_\tau, Y_\tau, Z_\tau),\ \tau \in [t,T]\}$ is a solution on $[t,T]$ with boundary conditions ξ and ϕ.

Proof. For $x \in \mathbb{R}^n$ and $\tau \in [s,T]$,

$$\begin{cases} X^1_\tau(x) = x + \displaystyle\int_t^\tau f_0(\sigma, X^1_\sigma(x), Z^1_\sigma(x))\, d\sigma + \int_t^\tau g(\sigma, X^1_\sigma(x))\, dW_\sigma, \\ Y^1_\tau(x) + \displaystyle\int_\tau^T Z^1_\sigma(x)\, dW_\sigma = \phi(X^1_T(x)) + \int_\tau^T l_0(\sigma, X^1_\sigma(x), Z^1_\sigma(x))\, d\sigma. \end{cases}$$

Since X_s is $\mathcal{F}_s$-measurable, we can replace x by X_s obtaining, for $\tau \in [s,T]$,

$$\begin{cases} X_\tau = X_s + \displaystyle\int_t^\tau f_0(\sigma, X_\sigma, Z_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X_\sigma)\, dW_\sigma, \\ Y_\tau + \displaystyle\int_\tau^T Z_\sigma\, dW_\sigma = \phi(X_T) + \int_\tau^T l_0(\sigma, X_\sigma, Z_\sigma)\, d\sigma. \end{cases} \tag{3.1}$$

By definition, for $\tau \in [t,s]$,

$$\begin{cases} X_\tau = \xi + \displaystyle\int_t^\tau f_0(\sigma, X_\sigma, Z_\sigma)\, d\sigma + \int_t^\tau g(\sigma, X_\sigma)\, dW_\sigma, \\ Y_\tau + \displaystyle\int_\tau^s Z_\sigma\, dW_\sigma = V(X_s) + \int_\tau^s l_0(\sigma, X_\sigma, Z_\sigma)\, d\sigma, \end{cases} \tag{3.2}$$

so that, in particular, $Y_s = V(X_s)$. Now the required result follows easily from (3.1) and (3.2). □

We are now ready to state one of our main results.

THEOREM 3.4. *Assume that the Hypotheses* 2.1 *and* 2.4 *hold. For every* $t \in [0,T]$ *and every* $\mathcal{F}_t$*-measurable* ξ *with* $\mathbb{E}|\xi|^2 < \infty$ *there exists a unique solution of the forward-backward system* (2.13) *on* $[t,T]$ *with boundary conditions* ξ *and* ϕ.

Proof. By Lemma 2.2 and our assumptions, there exists a constant L such that

$$|V(t,x) - V(t,x')| \le L|x - x'|, \tag{3.3}$$

and in particular $|\phi(x) - \phi(x')| \le L|x - x'|$ for every $t \in [0,T]$, $x, x' \in \mathbb{R}^n$. We choose δ as in Proposition 3.1.

Proof of existence. Step 1. If $t \ge T - \delta$, then by Proposition 3.1 the required solution exists.

In particular, for every $x \in \mathbb{R}^n$, the process $\{(X_\tau(T-\delta,x), Y_\tau(T-\delta,x), Z_\tau(T-\delta,x)),\ \tau \in [T-\delta,T]\}$, introduced above, is well defined. By Proposition 2.7, or by Remark 3.2, we have $Y_{T-\delta}(T-\delta,x) = V(T-\delta,x)$.

Step 2. Assume now that $T - 2\delta \le t < T - \delta$. We note that (3.3) holds for $t = T - \delta$ and consequently, by Proposition 3.1, there exists a solution (X,Y,Z) on $[t, T-\delta]$ with boundary conditions ξ and $V(T-\delta,\cdot)$. By Lemma 3.3, if we define, for $\tau \in (T-\delta, T]$,

$$X_\tau = X_\tau(T-\delta, X_{T-\delta}), \quad Y_\tau = Y_\tau(T-\delta, X_{T-\delta}), \quad Z_\tau = Z_\tau(T-\delta, X_{T-\delta}),$$

then the process $\{(X_\tau, Y_\tau, Z_\tau),\ \tau \in [t,T]\}$ is a solution on $[t,T]$ with boundary conditions ξ and ϕ.

In particular, choosing $t = T - 2\delta$ and $\xi = x \in \mathbb{R}^n$, the constructed solution is denoted $\{(X_\tau(T-2\delta,x), Y_\tau(T-2\delta,x), Z_\tau(T-2\delta,x)),\ \tau \in [T-2\delta,T]\}$, in agreement with the previous notation. By Proposition 2.7 or by Remark 3.2, we have $Y_{T-2\delta}(T-2\delta,x) = V(T-2\delta,x)$. □

Conclusion. If $T - 3\delta \le t < T - 2\delta$, we can repeat the above construction and, after a finite number of steps, we obtain the required solution for arbitrary $t \in [0,T]$.

Proof of uniqueness. If $t \ge T - \delta$, then the solution is unique by Proposition 3.1.

Assume now $T - 2\delta \le t < T - \delta$ and let (X,Y,Z) and (X',Y',Z') be solutions on $[t,T]$ with boundary conditions ξ and ϕ.

For every $x \in \mathbb{R}^n$, we consider again the process $\{(X_\tau(T-\delta,x), Y_\tau(T-\delta,x), Z_\tau(T-\delta,x)),\ \tau \in [T-\delta,T]\}$, introduced above, and we define, for $\tau \in [T-\delta,T]$,

$$\widetilde{X}_\tau = X_\tau(T-\delta, X_{T-\delta}), \quad \widetilde{Y}_\tau = Y_\tau(T-\delta, X_{T-\delta}), \quad \widetilde{Z}_\tau = Z_\tau(T-\delta, X_{T-\delta}).$$

It is easy to see that $(\widetilde{X}, \widetilde{Y}, \widetilde{Z})$ is a solution on $[T-\delta,T]$ with boundary conditions $X_{T-\delta}$ and ϕ. Since we have already proved uniqueness on the interval $[T-\delta,T]$, it follows that it is a modification of $\{(X_\tau, Y_\tau, Z_\tau),\ \tau \in [T-\delta,T]\}$, and, in particular, $Y_{T-\delta} = \widetilde{Y}_{T-\delta}$. Since by Proposition 2.7 we have $Y_{T-\delta}(T-\delta,x) = V(T-\delta,x)$, we conclude that $Y_{T-\delta} = V(T-\delta, X_{T-\delta})$. It follows easily that the process $\{(X_\tau, Y_\tau, Z_\tau),\ \tau \in$

$[t, T-\delta]\}$ is a solution on $[t, T-\delta]$ with boundary conditions ξ and $V(T-\delta, \cdot)$. Repeating this argument, we also conclude that the process $\{(X'_\tau, Y'_\tau, Z'_\tau),\ \tau \in [t, T-\delta]\}$ is also solution on $[t, T-\delta]$ with the same boundary conditions ξ and $V(T-\delta, \cdot)$. We note that (3.3) holds for $t = T-\delta$ and, consequently, by Proposition 3.1, the two solutions coincide on $[t, T-\delta]$. Since $\{(X_\tau, Y_\tau, Z_\tau),\ \tau \in [T-\delta, T]\}$ and $\{(X'_\tau, Y'_\tau, Z'_\tau),\ \tau \in [T-\delta, T]\}$ are clearly solutions on $[T-\delta, T]$ with the same boundary conditions $X_{T-\delta} = X'_{T-\delta}$ and ϕ, and since uniqueness holds on $[T-\delta, T]$, it follows that the two solutions also coincide on this interval and that uniqueness holds on $[t, T]$.

If $T - 3\delta \le t < T - 2\delta$, we can repeat the above argument and, after a finite number of steps, we obtain the required uniqueness property for arbitrary $t \in [0, T]$.

In the following, for every $(t, x) \in [0, T] \times \mathbb{R}^n$, we denote by $\{(X_\tau(t, x), Y_\tau(t, x), Z_\tau(t, x)),\ \tau \in [t, T]\}$ the (global) solution on $[t, T]$ with boundary conditions x and ϕ, constructed in Theorem 3.4. It follows from the construction that for fixed t, the mapping $(\omega, \tau, x) \to (X_\tau(t, x, \omega),\ Y_\tau(t, x, \omega),\ Z_\tau(t, x, \omega))$ is measurable with respect to $\mathcal{P}_{[t,T]} \times \mathcal{B}(\mathbb{R}^n)$.

Now we consider the implications of Theorem 3.4 on the existence of optimal controls. The following theorem is an immediate consequence of Theorem 3.4 and Proposition 2.7.

THEOREM 3.5. *Assume that the Hypotheses* 2.1 *and* 2.4 *hold and, for every* $t \in [0, T]$ *and* $x \in \mathbb{R}^n$, *let us denote by* $\{(X_\tau(t, x), Y_\tau(t, x), Z_\tau(t, x)),\ \tau \in [t, T]\}$ *the solution of the forward-backward system* (2.13) *on* $[t, T]$ *with boundary conditions* x *and* ϕ. *Setting*

$$u^*_\tau = \gamma(\tau, X_\tau(t, x), Z_\tau(t, x)), \qquad \tau \in [t, T], \tag{3.4}$$

the process $u^*_\cdot$ *is then an optimal control for the control problem starting from* x *at time* t. *Denoting by* X^{u^*} *the corresponding solution, then,* $\mathbb{P}$*-a.s.,* $X^{u^*}_\tau = X_\tau(t, x)$ *for* $\tau \in [t, T]$. *Finally, the optimal cost* $V(t, x) = J(t, x, u^*_\cdot)$ *is equal to* $J^*(t, x)$ *and to* $Y_t(t, x)$.

We can now state a uniqueness result.

THEOREM 3.6. *Assume that the Hypotheses* 2.1 *and* 2.4 *hold and suppose that, for any given* $t \in [0, T]$, $x \in \mathbb{R}^n$, $z \in \mathbb{R}^d$, *the infimum in* (2.2) *is attained at a unique point.*

Then, for any $x \in \mathbb{R}^n$ *and* $t \in [0, T]$, *there exists a unique optimal control, up to modification.*

Proof. Let $u_\cdot$ be an optimal control for the control system starting from x at time t. By Theorem 3.5 the optimal cost is $J^*(t, x) = J(t, x, u_\cdot)$ and therefore, by Corollary 2.6, we have, $\mathbb{P}$-a.s.,

$$u_\tau = \gamma(\tau, X^u_\tau, Z^u_\tau) \qquad \text{for a.e. } \tau \in [t, T],$$

where X^u is the solution of (1.1) corresponding to $u_\cdot$, and (Y^u, Z^u) denotes the solution of (2.8). Corollary 2.6 also states that (X^u, Y^u, Z^u) satisfies the systems (2.12), and consequently (X^u, Y^u, Z^u) is a solution of the forward-backward system (2.13) on $[t, T]$ with boundary conditions x and ϕ. By uniqueness it must coincide with $\{(X_\tau(t, x), Y_\tau(t, x),\ Z_\tau(t, x)),\ \tau \in [t, T]\}$. Thus

$$u_\tau = \gamma(\tau, X_\tau(t, x), Z_\tau(t, x)), \qquad \mathbb{P}\text{-a.s., for a.e. } \tau \in [t, T],$$

and therefore $u_\cdot$ is a modification of the optimal control $u^*_\cdot$ introduced in Theorem 3.5. ∎

Remark 3.7. It follows from Lemma 3.3 and from uniqueness of the forward-backward system (2.13) that for $x \in \mathbb{R}^n$ and $0 \le t \le s \le \tau \le T$ we have, $\mathbb{P}$-a.s.,

$$Y_\tau(t,x) = Y_\tau(s, X_s(t,x)). \tag{3.5}$$

Letting u^* be defined as in (3.4) and recalling that $X_\tau^{u^*} = X_\tau(t,x)$ the second equation in (2.13) gives

$$\mathbb{E}Y_\tau(t,x) = Y_t(t,x) - \mathbb{E}\int_t^\tau l(\sigma, X_\sigma^{u^*}, u_\sigma^*)d\sigma.$$

By Proposition 2.7 we have $Y_\tau(\tau,x) = V(\tau,x)$ and setting $s=\tau$ in (3.5) we obtain

$$V(t,x) = \mathbb{E}\left[V(\tau, X_\tau^{u^*}) + \int_t^\tau l(\sigma, X_\sigma^{u^*}, u_\sigma^*)d\sigma\right].$$

By standard techniques (see [7] or [23]) we have that for an arbitrary admissible control u

$$V(t,x) \le \mathbb{E}\left[V(\tau, X_\tau^u) + \int_t^\tau l(\sigma, X_\sigma^u, u_\sigma)d\sigma\right].$$

So we immediately obtain the usual dynamic programming principle:

$$\begin{aligned} V(t,x) &= \min_u \mathbb{E}\left[V(\tau, X_\tau^u) + \int_t^\tau l(\sigma, X_\sigma^u, u_\sigma)d\sigma\right] \\ &= \mathbb{E}\left[V(\tau, X_\tau^{u^*}) + \int_t^\tau l(\sigma, X_\sigma^{u^*}, u_\sigma)d\sigma\right], \end{aligned}$$

where the minimum is over all admissible controls.

4. The feedback law and the closed-loop equation. Throughout this section, Hypotheses 2.1 and 2.4 are still in force.

The following lemma introduces a function ζ which will enter the definition of the optimal feedback law. In its formulation we refer to the process $\{(\widetilde{X}_\tau(t,x), \widetilde{Y}_\tau(t,x), \widetilde{Z}_\tau(t,x)),\ \tau \in [t,T]\}$, solution of (2.4)–(2.5) for given $t \in [0,T]$ and $x \in \mathbb{R}^n$.

LEMMA 4.1. *There exists a Borel measurable function $\zeta : [0,T] \times \mathbb{R}^n \to \mathbb{R}^d$ such that, for every $t \in [0,T]$, $x \in \mathbb{R}^n$, $\mathbb{P}$-a.s.,*

$$\zeta(\tau, \widetilde{X}_\tau(t,x)) = \widetilde{Z}_\tau(t,x) \qquad \text{for a.e. } \tau \in [t,T].$$

ζ depends only on the functions f, g, ϕ, ψ (not on the particular probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ nor on the Wiener process $\widetilde{W}$).

Proof. This is a kind of result which is familiar in the theory of backward stochastic differential equations (see, e.g., [2, Theorem 28]) so we will only sketch the proof. We first note that, as a consequence of uniqueness for the system (2.4)–(2.5), for $x \in \mathbb{R}^n$ and $0 \le t \le s \le \tau \le T$ we have, $\mathbb{P}$-a.s.,

$$\widetilde{X}_\tau(t,x) = \widetilde{X}_\tau(s, \widetilde{X}_s(t,x)), \qquad \widetilde{Y}_\tau(t,x) = \widetilde{Y}_\tau(s, \widetilde{X}_s(t,x)). \tag{4.1}$$

In particular, the family $\widetilde{X} = \{\widetilde{X}_\tau(t,x),\ 0 \le t \le \tau \le T,\ x \in \mathbb{R}^n\}$ is a Markov process. Let us denote by $\widetilde{W}^i$ and $\widetilde{Z}^i$ the ith components of $\widetilde{W}$ and $\widetilde{Z}$. For $0 \le t \le T$ and $x \in \mathbb{R}^n$ we denote by $\{A^i(\tau,t,x), \tau \in [t,T]\}$ the joint quadratic variation process

between $\{\widetilde{Y}_\tau(t,x), \tau \in [t,T]\}$ and $\{\widetilde{W}^i_\tau, \tau \in [t,T]\}$. It follows from the backward equation in the system (2.13) that

$$A^i(\tau,t,x) = \int_t^\tau \widetilde{Z}^i_s(t,x)\, ds,$$

which implies in particular that $A^i(\cdot,t,x)$ is absolutely continuous with respect to the Lebesgue measure. Next we note the following equality, which is a consequence of the definition of A^i and (4.1): for $x \in \mathbb{R}^n$ and $0 \le t \le s \le \tau \le T$ we have, $\mathbb{P}$-a.s.,

$$A^i(\tau,t,x) = A^i(s,t,x) + A^i(\tau,s,X_s(t,x)).$$

If $\widetilde{X}$ is a Markov process homogeneous in time, then the above relation means that A^i is an additive functional of $\widetilde{X}$, absolutely continuous with respect to the Lebesgue measure, and the existence of ζ is classical (see, e.g., Theorem 66.2 in [22]). In the general case the proof is similar. □

COROLLARY 4.2. *Let $\{(X_\tau(t,x), Y_\tau(t,x), Z_\tau(t,x)), \tau \in [t,T]\}$ denote the solution of the forward-backward system* (2.13) *on $[t,T]$ with boundary conditions x and ϕ. Then, $\mathbb{P}$-a.s.,*

$$\zeta(\tau, X_\tau(t,x)) = Z_\tau(t,x) \qquad \text{for a.e. } \tau \in [t,T].$$

Proof. During the proof of Proposition 2.7 we showed that (X,Y,Z) solves (2.7)–(2.8), which coincides with (2.4)–(2.5) (with different, suitable probability measure and Wiener process). The claim then follows from Lemma 4.1. □

We define a Borel measurable function $\underline{u} : [0,T] \times \mathbb{R}^n \to \mathcal{U}$ setting

$$\underline{u}(t,x) = \gamma(t,x,\zeta(t,x)), \qquad t \in [0,T],\ x \in \mathbb{R}^n,$$

and, for $t \in [0,T]$ and $x \in \mathbb{R}^n$, we introduce the so-called *closed-loop* equation:

$$\begin{aligned} \overline{X}_\tau = \ & x + \int_t^\tau f(\sigma, \overline{X}_\sigma)\, d\sigma + \int_t^\tau g(\sigma, \overline{X}_\sigma) r(\sigma, \overline{X}_\sigma, \underline{u}(\sigma, \overline{X}_\sigma))\, d\sigma \\ & + \int_t^\tau g(\sigma, \overline{X}_\sigma)\, dW_\sigma, \qquad \tau \in [t,T]. \end{aligned} \tag{4.2}$$

Since r is bounded, it is meaningful to look for a continuous $(\mathcal{F}_t)$-adapted solution of this equation. However, it is difficult to prove directly any existence or uniqueness result for (4.2) in its strong form, since regularity properties of ζ, and hence of $\underline{u}$, are not immediate. Nevertheless, we have the following result.

THEOREM 4.3. *Assume that the Hypotheses* 2.1 *and* 2.4 *hold. For every $t \in [0,T]$ and $x \in \mathbb{R}^n$, let $u^*_\cdot$ be the optimal control introduced in* (3.4). *Then $u^*_\cdot$ is related to the corresponding solution X^{u^*} by the* feedback relation

$$u^*_\tau = \underline{u}(\tau, X^{u^*}_\tau), \qquad \mathbb{P}\text{-a.s. for a.e. } \tau \in [t,T].$$

In particular, X^{u^*} solves the closed-loop equation.

Proof. We have, $\mathbb{P}$-a.s., for almost every $\tau \in [t,T]$,

$$u^*_\tau = \gamma(\tau, X_\tau(t,x), Z_\tau(t,x)) = \gamma(\tau, X_\tau(t,x), \zeta(\tau, X_\tau(t,x))) = \underline{u}(\tau, X_\tau(t,x)).$$

Indeed, the first and third equalities follow from the definition of $u^*_\cdot$ and $\underline{u}$, and the second equality follows from Corollary 4.2. Since Theorem 3.5 states that $\{X_\tau(t,x),\ \tau \in [t,T]\}$ coincides with X^{u^*}, the result follows immediately. □

THEOREM 4.4. *Assume that the Hypotheses* 2.1 *and* 2.4 *hold. For given* $t \in [0,T]$ *and* $x \in \mathbb{R}^n$*, assume that there exists a continuous,* $(\mathcal{F}_t)$*-adapted solution* $\overline{X}$ *of the closed-loop equation. Then the process defined by*

$$\overline{u}_\tau = \underline{u}(\tau, \overline{X}_\tau), \qquad \tau \in [t,T], \tag{4.3}$$

is an optimal control, and $\overline{X}$ *is the solution corresponding to* $\overline{u}_\cdot$*.*

Proof. Let $\overline{u}$ be given by (4.3) and let $X^{\overline{u}}$, $W^{\overline{u}}$, $Y^{\overline{u}}$, $Z^{\overline{u}}$ be defined as in (1.1), (2.6), (2.7)–(2.8), respectively. Then by definition $X^{\overline{u}}_\tau = \overline{X}_\tau$. Moreover, by Lemma 4.1, $Z^{\overline{u}}_\tau = \zeta(\tau, \overline{X}_\tau)$ for almost every $\tau \in [t,T]$. Joining the above relations with the definition of the map $\underline{u}$, we get $\overline{u}_\tau = \gamma(\tau, X^{\overline{u}}_\tau, Z^{\overline{u}}_\tau)$ and the claim follows by Corollary 2.6. □

THEOREM 4.5. *Assume that the Hypotheses* 2.1 *and* 2.4 *hold and suppose that, for any given* $t \in [0,T]$*,* $x \in \mathbb{R}^n$*,* $z \in \mathbb{R}^d$*, the infimum in* (2.2) *is attained at a unique point.*

Then, for every $t \in [0,T]$ *and* $x \in \mathbb{R}^n$*, there exists a unique continuous* $(\mathcal{F}_t)$*-adapted solution* $\overline{X}$ *of the closed-loop equation* (4.2)*. The process defined by*

$$\overline{u}_\tau = \underline{u}(\tau, \overline{X}_\tau), \qquad \tau \in [t,T],$$

is the unique optimal control and $\overline{X}$ *is the solution corresponding to* $\overline{u}_\cdot$*.*

Proof. Assume that $\overline{X}$ is a solution of the closed-loop equation. Then, by Theorem 4.4, the control $\overline{u}_\cdot$ is optimal, and $\overline{X} = X^{\overline{u}}$ is the solution of (1.1) corresponding to $\overline{u}_\cdot$. Since, by Theorem 3.6, under our assumptions the optimal control is unique, $\overline{u}_\cdot$ must coincide with the (optimal) control $u^*_\cdot$ introduced in (3.4), and consequently $\overline{X}$ must coincide with X^{u^*}.

The existence of a solution of the closed-loop equation follows from Theorem 4.3. □

REFERENCES

[1] F. ANTONELLI, *Backward-forward stochastic differential equations*, Ann. Appl. Probab., 3 (1993), pp. 777–793.

[2] V. BALLY, E. PARDOUX, AND L. STOICA, *Backward Stochastic Differential Equations Associated to a Symmetric Markov Process*, Tech. rep. 4454, INRIA Rocquencourt, Le Chesnay, France, 2002.

[3] S. CERRAI, *Second Order PDE's in Finite and Infinite Dimensions. A Probabilistic Approach*, Lecture Notes in Math. 1762, Springer-Verlag, Berlin, 2001.

[4] F. DELARUE, *On the existence and uniqueness of solutions to FBSDEs in a non-degenerate case*, Stochastic Process. Appl., 99 (2002), pp. 209–286.

[5] N. EL KAROUI, D. HU NGUYEN, AND M. JEANBLANC-PIQUÉ, *Compactification methods in the control of degenerate diffusion: Existence of an optimal control*, Stochastics, 20 (1987), pp. 169–219.

[6] N. EL KAROUI, S. PENG, AND M. C. QUENEZ, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.

[7] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Appl. Math. 25, Springer-Verlag, New York, 1993.

[8] U. G. HAUSSMANN AND W. SUO, *Existence of singular optimal control laws for stochastic differential equations*, Stochastics Stochastics Rep., 48 (1994), pp. 249–272.

[9] S. HAMADÈNE, *Backward forward SDE's and stochastic differential games*, Stochastic Process. Appl., 77 (1998), pp. 1–15.

[10] S. HAMADÈNE AND J.-P. LEPELTIER, *Zero sum differential games and backward equations*, Systems Control Lett., 24 (1995), pp. 259–263.

[11] S. HAMADÈNE, J.-P. LEPELTIER, AND S. PENG, *BSDEs with continuous coefficients and stochastic differential games*, in Backward Stochastic Differential Equations, Pitman Res. Notes Math. Ser. 364, N. El Karoui, L. Mazliak, eds., Longman, Harlow, UK, 1997, pp. 115–128.

[12] U. G. Haussmann and J.-P. Lepeltier, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.

[13] Y. Hu and S. Peng, *Solution of forward-backward stochastic differential equations*, Probab. Theory Related Fields, 103 (1995), pp. 273–283.

[14] Y. Hu and J. Yong, *Forward-backward stochastic differential equations with nonsmooth coefficients*, Stochastic Process. Appl., 87 (2000), pp. 93–106.

[15] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-Linear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[16] J. Ma, P. Protter, and J. Yong, *Solving forward-backward stochastic differential equations explicitly—a four step scheme*, Probab. Theory Related Fields, 98 (1994), pp. 339–359.

[17] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Math. 1702, Springer-Verlag, Berlin, 1999.

[18] E. Pardoux, *Backward stochastic differential equations and viscosity solutions of systems of semilinear parabolic and elliptic PDEs of second order*, in Stochastic Analysis and Related Topics, the Geilo Workshop 1996, Progr. Probab. 42, L. Decreusefond, J. Gjerde, B. Øksendal, and A. S. Üstünel, eds., Birkhäuser Boston, Boston, 1998, pp. 79–127.

[19] E. Pardoux, *BSDE's, weak convergence and homogeneization of semilinear PDE's*, in Nonlinear Analysis, Differential Equations and Control, F. H. Clarke and R. J. Stern, eds., Kluwer, Dordrecht, The Netherlands, 1999, pp. 503–549.

[20] E. Pardoux and S. Tang, *Forward-backward stochastic differential equations and quasilinear PDEs*, Probab. Theory Related Fields, 114 (1999), pp. 123–150.

[21] S. Peng and Z. Wu, *Fully coupled forward-backward stochastic differential equations and applications to optimal control*, SIAM J. Control Optim., 37 (1999), pp. 825–843.

[22] M. Sharpe, *General Theory of Markov Processes*, Pure Appl. Math. 133, Academic Press, Boston, 1988.

[23] J. Yong and X. Y. Zhou, *Stochastic Controls*, Appl. Math. 43, Springer-Verlag, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 831–855

NONSQUARE SPECTRAL FACTORS OF NONLINEAR CONTROL SYSTEMS IN TERMS OF INNER-INNER FACTORIZATIONS*

MARK A. PETERSEN†

Abstract. This paper considers nonsquare spectral factors of nonlinear input affine state space control systems in continuous time. More specifically, we obtain a parametrization of nonsquare spectral factors in terms of a special class of inner-inner factorizations. Explicit formulas for different classes of spectral factors and related inner systems in terms of solutions of Hamilton–Jacobi equations are provided. Furthermore, the important notion of a minimal inner embedding for nonlinear control systems is introduced and used to derive several relationships between inner systems.

Key words. spectral factorization, nonlinear control systems, inner systems

AMS subject classifications. Primary, 47A68, 93C10; Secondary, 34A34

DOI. 10.1137/S0363012902409568

1. Introduction and preliminaries. The problem of finding spectral factors for nonlinear control systems has attracted some recent attention (see [1, 2, 3, 4, 5, 6, 14, 20, 27, 32]). Our main aim in this paper is to treat the (nonsquare) spectral factorization problem for nonlinear control systems that are one-sided invertible. In particular, we establish a parametrization of nonsquare minimal spectral factors that are stable in terms of a special class of inner-inner factorizations of inner systems (see [3, 20, 27, 32] for the nonlinear case, see [11, 12, 13, 21, 22] for the linear case).

Our motivation for understanding the spectral factorization problem referred to above is that it has important applications in systems and control theory. For instance, several important connections with the control of mechanical systems that may not be invertible have been identified (see [17] and [32]). In this regard, it is important to note that our paper discusses classes of nonlinear control systems that are one-sided invertible. Furthermore, spectral factorization of nonlinear control systems plays an important part in H_∞-control (see [5]) and chemical process control (see [10] and [35]). Also, in economics, the understanding of realizations of financial models with multidimensional state spaces can be simplified by considering the dynamics of financial systems of reduced state space dimension and certain disturbance errors (see [24] and [28]). In the above, we recall that spectral factorization is related to inner-outer factorization (as will be demonstrated below) with the inner factor often sharing asymptotic behavior with the system that is being factored. In many cases, the inner factor may have more favorable properties and its control may form a basis for the control of the system being factored. Another reason for considering the spectral factorization problem for nonlinear control systems is its relation with stochastic control via stochastic realization theory, although this connection may currently not be as apparent as for the linear case. In the problem of linear stochastic realization, any minimal factor W of the spectral density Δ determines a minimal realization for the output process y. If Δ is of size $m \times m$ and W is of size $m \times p$, then the number p actually determines the space in which the driving white noise process is taken. Usually, if one chooses W to be square, this space is also taken to be as small as possible.

*Received by the editors June 14, 2002; accepted for publication (in revised form) January 4, 2004; published electronically September 18, 2004.
http://www.siam.org/journals/sicon/43-3/40956.html
†Department of Mathematics and Applied Mathematics, North-West University, Potchefstroom Campus, Potchefstroom x6001, South Africa (wskmap@puk.ac.za).

However, the stochastic realization problem does not actually require this, and hence nonsquare minimal spectral factors of Δ are of interest. A further practical reason for understanding the nonsquare spectral factorization problem is multichannel signal transmission. For instance, a situation that arises in the mobile phone industry is that a mobile phone has to decode signals that reach it from surrounding transmitters that themselves have to retain the capacity to emit signals destined for other phones.

Next, we make clear the additional value of the current paper when compared with recent research on the subject of spectral factorization of nonlinear control systems. In the first instance, this paper extends results obtained for nonlinear, minimal, square, stable spectral factors by Ball and Petersen in [3] and Ball and van der Schaft in [6] to the nonsquare case. In [3] we obtained a natural extension of the idea of an equivalence between minimal square spectral factors of a given spectral density, invariant subspaces of an associated Hamiltonian matrix, solutions of an appropriate algebraic Riccati equation, and minimal unitary left divisors of a certain unitary function on the imaginary line to a nonlinear setting. More specifically, we established a bijective correspondence between nonlinear, minimal, square, stable spectral factors, invariant Lagrangian submanifolds, solutions of Hamilton–Jacobi equations, and minimal right inner divisors. The scheme in the current paper is more complicated since it treats the situation where the nonlinear, minimal, stable spectral factors may be nonsquare. In contrast to [3], the work in this paper entails studying a slightly different type of Hamilton–Jacobi equation and a parametrization of the spectral factors in terms of a finer inner-inner factorization that, in addition, involves coprime factors. [6] is different from our paper in that the other work establishes a relationship between the inner-outer and spectral factorization problems and makes no attempt to discuss their connection with inner-inner factorizations. In the aforementioned contribution by Ball and van der Schaft, a nonlinear state space system is expressed as the cascade connection of an inner (lossless) system and a stable minimum phase (outer) system that is found to be a solution of an associated nonlinear spectral factorization problem. To the best of our knowledge, the first attempt to understand nonlinear, nonsquare, stable, spectral factors was made in [25] and the related conference paper [26]. In these contributions we obtain a parametrization of nonsquare spectral factors in terms of an invariant Lagrangian submanifold and associated solutions of a Hamilton–Jacobi inequality that is a nonlinear analogue of the bounded real lemma and the control algebraic Riccati inequality. By way of an application, we discussed an alternative characterization of minimum and maximum phase spectral factors and introduced the notion of a rigid nonlinear system. Our current paper makes use of some of this information to establish an equivalence between nonsquare spectral factors and coprime inner-inner factorizations. It is important to note that in the present paper we make use of a more general nonsquare spectral factor and associated Hamilton–Jacobi equation than in [25]. This is done in order to derive explicit formulas for the spectral factors comparable to those that are determined in the linear case. Moreover, the results established in this paper may be regarded as natural nonlinear analogues of those obtained in [11, 12, 13, 16, 21, 22, 23]. In particular, the explicit formulas for the spectral factors derived here bear a resemblance to those in [13] and [22], where the formulas are written in terms of solutions of algebraic Riccati equations.

We provide a brief description of the class of nonlinear control systems that we study. We consider a smooth nonlinear input-affine system

$$\Sigma\colon \begin{cases} \dot{x} = a(x) + b(x)u, & u \in \mathbf{R}^{\mathbf{m}}, \\ y = c(x) + d(x)u, & y \in \mathbf{R}^{\mathbf{p}}, \end{cases} \tag{1}$$

where $a : \mathbf{R^n} \to \mathbf{R^n}, b : \mathbf{R^n} \to \mathbf{R^{n\times m}}, c : \mathbf{R^n} \to \mathbf{R^p}$, and $d : \mathbf{R^n} \to \mathbf{R^{p\times m}}$ are smooth functions. The fact that Σ is assumed to be input affine will result in explicit formulas that resemble those obtained in the linear case. We suppose that $p \geq m$ and that $d(x)$ and $d(x)^T$ are injective for all $x = (x_1, \ldots, x_n) \in \mathbf{R^n}$ that are local coordinates for the n-dimensional state space manifold $\mathcal{X}$, with globally asymptotically stable equilibrium $x_0 = 0$ for $u = 0$ (so $a(x_0) = 0$ and $c(x_0) = 0$). In this case, we have that

$$F(x) := d(x)d(x)^T \tag{2}$$

is invertible for each x. We suppose that $x \to a(x) + b(x)u$ is *complete* for each $u \in \mathbf{R^m}$. This means that there exists a unique solution of $\dot{x} = a(x) + b(x)u$ for all $t > 0$ for any initial condition $x(0) = x_1 \in \mathcal{X}$. Hence, given any initial condition $x_1 \in \mathcal{X}$, the system of equations (1) induces a well-defined causal input-output map $T_\Sigma^{x_1}$, from piecewise constant input signals $u(t)$ defined on the nonnegative real line $\mathbf{R}^+$ to smooth output functions $y(t)$ defined for all $t \geq 0$. Operators $T_\Sigma^{x_1}$ arising in this way from a system of equations (1) are automatically *causal*, i.e., $T_\Sigma^{x_1}[u](t) = T_\Sigma^{x_1}[u'](t)$ for $t \leq T$ whenever $u(t) = u'(t)$ for $t \leq T$. We assume that $T_\Sigma^{x_1}$ extends by continuity to define a well-defined map on $L^m_{2,e}[0,\infty)$ (the space of measurable $\mathbf{R^m}$-valued functions $u(\cdot)$ on $[0,\infty)$ such that $\int_0^T \|u(t)\|^2 dt < \infty$ for all $T < \infty$). We consider only systems Σ that are *stable*. In other words, at the systems level, the vector field $x \to a(x)$ is globally asymptotically stable while $T_\Sigma^{x_1}$ is a diffeomorphism of $L_2^m[0,\infty)$ for each $x_1 \in \mathcal{X}$ at the input-output level. We call the system of equations (1) a (state space) *realization* of the input-output operator $T_\Sigma\colon L^m_{2,e}[0,\infty) \to L^p_{2,e}[0,\infty)$. If the dimension n of the state manifold $\mathcal{X}$ is as small as possible among all possible state space realizations of the given input-output operator $T_\Sigma^{x_1}$, we say that the realization Σ is *minimal*. Minimality with regard to realization and factorization plays a indispensable part at various levels in the ensuing discussion. Although the existence of realizations of input-output maps is important, the uniqueness of such realizations may also be required. We know that the uniqueness of any realization of a given input-output map is not always guaranteed. However, if we consider the minimal case, uniqueness of the realizations may become a definite possibility (see, for instance, Propositions 1 and 4). In addition, the major objective of our paper is to characterize *minimal*, stable, spectral factors and provide explicit formulas for them. Here, the assumption that the realizations for most of the nonlinear systems are *minimal* makes computations much easier. Furthermore, our handling of *minimal* inner-inner factorization throughout this paper is informed by the contributions made in [7, 8] and [14], where *minimal*, cascade factorization of nonlinear systems is discussed.

Next, we consider two types of extensions of Σ given by (1) that play a part in the analysis in what follows. We recall for Σ, where $a : \mathbf{R^n} \to \mathbf{R^n}$, $b : \mathbf{R^n} \to \mathbf{R^{n\times m}}$, $c : \mathbf{R^n} \to \mathbf{R^p}$, and $d : \mathbf{R^n} \to \mathbf{R^{p\times m}}$, that if $m < p$, then Σ is a nonsquare nonlinear system, and if $m = p$, then Σ is square. Thus, a requirement for Σ to be square is that m should be increased in order to equal p. This can be accomplished by introducing the concept of a *zero extension*. We say that Σ^e is a zero extension of Σ if it is the square system that is obtained from a (possibly) nonsquare Σ by augmenting b and d with an appropriate null function so that $m = p$. Of course, this notion can also be extended to the input-output map $T_\Sigma\colon L^m_{2,e}[0,\infty) \longmapsto L^p_{2,e}[0,\infty)$ introduced earlier, where for the zero extension we require that $m = p$. The *Hamiltonian extension* of Σ (where Σ is given as in (1)) has the form

$$
(3)\qquad \begin{cases} \dot{x} = a(x) + b(x)u, \\ \dot{p} = -\left[\dfrac{\partial a}{\partial x}(x) + \dfrac{\partial b}{\partial x}(x)u\right]^T p \\ \qquad\quad -\left[\dfrac{\partial c}{\partial x}(x) + \dfrac{\partial d}{\partial x}(x)u\right]^T u_a, \\ y = c(x) + d(x)u, \\ y_a = b^T(x)p + d^T(x)u_a \end{cases}
$$

(see [9]), where $u, y_a \in \mathbf{R^m}$ and $u_a \in \mathbf{R^p}$. This system has state space equal to the *cotangent bundle* of the state space manifold $\mathcal{X}$, denoted by $T^*\mathcal{X}$, with *natural* local coordinates $(x,p) = (x_1,\dots,x_n,\ p_1,\dots,p_n)$, inputs equal to (u,u_a), and outputs equal to (y,y_a). If we impose the interconnection law $u_a = y$ in (3), we get the Hamiltonian system of the form

$$
(4)\qquad \Phi = [D\Sigma]^T \circ \Sigma : \begin{cases} \dot{x} = \dfrac{\partial H}{\partial p}(x,p,u), \\ \dot{p} = -\dfrac{\partial H}{\partial x}(x,p,u), \\ y_a = \dfrac{\partial H}{\partial u}(x,p,u), \end{cases}
$$

with the Hamiltonian function $H(x,p,u)$ given by

$$
H(x,p,u) = p^T[a(x) + b(x)u] + \frac{1}{2}[c(x) + d(x)u]^T[c(x) + d(x)u].
$$

Here the state space is $T^*\mathcal{X}$, the inputs $u \in \mathbf{R^m}$, and the outputs $y_a \in \mathbf{R^m}$. We introduce the notion of the Hamiltonian extension of a nonlinear system in (3) to transform the (all-pass) inner-outer factorization problem into a nonlinear spectral factorization problem (see [32] for more details). In this regard, the Σ appearing in (4) is known as a *spectral factor* of Φ. In this paper we discuss a large variety of spectral factors that, for instance, may be minimal, stable, antistable, minimum phase, or maximum phase. Furthermore, Φ in (4) is *weakly coercive* if the spectral factor Σ is at least one-sided invertible. More specifically, in what follows, we consider right-sided inverses of various nonlinear spectral factors.

The main achievements of the analysis in this paper are listed below.

1. In Theorem 3 of section 3 we suggest a method of extending column- and row-rigid systems to inner systems.
2. We derive essentially unique factorizations that involve zero-extended stable and antistable, minimum and maximum phase spectral factors (Proposition 4, section 4). We recall that the term *essentially unique factorization* refers to the situation where uniqueness pertains to the underlying factorization structure only, with any variation in the form not affecting the mathematical meaning.
3. A necessary and sufficient condition for the existence of a minimal stable spectral factor is established in Theorem 6 of section 4.

4. We deduce certain factorization and coprimeness properties for inner systems (Theorem 7, section 4) that play a pivotal role in the rest of the paper. This result is dependent on the internal-external and external-internal factorizations introduced in Definition 5 of section 4.
5. We find necessary and sufficient conditions for the existence of an internal spectral factor in terms of an inner extension whose formula is given explicitly (Proposition 9, section 4). In order to accomplish this we introduce internal and external spectral factors in Definition 8 of section 4.
6. In Definition 12 we introduce the notion of a coprime inner-inner factorization. In addition, we prove the existence of such a factorization in Proposition 13.
7. We verify that a bijective correspondence exists between minimal, external, decomposable, nonsquare spectral factors and coprime inner-inner factorizations (Theorem 14, section 4).
8. Finally, in Theorem 15 of section 4, we make use of the preceding analysis to parametrize minimal, decomposable, nonsquare spectral factors in terms of coprime inner-inner factorizations.

Although many of the concepts mentioned in this list have not yet been defined, a comment on notation is in order. Throughout the paper, Σ denotes a general (square or nonsquare) nonlinear system that may be a spectral factor. On the other hand, R is a nonsquare spectral factor that is the starting point for our analysis in the sense that subsequent explicit formulas for various types of spectral factors and inner systems are written in terms of its components and associated Hamilton–Jacobi equation. We denote the minimal, stable, minimum and maximum phase spectral factors by Σ_- and Σ_+, respectively, while Σ_-^e (resp., Σ_+^e) is the notation used to denote a zero extension of Σ_- (resp., Σ_+). Minimal, antistable, minimum and maximum phase spectral factors are denoted by $\overline{\Sigma}_-$ and $\overline{\Sigma}_+$, respectively. Θ_c and Θ_r is the notation used for column-rigid and row-rigid systems, respectively. S denotes a minimal, stable spectral factor that may be decomposable. Φ is the symbol used to denote the nonlinear system that undergoes spectral factorization of the type given in (4). Various types of inner systems are introduced in what follows and are denoted by a Θ in combination with tildes, overlines, and sub- and superscripts.

The rest of the paper is arranged in the following manner. Section 2 provides a cursory exposition on minimum and maximum phase spectral factors. The third section considers various issues related to nonlinear inner systems. Section 4 contains the main results of the paper on the parametrization of minimal stable nonsquare spectral factors in terms of coprime inner-inner factorizations. An illustrative example is given in the fifth section, while we make conclusive remarks and comments on possible future research in section 6.

2. Minimum and maximum phase spectral factors. In this section, we consider a special class of spectral factors, namely, those that are maximum and minimum phase. In particular, we consider their connection with column- and row-rigid systems. Our starting point is the nonsquare, stable nonlinear system R of the form

$$R\colon \begin{cases} \dot{x} = a(x) + b(x)u, \\ y = \begin{pmatrix} c_1(x) \\ c_2(x) \end{pmatrix} + \begin{pmatrix} d(x) \\ 0 \end{pmatrix} u, \end{cases} \tag{5}$$

which provides a spectral factorization $\Phi = [D\Sigma]^T \circ \Sigma = [DR]^T \circ R$ for Φ. From [25] we know that there is a bijective correspondence between R in (5) and the set of triples

$(\mathcal{X}, P, c_2)$. Here, $\mathcal{X}$ is an invariant Lagrangian submanifold, P is a smooth solution of the Hamilton–Jacobi inequality

$$-P_x(x)[a(x) - b(x)d^T(x)F^{-1}(x)c(x)] + P_x(x)b(x)d(x)^T(F^{-1}(x))^2d(x)b(x)^TP_x(x)^T$$
$$-P_x(x)a(x) + c(x)^TF^{-1}(x)d(x)b(x)^TP_x(x)^T \leq 0,$$

$c_1(x) = c(x) + F^{-1}(x)d(x)b(x)^TP_x(x)^T$, and c_2 satisfies the equation

(6)
$$-P_x(x)[a(x) - b(x)d^T(x)F^{-1}(x)c(x)] + P_x(x)b(x)d(x)^T(F^{-1}(x))^2d(x)b(x)^TP_x(x)^T$$
$$-P_x(x)a(x) + c(x)^TF^{-1}(x)d(x)b(x)^TP_x(x)^T + c_2(x)^Tc_2(x) = 0.$$

Moreover, if P is any solution of (6), then there exists a map X such that

$$c_2(x) = -X_x(x)^TP_x(x)^T. \tag{7}$$

Next, we discuss *minimum* and *maximum phase* nonlinear systems that may be given explicitly in terms of R in (5) and its associated Hamilton–Jacobi equation (6). We recall that the *output nulling (or zero) dynamics* of a nonlinear system is the set of all system trajectories $x(\cdot)$ generated by some input trajectory $u(\cdot)$ such that $y(\cdot)$ is identically zero. Under some regularity conditions and the assumption that an appropriate function f exists, the zero dynamics can be computed as

$$\dot{x} = f(x, u), \quad x \in \mathcal{N}^* \subset \mathcal{X}. \tag{8}$$

Under additional conditions, it will actually be a system without inputs:

$$\dot{x} = f(x), \quad x \in \mathcal{N}^* \subset \mathcal{X}.$$

In this case the system is minimum phase if $f(x)$ is asymptotically stable and maximum phase if $f(x)$ is antistable. For (8) we define the system to be minimum phase if there exists $u = \alpha(x)$ such that the system is stable and maximum phase if there exists $u = \alpha(x)$ such that the system is antistable.

Column- and row-rigidness was introduced in [25] (see also [20] and [27]) and forms an integral part of the analysis in what follows. In what follows, we shall denote by Θ_c a column-rigid system and by Θ_r a row-rigid system. Moreover, we consider the Hamiltonian system $[D\Theta_c]^T \circ \Theta_c$ (Hamiltonian extension with $u_a = y$) with Hamiltonian

$$H(x, p, u) = p^T[a(x) + b(x)u] + \frac{1}{2}[c(x) + u]^T[c(x) + u].$$

We investigate the observability function P^o (see [29] for more details) defined as the solution of

$$P_x^o(x)a(x) + \frac{1}{2}c(x)^Tc(x) = 0$$

and define new coordinates $\overline{p} = p - P_x^o(x)$. It follows that

$$H(x, \overline{p}, u) = \overline{p}^T[a(x) + b(x)u] + u^T[b(x)^TP_x^o(x)^T + c(x)] + \frac{1}{2}u^Tu.$$

In addition, if P^o satisfies $P_x^o(x)b(x) + c(x)^T = 0$, then the submanifold $\overline{p}$ is an invariant manifold, with $[D\Theta_c]^T \circ \Theta_c$ restricted to this manifold being given by the static

input-output identity map $u \longmapsto y_a = u$. Here, the system Θ_c is said to be *column-rigid*. For the related concept of *row-rigidness*, we investigate the Hamiltonian system $[D\Theta_r] \circ \Theta_r^T$ (Hamiltonian extension with $u = y_a$) with Hamiltonian

$$H(x, p, u_a) = p^T a(x) + \frac{1}{2} p^T b(x) b(x)^T p + p^T b(x) u_a + c(x)^T u_a + \frac{1}{2} u_a^T u_a.$$

Consider the controllability function P^c (see [29] for more details) defined as the solution of

$$P_x^c(x) a(x) + \frac{1}{2} P_x^c(x) b(x) b(x)^T P_x^c(x)^T = 0,$$

where P^c is the solution such that $a(x) + b(x) b(x)^T P_x^c(x)^T$ is antistable. Furthermore, define canonical coordinates $\overline{p} = p - P_x^c(x)$. Here we have

$$\begin{aligned} H(x, \overline{p}, u_a) = {} & \overline{p}^T a(x) + \frac{1}{2} \overline{p}^T b(x) b(x)^T \overline{p} + P_x^c(x)^T [b(x) b(x)^T \overline{p} - b(x) u_a] \\ & + c(x)^T u_a + \frac{1}{2} u_a^T u_a. \end{aligned}$$

For the situation where P^c satisfies $P_x^c(x) b(x) + c(x)^T = 0$, the submanifold $\overline{p}$ is invariant, and the system $[D\Theta_r] \circ \Theta_r^T$ restricted to this manifold is given by the static input-output identity map $u_a \longmapsto y = u_a$. In this case, the system Θ_r is *row-rigid*.

In the following proposition, we collect facts that are pertinent to our investigation from the series of papers [20, 25] and [27] involving minimum and maximum phase spectral factors and their connection with row- and column-rigid systems.

PROPOSITION 1. *Suppose that weakly coercive* $\Phi = [D\Sigma]^T \circ \Sigma = [DR]^T \circ R$ *as in* (4), *with* Σ *and* R *given by* (1) *and* (5), *respectively. A minimal realization of the stable, minimum phase (outer) spectral factor* Σ_- *is given by*

$$\Sigma_- : \begin{cases} \dot{x} = a(x) + b(x) u, \\ y = c_1(x) + F^{-1}(x) d(x) b(x)^T P_x^-(x)^T + d(x) u, \end{cases} \tag{9}$$

where F *is given by* (2) *and* P^- *is the smooth solution of the Hamilton–Jacobi equation*

$$\begin{aligned} & P_x^-(x)[a(x) - b(x) d(x)^T F^{-1}(x) c_1(x)] \\ & \quad - \frac{1}{2} P_x^-(x) b(x) d(x)^T (F^{-1}(x))^2 d(x) b(x)^T P_x^-(x)^T + \frac{1}{2} c_2(x)^T c_2(x) = 0, \end{aligned} \tag{10}$$

with $P^-(0) = 0$ *and where the stability side condition*

$$a(x) - b(x) d^T(x) F^{-1}(x) c_1(x) \tag{11}$$

is Lyapunov stable. A minimal realization of the stable, maximum phase spectral factor Σ_+ *is given by*

$$\Sigma_+ : \begin{cases} \dot{x} = a(x) + b(x) u, \\ y = c_1(x) + F^{-1}(x) d(x) b(x)^T P_x^+(x)^T + d(x) u, \end{cases} \tag{12}$$

where F *is given by* (2) *and* P^+ *is the smooth solution of the Hamilton–Jacobi equation*

$$\begin{aligned} & P_x^+(x)[a(x) - b(x) d(x)^T F^{-1}(x) c_1(x)] \\ & \quad - \frac{1}{2} P_x^+(x) b(x) d(x)^T (F^{-1}(x))^2 d(x) b(x)^T P_x^+(x)^T + \frac{1}{2} c_2(x)^T c_2(x) = 0, \end{aligned} \tag{13}$$

with $P^+(0) = 0$ *and where the antistability side condition*

$$a(x) - b(x)d^T(x)F^{-1}(x)c_1(x) \tag{14}$$

is antistable. Minimal realizations for the right inverses of Σ_- *and* Σ_+ *are*

$$\Sigma_-^{-R}\colon \begin{cases} \dot{x} = a(x) - b(x)d(x)^T F^{-1}(x)c_1(x) \\ \qquad - b(x)d(x)^T(F^{-1}(x))^2 d(x)b(x)^T P_x^-(x)^T \\ \qquad + b(x)d(x)^T F^{-1}(x)y, \\ u = -d(x)^T F^{-1}(x)c_1(x) \\ \qquad - d(x)^T(F^{-1}(x))^2 d(x)b(x)^T P_x^-(x)^T \\ \qquad + d(x)^T F^{-1}(x)y \end{cases} \tag{15}$$

and

$$\Sigma_+^{-R}\colon \begin{cases} \dot{x} = a(x) - b(x)d^T(x)F^{-1}(x)c_1(x) \\ \qquad - b(x)d(x)^T(F^{-1}(x))^2 d(x)b(x)^T P_x^+(x)^T \\ \qquad + b(x)d^T(x)F^{-1}(x)y, \\ u = -d^T(x)F^{-1}(x)c_1(x) \\ \qquad - d(x)^T(F^{-1}(x))^2 d(x)b(x)^T P_x^+(x)^T \\ \qquad + d(x)^T F^{-1}(x)y, \end{cases}$$

respectively. If Σ_- *is the stable, minimum phase* (*outer*) *spectral factor of* Φ *given by* (9)*, then there exists an essentially unique minimal column-rigid system* Θ_c *for which*

$$R = \Theta_c \circ \Sigma_-, \tag{16}$$

with

$$\Theta_c\colon \begin{cases} \dot{x} = a(x) - b(x)d^T(x)F^{-1}(x)c_1(x) \\ \qquad - b(x)d(x)^T(F^{-1}(x))^2 d(x)b(x)^T P_x^-(x)^T \\ \qquad + b(x)d^T(x)F^{-1}(x)y, \\ y_a = \begin{pmatrix} -F^{-1}(x)d(x)b(x)^T P_x^-(x)^T \\ c_2(x) \end{pmatrix} + \begin{pmatrix} I \\ 0 \end{pmatrix} u. \end{cases} \tag{17}$$

If Σ_+ *is the stable, maximum phase spectral factor of* Φ *given by* (12)*, then there exists an essentially unique minimal row rigid system* Θ_r *for which*

$$\Sigma_+ = \Theta_r \circ R, \tag{18}$$

with

$$\Theta_r\colon \begin{cases} \dot{x} = -a(x)^T + c_1(x)^T F^{-1}(x)d(x)b(x)^T \\ \qquad + P_x^+(x)b(x)d(x)^T(F^{-1}(x))^2 d(x)b(x)^T \\ \qquad + (P_x^+(x)b(x)d(x)^T F^{-1}(x) - c_2(x)^T)u, \\ y = F^{-1}(x)d(x)b(x)^T + (I \quad 0)u. \end{cases} \tag{19}$$

Proof. The proof of the various components of the result can be found in [20, 25] or [27]. □

3. Nonlinear inner systems. In this section we lay down the structure of nonlinear inner systems that will be useful in proving the main results in this paper. In particular, we prove in subsection 3.2 that the column- and row-rigid systems introduced earlier may be extended to inner systems in a minimal way. Moreover, explicit formulas for these extended inner systems in terms of solutions of Hamilton–Jacobi equations are given. Subsection 3.3 outlines an approach to inner-inner factorization developed by Ball and Petersen in the recent contribution [3].

3.1. Basic properties. First, we provide a brief description of a *nonlinear inner system*. We assume that j is any $m \times m$ signature matrix ($j = j^* = j^{-1}$) and J is any $p \times p$ signature matrix ($J = J^* = J^{-1}$).

DEFINITION 2. *Nonlinear* Θ *is* (j, J)-inner (*or* (j, J)-stable conservative) *if*

- $x \to a(x)$ *is stable* (*with respect to assumed equilibrium point* $x = 0$) *and*
- *there is a nonnegative-valued storage function* $P(x)$ *with* $P(0) = 0$ *such that*

$$P(x(t_2)) - P((x(t_1)) = \frac{1}{2}\int_{t_1}^{t_2} [u(t)^T j u(t) - y(t)^T J y(t)]dt \tag{20}$$

over all trajectories $(u(t), x(t), y(t))$ *of the system. This is true for all* $t_1 \le t_2$ *and* $u(\cdot)$, *with* $x(t_2)$ *denoting the state at time* t_2 *that originates from the initial state* $x(t_1)$ *at time* t_1 *and input* $u(\cdot)$ *on the time interval* $[t_1, t_2]$.

Alternatively, Θ *is said to be* (j, J)-inner *if it is lossless with respect to the* L_2*-gain supply rate*

$$s(u, y) = \frac{1}{2}u^T j u - \frac{1}{2}y^T J y.$$

The above characterization of nonlinear (j, J)-inner systems was achieved within the dissipative systems framework of Hill and Moylan [15] and Willems [34]. Here, the dissipation equality in (20) may be derived from a state space implementation of the L_2-gain condition in the formulation of the nonlinear H_∞-problem (see [31, 32]). Note that the function defined in (20) may also be thought of as a Lyapunov function (see [15]). If P is assumed to be smooth, the energy balance relation (20) can be expressed as

$$P_x(x)[a(x) - b(x)c(x)] = \frac{1}{2}u^T j u - \frac{1}{2}[c(x) + d(x)u]^T J[c(x) + d(x)u], \quad P(0) = 0,$$

for all x and u, or equivalently, in infinitesimal form as

$$\begin{aligned} P_x(x)b(x) + c(x)^T J d(x) &= 0, \\ P_x(x)a(x) + \frac{1}{2}c(x)^T J c(x) &= 0, \\ d(x)^T J d(x) &= j. \end{aligned} \tag{21}$$

In fact, realizations for nonlinear invertible (j, J)-inner systems may be expressed in terms of smooth solutions of Hamilton–Jacobi equations as follows:

$$\Theta : \begin{cases} \dot{x} = a(x) + b(x)u, & u \in \mathbf{R^m}, \\ y_a = -b(x)^T P_x(x)^T + u, & y_a \in \mathbf{R^p}, \end{cases} \tag{22}$$

where P is a solution of the Hamilton–Jacobi equation

$$P_x(x)[a(x) - b(x)c(x)] - \frac{1}{2}P_x(x)b(x)b(x)^T P_x(x)^T = 0, \quad P(0) = 0,$$

with $P_x(x) = (\frac{\partial P}{\partial x_1}(x), \ldots, \frac{\partial P}{\partial x_n}(x))$. In the linear case, the function Θ_{linear} is said to be (j, J)-*inner* if it has the property that

$$\Theta^*_{\text{linear}} J \Theta_{\text{linear}} = \Theta_{\text{linear}} J \Theta^*_{\text{linear}} = j.$$

In subsequent discussions, we assume that $j = J = I$. We say that the inner system Θ_1 *divides* Θ_2 *on the right (left)* if the interconnection of Θ_1^T and Θ_2 (Θ_2 and Θ_1^T) is stable. Given two inner systems Θ_1 and Θ_2, we denote the *greatest common right (left) divisor* by $\Theta_1 \wedge_R \Theta_2$ ($\Theta_1 \wedge_L \Theta_2$). Also, we denote the *least common right (left) multiple* by $\Theta_1 \vee_R \Theta_2$ ($\Theta_1 \vee_L \Theta_2$). Two inner systems are *right (left) coprime* if their greatest common right (left) divisor is the identity (see [18] and [19] for more details).

3.2. Minimal inner embeddings. In this subsection, we study the embedding of inner and rigid systems in minimal inner systems. In particular, in the next result, we extend a column-rigid system Θ_c given by (17) in a minimal way in order to obtain a system that is inner. Similarly, we extend a row-rigid system Θ_r given by (19) minimally to an inner system.

THEOREM 3. *The column-rigid system* Θ_c *from* (17) *can be extended minimally to the inner system*

$$\begin{cases} \dot{x} = a(x) - b(x)d(x)^T F^{-1}(x)c_1(x) \\ \qquad - b(x)d(x)^T (F^{-1}(x))^2 d(x)b(x)^T P_x^-(x)^T \\ \qquad + (b(x)d(x)^T F^{-1}(x) X_x^-(x))u, \\ y_a = \begin{pmatrix} -F^{-1}(x)d(x)b(x)^T P_x^-(x)^T \\ c_2(x) \end{pmatrix} + \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} u, \end{cases} \tag{23}$$

where F *is given by* (2) *and* P^- *is a smooth solution of* (10) *with stability side condition* (11) *and where* X^- *satisfies* (7). *Moreover,* Θ_r *given as in* (19) *may be extended to the inner system*

$$\begin{cases} \dot{x} = -a(x)^T + c_1(x)^T F^{-1}(x)d(x)b(x)^T \\ \qquad + P_x^+(x)b(x)d(x)^T (F^{-1}(x))^2 d(x)b(x)^T \\ \qquad + (P_x^+(x)b(x)d(x)^T F^{-1}(x) - c_2(x)^T)u, \\ y = \begin{pmatrix} F^{-1}(x)d(x)b(x)^T \\ X_x^+(x)^T \end{pmatrix} + \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} u, \end{cases} \tag{24}$$

where F *is given by* (2) *and* P^+ *is a smooth solution of* (13) *with stability side condition* (14) *and where* X^+ *satisfies* (7).

Proof. It is clear that any minimal inner extension of column-rigid Θ_c given by (17) will be of the form (23). Next, we have to establish that (23) is, in fact, inner. In order to accomplish this, we rewrite (23) as

$$\begin{cases} \dot{x} = f(x) + g(x)u, \\ y_a = -g(x)^T Q_x(x)^T + u, \end{cases}$$

where Q is a solution of the Hamilton–Jacobi equation

$$Q_x(x)[f(x)-g(x)h(x)]-\frac{1}{2}Q_x(x)g(x)g(x)^TQ_x(x)^T=0,\quad Q(0)=0.$$

It remains to show that $Q=P^-$ is a solution of this equation. Indeed, it is clear that

$$\begin{aligned}&P_x^-(x)[a(x)-b(x)d(x)^TF^{-1}(x)c_1(x)]+c_2(x)^Tc_2(x)\\&\quad-\frac{1}{2}[P_x^-(x)b(x)d(x)^T(F^{-1}(x))^2d(x)b(x)^TP_x^-(x)^T+c_2(x)^Tc_2(x)]\\&\quad=P_x^-(x)[a(x)-b(x)d(x)^TF^{-1}(x)c_1(x)]\\&\quad-\frac{1}{2}P_x^-(x)b(x)d(x)^T(F^{-1}(x))^2d(x)b(x)^TP_x^-(x)^T+\frac{1}{2}c_2(x)^Tc_2(x)=0.\end{aligned}$$

As was the case in the above, any minimal, inner extension of row-rigid Θ_r given by (19) will be of the form (24). The remainder of the proof of the second part can be obtained from the observation that

$$\begin{aligned}&\frac{1}{2}P_x^+(x)b(x)d(x)^T(F^{-1}(x))^2d(x)b(x)^TP_x^+(x)^T-\frac{1}{2}c_2(x)^Tc_2(x)\\&\quad=[-a(x)^T+c_1(x)^TF^{-1}(x)d(x)b(x)^T]P_x^+(x)^T\\&\quad=P_x^+(x)[a(x)-b(x)d(x)^TF^{-1}(x)c_1(x)].\end{aligned}$$

This computation is made possible by the symmetry in the Hamilton–Jacobi equation. □

The extensions from rigid to minimal inner systems referred to in Theorem 3 are known as *minimal inner embeddings*. A special subclass of these embeddings are the extensions from inner to minimal inner systems. The extension from the inner system

$$\begin{cases}\dot{x}=a(x)-b(x)d^T(x)F^{-1}(x)c_1(x)\\\qquad-b(x)d(x)^T(F^{-1}(x))^2d(x)b(x)^TQ_x^-(x)^T\\\qquad+b(x)d^T(x)F^{-1}(x)u,\\y=-F^{-1}(x)d(x)b(x)^TQ_x^-(x)^T+u,\end{cases}$$

with corresponding Hamilton–Jacobi equation

$$\begin{aligned}&Q_x^-(x)[a(x)-b(x)d(x)^TF^{-1}(x)c_1(x)]\\&\quad-\frac{1}{2}Q_x^-(x)b(x)d(x)^T(F^{-1}(x))^2d(x)b(x)^TQ_x^-(x)^T=0,\end{aligned}$$

to the minimal inner system (23) is an example of such an embedding. The latter type of inner embedding will play a prominent role in the discussion in what follows.

3.3. Minimal inner factorizations of nonlinear inner systems. Our approach to the factorization of nonlinear inner systems is dependent on the adaptation of the cascade factorization of general nonlinear systems to nonlinear inner systems. This approach was developed in the recent paper by Ball and Petersen [3] (see also [30] for an earlier work) and was central to the solution of the problem of parametrizing square solutions of the nonlinear spectral factorization problem in terms of minimal

inner-inner factorizations of nonlinear systems. As this approach will underpin our analysis, the highlights of this scheme are outlined below.

Suppose that the state manifold has a foliation $x = (x_1, x_2)$ (a diffeomorphism of the state space $\mathcal{X}$ with the Cartesian product $\mathcal{X}_1 \times \mathcal{X}_2$ of two manifolds $\mathcal{X}_1$ and $\mathcal{X}_2$), so that (x_1, x_2) becomes a new system of coordinates for Θ. Then, in terms of these coordinates, we write the state space equations (22) in the form

$$\Theta\colon \begin{cases} \dot{x}_1 = \mathbf{a}_1(x_1, x_2) + \mathbf{b}_1(x_1, x_2)u, \\ \dot{x}_2 = \mathbf{a}_2(x_1, x_2) + \mathbf{b}_2(x_1, x_2)u, \\ \ \ y = \mathbf{c}(x_1, x_2) + u. \end{cases}$$

In this case the inverse system has the form

$$\Theta^{-1}\colon \begin{cases} \dot{x}_1 = \mathbf{a}_1^{\times}(x_1, x_2) + \mathbf{b}_1(x_1, x_2)y, \\ \dot{x}_2 = \mathbf{a}_2^{\times}(x_1, x_2) + \mathbf{b}_2(x_1, x_2)y, \\ \ \ u = -\mathbf{c}(x_1, x_2), \end{cases}$$

where $\mathbf{a}^{\times}(x_1, x_2) = \mathbf{a}(x_1, x_2) - \mathbf{b}(x_1, x_2)\mathbf{c}(x_1, x_2)$, or

$$\begin{pmatrix} \mathbf{a}_1^{\times}(x_1, x_2) \\ \mathbf{a}_2^{\times}(x_1, x_2) \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1(x_1, x_2) - \mathbf{b}_1(x_1, x_2)\mathbf{c}(x_1, x_2) \\ \mathbf{a}_2(x_1, x_2) - \mathbf{b}_2(x_1, x_2)\mathbf{c}(x_1, x_2) \end{pmatrix}. \tag{25}$$

The foliation $x = (x_1, x_2)$ is known as an *inner product-coordinate pair* for the inner system Θ with storage function P (see (21) and (26)) if the functions $P(x_1, x_2)$, $\mathbf{b}_j(x_1, x_2)$, and $\mathbf{a}_1(x_1, x_2)$ and $\mathbf{a}_2^{\times}(x_1, x_2)$ have the form

$$\begin{aligned} P(x_1, x_2) &= P_1(x_1) + P_2(x_2), \\ \mathbf{b}_1(x_1, x_2) &= b_1(x_1), \\ \mathbf{b}_2(x_1, x_2) &= b_2(x_2), \\ \mathbf{a}_1(x_1, x_2) &= a_1(x_1), \\ \mathbf{a}_2^{\times}(x_1, x_2) &= a_2^{\times}(x_2) \end{aligned} \tag{26}$$

for some single-variable functions $P_1(x_1), P_2(x_2), b_1(x_1), b_2(x_2), a_1(x_1)$, and $a_2^{\times}(x_2)$. From (25) and the first three equations in (26), we see that

$$\begin{aligned} \mathbf{a}_2^{\times}(x_1, x_2) &= \mathbf{a}_2(x_1, x_2) + b_2(x_2)(b_1(x_1)^T P_{1,x_1}(x_1)^T + b_2(x_2)^T P_{2,x_2}(x_2)^T) \\ &= [\mathbf{a}_2(x_1, x_2) + b_2(x_2)b_1(x_1)^T P_{1,x_1}(x_1)^T] + b_2(x_2)b_2(x_2)^T P_{2,x_2}(x_2)^T. \end{aligned}$$

Thus, we see that the last two equations in (26), in the definition of inner product-coordinate pairs, can be replaced with

$$\begin{aligned} \mathbf{a}_2(x_1, x_2) = a_2(x_2) - b_2(x_2)b_1(x_1)^T P_{1,x_1}(x_1)^T \\ \text{for a single-variable function } a_2(x_2), \end{aligned}$$

with $a_2(x_2) = a_2^{\times}(x_2) - b_2(x_2)b_2(x_2)^T P_{2,x_2}(x_2)^T$. These conventions are used to show that if we suppose that Θ is a nonlinear inner system with storage function P as in (21) and (26), then factorizations $\Theta = \Theta_2 \circ \Theta_1$, with Θ_2 and Θ_1 inner, are in bijective

correspondence with inner product-coordinate pairs $x = (x_1, x_2)$. Indeed, given an inner product-coordinate pair as in (26), define systems Θ_1 and Θ_2 by

$$\Theta_1 : \begin{cases} \dot{x}_1 = a_1(x_1) + b_1(x_1)u_1, \\ y_1 = -b_1(x_1)^T P_{1,x_1}(x_1)^T + u_1, \end{cases}$$

$$\Theta_2 : \begin{cases} \dot{x}_2 = a_2(x_2) + b_2(x_2)u_2, \\ y_2 = -b_2(x_2)^T P_{2,x_2}(x_2)^T + u_2, \end{cases}$$

with $a_2(x_2) = a_2^{\times}(x_2) - b_2(x_2)b_2(x_2)^T P_{2,x_2}(x_2)^T$. Then Θ_1 and Θ_2 are inner systems with storage functions $P_1(x_1)$ and $P_2(x_2)$, respectively, and we recover Θ as $\Theta = \Theta_2 \circ \Theta_1$. The same analysis as in the above can, of course, be done for the minimal inner extensions of Θ_1 and Θ_2.

4. Parametrizations of minimal stable nonsquare spectral factors. The main results of the paper are contained in this section. In the subsequent analysis, subsection 4.1 addresses some of the issues related to extremal spectral factors, while internal and external spectral factors are described in subsection 4.2. In subsection 4.3 the important notion of a coprime inner-inner factorization is introduced. Finally, the set of all minimal stable nonsquare spectral factors is parametrized in terms of the aforementioned inner-inner factorization in subsection 4.4.

Next, we state the problem that will be solved in the ensuing discussion.

Problem. Given a stable, nonlinear, input-affine control system, parametrize all nonsquare, stable, nonlinear, input-affine spectral factors in terms of a special class of inner-inner factorizations.

4.1. Extremal spectral factors. The next proposition discusses minimal stable and minimal antistable spectral factors of Φ and their zero extensions.

PROPOSITION 4. *Let Σ_- and Σ_+ be the stable minimum and maximum phase spectral factors of Φ given by* (9) *and* (12)*, respectively. Also, assume that $\overline{\Sigma}_-$ and $\overline{\Sigma}_+$ are the corresponding antistable minimum and maximum phase spectral factors of Φ, respectively.*

1. *Assume that Σ_-^e and Σ_+^e are the zero-extended, stable, minimum and maximum phase spectral factors of Φ, respectively. Given any minimal stable spectral factor S of Φ, there exist, essentially unique, inner systems Θ' and Θ'' for which*

$$\begin{cases} \Sigma_+^e = \Theta' \circ S, \\ S = \Theta'' \circ \Sigma_-^e. \end{cases} \tag{27}$$

2. *Suppose that $\overline{\Sigma}_-^e$ and $\overline{\Sigma}_+^e$ are the zero-extended, antistable, minimum and maximum phase spectral factors of Φ, respectively. Given any minimal antistable spectral factor $\overline{S}$ of Φ, there exist, essentially unique, inner systems $\overline{\Theta}'$ and $\overline{\Theta}''$ for which*

$$\begin{cases} \overline{\Sigma}_+^e = \overline{\Theta}' \circ \overline{S}, \\ \overline{S} = \overline{\Theta}'' \circ \overline{\Sigma}_-^e. \end{cases}$$

Proof.

1. First we verify the properties of the inner system Θ'' in the second equation of (27). We know from Proposition 1 that there exists an essentially unique column-rigid system Θ_c for which $S = \Theta_c \circ \Sigma_-$. Furthermore, the existence of Θ'' as a minimal inner extension of Θ_c is evident from Theorem 3. Since Σ_- and Σ_-^e are spectral factors of Φ, we have that

$$\begin{aligned}\Phi(u) &= [DS(u)]^T \circ S(u)\\ &= [D(\Theta_c \circ \Sigma_-)(u)]^T \circ (\Theta_c \circ \Sigma_-)(u)\\ &= [D(\Theta'' \circ \Sigma_-^e)(u)]^T \circ (\Theta'' \circ \Sigma_-^e)(u),\end{aligned}$$

with Σ_-^e extended outer. Conversely, let $S = \Theta_*'' \circ \Sigma_-^e$, with Θ_*'' being a minimal inner extension of the column-rigid system Θ_{c*}. Then $S = \Theta_c \circ \Sigma_- = \Theta_{c*} \circ \Sigma_-$ which, by the right invertibility of Σ_-, implies that

$$[D\Theta_{c*}(u)]^T \circ \Theta_{c*}(u) = [D\Theta_c(u)]^T \circ \Theta_c(u) = u.$$

It follows by minimality that $\Theta_{c*}(u) = \Theta_c(u)K$ for some inner system K. If Θ_{c*} and Θ_c have the same state space dimension, then K is a constant system. From this we may deduce that Θ'' is an essentially unique inner system with the property that $S = \Theta'' \circ \Sigma_-^e$. The properties of the inner system Θ' in the first equation of (27) can be verified in a similar manner.

2. The proof of the second part will be analogous to the first.

The result follows from the discussion above. ∎

Note that the factorization in the second equation of (27) is, in fact, an *inner-outer* factorization. The rigid and inner systems described in Proposition 4 can be factorized further. In this regard, we consider for row-rigid Θ_r and column-rigid Θ_c the finer factorizations

$$\begin{cases}\Theta_r = \Theta_r^1 \circ \Theta_r^2,\\ \Theta_c = \Theta_c^2 \circ \Theta_c^1,\end{cases} \tag{28}$$

where Θ_r^1 and Θ_c^1 are inner and Θ_r^2 and Θ_c^2 are right and left outer. Next, we introduce an important type of factorization.

Definition 5. *We call the equations occurring in* (28) *the* internal-external *and* external-internal factorizations *of Θ_r and Θ_c, respectively.*

Similarly, for Θ' and Θ'' we consider finer factorizations

$$\begin{cases}\Theta' = \Theta_1' \circ \Theta_2',\\ \Theta'' = \Theta_2'' \circ \Theta_1'',\end{cases} \tag{29}$$

where Θ_1' and Θ_1' are inner and Θ_2' and Θ_2'' are right and left outer. Likewise, we call (29) the *internal-external* and *external-internal factorizations* of Θ' and Θ'', respectively.

Theorem 6. *Suppose that Σ_+^e is an extended, maximum phase, stable spectral factor of the nonlinear system Φ. Then S is a minimal, stable, spectral factor of Φ if and only if there exists an inner system Θ' such that*

$$[DS(u)]^T \circ S(u) = [D(\Theta'^T \circ \Sigma_+^e)(u)]^T \circ (\Theta'^T \circ \Sigma_+^e)(u), \tag{30}$$

where we restrict the state manifold to the diagonal (x, x).

Proof. We prove the forward assertion by considering Proposition 4 and the properties of Σ_+^e. Indeed, from the first part of Proposition 4, if S is a minimal, stable spectral factor and Σ_+^e is the extended, maximum phase, stable spectral factor, then

$$\begin{aligned}[DS(u)]^T \circ S(u) &= [D\Sigma_+^e(u)]^T \circ \Sigma_+^e(u) \\ &= [D\Sigma_+^e(u)]^T \circ [D\Theta'^T(u)]^T \circ \Theta'^T(u) \circ \Sigma_+^e(u) \\ &= [D(\Theta'^T \circ \Sigma_+^e)(u)]^T \circ (\Theta'^T \circ \Sigma_+^e)(u)\end{aligned}$$

for some inner system Θ'. Conversely, assume that (30) holds for some inner system Θ'. Then clearly S is a stable spectral factor. The minimality of the realization S follows from the fact that in general

$$\begin{aligned}&\text{dimension of state manifold associated with } S \\ &\leq \text{dimension of state manifold associated with } \Sigma_+^e,\end{aligned}$$

which implies that

$$\begin{aligned}&2 \times \text{dimension of state manifold associated with } \Sigma_+^e \\ &= \text{dimension of state manifold associated with } \Phi \\ &\leq 2 \times \text{dimension of state manifold associated with } S \\ &\leq 2 \times \text{dimension of state manifold associated with } \Sigma_+^e.\end{aligned}$$

As a result of this, we have that

$$\begin{aligned}&\text{dimension of state manifold associated with } \Phi \\ &= 2 \times \text{dimension of state manifold associated with } S \\ &= 2 \times \text{dimension of state manifold associated with } \Sigma_+^e\end{aligned}$$

and S is a stable, spectral factor that is minimal. □

The next result provides an important first step for a parametrization of the set of all nonsquare minimal stable spectral factors. This result will be used extensively in the proofs of the main theorems in subsection 4.4, where the parametrization will be described explicitly.

THEOREM 7. *Suppose that Φ is a weakly coercive nonlinear control system with minimal nonsquare stable spectral factor S. Let Θ_+ and Θ_- be the inner systems with the properties*

$$\begin{cases} \Sigma_+ = \Theta_+ \circ \Sigma_-, \\ \overline{\Sigma}_+ = \Theta_- \circ \overline{\Sigma}_-. \end{cases} \tag{31}$$

Assume that $\Theta', \Theta'', \overline{\Theta}'$, and $\overline{\Theta}''$ are determined by Proposition 4, with internal-external factorizations

$$\Theta' = \Theta'_1 \circ \Theta'_2, \ \overline{\Theta}' = \overline{\Theta}'_1 \circ \overline{\Theta}'_2 \tag{32}$$

and the external-internal factorization

$$\Theta'' = \Theta''_2 \circ \Theta''_1, \ \overline{\Theta}'' = \overline{\Theta}''_2 \circ \overline{\Theta}''_1. \tag{33}$$

In this case we have that

1. (a) *Θ'_1 is the greatest common left inner factor of Θ' and Θ^e_+ and Θ''_1 is the greatest common right inner factor of Θ'' and Θ^e_+. $\overline{\Theta}'_1$ is the greatest common left inner factor of $\overline{\Theta}'$ and $\overline{\Theta}^e_+$ and $\overline{\Theta}''_1$ is the greatest common right inner factor of $\overline{\Theta}''$ and $\overline{\Theta}^e_+$.*
   (b) *$\overline{\Theta}''_1$ is the greatest common right inner factor of $\overline{\Theta}''$ and $\overline{\Theta}^e_-$ and Θ''_1 is the greatest common right inner factor of $\overline{\Theta}''$ and $\overline{\Theta}^e_+$.*
2. *$\Theta' \circ \Theta''$ is an inner extension of Θ_+ and $\overline{\Theta}' \circ \overline{\Theta}''$ is an inner extension of Θ_-.*
3. *There exists a unique inner system Θ such that*

$$\Theta^e_+ = \Theta'_1 \circ \Theta \circ \Theta''_1.$$

Likewise for $\overline{\Theta}^e_+$, we have for the factorization

$$\overline{\Theta}^e_+ = \overline{\Theta}'_1 \circ \overline{\Theta} \circ \overline{\Theta}''_1 \tag{34}$$

that $\overline{\Theta}$ exists and is unique.
4. (a) *Θ and Θ'_2 are left coprime while Θ and Θ''_2 are right coprime.*
   (b) *$\overline{\Theta}$ and $\overline{\Theta}'_2$ are left coprime while $\overline{\Theta}$ and $\overline{\Theta}''_2$ are right coprime.*

Proof.

1. (a) We recall from $\Theta_r = \Theta^1_r \circ \Theta^2_r$ in (28) that Θ^2_r is outer. Also, from (16) and (18) we know that $\Sigma_+ = \Theta_r \circ \Theta_c \circ \Sigma_-$. Hence it follows from the first equation in (31) that $\Theta_+ = \Theta_r \circ \Theta_c$. Furthermore, Θ^1_r is a left inner factor of Θ_+ and the inner extension of Θ^1_r is a common left inner factor of Θ' and Θ^e_+. Any left inner factor of Θ^e_+ is, up to a constant right unitary factor, an inner extension of Θ_r. Since Θ^2_r is outer, it is clear that Θ'_1 is the greatest common left inner factor of Θ' and Θ^e_+. The proof of the second assertion is similar to the one for the above.
   (b) The proof of part (b) is similar to that for part (a).
2. First, recall from the proof of part 1 that $\Theta_+ = \Theta_r \circ \Theta_c$. From this equality and the fact that Θ' is an inner extension of Θ_r and Θ'' is an inner extension of Θ_c, we must have that $\Theta' \circ \Theta''$ is an inner extension of Θ_+. That $\overline{\Theta}'\overline{\Theta}''$ is an inner extension of Θ_- can be proven in a similar manner.
3. Using the extended version of (31) and the factorizations (32) and (33), we compute that

$$\begin{aligned}\Sigma^e_+ &= \Theta' \circ \Theta'' \circ \Sigma^e_- \\ &= \Theta'_1 \circ \Theta'_2 \circ \Theta''_2 \circ \Theta''_1 \circ \Sigma^e_- \\ &= \Theta'_1 \circ \Theta \circ \Theta''_1 \circ \Sigma^e_-.\end{aligned}$$

   In addition, we have that

$$\Sigma^e_+ = \Theta^e_+ \circ \Sigma^e_-$$

   so that $\Theta_+ = \Theta^1_r \circ \widehat{\Theta} \circ \Theta^1_c$, where $\widehat{\Theta} = \Theta^2_r \circ \Theta^2_c$. Clearly, from Propositions 1 and 4, it follows that $\Theta = \Theta'_2 \circ \Theta''_2$ is unique.
4. (a) From part 1 of the statement of this theorem, we know that Θ'_1 is the greatest common left inner factor of Θ' and Θ^e_+. Hence we have that Θ'_2 and $\Theta \circ \Theta''_1$ are left coprime. Also, from the above, it is clear that Θ'_2 and Θ are left coprime. The fact that Θ''_2 and Θ are right coprime may be proved in a similar manner.

(b) The proof of this part may be achieved by using techniques similar to those in (a) above. □

4.2. Internal and external spectral factors. In order to parametrize spectral factors in terms of inner-inner factorizations, we study two particular types of spectral factors, namely, those that are internal and those that are external. The definitions of these classes of spectral factors are provided below.

DEFINITION 8. *Let S be a minimal, stable spectral factor and let Θ' and Θ'' be the inner systems characterized in Proposition* 4. *Then we say the following:*

1. *S is an* internal spectral factor *if we have*

$$\Theta_+^e = \Theta' \circ \Theta''. \tag{35}$$

2. *S is an* external spectral factor *if we have that*
   (a) *Θ' and Θ_+^e are left coprime and*
   (b) *Θ'' and Θ_+^e are right coprime.*

The analysis of general spectral factors in subsection 4.4 will depend on the properties of both the external and internal spectral factors. In the next result, we determine necessary and sufficient conditions for the existence of an internal spectral factor and provide explicit formulas for Θ_+ and its minimal inner extension Θ_+^e as referred to in Theorem 7 and (35) of Definition 8.

PROPOSITION 9. *Let S be a minimal, stable spectral factor and let Θ' and Θ'' be the inner systems characterized in Proposition* 4. *Then S is an internal spectral factor if and only if we have that*

$$\Theta_+^e = \Theta' \circ \Theta'' \colon \begin{cases} \dot{x} = a(x) - b(x)d(x)^T F^{-1}(x)c_1(x) \\ \qquad - b(x)d(x)^T (F^{-1}(x))^2 d(x)b(x)^T P_x^-(x)^T \\ \qquad + (b(x)d(x)^T F^{-1}(x) X_x^-(x))u, \\ y_a = \begin{pmatrix} -F^{-1}(x)d(x)b(x)^T (P_x^-(x) - P_x^+(x))^T \\ c_2(x) \end{pmatrix} \\ \qquad + \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} u \end{cases} \tag{36}$$

is an inner extension of the minimal inner system

$$\Theta_+ =: \begin{cases} \dot{x} = a(x) - b(x)d^T(x)F^{-1}(x)c_1(x) \\ \qquad - b(x)d(x)^T (F^{-1}(x))^2 d(x)b(x)^T P_x^-(x)^T \\ \qquad + b(x)d^T(x)F^{-1}(x)u, \\ y = -F^{-1}(x)d(x)b(x)^T (P_x^-(x) - P_x^+(x))^T + u, \end{cases} \tag{37}$$

where Θ_+ is the inner factor of Σ_+ in $\Sigma_+ = \Theta_+ \circ \Sigma_-$ after restriction of the state space to the diagonal (x, x). Furthermore, $(P^- - P^+)$ is the solution of the Hamilton–Jacobi equation

$$Z_x(x)[a(x) - b(x)c(x)] - \frac{1}{2} Z_x(x)b(x)b(x)^T Z_x(x)^T = 0, \quad Z(0) = 0.$$

Moreover, up to a right unitary factor for Θ' and a left factor for Θ'', the factoriza-

tions of the inner extension (36) *are in a bijective correspondence with inner-inner factorizations of* Θ_+*—that is, factorizations* $\Theta_+ = \Theta_1 \circ \Theta_2$.

Proof. By the discussion in subsection 3.2, the equation $\Sigma_+^e = \Theta' \circ S$ from Proposition 4, and the fact that Σ_+^e is a zero extension of Σ_+, we know that Θ' is a minimal inner extension of some inner system Θ_1. Similarly, for $S = \Theta'' \circ \Sigma_-^e$, we know that an inner system Θ_2 exists that may be minimally extended to Θ''. In this case, the discussion on inner-inner factorizations in subsection 3.3 may be used to obtain the equality $\Theta_+ = \Theta_1 \circ \Theta_2$ (compare Theorem 3.4 in [3]).

The explicit formula $\Theta_+ = \Sigma_+ \circ \Sigma_-^{-R}$ arises from computing the product of the stable maximum phase spectral factor Σ_+ in (12) and the right inverse of the stable minimum phase spectral factor Σ_-^{-R} from (15) (compare with Proposition 3.3 in [3]). ∎

Remark 10. It is always possible to obtain explicit formulas for the inner factors Θ' and Θ'' of the inner-inner factorization $\Theta' \circ \Theta''$ in (36) by utilizing the scheme suggested in subsection 3.3.

The following corollary to Proposition 9 states that any minimal, stable spectral factor is, in fact, internal.

COROLLARY 11. *If* S *is a minimal, stable spectral factor, then it is internal.*

Proof. From the proof of Proposition 9, we have that Θ_1 must be a left factor of Θ_+ and hence S must be internal. ∎

4.3. Coprime inner-inner factorization. In this subsection we discuss a special type of factorization associated with inner systems, namely, *coprime inner-inner factorizations*. In order to define this concept, we introduce the inner systems

$$V : \begin{cases} \dot{x}_V = a_V(x_V) + b_V(x_V)u_V, & u_V \in \mathbf{R^{m_0}}, \\ y_V = c_V(x_V) + u_V, & y_V \in \mathbf{R^{m_0}}, \end{cases} \tag{38}$$

with $a_V : \mathbf{R^n} \to \mathbf{R^n}, b_V : \mathbf{R^n} \to \mathbf{R^{n \times m_0}}$, and $c_V : \mathbf{R^n} \to \mathbf{R^{m_0}}$, and

$$W : \begin{cases} \dot{x}_W = a_W(x_W) + b_W(x_W)u_W, & u_W \in \mathbf{R^{m-m_0}}, \\ y_W = c_W(x_W) + u_W, & y_W \in \mathbf{R^{m-m_0}}, \end{cases} \tag{39}$$

where $a_W : \mathbf{R^n} \to \mathbf{R^n}, b_W : \mathbf{R^n} \to \mathbf{R^{n \times (m-m_0)}}$, and $c_W : \mathbf{R^n} \to \mathbf{R^{m-m_0}}$.

DEFINITION 12. *Let* V *and* W *be inner systems given by the general formulas* (38) *and* (39)*, respectively. We say that a factorization*

$$\widetilde{\Theta} = \Theta' \circ \Theta''$$

into the product of two inner systems is a coprime inner-inner factorization *if* Θ' *is left coprime and* Θ'' *is right coprime with both the inner extensions* V^e *and* W^e *of* V *and* W*, respectively.*

The result below postulates the existence of a coprime inner-inner factorization of an appropriate extension of the inner system Θ_+ given by (37).

PROPOSITION 13. *Let* P^- *be a smooth solution of* (10) *with stability side condition* (11) *and* X^- *satisfying* (7)*. Furthermore, assume that* P^+ *is a smooth solution of* (13) *with stability side condition* (14) *and* X^+ *satisfying* (7)*. Let* Θ_+ *be an inner system*

given as in (37). *Then there exists a coprime inner-inner factorization*

$$\Theta' \circ \Theta'' : \begin{cases} \dot{x} = a(x) - b(x)d^T(x)F^{-1}(x)c_1(x) \\ \qquad - b(x)d(x)^T(F^{-1}(x))^2 d(x)b(x)^T P_x^-(x)^T \\ \qquad + (b(x)d(x)^T F^{-1}(x)X_x^-(x))u, \\ y = \begin{pmatrix} -F^{-1}(x)d(x)b(x)^T(P_x^-(x) - P_x^+(x))^T \\ \qquad -(X_x^-(x) - X_x^+(x))^T P_x^-(x)^T \end{pmatrix} \\ \qquad + \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} u. \end{cases} \tag{40}$$

Proof. We start by choosing inner systems V and W that appear in Definition 12 as

$$V : \begin{cases} \dot{x} = a(x) - b(x)d^T(x)F^{-1}(x)c_1(x) \\ \qquad - b(x)d(x)^T(F^{-1}(x))^2 d(x)b(x)^T Q_x^-(x)^T \\ \qquad + b(x)d^T(x)F^{-1}(x)u, \\ y = -F^{-1}(x)d(x)b(x)^T Q_x^-(x)^T + u, \end{cases}$$

with corresponding Hamilton–Jacobi equation

$$Q_x^-(x)[a(x) - b(x)d(x)^T F^{-1}(x)c_1(x)] \\ -\frac{1}{2}Q_x^-(x)b(x)d(x)^T(F^{-1}(x))^2 d(x)b(x)^T Q_x^-(x)^T = 0,$$

and as

$$W : \begin{cases} \dot{x} = -a(x)^T + c_1(x)^T F^{-1}(x)d(x)b(x)^T \\ \qquad + Q_x^+(x)b(x)d(x)^T(F^{-1}(x))^2 d(x)b(x)^T \\ \qquad + Q_x^+(x)b(x)d(x)^T F^{-1}(x)u, \\ y = F^{-1}(x)d(x)b(x)^T + u, \end{cases}$$

with corresponding Hamilton–Jacobi equation

$$Q_x^+(x)[a(x) - b(x)d(x)^T F^{-1}(x)c_1(x)] \\ -\frac{1}{2}Q_x^+(x)b(x)d(x)^T(F^{-1}(x))^2 d(x)b(x)^T Q_x^+(x)^T = 0,$$

respectively. The rest of the proof may be outlined as follows. We obtain explicit formulas for the inner factors Θ' and Θ'' of the inner-inner factorization $\Theta' \circ \Theta''$ in (40) by utilizing the scheme suggested in subsection 3.3. Next, from subsection 3.2, we extend V and W to minimal inner systems V^e and W^e, respectively. Then by Theorem 7 we have that Θ' is left coprime and Θ'' is right coprime with both V^e and W^e. ∎

We note that the explicit formulas for $\Theta' \circ \Theta''$ in (36) of Proposition 9 and (40) of Proposition 13 have the same form since it was established in (7) that there exists a map X such that $c_2(x) = -X_x(x)^T P_x(x)^T$ (see Theorem 3 of [25]).

4.4. Parametrizing nonsquare spectral factors in terms of coprime inner-inner factorizations. In this section, we establish a bijective correspondence between coprime inner-inner factorizations of inner systems and a certain subclass of minimal, stable spectral factors S for Φ. This special class arises as follows.

Our description of the class of *decomposable* spectral factors will follow [3] rather closely. Let $P(x)$ be a solution of the Hamilton–Jacobi equation (6) and let $P^-(x)$ be the solution also satisfying the stabilizing side condition (11). As in subsection 3.3, we suppose that the state space manifold $\mathcal{X}$ has a decomposition $x = (x_1, x_2)$. We start with the extended outer factor Σ_-^e, where $S = \Theta'' \circ \Sigma_-^e$ from Proposition 4. In the language of [3], the outer factor Σ_-^e will assume the form

$$\Sigma_-^e : \begin{cases} \dot{x}_1 = \mathbf{a}_1(x_1, x_2) + \mathbf{b}_1(x_1, x_2)u, \\ \dot{x}_2 = \mathbf{a}_2(x_1, x_2) + \mathbf{b}_2(x_1, x_2)u, \\ \quad y = \mathbf{c}(x_1, x_2) + u, \end{cases}$$

and we write $P(x) = P(x_1, x_2), P^-(x) = P^-(x_1, x_2)$. We assume that $P^-(x_1, x_2)$, $P(x_1, x_2), \mathbf{a}_1, \mathbf{a}_2, \mathbf{b}_1$, and $\mathbf{b}_2$ have the special form

$$\begin{aligned} P^-(x_1, x_2) &= P_1(x_1) + P_2(x_2), \\ P(x_1, x_2) &= P_2(x_2), \\ \mathbf{b}_1(x_1, x_2) &= b_1(x_1), \mathbf{b}_2(x_1, x_2) = b_2(x_2), \\ a^{U_1}(x_1) &:= \mathbf{a}_1(x_1, x_2) \\ &\quad - b_1(x_1)\mathbf{c}(x_1, x_2) - b_1(x_1)b_1(x_1)^T P_{1,x_1}(x_1)^T \\ &\quad - b_1(x_1)b_2(x_2)^T P_{2,x_2}(x_2)^T \\ &\quad \text{(a function only of } x_1\text{)}, \\ a^{U_2}(x_2) &:= \mathbf{a}_2(x_1, x_2) \\ &\quad - b_2(x_2)\mathbf{c}(x_1, x_2) - b_2(x_2)b_2(x_2)^T P_{2,x_2}(x_2)^T \\ &\quad \text{(a function only of } x_2\text{)} \end{aligned} \tag{41}$$

for some single-variable functions $P_1(x_1), P_2(x_2), b_1(x_1), b_2(x_2), a^{U_1}(x_1)$, and $a^{U_2}(x_2)$. Whenever $P(x)$ is a solution of the Hamilton–Jacobi equation for which there is a decomposition $x = (x_1, x_2)$ for which all the relations (41) are satisfied, we shall say that P is a *decomposable solution* of the Hamilton–Jacobi equation (6), and that the associated minimal, stable factor S of Φ given earlier is a *decomposable minimal, stable, spectral factor*.

The following result gives an explicit correspondence between decomposable, minimal, stable, spectral factors S of Φ and coprime inner-inner factorizations $\Theta' \circ \Theta''$.

THEOREM 14. *Let P^- be a smooth solution of* (10) *with stability side condition* (11) *and X^- satisfying* (7). *Furthermore, assume that P^+ is a smooth solution of* (13) *with stability side condition* (14) *and X^+ satisfying* (7). *A bijective correspondence exists between minimal, external, decomposable, nonsquare spectral factors of Φ and coprime inner-inner factorizations $\Theta' \circ \Theta''$ of the form given by* (40).

Proof. Let S be a minimal stable, nonsquare external spectral factor of Φ. By Theorem 7 and Proposition 13, we know that the factorization $\Theta' \circ \Theta''$ exists. From Definition 12, this factorization is also coprime inner-inner.

In order to prove the converse statement, let $\Theta' \circ \Theta''$ be any coprime inner-inner factorization. The existence of such factorizations is guaranteed by Proposition 13. Next, define $S = \Theta'' \circ \Sigma_-^e$. It is clear that S is a stable spectral factor of Φ—since $\Sigma_+^e = \Theta' \circ S$, it has to be minimal. Also, it is immediately apparent that the correspondence between S and coprime inner-inner factorizations $\Theta' \circ \Theta''$ is bijective. □

A complete parametrization of the set of all nonsquare minimal stable decomposable spectral factors is given in the following theorem.

THEOREM 15. *There is a bijective correspondence between the set of all minimal, stable, decomposable, nonsquare spectral factors S of Φ and factorizations of the form*

$$\Theta_1'\Theta_2'\Theta_2''\Theta_1'' : \begin{cases} \dot{x} = a(x) - b(x)d^T(x)F^{-1}(x)c_1(x) \\ \qquad - b(x)d(x)^T(F^{-1}(x))^2 d(x)b(x)^T P_x^-(x)^T \\ \qquad + (b(x)d(x)^T F^{-1}(x)X_x^-(x))u, \\ y = \begin{pmatrix} -F^{-1}(x)d(x)b(x)^T(P_x^-(x) - P_x^+(x))^T \\ \qquad -(X_x^-(x) - X_x^+(x))^T P_x^-(x)^T \end{pmatrix} \\ \qquad + \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} u, \end{cases}$$

with $\Theta_+ = \Theta_r \circ \Theta_c$ given by (37), *P^- and P^+ satisfying* (10) *and* (13), *respectively, and X^- and X^+ satisfying* (7). *Furthermore, we have that $\Theta_2' \circ \Theta_2''$ is a coprime inner-inner factorization.*

Proof. The proof of this result follows directly from Theorems 7 and 14 above. □

5. An illustrative example. By way of illustration we discuss how the robust stabilization problem that was treated in [33] (see also [6]) by using the certainty equivalence principle can be solved in terms of concepts related to the spectral factorization problem discussed in this paper. Before we present the example, we provide some background from nonlinear H_∞-control that will prove useful. Consider the nonlinear plant $P : \binom{w}{u} \to \binom{z}{y}$ with state space realization

$$P : \begin{cases} \dot{x} = A(x) + B_1(x)w + B_2(x)u, & w \in \mathbf{R^{n_w}}, \quad u \in \mathbf{R^{n_u}}, \\ z = C_1(x) + D_{12}(x)u, & z \in \mathbf{R^{n_z}}, \\ y = C_2(x) + D_{21}(x)w, & y \in \mathbf{R^{n_y}}, \end{cases} \tag{42}$$

where w is a reference and/or disturbance, u is the control, z is the error, and y is a measurement. The nonlinear H_∞-problem is to design a dynamic compensator $K : y \to u$ of the form

$$K : \begin{cases} \dot{x}_K = A_K(x_K) + B_K(x_K)y, \\ u = C_K(x_K) + D_K(x_K)y, \end{cases} \tag{43}$$

with the properties that (42) and (43) are internally stable and the closed-loop input-output map with $x(0) = 0$ and $x_K(0) = 0$ has L_2-gain at most γ, i.e.,

$$\|z\|_2 \leq \gamma \|w\|_2$$

for all $w \in L_2^{n_w}(\mathbf{R}^+)$. In the above, the assumption is that

$$A(0) = 0,\ C_1(0) = 0,\ C_2(0) = 0,\ A_K(0) = 0,\ C_K(0) = 0,$$

and all functions $A, B_1, B_2, C_1, C_2, D_{12}$, and D_{21} are smooth. Internal stability may be taken in either the input-output sense or the internal state space sense.

From [6] (see also [5]), we have that, under the additional assumptions that $D_{21}(x)$ is square and invertible for all x and $A(x) - B_1(x)D_{21}(x)^{-1}C_2(x)$ is asymptotically stable, the disturbance feedforward case of the nonlinear H_∞-control problem mentioned earlier can be reduced to a spectral factorization problem via an inner-outer factorization problem. This problem can, in turn, be reduced to solving a Hamilton–Jacobi equation with stabilizing side condition. More specifically, solutions of the type given by K in (43) of the disturbance feedforward case of the nonlinear H_∞-control problem for the plant P in (42) can be related to an inner-outer factorization of the form

$$G = \Omega \circ \Sigma_-.$$

Here G has the same trajectories as P and leads to a Hamilton–Jacobi equation with stabilizing side condition, Ω is an inner system, and Σ_- is a stable minimum phase (outer) spectral factor that is at least right-sided invertible (see Proposition 1). As is well known, it may be difficult to find an explicit solution for the aforementioned Hamilton–Jacobi equation. However, in the example below, we demonstrate that with additional assumptions on the plant P, the associated Hamilton–Jacobi equation may be solved explicitly.

For the ensuing example, we consider the zero-state detectable system

$$\Omega : \begin{cases} \dot{x} = q(x) + r(x)u, & q(0) = 0, \quad u \in \mathbf{R^m}, \\ y = s(x), & s(0) = 0, \quad y \in \mathbf{R^m}, \end{cases} \tag{44}$$

where q and s are smooth functions and x is contained in the state manifold $\mathcal{X}$. We assume that Ω is an inner system with the property that there exists a storage function $H : \mathcal{X} \to \mathbf{R}$ with $H(0) = 0$ and $H(x) > 0$ for $x > 0$ such that

$$\frac{d}{dt}\{H(x(t))\} = u(t)^T y(t) \tag{45}$$

over all trajectories $(u(t), x(t), y(t))$ of Ω or, equivalently,

$$H_x(x)q(x) = 0,\ H_x(x)r(x) = s(x)^T.$$

The perturbation Ω_p of Ω that comes from the normalized kernel representation of Ω (see [33]) has the form

$$P\colon \begin{cases} \dot{x} = q(x) + r(x)u + r(x)w, w \in \mathbf{R^m}, \\ z = \begin{pmatrix} u \\ y \end{pmatrix}, \\ y = s(x)^T H_x(x)^T + w. \end{cases} \tag{46}$$

We are now in a position to state the robust stabilization problem (see [33]) that we will solve by using a spectral factorization approach.

Problem. Construct a measurement feedback law $K : y \to u$ to minimize the L_2-gain from w to z for the resulting closed-loop system.

First, we introduce a related system G that has the same trajectories as P in (46) but with inputs equal to (u, y) and outputs equal to (z, w). In this case, G may be given by

$$G: \begin{cases} \dot{x} = [q(x) - r(x)r(x)^T H_x(x)^T] + r(x)u + r(x)w, \\ z = \begin{pmatrix} u \\ y \end{pmatrix}, \\ \widetilde{w} = -\gamma s(x)^T H_x(x)^T + \gamma y. \end{cases} \tag{47}$$

With G in (47), we can associate the Hamilton–Jacobi equation

$$P_x(x)[q(x) - (1-\gamma^2)^{-1} r(x)r(x)^T H_x(x)^T] - \frac{1}{2}[(1-\gamma^2)^{-1} + 1] \times P_x(x)r(x)r(x)^T P_x(x)^T - \frac{1}{2}\gamma^2(1-\gamma^2)^{-1} H_x(x)r(x)r(x)^T H_x(x)^T = 0$$

with stability side condition

$$q(x) - (1-\gamma^2)^{-1} r(x)r(x)^T H_x(x)^T - (2-\gamma^2)(1-\gamma^2)^{-1} r(x)r(x)^T P_x(x)^T$$

as being asymptotically stable. Next, we note that (48) has the solution

$$P(x) = \gamma^2(\gamma^2 - 2)^{-1} H(x)$$

for $\gamma > \sqrt{2}$. From this, it is possible to find an explicit formula for a stable minimum phase (outer) spectral factor Σ_- (compare Σ_- in (9); see details in [6]) in the inner-outer factorization $G = \Omega \circ \Sigma_-$ given by

$$\Sigma_-: \begin{cases} \dot{x} = [q(x) - r(x)r(x)^T H_x(x)^T] + r(x)u + r(x)w, \\ \begin{pmatrix} u' \\ y' \end{pmatrix} = \overline{d}(x) \left[\begin{pmatrix} \gamma^2(\gamma^2-2)^{-1} r(x)^T H_x(x)^T \\ \gamma^2(\gamma^2-2)^{-1}(\gamma^2-1)^{-1} r(x)^T H_x(x)^T \end{pmatrix} + \begin{pmatrix} u \\ y \end{pmatrix} \right], \end{cases}$$

where $\overline{d}(x) = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{\gamma^2 - 1} \end{pmatrix}$. In [6] it was found that the solution K_c (also known as the central compensator) of the robust stabilization problem may be deduced from this outer spectral factor Σ_- by putting $u' = 0$. In fact, an explicit formula for this controller, which corresponds exactly to the one in [33], is given by

$$K_c: \begin{cases} \dot{x} = q(\widehat{x}) - r(\widehat{x})r(\widehat{x})^T H_x(\widehat{x})^T \\ \qquad - \gamma^2(\gamma^2-2)^{-1} r(\widehat{x})r(\widehat{x})^T H_x(\widehat{x})^T + r(\widehat{x})y \\ \quad = [q(\widehat{x}) - \gamma^2(\gamma^2-2)^{-1} r(\widehat{x})r(\widehat{x})^T H_x(\widehat{x})^T] \\ \qquad + r(\widehat{x})[y - r(\widehat{x})^T H_x(\widehat{x})^T]y, \\ u = -\gamma^2(\gamma^2-2)^{-1} r(\widehat{x})r(\widehat{x})^T H_x(\widehat{x})^T. \end{cases}$$

6. Conclusions and future directions. This paper enables us to obtain nonlinear analogues of results in [11, 12, 13, 21] and [22], where state space formulas for nonsquare spectral factors of given size and related unitary or inner functions were derived. We use the Ball–Helton–van der Schaft nonlinear factorization theory [1, 2, 6, 14] and several recent contributions [3, 20, 25, 26, 27] to establish the equivalence between stable nonsquare spectral factors and coprime inner-inner factorizations of inner systems. In addition, we demonstrate the utility of the spectral factorization problem discussed in the above by applying it to issues that arise in the context of the nonlinear H_∞-control problem. More specifically, we investigate how the robust stabilization problem discussed in [33] (see also [6]) can be solved in terms of concepts related to the spectral factorization problem.

Several interesting open problems involving nonlinear spectral factorization (and the related inner-outer factorization) remain. For instance, a characterization of the relationship between the spectral factorization problem for nonlinear control systems and the control of mechanical systems requires further investigation (see [17] and [32]). Another issue that requires attention in the future is the solvability of the Hamilton–Jacobi equations. Here we assume throughout that the solutions of said equations are smooth. If this assumption is removed, we will probably have to work with viscosity-type solutions. Furthermore, we would like to obtain analogues of the results presented in this paper in a nonminimal setting. We suspect that in such a situation, it may still be true that inner systems may be expressed in terms of the solutions of Hamilton–Jacobi equations. A consideration of spectral factors that are nonstable, where a functional Hamilton–Jacobi equation may be used instead of the one used in earlier discussions appears to be another interesting topic to be researched. The establishment of further connections between nonlinear systems and control theory and financial mathematics (see [24] and [28]) seems to be a fertile area of endeavor related to our discussions in this paper. For instance, it has recently been discovered that it is possible to attach an economic interpretation to the state space associated with (minimal) realizations of certain input-affine interest rate models. As a result, standard factorization theory has a significant part to play in advancing that field of research.

REFERENCES

[1] J. A. Ball and J. W. Helton, *Factorization and general properties of nonlinear Toeplitz operators*, Oper. Theory Adv. Appl., Birkhäuser-Verlag, Basel, 41 (1989), pp. 25–41.

[2] J. A. Ball and J. W. Helton, *Inner-outer factorization of nonlinear operators*, J. Funct. Anal., 104 (1992), pp. 363–413.

[3] J. A. Ball and M. A. Petersen, *Nonlinear minimal square spectral factorization*, Internat. J. Control, 76 (2003), pp. 1233–1247.

[4] J. A. Ball, M. A. Petersen, and A. J. van der Schaft, *Inner-outer factorization for nonlinear noninvertible control systems*, IEEE Trans. Automat. Control, 49 (2004), pp. 483–492.

[5] J. A. Ball and M. Verma, *Factorization and feedback stabilization for nonlinear systems*, Systems Control Lett., 23 (1991), pp. 187–196.

[6] J. A. Ball and A. J. van der Schaft, *J-inner-outer factorization, J-spectral factorization and robust control for nonlinear systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 379–392.

[7] A. Ben-Artzi and J. W. Helton, *Riccati Partial Differential Equations for Factoring Nonlinear Systems*, preprint.

[8] A. Ben-Artzi and J. W. Helton, *The Riccati PDEs associated with invariant distributions and minimal factorization of systems*, in Proceedings of the American Control Conference, San Diego, CA, 1 (1990), pp. 993–995.

[9] P. E. Crouch and A. J. van der Schaft, *Variational and Hamiltonian Control Systems*, Lecture Notes in Control and Inform. Sci. 101, Springer-Verlag, Berlin, 1987.

[10] F. J. Doyle, F. Allgöwer, and M. Morari, *A normal form approach to approximate input-output linearization for maximum phase nonlinear systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 305–309.

[11] P. A. Fuhrmann, *On the characterization and parametrization of minimal spectral factors*, J. Math. Systems Estim. Control, 5 (1995), pp. 383–444.

[12] P. A. Fuhrmann and A. Gombani, *On a Hardy space approach to the analysis of spectral factors*, Internat. J. Control, 71 (1998), pp. 277–357.

[13] P. A. Fuhrmann and A. Gombani, *On the Lyapunov equation, coinvariant subspaces and some problems related to spectral factorizations*, Internat. J. Control, 73 (2000), pp. 1129–1159.

[14] J. W. Helton, *Factorization of nonlinear systems*, in The Gohberg Anniversary Collection, Vol. II, Oper. Theory Adv. Appl. 41, Birkhäuser-Verlag, Basel, 1989, pp. 311–328.

[15] D. J. Hill and P. J. Moylan, *Dissipative dynamical systems: Basic input and state properties*, J. Franklin Inst., 309 (1980), pp. 322–357.

[16] L. Lerer, M. A. Petersen, and A. C. M. Ran, *Existence of minimal nonsquare J-symmetric factorizations for self-adjoint rational matrix functions*, Linear Algebra Appl., 379 (2003), pp. 159–178.

[17] R. Lozano, B. Brogliato, O. Egeland, and B. Mascke, *Dissipative Systems Analysis and Control: Theory and Applications*, Comm. Control Eng. Ser., Springer-Verlag, London, 2000.

[18] A. D. B. Paice, *Stabilization and Identification of Nonlinear Systems*, doctoral dissertation, Australian National University, Canberra, 1992.

[19] A. D. B. Paice and A. J. van der Schaft, *The class of stabilizing nonlinear plant controller pairs*, IEEE Trans. Automat. Control, 41 (1996), pp. 634–645.

[20] M. A. Petersen, *On nonlinear (j, J)-inner systems*, in Proceedings of the IEEE's 3rd International Conference on Control Theory and Applications, Pretoria, South Africa, 2001, pp. 231–235.

[21] M. A. Petersen and A. C. M. Ran, *Nonsquare minimal spectral factors*, Linear Algebra Appl., 351/352 (2002), pp. 553–565.

[22] M. A. Petersen and A. C. M. Ran, *Nonsquare spectral factors via factorizations of unitary functions*, Linear Algebra Appl., 351/352 (2002), pp. 567–583.

[23] M. A. Petersen and A. C. M. Ran, *Minimal nonsquare J-spectral factorization, generalized Bezoutians and common zeros of rational matrix functions*, Integral Equations Operator Theory, 47 (2003), pp. 197–216.

[24] M. A. Petersen, H. Raubenheimer, F. van der Walt, and H. van Rooy, *Stochastic controllability of linear interest rate models*, Oper. Theory Adv. Appl., 149 (2004), pp. 501–523.

[25] M. A. Petersen and A. J. van der Schaft, *Nonsquare spectral factorization for nonlinear control systems*, IEEE Trans. Automat. Control, to appear.

[26] M. A. Petersen and A. J. van der Schaft, *On a connection between nonlinear nonsquare spectral factors and Hamilton–Jacobi equations*, in Nonlinear Control Systems 3, A. B. Kurzhanski and A. L. Fradkov, eds., IFAC, 2002, pp. 1493–1499.

[27] M. A. Petersen and A. J. van der Schaft, *On nonlinear inner systems and connections with control theory*, in Proceedings of the 15th IFAC World Congress, Barcelona, Spain, 2002, pp. 1656–1661.

[28] M. A. Petersen, H. Raubenheimer, and J. H. van Schuppen, *Control of pension funds with stochastic control theory*, submitted.

[29] J. M. A. Scherpen, *Balancing for nonlinear systems*, Systems Control Lett., 21 (1993), pp. 143–153.

[30] J. M. A. Scherpen and A. J. van der Schaft, *The normalized left and right coprime factorization and balancing for nonlinear systems*, Internat. J. Control, 60 (1994), pp. 1193–1222.

[31] A. J. van der Schaft, *L_2-gain analysis of nonlinear systems and nonlinear state feedback H_∞-control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.

[32] A. J. van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*, Lecture Notes in Control and Inform. Sci. 218, Springer-Verlag, London, 1996.

[33] A. J. van der Schaft, *Robust stabilization of nonlinear systems via stable kernel representations with L_2 gain bounded uncertainty*, Systems Control Lett., 24 (1995), pp. 75–81.

[34] J. C. Willems, *Least squares stationary optimal control and the algebraic Ricatti equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

[35] R. A. Wright and C. Kravaris, *Nonminimum-phase compensation for nonlinear processes*, AIChE J., 38 (1992), pp. 26–40.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 856–871

BACKSTEPPING WITH BOUNDED FEEDBACKS FOR TIME-VARYING SYSTEMS*

FREDERIC MAZENC† AND SAMUEL BOWONG‡

Abstract. A family of time-varying nonlinear systems is globally uniformly asymptotically stabilized by bounded feedbacks constructed through a new extension of the backstepping approach. Explicit expressions of control laws and Lyapunov functions are given.

Key words. backstepping, bounded feedback, time-varying system

AMS subject classifications. 93D05, 93D15, 93D20

DOI. 10.1137/S0363012902408733

1. Introduction. One of the most popular nonlinear techniques of design of control laws is the backstepping approach. The multiple advantages offered by it are well known. Observe in particular that this technique yields a wide family of globally asymptotically stabilizing control laws, and it allows one to address robustness issues and to solve adaptive problems. However, for a long time, it was a widely held belief that this technique could not be used to solve the problem of designing feedbacks bounded in norm, which in many practical situations should be addressed: For instance, the possibility of actuator saturation or constraints on actuators imposes bounded input. But it turns out that, as a matter of fact, the backstepping approach can be adapted to the problem of designing bounded feedbacks. In three recent works [22, 2, 10], it is shown that for some time-invariant systems (an n-dimensional chain of integrators, for instance), bounded stabilizing feedbacks can be constructed by applying new versions of this technique: The approach of [22, 2] mainly relies on the nested saturation control laws proposed in [18, 20], and the approach of [10] mainly relies on the determination of a particular family of control Lyapunov functions. However, for families of time-varying systems, no bounded backstepping method has ever been developed, and the main results of [22, 2] and [10] cannot be straightforwardly extended.

In the present work, we address the problem of constructing globally uniformly asymptotically stabilizing differentiable bounded feedbacks and accompanying strict Lyapunov functions, using the backstepping approach for time-varying systems of the following form:

$$\begin{cases} \dot{x} = f(t,x) + g(t,x)z, \\ \dot{z} = p(t)(u + b(t,x,z)) \end{cases} \tag{1}$$

with $x \in R^{n_x}$, $z \in R$, where $u \in R$ is the input, $p(t)$ is a bounded function of t, and $f(t,x)$ and $b(t,x,z)$ satisfy $f(t,0) = 0$, $b(t,0,0) = 0$ for all t.

In the particular case where $p(t)$ is a continuous function larger (resp., smaller) than a strictly positive real number (resp., a strictly negative real number), then a

*Received by the editors June 4, 2002; accepted for publication (in revised form) February 9, 2004; published electronically September 18, 2004.
http://www.siam.org/journals/sicon/43-3/40873.html
†Projet MERE INRIA-INRA, UMR LASB, INRA 2, pl. Viala, 34 060 Montpellier, France (mazenc@helios.ensam.inra.fr).
‡INRIA Lorraine, Projet CONGE, ISGMP Bât. A, Ile du Saulcy, 57 045 Metz Cedex 01, France (bowong@loria.fr).

Lyapunov design of bounded feedbacks can be carried out, for instance, by combining the results of [23, 24] and [22]. But when $p(t)$ is a time-varying function which is neither strictly positive nor strictly negative, then the construction of globally uniformly stabilizing feedbacks and accompanying strict Lyapunov functions for systems (1) is a challenging open problem: To the best of our knowledge, no technique of construction of this type of Lyapunov functions is available in the literature, even in the case where the systems (1) are stabilized by unbounded control laws. We want to emphasize that in the present paper, we *will not impose* on $p(t)$ to be a function which is never equal to zero: We will only assume that $p(t)$ satisfies a persistency of excitation property and is of class C^1. Observe that the study of nonlinear time-varying systems is motivated in particular by the fact that a tracking problem for a nonlinear system can be reformulated as a stabilization problem for the time-varying error system. Through the family of chained form nonholonomic systems, we will show in section 4 how tracking problems for nonlinear systems may lead to the study of systems of the form (1), where $p(t)$ is a function which takes positive and negative values, and how, by applying the main result of the present work repeatedly, one can solve the open problem of determining explicit expressions of globally uniformly asymptotically and locally exponentially stabilizing bounded feedbacks and of accompanying strict Lyapunov functions for time-varying chains of integrators, which in turn implies that one can solve the problem of constructing globally uniformly asymptotically and locally exponentially stabilizing bounded feedbacks and accompanying strict Lyapunov functions for error equations of systems in chained form.

The approach we propose relies extensively on two results. On the one hand, we exploit the family of changes of coordinates used in [10] to obtain explicit expressions of globally uniformly asymptotically stabilizing bounded feedbacks. On the other hand, we construct explicitly strict Lyapunov functions using the main result of [9].

Observe that the strict Lyapunov functions (which at the same time are control Lyapunov functions) we will construct are not just tools enabling us to establish the asymptotic stability of the closed-loop system: The knowledge of continuously differentiable strict Lyapunov functions can be of great help. The potential benefits they offer are so numerous that they cannot be exhaustively enumerated. However, observe in particular the following:

- Recent advances in stabilization of nonlinear delay systems are based on the knowledge of continuously differentiable Lyapunov functions (see in particular [21, 3, 11]).
- Lyapunov functions are known to be very efficient tools for robustness analysis: For example, many proofs of nonlinear disturbance-to-state L^p stability properties rely on Lyapunov functions (see [19, 8]). Moreover, the control Lyapunov function–based theory has provided control designs with guaranteed robustness to different types of disturbances, including deterministic [1] and stochastic [6], as well as the robustness to unmodeled dynamics [13, 15].
- When a control Lyapunov function satisfying the small control property is available, one can apply universal formulas, in particular the one proposed in [16], and obtain that way the expression of an asymptotically stabilizing feedback which is optimal with respect to the control Lyapunov function as optimal value function.

The expressions of the bounded control law and of the Lyapunov function we propose are far from being the only possible expressions that can be obtained; many different formulas can be determined. Moreover, many extensions of the result can

be proved; we have briefly mentioned some of them in the discussion of the main result given in section 3 and in the concluding remarks of section 5. For the sake of clarity, we have chosen to restrict ourselves to the systems (1); the control design can be easily carried out for them. However, it is worth noting that the key ideas of our approach can be used in several contexts beyond the scope of the present work. In particular, they can be utilized to solve the problem of constructing bounded feedbacks for systems of the form

$$\begin{cases} \dot{x} = f(t,x,z) + h(t,x,z,u)u, \\ \dot{z} = p(t)u + b(t,x,z), \end{cases} \tag{2}$$

which, due to the term $h(t,x,z,u)u$, are not in feedback form.

The paper is organized as follows. In section 2, a technical lemma is given. In section 3 the main result is stated and proved. The technique is applied to an illustrative example in section 4. Concluding remarks in section 5 end the work.

Preliminaries.

1. The argument of the functions will be omitted whenever no confusion can arise from the context.

2. We assume throughout the paper that the functions encountered are sufficiently smooth.

3. For a real-valued C^1 function $k(\cdot)$, we denote by $k'(\cdot)$ its first derivative.

4. $|x| = \sqrt{x^\top x}$ stands for the Euclidean norm of vector $x \in R^{n_x}$.

5. A function $k(\cdot) : R_{\geq 0} \to R_{\geq 0}$ is of class $\mathcal{K}_\infty$ if it is continuous, zero at zero, strictly increasing, and unbounded.

6. By S, we denote the set of the functions $\sigma : R \to R$ such that

(a) $\sigma(s)$ is a bounded function,

(b) $s\sigma(s)$ is positive definite,

(c) $s\sigma(s) \leq s^2$,

(d) $\sigma'(s)$ is nonnegative and bounded and $\sigma'(0) = 1$.

7. A function $V(\cdot)$ is a strict Lyapunov function for the time-varying system

$$\dot{\chi} = \varphi(t,\chi)$$

if there exists a positive definite function $W(\chi)$ such that, for all t and χ,

$$\frac{\partial V}{\partial \chi}(t,\chi)\varphi(t,\chi) + \frac{\partial V}{\partial t}(t,\chi) \leq -W(\chi)$$

and there exist two functions $\Gamma_1(\cdot), \Gamma_2(\cdot)$ of class $\mathcal{K}_\infty$ such that, for all t and χ,

$$\Gamma_1(|\chi|) \leq V(t,\chi) \leq \Gamma_2(|\chi|).$$

2. Technical result. In this section, we give a technical result which will be used in the next section to prove the main result of the work. We construct a strict Lyapunov function for the one-dimensional time-varying system

$$\dot{\xi} = -q(t)\sigma(\xi), \tag{3}$$

where $q(t)$ is a nonnegative function of class C^1 such that

$$0 \leq q(t) \leq \delta_1 \qquad \forall t, \tag{4}$$

$$\int_t^{t+T} q(s)ds \geq \delta_2 > 0 \qquad \forall t, \tag{5}$$

where δ_1, δ_2, and T are positive real numbers and where $\sigma(\cdot)$ belongs to the set S defined in the preliminaries. We carry out the construction by adapting the approach of [9] to the case where $q(t)$ is not necessarily a periodic function of t but satisfies the persistency of excitation condition (5). First, observe that the property $0 \leq s\sigma(s) \leq s^2$ implies that the function $\frac{|\sigma(s)|}{|s|}$ is bounded. Moreover, $\sigma'(\cdot)$ is bounded and (4) is satisfied. It follows that

$$M := \sup_{t\in R}\left[T + \left|\int_t^{t+T}(s-t-T)q(s)ds\right| \sup_{s\in R, s\neq 0}\left(\frac{|\sigma(s)| + |s\sigma'(s)|}{|s|}\right)\right] \tag{6}$$

is finite and positive. We are now in position to give a technical lemma.

LEMMA 2.1. *The function*

$$\nu(t,\xi) := (M+1)\xi^2 + \left(\int_t^{t+T}(s-t-T)q(s)ds\right)\xi\sigma(\xi) \tag{7}$$

is a strict Lyapunov function for the system (3).

REMARK 1. *When $q(t)$ is a periodic function, then $\nu(t,\xi)$ is a periodic function of t as well, and these functions have the same period.*

Proof. The derivatives of the functions

$$R_1(t,\xi) := \left(\int_t^{t+T}(s-t-T)q(s)ds\right)\xi\sigma(\xi), \qquad R_2(\xi) := \frac{1}{2}\xi^2 \tag{8}$$

along the trajectories of (3) satisfy

$$\begin{aligned}\dot{R}_1 &= \left[-\int_t^{t+T} q(s)ds + Tq(t)\right]\xi\sigma(\xi)\\ &\quad - \left[\int_t^{t+T}(s-t-T)q(s)ds\right][\sigma(\xi) + \xi\sigma'(\xi)]\,q(t)\sigma(\xi),\\ \dot{R}_2 &= -q(t)\xi\sigma(\xi).\end{aligned} \tag{9}$$

Further, since

$$\nu(t,\xi) = 2(M+1)R_2(\xi) + R_1(t,\xi), \tag{10}$$

it follows from (9), (6), and (5) that

$$\begin{aligned}\dot{\nu}(t,\xi) &\leq -\left(\int_t^{t+T} q(s)ds\right)\xi\sigma(\xi) + Mq(t)\xi\sigma(\xi) - 2(M+1)q(t)\xi\sigma(\xi)\\ &\leq -\delta_2\xi\sigma(\xi) < 0 \qquad \forall \xi \neq 0.\end{aligned} \tag{11}$$

Moreover, (7) and (6) imply that

$$(M+1)\xi^2 - M\xi^2 \leq \nu(t,\xi) \leq (M+1)\xi^2 + M\xi^2, \tag{12}$$

which results in

$$\xi^2 \leq \nu(t,\xi) \leq (2M+1)\xi^2. \tag{13}$$

According to (11) and (13), the function $\nu(t,\xi)$ is a strict Lyapunov function for the system (3). This concludes the proof. ∎

3. Main result. In this section, we state and prove the main result of the paper. Consider the nonlinear time-varying system (1). We introduce a set of assumptions.

ASSUMPTION A1. *The functions $p(t)$ and $\dot{p}(t)$ are bounded in norm by a positive real number P and two positive numbers T and γ such that, for all t,*

$$\int_t^{t+T} p(s)^2 ds \geq \gamma > 0 \tag{14}$$

are known.

ASSUMPTION A2. *Let ε be a positive real number and n a nonnegative integer. A Lyapunov function $V(t,x)$ such that*

$$\alpha_1(|x|) \leq V(t,x) \leq \alpha_2(|x|), \qquad \left|\frac{\partial V}{\partial x}(t,x)\right| \leq \alpha_3(|x|), \tag{15}$$

where the $\alpha_i(\cdot)$'s are functions of class $\mathcal{K}_\infty$, a positive definite function $W(x)$, and a feedback $z_s(t,x) := p(t)^{n+2}\mu_s(t,x)$, bounded in norm by ε such that $\mu_s(t,0) = 0$ and

$$\frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)[f(t,x) + g(t,x)z_s(t,x)] \leq -W(x), \tag{16}$$

are known. Moreover, the functions

$$|\mu_s(t,x)|, \quad \left|\frac{\partial \mu_s}{\partial t}(t,x)\right|, \quad \left|\frac{\partial \mu_s}{\partial x}(t,x)f(t,x)\right|, \quad \left|\frac{\partial \mu_s}{\partial x}(t,x)g(t,x)\right|, \quad |b(t,x,z)| \tag{17}$$

are bounded.

ASSUMPTION A3. *A real-valued function $\zeta(\cdot)$ such that $\zeta(s) > 0$ for all $s \neq 0$ and $\int_0^r \zeta(s)ds$ is of class $\mathcal{K}_\infty$, a function $\alpha_4(\cdot)$ of class $\mathcal{K}_\infty$, and a nonnegative function $\beta(\cdot)$ such that the inequalities*

$$\zeta(V(t,x))\left|\frac{\partial V}{\partial x}(t,x)g(t,x)\right|^2 \leq \frac{1}{2}W(x), \tag{18}$$

$$|f(t,x)| \leq \alpha_4(|x|)\ ,\ |g(t,x)| \leq \beta(|x|) \tag{19}$$

are satisfied for all t, x are known.

ASSUMPTION A3′. *The function $W(x)$ is such that, for a real number $c_1 > 0$,*

$$W(x) \geq c_1|x|^2 \qquad \forall x : |x| \leq 1. \tag{20}$$

THEOREM 3.1. *Assume that the system* (1) *satisfies Assumptions* A1, A2, *and* A3. *Then the system* (1) *is globally uniformly asymptotically stabilizable by a bounded feedback $u_s(t,x,z)$ such that, for all t, $u_s(t,0,0) = 0$. For the corresponding closed-loop system, a strict Lyapunov function can be constructed. This Lyapunov function belongs to the family of functions of the form*

$$U(t,x,z) = l(V(t,x)) + k(\nu(t,\Omega(z) - z_s(t,x))) \tag{21}$$

with

$$\nu(t,Z) = (M+1)Z^2 + \left(\int_t^{t+T}(s-t-T)p(s)^{2m}ds\right)Z\sigma(Z), \tag{22}$$

where m is a positive integer, $l(\cdot), k(\cdot)$ are functions of class $\mathcal{K}_\infty$, and $\Omega(\cdot)$ is a real-valued function zero at zero such that $\Omega'(z) \geq 1$ for all z. If in addition Assumption A3′ *is satisfied, the system* (1) *is globally uniformly asymptotically and locally exponentially stabilizable by a bounded feedback $u_s(t,x,z)$ such that, for all t, $u_s(t,0,0) = 0$, and a strict Lyapunov function for the corresponding closed-loop system with a derivative along the trajectories upper bounded on a neighborhood of the origin by a negative definite quadratic function of (x,z) can be constructed. This Lyapunov function belongs to the family of functions* (21).

Discussion of Theorem 3.1.

- All the real-valued periodic functions of class C^1 which are not identically equal to zero satisfy Assumption A1. In the particular case where, for all t, $p(t) > 0$ or $p(t) < 0$, a simpler proof than the one we shall give can be carried out by taking advantage of the change of feedback $v = p(t)(u + b(t,x,z))$. But assuming that, for all t, $p(t) > 0$ or $p(t) < 0$ is very restrictive.
- The boundedness property of the functions in (17) in Assumption A2 and the growth property in Assumption A3 are not surprising assumptions: In the time-invariant case, similar assumptions have been imposed (see [2, 10]). Due to the finite escape time phenomenon, they cannot be removed without being replaced by other assumptions.
- Assumption A3′ ensures that the feedback $z_s(t,x)$ not only globally uniformly asymptotically stabilizes the origin of x-subsystem of (1) but also locally exponentially stabilizes it.
- In the formula of the stabilizing feedback we shall construct (see (27)), the function $V(t,x)$ is not involved: So it turns out that the control design strategy we propose can be applied even when the function $V(t,x)$ is not accurately known.
- An important issue is whether or not Theorem 3.1 can be applied recursively. In general, it appears that the assumptions will not be satisfied repeatedly because the presence of $b(t,x,z)$ in the expression of the control law we will construct (see (27)) typically prevents $u_s(t,x,z)$ and its derivatives along the trajectories from vanishing with $p(t)$. However, in particular cases, Theorem 3.1 can be applied recursively. Basically, this can be done for systems of the form

$$\begin{cases} \dot{x} = f(t,x,z_1), \\ \dot{z}_1 = p_1(t)z_2 + b_1(t,x,z), \\ \quad \vdots \\ \dot{z}_n = p_n(t)u + b_n(t,x,z), \end{cases} \tag{23}$$

  when the $b_i(t,x,z)$'s are identically equal to zero or when, roughly speaking, they are sufficiently "small": Indeed, since one can construct explicitly a globally uniformly asymptotically stabilizing feedback with an accompanying strict Lyapunov function for a system (23) in absence of the $b_i(t,x,z)$'s, one can take advantage of these tools to determine in a second step how "small" the terms $b_i(t,x,z)$, regarded as disturbances, should be for not destroying the stability properties of the system stabilized by the control law constructed in their absence. It is quite clear that the choice of the feedback at each step plays an important role in this approach. In particular at each step the control law must anticipate the $p_i(t)$'s that follow: One understands from the mechanism of the control design used to prove Theorem 3.1 that a possible strategy

of design for the system (23) consists in repeatedly constructing feedbacks, which for convenience we denote $z_{i,f}(t,x,z_1,\dots,z_{i-1})$ for $i=2$ to $n+1$, such that $z_{i,f}(t,x,z_1,\dots,z_{i-1}) = p_i(t)^2\dots p_n(t)^{n-i+2}\psi_i(t,x,z_1,\dots,z_{i-1})$, where $\psi_i(t,x,z_1,\dots,z_{i-1})$'s are sufficiently smooth functions. We will not present a rigorous and complete study of this problem; it would require pages of simple but lengthy calculations which can be inferred from the ideas of the proof of Theorem 3.1. For the sake of simplicity, we restrict ourselves to illustrating the possibility of applying the approach repeatedly by solving in section 4 the problem of stabilizing a three-dimensional chain of integrators with time-varying coefficients.

Proof of Theorem 3.1.

Step 1: *New variable.* We introduce the variable

$$Z := \Omega(z) - z_s(t,x), \tag{24}$$

where $\Omega(\cdot)$ is the function present in (21). In addition to the properties $\Omega(0)=0$ and $\Omega'(z)\geq 1$ for all z, we require that this function be such that

(a) $\Omega'(z)=1$ when $|z|\leq 2\varepsilon$,

(b) $\Omega'(z)\geq |z|$ when $|z|\geq 2\varepsilon+1$.

Its time derivative satisfies

$$\begin{aligned}\dot{Z} &= \Omega'(z)p(t)(u+b(t,x,z)) - \frac{\partial z_s}{\partial t}(t,x) - \frac{\partial z_s}{\partial x}(t,x)[f(t,x)+g(t,x)z]\\ &= \Omega'(z)p(t)(u+b(t,x,z)) + p(t)\lambda(t,x,z)\end{aligned} \tag{25}$$

with

$$\begin{aligned}\lambda(t,x,z) &= -(n+2)\dot{p}(t)p(t)^n\mu_s(t,x) - p(t)^{n+1}\frac{\partial \mu_s}{\partial t}(t,x)\\ &\quad - p(t)^{n+1}\frac{\partial \mu_s}{\partial x}(t,x)[f(t,x)+g(t,x)z].\end{aligned} \tag{26}$$

We choose for u

$$u = u_s(t,x,z) := -b(t,x,z) - \frac{p(t)^{2m-1}\sigma(Z)+\lambda(t,x,z)}{\Omega'(z)}, \tag{27}$$

where m is a positive integer and where $\sigma(\cdot)$ is a function belonging to the set S defined in the preliminaries. Such a choice of feedback yields

$$\dot{Z} = -p(t)^{2m}\sigma(Z). \tag{28}$$

One can check readily that Assumption A1 and the properties of $\sigma(\cdot)$ imply that this system is globally uniformly asymptotically and locally exponentially stable. Our objective is now to construct a strict Lyapunov function for the system (1) in closed-loop with (27) by exploiting the stability properties of (28).

Step 2: *Strict Lyapunov function for the system* (28). Using Young's inequality, one can check readily that Assumption A1 implies that for all positive integer m one can find a positive real number γ_m such that, for all t,

$$\int_t^{t+T} p(s)^{2m}ds \geq \gamma_m > 0. \tag{29}$$

Moreover, $p(t)$ is bounded in norm. It follows that Lemma 2.1 applies to the system (28): The function defined in (22), where

$$M = \sup_{t\in R}\left[T + \left|\int_t^{t+T}(s-t-T)p(s)^{2m}ds\right| \sup_{s\in R, s\neq 0}\left(\frac{|\sigma(s)| + |s\sigma'(s)|}{|s|}\right)\right], \tag{30}$$

is a strict Lyapunov function for the system (28). Its time derivative along the trajectories of (28) satisfies

$$\dot\nu(t,Z) \leq -\left(\int_t^{t+T} p(s)^{2m}ds\right) Z\sigma(Z) \leq -\gamma_m Z\sigma(Z) < 0 \qquad \forall Z \neq 0. \tag{31}$$

Step 3: *Strict Lyapunov function for the system* (1). We construct a strict Lyapunov function for the system (1) in closed-loop with the feedback (27) by using a combination of the Lyapunov functions $V(t,x)$ and $\nu(t,Z)$. This construction is reminiscent of the constructions of Lyapunov functions presented in [17, 12]. Consider a function $U(t,x,z)$, belonging to the family of functions (21), and require that the function $k(\cdot)$ be such that $k'(s) \geq 1$ for all $s \geq 0$. Thanks to Assumption A2, the properties satisfied by the functions $\Omega(\cdot), k(\cdot), l(\cdot)$, and Lemma 2.1, one can prove that there exist two functions $\gamma_1(\cdot)$, $\gamma_2(\cdot)$ of class $\mathcal{K}_\infty$ such that

$$\gamma_1(|(x,z)|) \leq U(t,x,z) \leq \gamma_2(|(x,z)|). \tag{32}$$

The derivative of $U(t,x,z)$ along the trajectories of the closed-loop system satisfies

$$\begin{aligned}\dot U &= l'(V(t,x))\dot V + k'(\nu(t,Z))\dot\nu \\ &\leq -l'(V(t,x))W(x) + l'(V(t,x))\frac{\partial V}{\partial x}(t,x)g(t,x)(z - z_s(t,x)) \\ &\quad - k'(\nu(t,Z))\gamma_m Z\sigma(Z).\end{aligned} \tag{33}$$

From the triangular inequality, the inequality

$$\begin{aligned}\dot U \leq\ & -l'(V(t,x))W(x) + l'(V(t,x))^2\left|\frac{\partial V}{\partial x}(t,x)g(t,x)\right|^2 \\ &+ \frac{1}{4}(z - z_s(t,x))^2 - k'(\nu(t,Z))\gamma_m Z\sigma(Z)\end{aligned} \tag{34}$$

can be deduced. According to Assumption A3, a possible choice for $l(\cdot)$ is $l(r) := \int_0^r \zeta(s)ds$, since this function is of class $\mathcal{K}_\infty$. Moreover, for such a choice, the inequality

$$\dot U \leq -\frac{1}{2}\zeta(V(t,x))W(x) + \frac{1}{4}(z - z_s(t,x))^2 - \gamma_m k'(\nu(t,Z))Z\sigma(Z) \tag{35}$$

is satisfied. Next, observe that

$$Z^2 = (\Omega(z) - z_s(t,x))^2 = (\Omega(z) - \Omega(z_s(t,x)))^2 \tag{36}$$

because $|z_s(t,x)| \leq \varepsilon$ and $\Omega(s) = s$ when $|s| \leq 2\varepsilon$. It follows that

$$Z^2 = \left(\int_{z_s(t,x)}^{z} \Omega'(s)ds\right)^2. \tag{37}$$

Since $\Omega'(s) \geq 1$ for all s, the inequality

$$Z^2 \geq (z - z_s(t,x))^2 \tag{38}$$

holds. Combining (38) and (35) yields

$$\dot{U} \leq -\frac{1}{2}\zeta(V(t,x))W(x) + \frac{1}{4}Z^2 - \gamma_m k'(\nu(t,Z))Z\sigma(Z). \tag{39}$$

Thanks to the inequalities (13) and the properties of $\sigma(\cdot)$, one can easily determine a function $k(\cdot)$ such that

$$\gamma_m k'(\nu(t,Z))Z\sigma(Z) \geq \frac{1}{2}Z^2. \tag{40}$$

This inequality leads to

$$\dot{U} \leq -\frac{1}{2}\zeta(V(t,x))W(x) - \frac{\gamma_m}{2}k'(\nu(t,Z))Z\sigma(Z) \leq -N(x,z), \tag{41}$$

where $N(x,z)$ is the positive definite function

$$N(x,z) := \inf_{t\in R}\left(\frac{1}{2}\zeta(V(t,x))W(x) + \frac{\gamma_m}{2}k'(\nu(t,Z))Z\sigma(Z)\right). \tag{42}$$

It follows that $U(t,x,z)$ is a strict Lyapunov function for the system (1) in closed-loop with the feedback (27). This implies that the system (1) in closed-loop with the feedback (27) is globally uniformly asymptotically stable.

Step 4: *Boundedness of the feedback* (27). Since $\Omega'(z) \geq 1$, $|p(t)| \leq P$, and $|\dot{p}(t)| \leq P$, the inequality

$$\begin{aligned} |u| \leq\ & |b(t,x,z)| + P^{2m-1}|\sigma(Z)| + (n+2)P^{n+1}|\mu_s(t,x)| + P^{n+1}\left|\frac{\partial \mu_s}{\partial t}(t,x)\right| \\ & + P^{n+1}\left|\frac{\partial \mu_s}{\partial x}(t,x)f(t,x)\right| + P^{n+1}\left|\frac{\partial \mu_s}{\partial x}(t,x)g(t,x)\right|\left|\frac{z}{\Omega'(z)}\right| \end{aligned} \tag{43}$$

is satisfied. On the one hand, Assumption A2 ensures that the functions $|b(t,x,z)|$, $|\mu_s(t,x)|$, $|\frac{\partial \mu_s}{\partial t}(t,x)|$, $|\frac{\partial \mu_s}{\partial x}(t,x)f(t,x)|$, and $|\frac{\partial \mu_s}{\partial x}(t,x)g(t,x)|$ are bounded. On the other hand, the functions $\sigma(\cdot)$ and $\Omega(\cdot)$ have been chosen such that $|\sigma(Z)|$ and $|\frac{z}{\Omega'(z)}|$ are bounded. It follows that the feedback (27) is bounded in norm.

Step 5: *The particular case where Assumption* A3′ *is satisfied.* When (20) holds, then, according to (40), the function $N(x,z)$ defined in (42) satisfies

$$N(x,z) \geq \inf_{t\in R}\left(\frac{c_1}{2}\zeta(V(t,x))|x|^2 + \frac{1}{4}Z^2\right) \qquad \forall (x,z) : |x| \leq 1. \tag{44}$$

Assumptions A2, A3, and A3′ ensure that there exists a positive real number c_2 such that

$$c_2\left|\frac{\partial V}{\partial x}(t,x)g(t,x)\right|^2 \leq \frac{c_1}{2}|x|^2 \leq \frac{1}{2}W(x) \qquad \forall (x,z) : |x| \leq 1. \tag{45}$$

It follows that, if necessary, $\zeta(\cdot)$ can be modified in such a way that $\zeta(0) > 0$ and (18) is satisfied. In that case, $c_3 = \inf_{t\in R,|x|\leq 1}(\frac{1}{2}\zeta(V(t,x)))$ is a positive real number and the property

$$N(x,z) \geq c_1c_3|x|^2 + \frac{1}{4}\inf_{t\in R}\left(Z^2\right) \qquad \forall (x,z) : |x| \leq 1 \tag{46}$$

is satisfied. Through lengthy but simple calculations, one can deduce from (46) that there exists a positive real number c_4 such that

$$N(x,z) \geq c_4(|x|^2 + |z|^2) \qquad \forall (x,z) : |(x,z)| \leq 1. \tag{47}$$

This implies that the system (1) in closed-loop with the feedback (27) is globally uniformly asymptotically stable and locally exponentially stable. This concludes the proof. □

4. Illustration: Time-varying chain of integrators. We will illustrate Theorem 3.1 by using it to construct for the particular three-dimensional chain of integrators with time-varying coefficients

$$\begin{cases} \dot{x}_1 = \cos(t)x_2, \\ \dot{x}_2 = \cos(t)x_3, \\ \dot{x}_3 = \cos(t)u \end{cases} \tag{48}$$

a globally uniformly asymptotically and locally exponentially stabilizing bounded state feedback and a strict Lyapunov function for the corresponding closed-loop system. Before doing that, we give a motivation for it.

4.1. Motivation: Systems in chained form. For the sake of simplicity, we will restrict our attention to the system (48). But it is worth noting that Theorem 3.1 can be successfully applied repeatedly to any time-varying chain of integrators

$$\begin{cases} \dot{x}_n = p_n(t)x_{n-1}, \\ \dot{x}_{n-1} = p_{n-1}(t)x_{n-2}, \\ \qquad \vdots \\ \dot{x}_1 = p_1(t)u, \end{cases} \tag{49}$$

where the $p_i(t)$'s are bounded functions with bounded first derivatives such that the product $p_1(t)\cdots p_n(t)$ satisfies Assumption A1.

One of the motivations for solving the problem of globally uniformly asymptotically stabilizing time-varying chains of integrators by bounded feedback arises from the tracking problem for systems in chained form under input saturation. The importance of this family of systems is well known: The kinematic model of several nonholonomic systems can be transformed into a system in chained form, and a lot of interest has been devoted to the stabilization and the tracking of these systems. In particular, in [4, 5, 7] the backstepping approach has been used to achieve for these systems the global tracking of trajectories. Let us briefly recall how. Systems in chained form of order n with two inputs (see, for instance, [14]) are described by the equations

$$\begin{cases} \dot{z}_n = z_{n-1}v_1, \\ \qquad \vdots \\ \dot{z}_3 = z_2 v_1, \\ \dot{z}_2 = v_2, \\ \dot{z}_1 = v_1. \end{cases} \tag{50}$$

Assume that the trajectory to be tracked satisfies

$$\begin{cases} \dot{z}_{n,r}(t) = z_{n-1,r}(t)v_{1,r}(t), \\ \qquad\vdots \\ \dot{z}_{3,r}(t) = z_{2,r}(t)v_{1,r}(t), \\ \dot{z}_{2,r}(t) = v_{2,r}(t), \\ \dot{z}_{1,r}(t) = v_{1,r}(t) \end{cases} \tag{51}$$

and is bounded. Then, after the change of feedbacks $v_1 = v_{1,r}(t)+u_1, v_2 = v_{2,r}(t)+u_2$, and denoting $v_{1,r}(t)$ simply by $p(t)$, the error equation is

$$\begin{cases} \dot{z}_{n,e} = p(t)z_{n-1,e} + z_{n-1}u_1, \\ \qquad\vdots \\ \dot{z}_{3,e} = p(t)z_{2,e} + z_2u_1, \\ \dot{z}_{2,e} = u_2, \\ \dot{z}_{1,e} = u_1, \end{cases} \tag{52}$$

where $z_{i,e} = (z_i - z_{i,r}(t))$ for all $i = 1$ to n. Assume that for the chain of integrators

$$\begin{cases} \dot{z}_{n,e} = p(t)z_{n-1,e}, \\ \qquad\vdots \\ \dot{z}_{3,e} = p(t)z_{2,e}, \\ \dot{z}_{2,e} = u_2 \end{cases} \tag{53}$$

there are a bounded control law $u_2(t, z_e)$ with $z_e = (z_{2,e}, \ldots, z_{n,e})$, a strict Lyapunov function $V_e(t, z_e)$, and a positive definite function $W_e(z_e)$ such that the derivative of $V_e(\cdot)$ along the trajectories of (53) in closed-loop with $u_2(t, z_e)$ satisfies

$$\dot{V}_e \leq -W_e(z_e). \tag{54}$$

Then the derivative of the function

$$U_e(t, z_e, z_{1,e}) = V_e(t, z_e) + \frac{1}{2}z_{1,e}^2 \tag{55}$$

along the trajectories of (52) in closed-loop with $u_2(t, z_e)$ and the bounded feedback

$$u_1(t, z_{1,e}, z_e) = -\frac{\frac{\partial V_e}{\partial z_{n,e}}(t, z_e)z_{n-1} + \cdots + \frac{\partial V_e}{\partial z_{3,e}}(t, z_e)z_2 + z_{1,e}}{1 + \left(\frac{\partial V_e}{\partial z_{n,e}}(t, z_e)z_{n-1}\right)^2 + \cdots + \left(\frac{\partial V_e}{\partial z_{3,e}}(t, z_e)z_2\right)^2 + z_{1,e}^2} \tag{56}$$

satisfies

$$\dot{U}_e \leq -W_e(z_e) - \frac{\left(\frac{\partial V_e}{\partial z_{n,e}}(t, z_e)z_{n-1} + \cdots + \frac{\partial V_e}{\partial z_{3,e}}(t, z_e)z_2 + z_{1,e}\right)^2}{1 + \left(\frac{\partial V_e}{\partial z_{n,e}}(t, z_e)z_{n-1}\right)^2 + \cdots + \left(\frac{\partial V_e}{\partial z_{3,e}}(t, z_e)z_2\right)^2 + z_{1,e}^2}. \tag{57}$$

One can check readily that this implies that $U_e(\cdot)$ is a strict Lyapunov function for the system (52) in closed-loop with the bounded feedbacks $u_1(t, z_{1,e}, z_e)$, $u_2(t, z_e)$.

Consequently, we have shown that the problem of determining globally uniformly asymptotically stabilizing bounded feedbacks and strict Lyapunov functions for error equations resulting from the problem of tracking a bounded trajectory of a system in chained form can be reduced to the problem of determining globally uniformly asymptotically stabilizing bounded feedbacks for time-varying chains of integrators (49) and accompanying strict Lyapunov functions.

4.2. Control design for the system (48). We begin the construction with a preliminary result which will be used throughout the remainder of the section.

Preliminary result. By applying Lemma 2.1, one can prove that the derivative of the function

$$V(t,x) = 80x^2 + \left(\int_t^{t+2\pi} (s-t-2\pi)\cos^6(s)ds\right)\frac{x^2}{\sqrt{1+x^2}} \tag{58}$$

along the trajectories of

$$\dot{x} = -\cos^6(t)\frac{x}{\sqrt{1+x^2}} \tag{59}$$

satisfies

$$\dot{V}(t,x) \le -\frac{5\pi}{8}\frac{x^2}{\sqrt{1+x^2}}. \tag{60}$$

Moreover, $V(t,x)$ is periodic of period 2π and

$$70x^2 \le V(t,x) \le 90x^2. \tag{61}$$

We are ready now to carry out the backstepping construction of a bounded feedback for (48) by applying repeatedly the main result of section 3.

The x_1-subsystem. According to the preliminary result, the time derivative of the function $V(t,x_1)$ along the trajectories of (48) satisfies

$$\begin{aligned}\dot{V} &\le -\frac{5\pi}{8}\frac{x_1^2}{\sqrt{1+x_1^2}} + \frac{\partial V}{\partial x_1}(t,x_1)\cos(t)\left[x_2 + \cos^5(t)\frac{x_1}{\sqrt{1+x_1^2}}\right] \\ &\le -\frac{5\pi}{8}\frac{x_1^2}{\sqrt{1+x_1^2}} + 180|x_1|\left|x_2 + \cos^5(t)\frac{x_1}{\sqrt{1+x_1^2}}\right|.\end{aligned} \tag{62}$$

The x_2-subsystem. Consider the variable

$$X_2 = \Omega_a(x_2) + \cos^5(t)\frac{x_1}{\sqrt{1+x_1^2}}, \tag{63}$$

where $\Omega_a(\cdot)$ is the odd function

$$\Omega_a(r) = \int_0^r \left(1 + \max\{0, 9(|s|-2)^3\}\right) ds. \tag{64}$$

An immediate calculation yields

$$\begin{aligned}\dot{X}_2 &= \Omega_a'(x_2)\cos(t)x_3 - 5\sin(t)\cos^4(t)\frac{x_1}{\sqrt{1+x_1^2}} + \cos^6(t)\frac{x_2}{(1+x_1^2)^{\frac{3}{2}}} \\ &= -\cos^6(t)\frac{X_2}{\sqrt{1+X_2^2}} + \Omega_a'(x_2)\cos(t)x_3 + \cos^6(t)\frac{X_2}{\sqrt{1+X_2^2}} \\ &\quad - 5\sin(t)\cos^4(t)\frac{x_1}{\sqrt{1+x_1^2}} + \cos^6(t)\frac{x_2}{(1+x_1^2)^{\frac{3}{2}}}.\end{aligned} \tag{65}$$

Considering x_3 as a fictitious input v and choosing for it

$$v(t,x_1,x_2) = \frac{-\cos^5(t)\frac{X_2}{\sqrt{1+X_2^2}} + 5\sin(t)\cos^3(t)\frac{x_1}{\sqrt{1+x_1^2}} - \cos^5(t)\frac{x_2}{(1+x_1^2)^{\frac{3}{2}}}}{\Omega_a'(x_2)}, \tag{66}$$

the dynamics (65) become

$$\dot{X}_2 = -\cos^6(t)\frac{X_2}{\sqrt{1+X_2^2}}. \tag{67}$$

The derivative of

$$U(t,x_1,X_2) = l(V(t,x_1)) + k(V(t,X_2)) \tag{68}$$

satisfies

$$\begin{aligned}
\dot{U} \leq & -\frac{5\pi}{8}l'(V(t,x_1))\frac{x_1^2}{\sqrt{1+x_1^2}} + 180l'(V(t,x_1))|x_1|\left|x_2 + \cos^5(t)\frac{x_1}{\sqrt{1+x_1^2}}\right| \\
& - \frac{5\pi}{8}k'(V(t,X_2))\frac{X_2^2}{\sqrt{1+X_2^2}} \\
\leq & -\frac{5\pi}{8}l'(V(t,x_1))\frac{x_1^2}{\sqrt{1+x_1^2}} + 180l'(V(t,x_1))|x_1||X_2| \\
& - \frac{5\pi}{8}k'(V(t,X_2))\frac{X_2^2}{\sqrt{1+X_2^2}} \\
\leq & -\frac{\pi}{2}l'(V(t,x_1))\frac{x_1^2}{\sqrt{1+x_1^2}} + \frac{2}{\pi}180^2\sqrt{1+x_1^2}l'(V(t,x_1))X_2^2 \\
& - \frac{5\pi}{8}k'(V(t,X_2))\frac{X_2^2}{\sqrt{1+X_2^2}}.
\end{aligned} \tag{69}$$

Choosing for $l(\cdot)$

$$l(r) = \sqrt{(80-4\pi^2)+r} - \sqrt{80-4\pi^2} \tag{70}$$

leads to

$$\begin{aligned}
\dot{U} \leq & -\frac{\pi}{2}l'(V(t,x_1))\frac{x_1^2}{\sqrt{1+x_1^2}} + \frac{180^2\sqrt{1+x_1^2}}{\pi\sqrt{(80-4\pi^2)+(80-4\pi^2)x_1^2}}X_2^2 \\
& - \frac{5\pi}{8}k'(V(t,X_2))\frac{X_2^2}{\sqrt{1+X_2^2}}.
\end{aligned} \tag{71}$$

Choosing for $k(\cdot)$

$$k(r) \;=\; 480\left[\left((80-4\pi^2)+r\right)^{\frac{3}{2}} - \left(80-4\pi^2\right)^{\frac{3}{2}}\right] \tag{72}$$

leads to

$$\dot{U} \leq -\frac{\pi}{2}l'(V(t,x_1))\frac{x_1^2}{\sqrt{1+x_1^2}} - \frac{\pi}{2}k'(V(t,X_2))\frac{X_2^2}{\sqrt{1+X_2^2}}. \tag{73}$$

The overall system. From the above analysis, it results that the system (48) is equivalent to the system

$$\begin{cases}
\dot{x}_1 \;=\; \cos(t)x_2, \\
\dot{X}_2 \;=\; -\cos^6(t)\dfrac{X_2}{\sqrt{1+X_2^2}} + \Omega_a'(x_2)\cos(t)(x_3 - v(t,x_1,x_2)), \\
\dot{x}_3 \;=\; \cos(t)u
\end{cases} \tag{74}$$

with $x_2 = \Omega_a^{-1}(X_2 - \cos^5(t)\frac{x_1}{\sqrt{1+x_1^2}})$. Using the inequality

$$\begin{aligned}(75)\qquad &\frac{(k^{-1})'(U(t,x_1,X_2))}{\sqrt{1+k^{-1}(U(t,x_1,X_2))}}\left[\left|\frac{\partial U}{\partial X_2}((t,x_1,X_2))\right| + \left|\frac{\partial U}{\partial x_1}((t,x_1,X_2))\right|\right]\\ &\qquad \le\ c\left(l'(V(t,x_1))\frac{x_1^2}{\sqrt{1+x_1^2}} + k'(V(t,X_2))\frac{X_2^2}{\sqrt{1+X_2^2}}\right),\end{aligned}$$

where c is a positive constant,[1] and observing that the function $\int_0^r \frac{(k^{-1})'(s)}{\sqrt{1+k^{-1}(s)}}ds$ is of class $\mathcal{K}_\infty$, one can prove that Assumptions A1, A2, A3, and A3′ of Theorem 3.1 are satisfied by the system (74) with $(x_1, X_2)^\top$ playing the role of x, x_3 playing the role of z, and $U(\cdot)$ playing the role of $V(\cdot)$. It follows that the construction of a globally uniformly asymptotically and locally exponentially stabilizing bounded feedback for the system (74) can be achieved. The last part of the section is devoted to this construction.

The function $v(t, x_1, x_2)$ defined in (66) satisfies

$$(76)\qquad |v(t,x_1,x_2)| \le \frac{6+|x_2|}{1+\max\{0,(|x_2|-2)^3\}} \le 10.$$

These inequalities lead one to consider the change of variable

$$(77)\qquad X_3 = \Omega_b(x_3) - v(t,x_1,x_2),$$

where $\Omega_b(\cdot)$ is the odd function defined as

$$(78)\qquad \Omega_b(r) = \int_0^r \left(1+\max\{0,21(|s|-20)^3\}\right)ds.$$

Its time derivative satisfies

$$(79)\qquad \dot{X}_3 = \Omega_b'(x_3)\cos(t)u - \dot{v}(t,x_1,x_2)$$

with

$$\begin{aligned}\dot{v} &= \cos^2(t)\zeta(t,x_1,x_2,x_3),\\ \zeta &= \frac{5\left[\sin(t)\cos^2(t)\left(\frac{X_2}{\sqrt{1+X_2^2}}+\frac{x_2}{(1+x_1^2)^{\frac{3}{2}}}\right)+(\cos^4(t)+3\sin^2(t))\frac{x_1}{\sqrt{1+x_1^2}}\right]}{1+\max\{0,9(|x_2|-2)^3\}}\\ &\quad + \frac{\cos^2(t)}{1+\max\{0,9(|x_2|-2)^3\}}\left[\frac{5\sin(t)}{(1+x_1^2)^{\frac{3}{2}}}x_2 - \frac{3\cos^2(t)x_2^2}{(1+x_1^2)^{\frac{5}{2}}} - \cos^3(t)\frac{x_3}{(1+x_1^2)^{\frac{3}{2}}}\right]\\ &\quad - \frac{\cos^4(t)\left[(1+\max\{0,9(|x_2|-2)^3\})x_3 - 5\sin(t)\cos^3(t)\frac{x_1}{\sqrt{1+x_1^2}}+\cos^5(t)\frac{x_2}{(1+x_1^2)^{\frac{3}{2}}}\right]}{(1+\max\{0,9(|x_2|-2)^3\})(1+X_2^2)^{\frac{3}{2}}}\\ &\quad + \cos(t)\frac{\cos^2(t)\frac{X_2}{\sqrt{1+X_2^2}} - 5\sin(t)\frac{x_1}{\sqrt{1+x_1^2}}+\cos^2(t)\frac{x_2}{(1+x_1^2)^{\frac{3}{2}}}}{(1+\max\{0,9(|x_2|-2)^3\})^2}\max\{0,27(|x_2|-2)^2\}.\end{aligned}\qquad(80)$$

[1] The explicit value of c can be determined through lengthy calculations. But an explicit value is useless; to carry out the proof, it is only required that we know that c exists.

One can check readily that the feedback

$$
\begin{aligned}
u &= \frac{-\cos^5(t)\frac{X_3}{\sqrt{1+X_3^2}} + \cos(t)\zeta(t,x_1,x_2,x_3)}{\Omega_b'(x_3)} \\
&= \frac{-\cos^5(t)\frac{X_3}{\sqrt{1+X_3^2}} + \cos(t)\zeta(t,x_1,x_2,x_3)}{1+\max\{0,21(|x_3|-20)^3\}}
\end{aligned} \tag{81}
$$

yields

$$
\dot{X}_3 = -\cos^6(t)\frac{X_3}{\sqrt{1+X_3^2}}. \tag{82}
$$

Moreover, the feedback u is bounded:

$$
\begin{aligned}
|u| &\le \frac{1+|\zeta(t,x_1,x_2,x_3)|}{1+\max\{0,21(|x_3|-20)^3\}} \\
&\le 25 + \frac{6|x_2|+x_2^2}{1+\max\{0,9(|x_2|-2)^3\}} + \frac{2|x_3|}{1+\max\{0,21(|x_3|-20)^3\}} \le 94.
\end{aligned} \tag{83}
$$

REMARK 2. *The feedback we have constructed is bounded in norm by* 94. *But for all* $\varepsilon > 0$, *one can modify the design in such a way that the resulting control law is bounded by* ε *instead of* 94.

5. Conclusion. We have proposed a new extension of the backstepping technique which applies to time-varying nonlinear systems and thereby can be utilized in particular for solving problems of global tracking. We have proposed families of bounded control laws. We have not explored all the possible extensions of the approach: We want to emphasize that the key ideas of the technique are even more important than the results themselves. Much remains to be done; robustness and disturbance attenuation issues and applications to the control design for systems with delay are some issues that may be pursued.

Acknowledgments. The authors are grateful to the referees and L. Praly for their insightful remarks and suggestions.

REFERENCES

[1] R. FREEMAN AND P. KOKOTOVIC, *Robust Control Nonlinear Systems,* Birkhäuser Boston, Cambridge, MA, 1996.

[2] R. FREEMAN AND L. PRALY, *Integrators backstepping for bounded controls and control rates,* IEEE Trans. Automat. Control, 43 (1998), pp. 258–262.

[3] M. JANKOVIC, *Control Lyapunov-Razumikhin functions and robust stabilization of time delay systems,* IEEE Trans. Automat. Control, 46 (2001), pp. 1048–1060.

[4] Z. P. JIANG AND H. NIJMEIJER, *A recursive technique for tracking control of nonholonomic systems in chained form,* IEEE Trans. Automat. Control, 42 (1999), pp. 265–279.

[5] Z. P. JIANG AND H. NIJMEIJER, *Tracking control of mobile robots: A case study in backstepping,* Automatica J. IFAC, 33 (1997), pp. 1393–1399.

[6] M. KRISTIC AND H. DENG, *Stabilization of Nonlinear Uncertain Systems,* Springer-Verlag, Berlin, 1998.

[7] E. LEFEBER, *Tracking Control of Nonlinear Mechanical Systems,* Ph.D. thesis, University of Twente, Twente, The Netherlands, 2000.

[8] W. LIU, Y. CHITOUR, AND E. SONTAG, *On finite-gain stabilizability of linear systems subject to input saturation,* SIAM J. Control Optim., 34 (1996), pp. 1190–1219.

[9] F. Mazenc, *Strict Lyapunov functions for time-varying systems,* Automatica J. IFAC, 39 (2003), pp. 349–353.

[10] F. Mazenc and A. Iggidr, *Backstepping with bounded feedbacks*, Systems Control Lett., 51 (2004), pp. 235–245.

[11] F. Mazenc and S. Niculescu, *Lyapunov stability analysis for nonlinear delay systems,* Systems Control Lett., 42 (2001), pp. 245–251.

[12] F. Mazenc and L. Praly, *Adding an integration and global asymptotic stabilization of feedforward systems,* IEEE Trans. Automat. Control, 41 (1996), pp. 1559–1578.

[13] L. Praly and Y. Wang, *Stabilization in spite of matched unmodeled dynamics and an equivalent definition of input-to-state stability,* Math. Control Signals Systems, 9 (1996), pp. 1–33.

[14] C. Samson, *Control of chained systems application to path following and time-varying point-stabilization of mobile robots,* IEEE Trans. Automat. Control, 40 (1995), pp. 64–77.

[15] R. Sepulchre, M. Jankovic, and P. V. Kokotovic, *Constructive Nonlinear Control,* Springer-Verlag, Berlin, 1996.

[16] E. Sontag, *A "universal" construction of Artstein's theorem on nonlinear stabilization,* Systems Control Lett., 13 (1989), pp. 117–123.

[17] E. Sontag and A. Teel, *Changing supply functions in input/state stable system,* IEEE Trans. Automat. Control, 40 (1995), pp. 1476–1478.

[18] A. Teel, *Using saturation to stabilize a class of single-input partially linear composite systems,* in Proceedings of the IFAC Nonlinear Control Systems Design Symposium, Bordeau, France, 1992, pp. 369–374.

[19] A. Teel, *On L_2 performance induced by feedbacks with multiple saturations,* ESAIM Control Optim. Calc. Var., 1 (1996), pp. 225–240.

[20] A. Teel, *Feedback Stabilization: Nonlinear Solutions to Inherently Nonlinear Problems,* Technical Report UCB/ERLM92/65, University of California, Berkeley, 1992.

[21] A. Teel, *Connections between Razumikhin-type theorems and the ISS nonlinear small gain theorems,* IEEE Trans. Automat. Control, 43 (1998), pp. 960–964.

[22] J. Tsinias, *Input to state stability properties of nonlinear systems and applications to bounded feedback stabilization using saturation,* ESAIM Control Optim. Calc. Var., 2 (1997), pp. 57–85.

[23] J. Tsinias, *Backstepping design for time-varying systems and application to partial-static feedback and asymptotic tracking,* Systems Control Lett., 39 (2000), pp. 219–227.

[24] J. Tsinias and I. Karafyllis, *ISS property for time-varying systems and application to partial-static feedback stabilization and asymptotic tracking,* IEEE Trans. Automat. Control, 44 (1999), pp. 2179–2184.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 872–898

LEARNING COMPLEXITY DIMENSIONS FOR A CONTINUOUS-TIME CONTROL SYSTEM*

PIRKKO KUUSELA†, DANIEL OCONE†, AND EDUARDO D. SONTAG†

Abstract. This paper takes a computational learning theory approach to a problem of linear systems identification. It is assumed that inputs are generated randomly from a known class consisting of linear combinations of k sinusoidals. The output of the system is classified at some single instant of time. The main result establishes that the number of samples needed for identification with small error and high probability, independently from the distribution of inputs, scales polynomially with n, the system dimension, and logarithmically with k.

Key words. linear systems identification, learning theory, VC dimension

AMS subject classifications. 68Q32, 68Q17, 93B30, 93C05

DOI. 10.1137/S0363012901384302

1. Introduction.

Systems identification may be regarded as an instance of the general problem of "learning" an unknown function. Computational learning theory (CLT), which provides a theory for understanding the complexity of a learning problem, can then be used to obtain new insight into identification; conversely, the input/output maps associated to systems theory supply interesting new families of classifiers to consider in CLT. The early paper by Ljung [16] already explored the connection between CLT and identification. Independently, the papers [6, 7] had already developed learning theory complexity results for discrete-time linear systems acting on finite-window data. However, continuous-time linear systems have not been much explored from the CLT viewpoint. The immediate problem is that even for finite-length inputs, the family of maps associated to a continuous-time linear system is not "learnable" in the precise mathematical sense defined in probably approximately correct (PAC) learning theory; for continuous-time systems, the Vapnik–Chervonenkis (VC) dimension, which measures the learning rate, is infinite. This was proved in [23] and, alternatively, can be derived by applying the discrete-time results from [6, 7] at higher and higher sampling rates. However, if we have prior information on the system we wish to learn, or if we are interested in less than the full input/output map, we can ask if the identification problem becomes learnable in the CLT setting, and if so, how many samples are needed. This is the issue motivating the work of this paper.

In practice, it is not necessary or feasible to learn how a linear system classifies every continuous-time input; it usually suffices to know how it acts on a sufficiently rich class. In this paper we consider continuous-time systems acting on linear combinations of k sinusoidals, which constitute a class of inputs used in assessing frequency responses. The parameter k will then measure the richness of the class. To keep things

*Received by the editors January 30, 2001; accepted for publication (in revised form) February 6, 2004; published electronically September 18, 2004.
http://www.siam.org/journals/sicon/43-3/38430.html

†Department of Mathematics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854-8019 (pirkko.kuusela@hut.fi, ocone@math.rutgers.edu, sontag@math.rutgers.edu). The first author is now at Networking Laboratory, Helsinki University of Technology, P.O. Box 3000, FIN-02015 HUT, Finland; research financed in part by grants from the Academy of Finland, the Finnish Academy of Sciences and Letters, and the Finnish Cultural Foundation. The third author was supported in part by U.S. Air Force grant F49620-01-1-0063 and NSF grant CCR-0206789.

as simple as possible, we focus our attention on two cases: (a) the map that gives the sign of the output, observed at a single instant of time, namely, zero-one classifiers; and (b) the full output observed at a single instant of time. For the learning theory framework, we have opted for the cleanest setting. In the training stage, inputs are generated randomly from our class of sinusoidals, and our object is to classify further randomly drawn inputs correctly with high probability.

Variations on the above set-up, such as allowing selection of inputs (called active learning) or modeling noise in the observations, can also be formulated. It is important to emphasize that, in this work, we ignore observations at time instants other than the last. This simplifies the problem and makes it easier to formulate our question in the framework of learning theory. However, a full treatment of the identification question in that framework will require further substantial research. One possibility is to see the additional information afforded by data at intermediate instants as "side information" available to the learner. In general, one expects side information to speed learning, so our results here will bound any learning rates that incorporate it. However, quantifying the advantage of side information in a general learning theory framework is a very challenging problem. A preliminary study of how side information can affect learning rates is carried out in [14, 15], in a simple model of learning intervals. One insight of this study is that the side information advantage very much depends on the probability distributions assumed on the training data and the structure of the problem. Because of this, the possibility of distribution free, model-independent side information improvement in learning rates, based on general complexity measures such as VC dimension, is not even clear; in any case, it is an open and difficult problem.

In summary, the system identification problem is posed as a noise-free parametric function identification problem with observations $\{(G_1, S_1), \ldots, (G_n, S_n)\}$, where $G_1, \ldots, G_n$ are independent random variables defining the input as a linear combination of basis input functions $\omega = (\omega_1, \ldots, \omega_k)^T$ and $S_i = \text{sign}(\Phi_\Sigma(G_i\omega; 1))$, where sign $z = 0$ if $z \leq 0$ and sign $z = 1$ if $z > 0$. Here $\Sigma(u; t)$ is the output of the linear system $\Sigma = (A, B, C, x^0)$ at time t when the input is u.

Our results show that the upper bound of samples needed to learn (i.e., identify) Σ in the above setting with a small error and high probability, independently of the distribution of G_i, scales logarithmically with the "bandwidth" k. The sample bound is analogous to the discrete-time case, in which k appeared as the length of the window employed. Also, we provide lower bounds on the number of samples needed. Hence the results can also be seen as unlearnability results where the difficulty arises from the richness of the input signals. Our second class of results concerns systems, with additional assumptions, in which the full output is observed at some time τ.

Our problem setting is selected so that it corresponds to the most standard learning scenario, which serves as the basic problem. In particular, our focus will be on calculating certain complexity dimensions that determine the number of samples needed to learn, given required accuracy and confidence. The CLT framework can accommodate various learning paradigms, and they share a number of common features. In particular, sample complexities are derived from complexity dimensions. Hence, the complexity dimension estimates from the basic problem can be utilized easily also in modified learning settings. We give more pointers to modified problems in the next section.

For papers combining learning theoretic ideas to control theory, see [11, 9] and references there as well as [12] for several results for nonlinear systems in discrete time. The reader is referred to [19] for results that apply to a class of nonlinear continuous-

time systems but which are formulated in terms of learning derivatives evaluated at a particular instant (as opposed to time data).

This paper is organized in a top-down fashion. We give definitions and main results in section 2, and in section 3 we state main upper and lower bounds for the complexity dimensions. After that we concentrate on proving the results; central techniques are discussed in section 4, and proofs are in sections 5 and 6. An example of a class with VC dimension k is given in section 7.

2. Definitions and statement of main results. The simplest learning setting is concept learning, in which there is some known concept class (e.g., "cars") and some target concept (e.g., "a sports car") we wish to learn from a sequence of N randomly chosen observations. Each observation is classified by some "oracle" that knows the target concept. After N classified observations we are required to form an estimate for the unknown target concept so that with high confidence, specified by parameter δ, the misclassification probability for a future unseen sample is smaller than a given level ϵ. The concept class is *learnable* if we can form an estimate that achieves any given confidence level, δ, and misclassification accuracy, ϵ, by taking enough observations. In this case, the number $s(\epsilon, \delta)$ of observations needed to achieve the confidence and misclassification levels is called the *sample complexity*. With this definition, learnability is equivalent to the finiteness of the VC dimension, which describes the "richness" of the concept class; VC dimension together with confidence parameter δ and accuracy parameter ϵ determine the sample complexity $s(\epsilon, \delta)$.

In the following subsections, we provide formal definitions, frame a linear system learning problem, and state some main results.

2.1. VC dimension and fat-shattering dimension. We begin by defining the key complexity dimension for this work.

Definition 2.1 (Vapnik–Chervonenkis dimension). *The richness of the collection $\mathcal{C}$ can be measured by its Vapnik–Chervonenkis (VC) dimension introduced in* [20]. *A set $S = \{x_1, \ldots, x_n\} \subseteq X$ is said to be* shattered *by $\mathcal{C}$ if, for every subset $B \subseteq S$, there exists a set $A \in \mathcal{C}$ such that $S \cap A = B$. The* VC dimension of $\mathcal{C}$, *denoted $VC(\mathcal{C})$, equals the largest integer n such that there exists a set of cardinality n that is shattered by $\mathcal{C}$.*

For example, in $\mathbb{R}^k$ the VC dimension of closed half-spaces through the origin is k [22]. Thus, if $\mathrm{VC}(\mathcal{C}) = d$, $\mathcal{C}$ is not rich enough to distinguish all subsets of any $d+1$ element set, but there is some d element set where subsets can be distinguished. Proving exact values of the VC dimension is hard, and typically one looks for upper and lower bounds for the VC dimension, as is also done in this paper.

For our purposes, it is more convenient to work with shattering in terms of dichotomies, i.e., Boolean-valued maps. We identify subsets of D with Boolean functions $\phi : D \to \{0, 1\}$. Similarly, each set $C \in \mathcal{C}$ gives rise to a Boolean function on X, and intersections $C \cap D$ are restrictions of functions to D. In this language, a subset $D \subset X$ is shattered by $\mathcal{F} := \{\phi;\ \phi : X \to \{0, 1\}\}$ if every dichotomy on D is a restriction to D of some $\phi \in \mathcal{F}$.

The VC dimension characterizes learnability of $\{0, 1\}$-valued functions, as formulated in section 2.2. For learning real-valued functions we look for a generalization of the VC dimension with similar properties. One such generalization is the pseudo-dimension. Unfortunately, pseudodimension does not share the property the VC dimension has; there are learnable function classes with infinite pseudodimension; see [21, p. 206] and [3].

DEFINITION 2.2 (pseudodimension with respect to a loss function). *Given a class of functions $\mathcal{F} : X \to Y$ and a loss function $L : Y \times Y \to [0, r]$, we introduce for each $f \in \mathcal{F}$ the function*

$$A_{f,L} : X \times Y \times \mathbb{R} \to \{0,1\}; \quad (x, y, \rho) \mapsto \operatorname{sign}(L(f(x), y) - \rho), \tag{1}$$

and let $A_{\mathcal{F},L}$ denote all such $A_{f,L}$ with $f \in \mathcal{F}$. The pseudodimension of $\mathcal{F}$ with respect to the loss function L, *$PD[\mathcal{F}, L]$, is defined as*

$$PD[\mathcal{F}, L] := VC(A_{\mathcal{F},L}).$$

Next we define the fat-shattering dimension that corresponds to shattering with fixed "margin" γ. Both the pseudodimension and the fat-shattering dimension can be used to bound certain covering numbers, and in this sense they act like the VC dimension. Moreover, the fat-shattering dimension gives upper and lower bounds for covering numbers of function classes, and the finiteness of the fat-shattering dimension can characterize learnability (see [1] and [2]).

DEFINITION 2.3 (fat-shattering dimension). *Let $\mathcal{F}$ be a set of real-valued functions. We say that a set of points X is* γ-shattered by $\mathcal{F}$ *if there are real numbers r_x indexed by $x \in X$ such that for all binary vectors b_x indexed by X, there is a function $f_b \in \mathcal{F}$ satisfying*

$$\begin{aligned} f_b(x) &\geq r_x + \gamma \quad \text{if } b_x = 1 \text{ and} \\ f_b(x) &\leq r_x - \gamma \quad \text{otherwise.} \end{aligned}$$

The fat-shattering dimension $\operatorname{fat}_\gamma(\mathcal{F})$ *is a function from positive real numbers to integers which maps a value γ to the size of the largest fat-shattered set if it is finite or infinity otherwise.*

The shattering dimension when the margin γ equals 0 *is called the* pseudodimension, *and it is denoted by $PD(\mathcal{F})$. Clearly, for all $\gamma > 0$, $\operatorname{fat}_\gamma(\mathcal{F}) \leq PD(\mathcal{F})$.*

2.2. Learning. In this section we discuss the learning setting more formally, beginning with a general introduction to classification problems. In section 2.3.1 we also indicate briefly modified learning settings that can be relevant in control problems.

Assume that a set X, to be called the *input space*, is given together with a collection $\mathcal{C}$ of mappings $X \to \{0,1\}$.[1] Let W be the set of all sequences

$$w = (u_1, \phi(u_1)), \ldots, (u_s, \phi(u_s))$$

over all $s \geq 1$, $(u_1, \ldots, u_s) \in X^s$, and let $\phi \in \mathcal{C}$. An *identifier* is a map $\psi : W \to \mathcal{C}$. The value of ψ on a sequence w above is denoted as ψ_w instead of $\psi(w)$. The "error" of ψ is the probability that ψ will misclassify a future sample. More formally, the *error of ψ with respect to a probability measure P on X, a $\phi \in \mathcal{C}$, and a sequence* $(u_1, \ldots, u_s) \in X^s$, is

$$\operatorname{Err}(P, \phi, u_1, \ldots, u_s) := P\{u \in X;\ \psi_w(u) \neq \phi(u)\}.$$

[1] The set X is assumed to be either countable or a Euclidean space, and the maps in $\mathcal{C}$ are assumed to be measurable. In addition, a mild regularity assumption called "permissibility" is needed so that all sets appearing below are measurable; for further discussion on the topic, see an appendix in [17]. In our context the measurability assumptions are satisfied.

The class $\mathcal{C}$ is said to be PAC *learnable* if there is some identifier ψ with the following property: For each *accuracy parameter* $\epsilon > 0$ and *confidence parameter* $\delta > 0$ there is some s so that, for every probability P and every $\phi \in \mathcal{C}$,

$$P^s\{(u_1, \ldots, u_s) \in X^s;\ \mathrm{Err}(P, \phi, u_1, \ldots, u_s) > \epsilon\} < \delta,$$

where P^s is the s-fold product of P. In the learnable case, the function $s(\epsilon, \delta)$ which provides the smallest s achieving for any positive ϵ and δ is called the *sample complexity*. It can be proved that learnability is equivalent to the finiteness of the *VC dimension* $\mathrm{VC}(\mathcal{C})$ of the class $\mathcal{C}$. Moreover, for learning algorithms that classify the observed samples correctly, the sample complexity is bounded by [18]

$$s(\epsilon, \delta) \leq \max\left\{ \frac{1}{\epsilon(1-\sqrt{\epsilon})} \left(2\mathrm{VC}(\mathcal{C}) \ln\left(\frac{6}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right), \frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right)\right\}.$$

In addition, there is a similar lower bound for the sample complexity.

Classification may be viewed as a problem of identifying systems with binary outputs. More generally, we introduce a problem of identification for systems having bounded outputs ($[0,1]$-valued, for technical reasons) via an L^1-error, following [4, 5] (for similar statements with L^2-error see [2]). Denote by $\mathcal{F}$ a class of mappings from X to $[0,1]$.

By definition, an *identifier* is a mapping from $\cup_{s\in\mathbb{N}} (X \times [0,1])^s$ to $[0,1]^X$. Such a map takes as data a sequence of labeled samples and produces a hypothesis. If h is a $[0,1]$-valued function defined on X and P is a probability measure over $X \times [0,1]$, we define the *error of h with respect to P* as

$$\mathrm{Er}_P(h) := \int_{X\times[0,1]} |h(u) - y|\, dP(u,y).$$

For $\epsilon > 0$ and $\delta > 0$ we say that an identifier ψ (ϵ, δ)*-learns in the agnostic sense with respect to $\mathcal{F}$ from s examples* if, for all distributions P on $X \times [0,1]$,

$$P^s\{w;\ \mathrm{Er}_P(\psi_w) \geq \inf_{f\in\mathcal{F}} \mathrm{Er}_P(f) + \epsilon\} < \delta.$$

Similarly, for $\epsilon > 0$ the function class $\mathcal{F}$ is said to be ϵ*-agnostically learnable* if there is a function $s_0 : (0,1) \to \mathbb{N}$ such that, for all $0 < \delta < 1$, there is an identifier ψ which (ϵ, δ)-learns in the above sense with s_0 samples. In addition, if the identifier always chooses a hypothesis from $\mathcal{F}$, we say that $\mathcal{F}$ is *properly ϵ-agnostically learnable*.

For learning $[0,1]$-valued functions, a sample complexity result may be stated in terms of the fat-shattering dimension. For $\epsilon > 0$ and $\delta > 0$ there is an identifier ψ that properly (ϵ, δ)-learns in the agnostic sense with respect to $\mathcal{F}$ from [4, 5]

$$\frac{4}{\alpha^2}\left(\frac{6d}{\ln 2}\ln\frac{7}{\alpha}\left(\frac{336e}{\alpha^3\ln 2}\ln\frac{7}{\alpha}\right) + \ln\frac{8}{\delta}\right) = O\left(\frac{1}{\alpha^2}\left(d\log^2\frac{1}{\alpha} + \log\frac{1}{\delta}\right)\right)$$

samples, where $0 < \alpha < \epsilon/4$ is chosen so that $d = \mathrm{fat}_{\epsilon/4-\alpha}(\mathcal{F})$ is finite. The quantity $\mathrm{fat}_\gamma(\mathcal{F})$ is called the *fat-shattering dimension* of the class $\mathcal{F}$, and it measures the richness of the class $\mathcal{F}$ with scale γ.

The sample complexity results show us that the difficulty of system identification in the learning theoretic setting can be analyzed by studying various complexity dimensions, and deriving bounds on the complexity dimension is the main focus of this paper.

2.3. Linear systems. In the context of learning we discuss continuous-time linear control systems

$$\dot{x} = Ax + Bu, \qquad x(0) = x^0, \qquad y = Cx, \tag{2}$$

where A, B, and C are $n \times n$, $n \times m$, and $p \times n$ real matrices, and the time interval is [0,1]. We study sign-observations (see [13] for related work in control theory)

$$\text{sign } y(1) = (\text{sign } y_1(1), \ldots, \text{sign } y_p(1))^T,$$

where sign $z = 0$ if $z \leq 0$, sign $z = 1$ if $z > 0$, and ^{T} stands for the transpose. For scalar observations this is a classification problem; each output is classified as either 0 or 1. The value of the final time plays no role in the results and is taken to be 1 for notational convenience.

Unlike the VC dimension associated to discrete-time linear systems [6, 12], the VC dimension of the classification problem for continuous-time control systems is unbounded [23], even when $n = 1$, and the identification problem is not learnable in the sense discussed earlier. Therefore, we restrict the class of admissible controls in order to achieve a bound for the VC dimension. We consider controls $u = (u_1, \ldots, u_m)$ such that $u = G\omega$, where G is an $m \times k$ matrix that parameterizes the control. The set of basis input functions $\Omega = \{\omega_1, \ldots, \omega_k\}$ is fixed. The bounds for the VC dimension or other complexity dimensions will depend on the properties of the set Ω.

For scalar inputs (i.e., $m = 1$) the VC dimension associated to the mapping from inputs G to scalar sign-observations is bounded by k, which in fact can be very large in applications.[2] However, by considering band-limited controls a better bound can be achieved. In this work we consider the set of basis input functions

$$\begin{aligned} \Omega = \Big\{ & \omega_1, \ldots, \omega_k; \ \omega_1, \ldots, \omega_k \text{ linearly independent and} \\ & \omega_j = t^{\ell_j} e^{\alpha_j t} \sin(\beta_j t) \text{ or } \omega_j = t^{\ell_j} e^{\alpha_j t} \cos(\beta_j t) \\ & \text{with } \ell_j \in \mathbb{N}, \alpha_j, \beta_j \in \mathbb{R}, j = 1, \ldots, k \Big\}, \end{aligned} \tag{3}$$

and let

$$\ell_{\max} = \max\{\ell_1, \ldots, \ell_k\}. \tag{4}$$

The results in this paper hold with straightforward modifications if the basis input functions $\omega_j, j = 1, \ldots, k$, are, for example, linear combinations of functions of the above form.

DEFINITION 2.4 (sign system concept class, $\mathcal{C}_{m,p}$). *Order the set of basis input functions Ω and denote $\omega = (\omega_1, \ldots, \omega_k)^T$. Let*

$$X_\Omega = \{G\omega : [0,1] \to \mathbb{R}^m; \ G \in \mathbb{R}^{mk}\},$$

and for each linear system $\Sigma = (A, B, C, x^0)$ of dimension n define the mapping $\Phi_\Sigma : X_\Omega \to \mathbb{R}^p$ by $\Phi_\Sigma(G\omega) = y(1)$, where $y(1)$ is the solution of Σ with control $u = G\omega$. Similarly, we define the mapping for sign-observations as

$$S_\Sigma : X_\Omega \to \{0,1\}^p, \quad G\omega \mapsto \text{sign}(\Phi_\Sigma(G\omega)).$$

[2] This bound is tight; we give an example of a function class Ω for which the associated VC dimension is indeed k (see section 7).

We call the class of above mappings the sign system concept class, $\mathcal{C}_{m,p} = \{S_\Sigma;\ \Sigma$ *linear system of dimension* $n\}$.

We formulate two theorems about bounding sample complexities as main results. They are immediate corollaries of learning complexity bounds proved in this paper.

THEOREM 2.5 (sample complexity for concept learning). *For sign systems concept class* $\mathcal{C}_{m,1}$ *with scalar observations, i.e.,* $p = 1$*, the sample complexity* $s(\epsilon, \delta)$ *for identifiers which agree with the observed sample can be bounded as*

$$s(\epsilon,\delta) \le \max\left\{ \frac{1}{\epsilon(1-\sqrt{\epsilon})}\left(2\,VC(\mathcal{C}_{m,1})\ln\left(\frac{6}{\epsilon}\right)+\ln\left(\frac{2}{\delta}\right)\right), \frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right)\right\},$$

where

$$VC(\mathcal{C}_{m,1}) \le 2(2mn^2+4n+1)\log_2[8e(8mn^2k(n+\ell_{\max})+1)(2nk+2(1+2k)^n)]$$

and $\ell_{\max}$ *is given by* (4).

Sketch of proof. The VC dimension bound is based on the observation that, due to the structure of the input, Ω_Σ can be written as a function of the parameters of Σ which is piecewise rational. (The parameterization is derived from the eigenstructure of matrix A.) The complexity upper bound utilizes the Goldberg–Jerrum bound. (Matching lower bounds are based on the notion of dual VC dimension and axis shattering.) ∎

In terms of n (the dimension of the state space) and k (the bandwidth), the upper bound for the VC dimension is of the form $O(n^3\log_2(nk))$. The next section states also a corresponding VC dimension lower bound, in terms of the bandwidth, of the form $O(\log(k))$, and, together with a lower bound for the sample complexity, this provides an estimate for the number of samples needed in learning. In particular, in a typical setting of fairly small system dimension n and large bandwidth k, the $\log k$ bound is a clear improvement over the linear bound given by elementary analysis.

For learning $[0,1]$-valued functions the role of the VC dimension is replaced by other complexity dimensions such as the pseudodimension or the fat-shattering dimension that give upper bounds on sample complexities in the corresponding learning paradigms. In the setting of learning $[0,1]$-valued functions we consider the time interval $[0,\tau]$, with $\tau > 1$ in order to show the impact of the final time on the sample complexity.

For the system

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx, \end{aligned} \tag{5}$$

we can write $Ce^{At}B = [\gamma_1, \ldots, \gamma_m]$, where each γ_i is a linear combination of n functions $\xi_1, \ldots, \xi_n$. Each ξ_i is of the form $t^\ell e^{at}\sin(bt)$ or $t^\ell e^{at}\cos(bt)$ with $\ell \in \{0, \ldots, n-1\}$ and $a+ib$ an eigenvalue of A. Assume that A has a fixed Jordan block structure, and let a_k+ib_k be an eigenvalue of A. We take $\alpha_{11}, \ldots, \alpha_{nm}, a_1, b_1, \ldots, a_r, b_r$ to be the system parameters, where $\gamma_i(t) = \sum_{j=1}^n \alpha_{ij}\xi_j(t)$ for $i = 1, \ldots, m$ and $a_1, b_1, \ldots, a_r, b_r$ are n eigenvalue parameters. For example, the eigenvalue parameters for a real 4×4 matrix A with eigenvalues $a_1 \pm b_1 i$, a_2, and a_3 would be a_1, b_1, a_2, and a_3. Similarly, the eigenvalue parameters with purely complex eigenvalues $a_1 \pm b_1 i$ and $a_2 \pm b_2 i$ would be a_1, b_1, a_2, and b_2, whereas real eigenvalue parameters would be listed as a_1, a_2, a_3, and a_4.

Let $U \subset \mathbb{R}^{mk}$ be a bounded set. Define a mapping $F : \lambda \times U \to \mathbb{R}$ such that $F(\lambda, u) = y(\tau)$, where $\tau \geq 1$ and $y(\tau)$ is a solution of (5) with system parameters λ and initial condition $x(0) = 0$.

DEFINITION 2.6. *Assume that the system $\dot{x} = Ax + Bu$, $y = Cx$, $x(0) = 0$ can be parameterized by $\lambda \in \mathbb{R}^{n(m+1)}$ as above and $\|\lambda\|_\infty = \max_{1 \leq i \leq n(m+1)} \lambda_i < 1$. Let $F(\lambda, u) = y(\tau)$ be the solution of (5) with system parameters λ and control $u = (u_1, \ldots, u_m) \in U = \{u = (u_1, \ldots, u_m);\ \int_0^\tau |u_i(t)|dt \leq M, i = 1, \ldots, m\}$. Denote $B_\infty^k(c) := \{x \in \mathbb{R}^k;\ \|x\|_\infty < c\}$ and define*

$$\mathcal{F}_B = \{F(\lambda, \cdot) \colon U \to \mathbb{R};\ \lambda \in B_\infty^{n(m+1)}(1)\}.$$

A suitable learning notion for the above function class $\mathcal{F}_B$ is the proper agnostic learning, defined formally in section 2.2. The related sample complexity result is as follows.

THEOREM 2.7 (sample complexity for proper agnostic learning). *Let $1/4 > \kappa > 0$ be an arbitrary real number; then the class $\mathcal{F}_B$, given in Definition 2.6, is properly agnostically learnable from*

$$O\left(\frac{1}{\epsilon^2}\left(\mathrm{fat}_{(1/4-\kappa)\epsilon}(\mathcal{F}_B)\log^2\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$$

samples, where

$$\mathrm{fat}_{(1/4-\kappa)\epsilon}(\mathcal{F}_B) \leq (m+1)n\log_2\left\lfloor \frac{n^2 m \tau^n e^\tau k M}{(1/4-\kappa)\epsilon} \right\rfloor,$$

and M is a constant satisfying

$$\int_0^\tau |u_i(\tau - t)|dt \leq kM$$

for all $i = 1, \ldots, m$. In the above, $\lfloor x \rfloor$ stands for the integer part of x. If the inputs are of the form $u = G\omega$, then also

$$\mathrm{fat}_{(1/4-\kappa)\epsilon}(\mathcal{F}_B) \leq 2(m+4)n\log_2(8e(nmk4(n+\ell_{\max})+1)(2nk+2(2k+1)^n)),$$

where $\ell_{\max}$ is as given by (3) and (4).

Sketch of proof. The proof is developed around a Lipschitz bound on Ω_Σ as a function of unknown parameters, which gives the upper term in the fat-shattering bound. The lower term in the bound is in turn a pseudodimension bound that can be derived from an associated VC dimension bound. □

2.3.1. Modified learning settings. In this paper we have opted for the standard setting in CLT in which inputs are random and observations are noiseless. However, CLT can accommodate various modified learning settings that typically share common features with the standard setting.

A paradigm in which a learner can select the samples to be classified is called active learning. The set of PAC learnable concept classes is not enlarged by active learning but, in general, fewer training samples are needed (concept classes that are "dense in themselves" make an exception) [8].

For dealing with the case of noisy observations the reader is referred to [21]. For example, it is shown that learning a concept class with a "noisy oracle" (that makes

a mistake with probability $\beta < 1/2$) to accuracy ϵ is the same as learning the same concept class to an accuracy $\epsilon/(1-2\beta)$ with a perfect oracle.

Finally, we have considered the case in which only the final state output is observed; i.e., observation is done only on a single instant of time. However, typically in control applications observations are made at multiple time instances. If one wishes to learn the mapping from inputs to final state outputs, but one can also see intermediate observations, can one learn faster by utilizing this additional information? This has motivated further research by the authors on "learning with side information" [14, 15]. The main problem is that in the control problem considered the samples become dependent, which poses a challenge in the theory of learning.

3. Complexity dimensions; main upper and lower bounds.

3.1. Bounds. We begin by stating bounds in the easiest learning setting, i.e., classifying the final state observations as either 0 or 1.

THEOREM 3.1 (VC dimension upper bound, $p = 1$). *The VC dimension of the sign system concept class $\mathcal{C}_{m,1}$ with scalar observations can be bounded as*

$$VC(\mathcal{C}_{m,1}) \leq 2(2mn^2 + 4n + 1)\log_2[8e(8mn^2k(n + \ell_{\max}) + 1)(2nk + 2(1 + 2k)^n)],$$

where $\ell_{\max}$ is given by (4).

In terms of n (the dimension of the state space) and k (the bandwidth) the upper bound is of the form $O(n^3 \log_2(nk))$.

All VC dimension upper bounds are based on the fact that input basis functions satisfy a certain rationality condition. Remark 5.3 indicates how the bound is formed when the input functions satisfy the more abstract rationality condition. In that case the degrees of the polynomials and the number of polynomial evaluations are different. However, in terms of n and k, the bound is of the same form. VC dimension or pseudodimension bounds stated in this paper can be modified for the rationality condition in the same way.

The lower bound for the VC dimension is in terms of n and k. It holds for linearly independent, continuous basis input functions, and, compared to upper bounds, no particular form of the functions is needed. The bound is obtained by imposing a specific structure on control systems, and a lower bound for a restricted class of control systems provides a lower bound for more general classes.

THEOREM 3.2 (VC dimension lower bound, $m = 1$, $p = 1$).

$$VC(\mathcal{C}_{1,1}) \geq \max\left\{m' \left\lfloor \log_2 \left\lfloor \frac{k}{m'} \right\rfloor \right\rfloor, m'\right\},$$

where $m' = \min\{n, k\}$.

In terms of k the upper and the lower bound match up to a constant. For n and k the lower bound is typically of the form $O(n \log_2(k/n))$. Note that if the system dimension n is small compared to the bandwidth k, the VC dimension upper and lower bounds in Theorems 3.1 and 3.2 become tighter, both being of the form $c \log_2 k$ (with different values of the constant c).

Extending the upper bounds to the case of vector-valued observations can be done in various ways based on the result obtained for scalar observations. For example, we may consider the p-dimensional output as bits representing a number in $\{0, \ldots, 2^p - 1\}$ and introduce a loss function for each $f \in \mathcal{C}_{m,p}$ as $L_{0\text{-}1,f}(z, a) = L_{0\text{-}1}(f(z), a) = 1$, when $f(z) \neq a$, and 0 otherwise. We define the VC dimension of the p-dimensional

observation as the VC dimension of the above class of loss functions. Modifying the argument used with scalar observations leads to a bound of the following form.

THEOREM 3.3 (VC dimension upper bound).

$$VC(\mathcal{C}_{m,p}) \le 2(2pmn^2 + 4n + p) \times \log_2\Big[8e\big(8mn^2k(n+\ell_{\max})+1\big)\big(2^p - 1 + 2p(2k+1)^n + 2nk\big)\Big],$$

where $\ell_{\max}$ *is given by* (4).

Next we state the main result concerning learnability of the actual input/output mapping, i.e., learning without taking the sign of the final state observation.

DEFINITION 3.4 (control system concept class, $\mathcal{G}_{p,L}$). *Let* $\widetilde{\mathcal{F}} = \{\Phi_\Sigma : X_\Omega \to \mathbb{R}^p;\ \Sigma$ *linear system of dimension* $n\}$, *and define the* control system concept class *as* $\mathcal{G}_{p,L} = A_{\widetilde{\mathcal{F}},L}$, *where* $A_{\widetilde{\mathcal{F}},L}$ *is given by* (1).

Methods for calculating upper bounds for the VC dimension readily extend to the case of pseudodimension with respect to loss that preserves the rationality structure of the output. A typical example is illustrated by the loss function,

$$L(z_1, z_2) = (z_1 - z_2)^2/(1 + (z_1 - z_2)^2), \tag{6}$$

and the following result.

THEOREM 3.5 (pseudodimension upper bound, $p = 1$).

$$PD(\mathcal{G}_{1,L}) \le 2\left(2mn^2 + 4n + 1\right)\log_2[16e(8mn^2k(n+\ell_{\max})+1)(2nk + 2(2k+1)^n)],$$

where the loss function L *is given by* (6) *and* $\ell_{\max}$ *is given by* (4).

This differs from the corresponding VC dimension bound only by the maximum degree of the polynomials, which is doubled. Extending this pseudodimension bound for p-dimensional observations can be done naturally by modifying the loss function. Lower bounds for the VC dimension are lower bounds for the pseudodimension as such.

The next results summarize upper bounds for the fat-shattering dimension. We begin by illustrating how the fat-shattering dimension can be bounded for Lipschitz functions in certain cases.

THEOREM 3.6 (fat-shattering bound). *Let* $F(\lambda, u) : \mathbb{R}^k \times U \to \mathbb{R}$ *be such that* $F(\cdot, u)$ *is Lipschitz with constant* L, *i.e.,* $|F(\lambda_1, u) - F(\lambda_2, u)| \le L\|\lambda_1 - \lambda_2\|$ *for all* $u \in U$. *For any subset* $B \subseteq \mathbb{R}^k$, *consider the following class of functions:*

$$\mathcal{F}_B = \{F(\lambda, \cdot) : U \to \mathbb{R};\ \lambda \in B\}.$$

Then

$$\mathrm{fat}_\gamma(\mathcal{F}_{B^k_\infty(C)}) \le k \log_2 \left\lfloor \frac{CL}{\gamma} \right\rfloor,$$

$$\mathrm{fat}_\gamma(\mathcal{F}_{\bar{B}^k_\infty(C)}) \le k \log_2 \left(1 + \left\lfloor \frac{CL}{\gamma} \right\rfloor\right),$$

$$\mathrm{fat}_\gamma(\mathcal{F}_{\bar{B}_{2,1}}) \le k \log_2 \left(C + \frac{L}{\gamma}\right),$$

where

$$B_\infty^k(C) = \{x \in \mathbb{R}^k;\ \|x\|_\infty = \max_{1\le i\le k} |x_i| < C\},$$
$$\bar{B}_\infty^k(C) = \{x \in \mathbb{R}^k;\ \|x\|_\infty = \max_{1\le i\le k} |x_i| \le C\},$$
$$\bar{B}_2^k(C) = \left\{x \in \mathbb{R}^k;\ \|x\|_2 = \sqrt{\sum\nolimits_{i=1}^k x_i^2} \le C\right\}.$$

THEOREM 3.7 (fat-shattering bound for a control system). *Assume that the system* $\dot{x} = Ax + Bu$, $y = Cx$, $x(0) = 0$, *can be parameterized by* $\lambda \in \mathbb{R}^{n(m+1)}$ *as in Definition* 2.6 *and* $\|\lambda\|_\infty < 1$. *Let* $F(\lambda, u) = y(\tau)$ *be the solution with system parameters* λ *and control* $u = (u_1, \dots, u_m) \in U = \{u = (u_1, \dots, u_m);\ \int_0^\tau |u_i(t)|dt \le M,\ i = 1, \dots, m\}$. *Define*

$$\mathcal{F}_B = \{F(\lambda, \cdot)\colon U \to \mathbb{R};\ \lambda \in B_\infty^k(1)\}.$$

Then

$$\mathrm{fat}_\gamma(\mathcal{F}_B) \le n(m+1)\log_2\left\lfloor \frac{n^2 m \tau^n e^\tau M}{\gamma}\right\rfloor.$$

4. Techniques for proving VC dimension results. Our main results are based on the fact that the basis input functions satisfy a certain rationality condition. In this section we first formulate this rationality condition, and then we summarize existing results that are used in proving upper and lower bounds for the complexity dimensions.

We recall briefly the control system setting. We study systems

$$\dot{x} = Ax + Bu, \quad x(0) = x^0, \quad y = Cx,$$
$$u = G\omega, \quad \text{and} \quad \omega_j \in \Omega \text{ for } j = 1, \dots, k,$$

with basis input functions

$$\Omega = \Big\{\omega_1, \dots, \omega_k; \omega_1, \dots, \omega_k \text{ linearly independent and}$$
$$\omega_j = t^{\ell_j} e^{\alpha_j t}\sin(\beta_j t) \text{ or } \omega_j = t^{\ell_j} e^{\alpha_j t}\cos(\beta_j t)$$
$$\text{with } \ell_j \in \mathbb{N}, \alpha_j, \beta_j \in \mathbb{R}, j = 1, \dots, k\Big\},$$

such that $\ell_{\max} = \max\{\ell_1, \dots, \ell_k\}$.

DEFINITION 4.1 (rationality condition (RAT)). *Let* n *be a positive integer. We say that a bounded function* $\omega : [0, 1] \to \mathbb{R}$ *satisfies the rationality condition relative to the class of* n*-dimensional systems if there exist* h *polynomial functions* $f_1, \dots, f_h : \mathbb{R}^4 \to \mathbb{R}$ *and* $2\gamma n$ *rational functions* $r_{ij\tilde{\ell}}$, $i \in \{1, 2\}$, $j \in \{1, \dots, \gamma\}$, *and* $\tilde{\ell} \in \{1, \dots, n\}$, *with no poles on subsets* $S_{ij\tilde{\ell}}$ *of* $\mathbb{R}^4$, *such that the following properties hold:*

1. *For each* $i \in \{1, 2\}$, $\tilde{\ell} \in \{1, \dots, n\}$, $\mathbb{R}^4$ *is a disjoint union of* $S_{i1\tilde{\ell}}, \dots, S_{i\gamma\tilde{\ell}}$.
2. *Each* $S_{ij\tilde{\ell}}$ *can be defined in terms of a Boolean expression involving* $[f_1 = 0]$, $\dots$, $[f_h = 0]$, *where we say that for functions* $f_1, \dots, f_h : \mathbb{R}^4 \to \mathbb{R}$, $[f_i = 0]$ *has value* 1 *if* $f_i(x_1, x_2, x_3, x_4) = 0$ *and* 0 *otherwise.*

3. *Let* $r_{i\tilde{\ell}} : \mathbb{R}^4 \to \mathbb{R}$, $i \in \{1,2\}$, $\tilde{\ell} \in \{1,\dots,n\}$, *be defined as*

$$r_{i\tilde{\ell}}(v) = \begin{cases} r_{i1\tilde{\ell}}(v) & \text{if } v \in S_{i1\tilde{\ell}}, \\ \vdots & \vdots \\ r_{i\gamma\tilde{\ell}}(v) & \text{if } v \in S_{i\gamma\tilde{\ell}}; \end{cases}$$

then for each $a, b \in \mathbb{R}$, *and for all* $\tilde{\ell} \in \{1,\dots,n\}$,

$$\int_0^1 t^{\tilde{\ell}-1} e^{at} \cos(bt)\omega(t)dt = r_{1\tilde{\ell}}(a, b, e^a \cos b, e^a \sin b),$$
$$\int_0^1 t^{\tilde{\ell}-1} e^{at} \sin(bt)\omega(t)dt = r_{2\tilde{\ell}}(a, b, e^a \cos b, e^a \sin b).$$

We denote by $d_{\max}$ the maximum degree of any polynomial (i.e., $f_1, \dots, f_h$, numerators and denominators of $r_{ij\tilde{\ell}}$'s) appearing in the rationality condition.

Remark 4.2. First, entries of e^{At} are functions of the form $t^s e^{at}\cos(bt)$ and $t^s e^{at}\sin(bt)$. Solving (2) involves convolutions of e^{At} and the basis input functions ω_j, and we require those to be rational functions.

Example 4.3. Let $\omega(t) = \sin(ct)$, with nonzero c. Then

$$\int_0^1 e^{at}\sin(bt)\omega(t)dt = \frac{1}{2}\int_0^1 e^{at}\cos((b-c)t)dt - \frac{1}{2}\int_0^1 e^{at}\cos((b+c)t)dt.$$

After integration this can be split into cases with no poles yielding

$$\int_0^1 e^{at}\sin(bt)\omega(t)dt = \begin{cases} \frac{p(a,b,e^a\cos b, e^a\sin b)}{(a^2+(b+c)^2)(a^2+(b-c)^2)} & \text{if } f_1 \neq 0, f_2 \neq 0, \\ \frac{b-\sin b\cos b}{2b} & \text{if } f_1 = 0, f_2 \neq 0, \\ \frac{\sin b\cos b - b}{2b} & \text{if } f_1 \neq 0, f_2 = 0, \end{cases}$$

where $f_1(a, b, e^a \sin b, e^a \cos b) = a^2 + (b+c)^2$, $f_2(a, b, e^a \sin b, e^a \cos b) = a^2 + (b-c)^2$, and $p(a, b, e^a \cos b, e^a \sin b)$ stands for the polynomial

$$\begin{aligned} &-4abc + 4abce^a \cos b \cos c - 2a^2 ce^a \cos c \sin b + 2b^2 ce^a \cos c \sin b \\ &\quad -2c^3 e^a \cos c \sin b - 2a^2 be^a \cos b \sin c - 2b^3 e^a \cos b \sin c + 2bc^2 e^a \cos b \sin c \\ &\quad +2a^3 e^a \sin b \sin c + 2ab^2 e^a \sin b \sin c + 2ac^2 e^a \sin b \sin c. \end{aligned}$$

LEMMA 4.4. *Each* $\omega_j \in \Omega$ *given by* (3) *satisfies the rationality condition. Further, the maximum degree of polynomials in (RAT) is at most* $4(n + \ell_{\max})$, *where* $\ell_{\max}$ *is given by* (4).

Review of VC dimension techniques. In the context of control theory it is sometimes easier to work with the dual VC dimension. Assume that a function $F : \Lambda \times X \to \{0,1\}$ is given. This induces two function classes

$$\mathcal{F} := \{F(\lambda, \cdot) : X \to \{0,1\};\ \lambda \in \Lambda\}$$

and

$$\mathcal{F}^* := \{F(\cdot, x) : \Lambda \to \{0,1\};\ x \in X\}.$$

The complexity dimension $\mathrm{VC}(\mathcal{F}^*)$ is called the *dual VC dimension* of $\mathcal{F}$, and it is related to $\mathrm{VC}(\mathcal{F})$ as follows [21]:

$$\mathrm{VC}(\mathcal{F}) \geq \lfloor \log_2 \mathrm{VC}(\mathcal{F}^*) \rfloor, \tag{7}$$

where $\lfloor x \rfloor$ is the integer part of x.

A sharper estimate can be obtained if Λ can be written as a product $\Lambda_1 \times \cdots \times \Lambda_n$. The following construction and result are due to DasGupta and Sontag [6]. We study in particular those dichotomies that are defined on "rectangular" subsets of Λ. Let $L = L_1 \times \cdots \times L_n$ be a subset of Λ such that for each i, $L_i \subset \Lambda_i$ is nonempty. Given any index $1 \leq \kappa \leq n$, a *κ-axis dichotomy* on L is any function $\delta : L \to \{0,1\}$ which depends only on the κth coordinate; i.e., there is some function $\phi : L_\kappa \to \{0,1\}$ so that $\delta(\lambda_1, \ldots, \lambda_n) = \phi(\lambda_\kappa)$ for all $(\lambda_1, \ldots, \lambda_n) \in L$. We say that a mapping is an axis dichotomy if it is a κ-axis dichotomy for some κ. A rectangular set L is said to be *axis-shattered* by $\mathcal{F}^*$ if every axis dichotomy is a restriction to L of some function of the form $F(\cdot, x) : \Lambda \to \{0,1\}$ for some $x \in X$.

THEOREM 4.5 (axis-shattering bound [6]). *If $L = L_1 \times \cdots \times L_n \subset \Lambda$ can be axis-shattered and each L_i has cardinality $r_i > 0$, then $VC(\mathcal{F}) \geq \lfloor \log_2(r_1) \rfloor + \cdots + \lfloor \log_2(r_n) \rfloor$.*

Upper bounds for VC dimensions of concept classes that are obtained by evaluating polynomial equalities and inequalities can be obtained in terms of the number and degrees of the polynomials.

THEOREM 4.6 (Goldberg–Jerrum bound [10]). *Given a function $F : \Lambda \times X \to \{0,1\}$ and the associated concept class $\mathcal{F} := \{F(\lambda, \cdot) : X \to \{0,1\};\ \lambda \in \Lambda\}$, suppose that $\Lambda = \mathbb{R}^\ell$ and $X = \mathbb{R}^k$. Let F be defined in terms of a Boolean formula involving at most s polynomial equalities and inequalities in $\ell + k$ variables, each polynomial being of degree at most d in λ for all $x \in \mathbb{R}^k$. Then, $VC(\mathcal{F}) \leq 2\ell \log_2(8eds)$.*

The Goldberg–Jerrum bound is based on a result showing that the number of sign assignments $\{-1, 0, 1\}$ to polynomials cannot grow too quickly.

THEOREM 4.7 (see [10]). *Suppose that $f_1, \ldots, f_m$ are polynomials of degree at most d in $n \leq m$ variables. Then the number of distinct vectors*

$$[\operatorname{sign} f_1(x), \ldots, \operatorname{sign} f_m(x)] \in \{-1, 0, 1\}^m$$

that can be generated by varying x over $\mathbb{R}^n$ is at most $\left((8edm)/n\right)^n$.

5. Proofs of VC dimension bounds.

5.1. An upper bound for the VC dimension with scalar observations. We begin this section by proving Lemma 4.4 stating that the input basis functions satisfy the rationality condition (RAT) and bounding the degrees of polynomials appearing in (RAT). As a proposition we formalize how control systems can be parameterized. After that, as a lemma, we develop an upper bound for the VC dimension induced by the control system (2) with its initial state fixed to be zero. Theorem 3.1 with an arbitrary initial condition is then a simple modification of the argument.

Proof of Lemma 4.4. If $\omega(t) = t^\ell e^{\alpha t} \sin(\beta t)$ or $\omega(t) = t^\ell e^{\alpha t} \cos(\beta t)$ with $\ell \leq \ell_{\max}$, then in place of

$$\int_0^1 t^{\tilde{\ell}} e^{at} \sin(bt)\omega(t)dt \quad \text{or} \quad \int_0^1 t^{\tilde{\ell}} e^{at} \cos(bt)\omega(t)dt, \tag{8}$$

by combining exponents and using sum formulae for sin and cos (see Example 4.3), it is enough to study terms of the form $\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \sin(\tilde{b}t)dt$ or $\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \cos(\tilde{b}t)dt$, where $\tilde{k} \in \{0, \dots, n + \ell_{\max} - 1\}$. In fact, each expression in (8) is one of the following types:

$$\frac{1}{2}\left(\int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \cos\left((b-\beta)t\right) dt - \int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \cos\left((b+\beta)t\right) dt\right),$$
$$\frac{1}{2}\left(\int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \cos\left((b+\beta)t\right) dt + \int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \cos\left((b-\beta)t\right) dt\right),$$
$$\frac{1}{2}\left(\int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \sin\left((b-\beta)t\right) dt - \int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \sin\left((b+\beta)t\right) dt\right),$$
$$\frac{1}{2}\left(\int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \sin\left((b+\beta)t\right) dt + \int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \sin\left((b-\beta)t\right) dt\right),$$

where $\tilde{a} = a + \alpha$. Because for $\tilde{k} > 0$

$$\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \sin(\tilde{b}t)dt = \frac{e^{\tilde{a}}}{\tilde{a}^2 + \tilde{b}^2}(\tilde{a}\sin\tilde{b} - \tilde{b}\cos\tilde{b}) - \frac{\tilde{k}}{\tilde{a}^2 + \tilde{b}^2}\int_0^1 t^{\tilde{k}-1} e^{\tilde{a}t}(\tilde{a}\sin(\tilde{b}t) - \tilde{b}\cos(\tilde{b}t))dt \tag{9}$$

and

$$\int_0^1 e^{\tilde{a}t} \sin(\tilde{b}t)dt = \frac{e^{\tilde{a}}(\tilde{a}\sin\tilde{b} - \tilde{b}\cos\tilde{b}) + \tilde{b}}{\tilde{a}^2 + \tilde{b}^2} \tag{10}$$

and similar formulae for $\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \cos(\tilde{b}t)dt$ hold, we see by induction that the numerator of $\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \sin(\tilde{b}t)dt$ is a polynomial of $\tilde{a}, \tilde{b}, e^{\tilde{a}}\cos\tilde{b}$, and $e^{\tilde{a}}\sin\tilde{b}$. By using sum formulae for sin and cos, the previous expression is in turn a polynomial of $a, b, e^a \cos b$, and $e^a \sin b$ because $\tilde{a} = a + \alpha$ and $\tilde{b} = b \pm \beta$ for some fixed α and β. By similar arguments, the denominator is a polynomial of a and b. Note that, for example, e^{α} equals a constant times e^a, so this process does not change the degrees of the polynomials.

Further, observe that the denominator of $\int_0^1 t^{\tilde{\ell}} e^{at} \sin(bt)\omega(t)dt$ consists of at most two products of variables a and b of the form $((a+\alpha)^2 + (b \pm \beta)^2)^{\tilde{\ell}+\ell+1}$, and similarly with the $\cos(bt)$ term. Let us index the basis input functions $\omega_1, \dots, \omega_k$ so that ω_κ has parameters α_κ and β_κ. Hence the functions f_i in (RAT), defining the subsets without poles, can be taken as

$$\{(a+\alpha_\kappa)^2 + (b-\beta_\kappa)^2, (a+\alpha_\kappa)^2 + (b+\beta_\kappa)^2 \,;\, \kappa = 1, \dots, k\}.$$

Furthermore, the sets $S_{ij\tilde{\ell}}$ are as simple as

$$\cup_{\kappa=1}^k \{\{(x_1,x_2,x_3,x_4);\ x_1 = -\alpha_\kappa, x_2 = -\beta_\kappa\} \cup \{(x_1,x_2,x_3,x_4);\ x_1 = -\alpha_\kappa, x_2 = \beta_\kappa\}\}$$

and

$$\mathbb{R}^4 \setminus \cup_{\kappa=1}^k \{\{(x_1,x_2,x_3,x_4);\ x_1 = -\alpha_\kappa, x_2 = -\beta_\kappa\} \cup \{(x_1,x_2,x_3,x_4);\ x_1 = -\alpha_\kappa, x_2 = \beta_\kappa\}\}.$$

We turn to estimating the maximum degree of polynomials appearing in (RAT). We already saw that functions f_i are polynomials of degree 2. Equations (9) and (10) show that the degree of numerator is not higher than the one of denominator. We claim that

$$\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \begin{cases} \sin(\tilde{b}t) \\ \cos(\tilde{b}t) \end{cases} dt = \frac{P(2(\tilde{k}+1))}{(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}+1}} \quad \text{for } \tilde{k} = 0, 1, \ldots,$$

where $P(2(\tilde{k}+1))$ stands for some polynomial in $\tilde{a}$, $\tilde{b}$, $e^{\tilde{a}}\sin(\tilde{b})$, and $e^{\tilde{a}}\cos(\tilde{b})$ of degree $2(\tilde{k}+1)$. Clearly, the claim is true for $\tilde{k} = 0$ by (10), and the inductive argument follows from (9). Assuming the claim is true for $\tilde{k}-1$, we get

$$\begin{aligned}
&\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \sin(\tilde{b}t) dt \\
&\quad = \frac{P(2)}{\tilde{a}^2+\tilde{b}^2} - \frac{\tilde{k}\tilde{a}}{\tilde{a}^2+\tilde{b}^2} \int_0^1 t^{\tilde{k}-1} e^{\tilde{a}t} \sin(\tilde{b}t) dt + \frac{\tilde{k}\tilde{b}}{\tilde{a}^2+\tilde{b}^2} \int_0^1 t^{\tilde{k}-1} e^{\tilde{a}t} \cos(\tilde{b}t) dt \\
&\quad = \frac{P(2)}{\tilde{a}^2+\tilde{b}^2} - \frac{P(1)}{(\tilde{a}^2+\tilde{b}^2)} \frac{P(2\tilde{k})}{(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}}} + \frac{P(1)}{(\tilde{a}^2+\tilde{b}^2)} \frac{P(2\tilde{k})}{(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}}} \\
&\quad = \frac{P(2)(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}} - 2P(2\tilde{k}+1)}{(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}+1}} = \frac{P(2(\tilde{k}+1))}{(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}+1}},
\end{aligned}$$

and similarly for the $\cos(\tilde{b}t)$ term, concluding the proof of the claim. As a corollary of the claim,

$$\int_0^1 t^{\tilde{\ell}} e^{at} \sin(bt)\omega(t) dt = \frac{P(2(\tilde{k}+1))}{(\tilde{a}^2+(b+\beta)^2)^{\tilde{k}+1}} + \frac{P(2(\tilde{k}+1))}{(\tilde{a}^2+(b-\beta)^2)^{\tilde{k}+1}},$$

where $\tilde{k} = \ell + \tilde{\ell}$ and $\tilde{a} = a + \alpha$.

Hence the maximum degree of denominators of expressions in (8) is $2(\tilde{k}+1) + 2(\tilde{k}+1) = 4(\tilde{k}+1)$ with $\tilde{k} \in \{0, \ldots, \ell_{\max} + n - 1\}$. Thus the maximum degree of polynomials appearing in the (RAT) is $4(n + \ell_{\max})$. □

The next proposition indicates how control systems are parameterized and later the concept or function classes associated to control systems are obtained by varying the parameter vector.

PROPOSITION 5.1. *Denote the basis input functions by* $\omega = (\omega_1, \ldots, \omega_k)^T$, *assume that each* ω_i, $i = 1, \ldots, k$, *satisfies the rationality condition (RAT), and let* $\Lambda = \mathbb{R}^{2pn^2m} \times \mathbb{R}^{4n} \times \mathbb{R}^p$. *Then there exists a mapping* $H : \Lambda \times \mathbb{R}^{mk} \to \mathbb{R}^p$ *(depending on* ω*) such that for each* $\Sigma = (A, B, C, x^0)$ *there exists a* $\lambda \in \Lambda$ *satisfying*

$$\Phi_\Sigma(G\omega) = H(\lambda, G) \quad \forall\ G \in \mathbb{R}^{mk}.$$

Proof. Given a system $\Sigma = (A, B, C, x^0)$,

$$\Phi_\Sigma(u) = y(1) = Ce^A x^0 + C \int_0^1 e^{A(1-t)} Bu(t) dt.$$

By an argument based on the real Jordan form of e^{At}, the entries of $e^{A(1-t)}$ are linear combinations of functions of the form $t^{\tilde{\ell}} e^{at}\cos(bt)$ and $t^{\tilde{\ell}} e^{at}\sin(bt)$, where $\tilde{\ell} \in \{0, \ldots, n-1\}$ and $a + ib$ is an eigenvalue of A. Hence we define the $2n$ functions $\xi_j(a, b, t) = t^{j-1} e^{at}\cos(bt)$, $\xi_{n+j}(a, b, t) = t^{j-1} e^{at}\sin(bt)$ for $j = 1, \ldots, n$.

By the rationality condition (RAT), for all $\ell = 1, \dots, 2n$,

$$\int_0^1 \xi_\ell(a,b,t)\omega_j(t)dt = \frac{\hat{P}_{\ell j}(a,b,e^a\cos b, e^a \sin b)}{\hat{Q}_{\ell j}(a,b,e^a\cos b, e^a \sin b)}$$

for all $a, b \in \mathbb{R}$ and where $\hat{P}_{\ell j}$ and $\hat{Q}_{\ell j}$ are piecewise polynomial expressions.

Let $H(\mathbf{A}, \mathbf{X}, h, G) = (H_1, \dots, H_p)^T$, where for $1 \le \kappa \le p$

$$H_\kappa(\mathbf{A}, \mathbf{X}, h, G) = \sum_{i=1}^{m}\sum_{r=1}^{n}\sum_{\ell=1}^{2n} \alpha_{ir\ell\kappa} \sum_{j=1}^{k} g_{ij} \frac{\hat{P}_{\ell j}(x_{r1}, x_{r2}, x_{r3}, x_{r4})}{\hat{Q}_{\ell j}(x_{r1}, x_{r2}, x_{r3}, x_{r4})} + h_\kappa$$

and

$$\mathbf{A} = (\alpha_{ir\ell\kappa})_{\substack{i=1,\dots,m\\ r=1,\dots,n\\ \ell=1,\dots,2n\\ \kappa=1,\dots,p}}, \qquad \mathbf{X} = (x_{r\eta})_{\substack{r=1,\dots,n\\ \eta=1,\dots,4}},$$
$$h = (h_1, \dots, h_p)^T, \qquad G = (g_{ij})_{\substack{i=1,\dots,m\\ j=1,\dots,k}}.$$

Next, we relate Φ_Σ and H and we write

$$Ce^{A(1-t)}B = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & & \vdots \\ \gamma_{p1} & \cdots & \gamma_{pm} \end{bmatrix}.$$

We list the eigenvalues of A as $a_r + ib_r$ for $r = 1, \dots, n$ and let $\xi_{r\ell}(t) = \xi_\ell(a_r, b_r, t)$ for $r = 1, \dots, n$ and $\ell = 1, \dots, 2n$. Then there exists some $(\alpha_{ir\ell\kappa})$ such that

$$\gamma_{\kappa i}(t) = \sum_{r=1}^{n}\sum_{\ell=1}^{2n} \alpha_{ir\ell\kappa}\xi_{r\kappa}(t). \tag{11}$$

Let $\lambda = (\mathbf{A}, \mathbf{X}, h)$, where $\mathbf{A}$ satisfies (11), $\mathbf{X} = (x_{r\eta})$, where $x_{r1} = a_r, x_{r2} = b_r, x_{r3} = e^{a_r}\cos b_r$ and $x_{r4} = e^{a_r}\sin b_r$, and $h = Ce^A x^0$. We claim that

$$H(\lambda, G) = y(1) = \Phi_\Sigma(G\omega) \quad \forall\, G \in \mathbb{R}^{mk}.$$

Note that the κth component of $\Phi_\Sigma(G\omega)$ is given by

$$\begin{aligned}
&\int_0^1 \sum_{i=1}^{m} \gamma_{\kappa i}(t)u_i(t)dt + h_\kappa \\
&= \sum_{i=1}^{m}\sum_{r=1}^{n}\sum_{\ell=1}^{2n} \alpha_{ir\ell\kappa} \int_0^1 \xi_{r\ell}(t) \sum_{j=1}^{k} g_{ij}\omega_j(t)dt + h_\kappa \\
&= \sum_{i=1}^{m}\sum_{r=1}^{n}\sum_{\ell=1}^{2n} \alpha_{ir\ell\kappa} \sum_{j=1}^{k} g_{ij} \int_0^1 \xi_\ell(a_r, b_r, t)\omega_j(t)dt + h_\kappa \\
&= \sum_{i=1}^{m}\sum_{r=1}^{n}\sum_{\ell=1}^{2n} \alpha_{ir\ell\kappa} \sum_{j=1}^{k} g_{ij} \frac{\hat{P}_{\ell j}(x_{r1}, x_{r2}, x_{r3}, x_{r4})}{\hat{Q}_{\ell j}(x_{r1}, x_{r2}, x_{r3}, x_{r4})} + h_\kappa \\
&= H_\kappa(\mathbf{A}, \mathbf{X}, h, G). \qquad \square
\end{aligned}$$

Next we take $p = 1$ and study the VC dimension of the sign system concept class, $\mathcal{C}_{m,1}$, where each control parameterized by G gives rise to $\operatorname{sign} y(1)$.

LEMMA 5.2. *The sign system concept class $\mathcal{C}_{m,1}$ with initial condition $x(0) = 0$ satisfies*

$$VC(\mathcal{C}_{m,1}) \leq 2\left(2mn^2 + 4n\right) \log_2[8e(8mn^2k(n + \ell_{\max}) + 1)(2nk + 2(1 + 2k)^n)],$$

where $\ell_{\max}$ is given by (4).

Proof. By Proposition 5.1, $y(1) = H(\lambda, G)$, where $\lambda \in \mathbb{R}^{2mn^2} \times \mathbb{R}^{4n}$ are considered as parameters. In fact, $y(1) = \frac{P}{Q}$, where P and Q denote piecewise polynomial functions. As in the statement of Goldberg–Jerrum bounds, we have a function $F : \Lambda \times \mathbb{R}^{mk} \to \{0, 1\}$ defined by $F(\lambda, G) = \operatorname{sign} H(\lambda, G)$. The concept class associated to the system identification problem is $\mathcal{F} := \{F(\lambda, \cdot) : \mathbb{R}^{mk} \to \{0, 1\};\ \lambda \in \Lambda\}$, where $\Lambda = \mathbb{R}^{2mn^2+4n}$. Before applying the Goldberg–Jerrum bound, we need to determine the possible degrees of P and Q with respect to the parameters.

The rationality condition implies that

$$\max_{\substack{i \leq \ell \leq n \\ 1 \leq j \leq k}} \left\{ \deg(\hat{P}_{\ell j}), \deg(\hat{Q}_{\ell j}) \right\} \leq d_{\max}.$$

Then

$$\frac{\widetilde{P}_{i\ell}}{\widetilde{Q}_{i\ell}} = \sum_{j=1}^{k} g_{ij} \frac{\hat{P}_{\ell j}}{\hat{Q}_{\ell j}},$$

so $\deg(\widetilde{Q}_{i\ell}) \leq kd_{\max}$ and $\deg(\widetilde{P}_{i\ell}) \leq kd_{\max}$. Note here that we are calculating the degree with respect to the system parameters, and the inputs g_{ij} do not contribute. By continuing in a similar fashion and combining r-summation to the ℓ-summation in Proposition 5.1, we write $P_i/Q_i = \sum_{\ell=1}^{2n^2} \alpha_{i\ell} \widetilde{P}_{i\ell} / \widetilde{Q}_{i\ell}$ to conclude that $\deg(Q_i) \leq 2n^2 \deg(\widetilde{Q}_{i\ell}) = 2n^2 kd_{\max}$ and $\deg(P_i) \leq 2n^2 kd_{\max} + 1$. Finally, $P/Q = \sum_{i=1}^{m} P_i/Q_i$ with $\deg(Q) \leq m2n^2 kd_{\max}$ and $\deg(P) \leq m2n^2 kd_{\max} + 1$.

Recall that with $p = 1$ and initial condition $x(0) = 0$, using the notation of Proposition 5.1,

$$y(1) = \sum_{i=1}^{m} \sum_{r=1}^{n} \sum_{\ell=1}^{2n} \alpha_{ir\ell} \sum_{j=1}^{k} g_{ij} \int_0^1 \xi_\ell(x_{r1}, x_{r2}, t) \omega_j(t) dt.$$

The proof of Lemma 4.4 indicates that the denominator of $\int_0^1 \xi_\ell(x_{r1}, x_{r2}, t) \omega_j(t) dt$ equals

$$((x_{r1} + \alpha_j)^2 + (x_{r2} + \beta_j)^2)^{z_{\ell j}} ((x_{r1} + \alpha_j)^2 + (x_{r2} - \beta_j)^2)^{z_{\ell j}},$$

where α_j, β_j are fixed parameters of the basis input function ω_j and $z_{\ell j} \in \mathbb{N}$.

By carrying out the summations we get $y(1) = P/Q$, where Q consists of powers of polynomials f_{ij1}, f_{ij2} with

$$f_{ij1}(\mathbf{A}, \mathbf{X}, G) = (x_{i1} + \alpha_j)^2 + (x_{i2} + \beta_j)^2,$$
$$f_{ij2}(\mathbf{A}, \mathbf{X}, G) = (x_{i1} + \alpha_j)^2 + (x_{i2} - \beta_j)^2,$$

and $i = 1, \ldots, n$, $j = 1, \ldots, k$.

Our final step before applying the Goldberg–Jerrum bound is finding out the number of polynomial inequalities s needed in the Boolean formula and evaluating the sign of the final state output. This is done by studying the number of different P/Q expressions without poles.

An upper bound for different P/Q expressions without poles can be obtained by applying Theorem 4.7 to $2nk$ polynomials f_{ij1}, f_{ij2}, $i = 1, \dots, n$ and $j = 1, \dots, k$, and viewing those as polynomials of $2n$ variables and each polynomial having degree 2. This gives the upper bound $(16ek)^{2n}$.

However, a more specific bound can be obtained in this problem. Note that varying x_{i1} and x_{i2} we can make at most one of the $2k$ polynomials f_{ij1}, f_{ij2}, $j = 1, \dots, k$, to be zero. For example, γ zeros among f_{ij1}, f_{ij2}, $i = 1, \dots, n$ and $j = 1, \dots, k$, can be obtained in $(2k)^\gamma \binom{n}{\gamma}$ ways, and the number of possible sign assignments is obtained by summing over γ yielding

$$\sum_{\gamma=0}^{n} (2k)^\gamma \binom{n}{\gamma} = (1+2k)^n.$$

Thus the number of P/Q expressions without poles is $(1+2k)^n$, which gives rise to $2(1+2k)^n$ polynomials.

Note that in order to write $\operatorname{sign} y(1)$ as a Boolean formula evaluating polynomial inequalities and equalities one also has to include the $2nk$ polynomials f_{ij1}, f_{ij2}, $i = 1, \dots, n$, $j = 1, \dots, k$. Values of these polynomials determine which P/Q expression is the valid one to determine $\operatorname{sign} y(1)$. The Boolean formula for $\operatorname{sign} y(1)$ can be given as a truth table involving polynomial inequalities of $2nk$ f_{ij1}, f_{ij2} expressions and $2(1+2k)^n$ different P and Q expressions.

Using Lemma 4.4 for bound on $d_{\max}$, we apply the Goldberg–Jerrum bound with $s = 2nk + 2(2k+1)^n$, $d = m2n^2k4(n+\ell_{\max})+1$, and $\ell = 2mn^2+4n$. □

A simple example of a piecewise polynomial function P/Q together with the decision table for the final output is provided in the appendix.

Remark 5.3. The VC dimension bound is modified for the more abstract rationality conditions as follows. Evaluating the sign of the output involves the evaluation of $2(8ed_{\max}2n^2kh/4n)^{4n} + 2n^2kh$ polynomials; $2n^2kh$ evaluations are needed to find an appropriate piece, and by Theorem 4.7 the maximum number of possible expressions of the type P/Q is bounded by $(8ed_{\max}2n^2kh/4n)^{4n}$. Applying the Goldberg–Jerrum bound with $s = 2(8ed_{\max}2n^2kh/4n)^{4n}+2n^2kh$, $d = m2n^2kd_{\max}+1$, and $\ell = 2mn^2+4n$ gives the result.

Proof of Theorem 3.1*, the VC dimension upper bound,* $p = 1$. By using the previous notation, $y = Ce^Ax^0 + C\int_0^1 e^{A(1-t)}Bu(t)dt$. Let $\widetilde{x} = Ce^Ax^0$. Then $y = \widetilde{x} + P/Q = (\widetilde{x}Q+P)/Q = \widetilde{P}/Q$. This has $2mn^2+4n+1$ parameters and $\deg(\widetilde{P}) \le m2n^2kd_{\max}+1$. □

5.2. Lower bounds for the VC dimension. The lower bounds for the VC dimension are developed for a single-input single-output system with initial state zero. The control is

$$u = \sum_{j=1}^{k} g_j\omega_j : [0,1] \to \mathbb{R}.$$

We derive lower bounds by fixing the structure of A, B, and C and using the dual VC dimension and axis shattering following the ideas of DasGupta and Sontag [6].

Lemmas 5.5, 5.6, and 5.7 given in this section together prove Theorem 3.2. These lower bounds are very general; we just assume that the input functions are continuous and linearly independent; thus no particular structure of input functions is required as in the upper bounds.

To make the next proof cleaner we formulate a part of it as a separate proposition. (The proposition is a standard fact and we omit the proof.)

PROPOSITION 5.4. *Let $\omega_j : [0,1] \to \mathbb{R}$, $j = 1, \dots, k$, be continuous and linearly independent. Then the functions*

$$h_j(\lambda) = \int_0^1 e^{\lambda t} \omega_j(t) dt, \quad j = 1, \dots, k,$$

are linearly independent.

LEMMA 5.5 (lower bound 1). *The sign system concept class $\mathcal{C}_{1,1}$ with scalar inputs and scalar outputs satisfies*

$$VC(\mathcal{C}_{1,1}) \geq m' \left\lfloor \log_2 \left\lfloor \frac{k}{m'} \right\rfloor \right\rfloor,$$

where $m' = \min\{n, k\}$.

Proof. Let $\omega_j(t)$, $j = 1, \dots, k$, be continuous and linearly independent. Let A have n distinct real eigenvalues $-\lambda_1, \dots, -\lambda_n$, and take B and C so that

$$Ce^{A(1-t)}B = \sum_{i=1}^{m'} e^{\lambda_i t},$$

where $m' = \min\{n, k\}$. Then the final output of the system is

$$y(1) = \int_0^1 Ce^{A(1-t)}B \sum_{j=1}^{k} g_j \omega_j(t) dt = \sum_{i=1}^{m'} \sum_{j=1}^{k} g_j \int_0^1 e^{\lambda_i t} \omega_j(t) dt.$$

Define $h_j(\lambda) = \int_0^1 e^{\lambda t} \omega_j(t) dt$. By Proposition 5.4 the h_j's are linearly independent and we can find $\lambda_1, \dots, \lambda_k$ such that the matrix

$$\begin{bmatrix} h_1(\lambda_1) & \cdots & h_k(\lambda_1) \\ \vdots & & \vdots \\ h_1(\lambda_k) & \cdots & h_k(\lambda_k) \end{bmatrix}$$

has rank k.

The control system with sign-observations gives the mapping $F : \mathbb{R}^{m'} \times \mathbb{R}^k \to \{0, 1\}$ by

$$(\lambda_1, \dots, \lambda_{m'}, g_1, \dots, g_k) \mapsto \operatorname{sign} \left[\sum_{i=1}^{m'} \sum_{j=1}^{k} g_j h_j(\lambda_i) \right].$$

We show that the mapping from parameters $\lambda_1, \dots, \lambda_{m'}$ to $\{0, 1\}$ can be axis-shattered. Let $L = \{\lambda_1, \dots, \lambda_k\}$ be so that $[h_j(\lambda_i)]_{i,j}$ has rank k. Denote by $L_1, \dots, L_{m'}$ disjoint subsets of L such that $|L_i| = \lfloor k/m' \rfloor$, and let $M = L \setminus \{\bigcup_{i=1}^{m'} L_i\}$. Next we want to interpolate in the points of L.

Fix s, $1 \le s \le m'$, and let $\phi : L_s \to \{0,1\}$ be any dichotomy. Next find $g_1, \dots, g_k$ such that

$$\begin{aligned} &\sum_{j=1}^{k} g_j h_j(\lambda_s) = \phi(\lambda_s) \qquad \forall\, \lambda_s \in L_s, \\ &\sum_{j=1}^{k} g_j h_j(\lambda) = 0 \qquad\quad \forall\, \lambda \in (L \cup M) \setminus L_s. \end{aligned} \tag{12}$$

Let $g_1^*, \dots g_k^*$ satisfy (12). (A unique solution exists because $[h_j(\lambda_i)]$ has rank k.) Then

$$F[\lambda_1, \dots, \lambda_{m'}, g_1^*, \dots, g_k^*] = \operatorname{sign}\left[\sum_{i=1}^{m'} \sum_{j=1}^{k} g_j^* h_j(\lambda_i)\right] = \phi(\lambda),$$

when $\lambda \in L_s$ and for all $(\lambda_1, \dots, \lambda_{m'}) \in L_1 \times \cdots \times L_{m'}$.

Let $\widetilde{\mathcal{F}} = \{F(\lambda_1, \dots, \lambda_{m'}, \cdot) : \mathbb{R}^k \to \{0,1\};\ (\lambda_1, \dots, \lambda_{m'}) \in \mathbb{R}^{m'}\}$. By the axis-shattering bound given in Theorem 4.5,

$$\mathrm{VC}(\widetilde{\mathcal{F}}) \ge m' \left\lfloor \log_2 \left\lfloor \frac{k}{m'} \right\rfloor \right\rfloor,$$

and thus $\mathrm{VC}(\mathcal{C}_{1,1}) \ge \mathrm{VC}(\widetilde{\mathcal{F}})$, where $\mathcal{C}_{1,1}$ is the control system concept class with $p = m = 1$. □

LEMMA 5.6 (lower bound 2). *If $k \le n$, then*

$$VC(\mathcal{C}_{1,1}) \ge k.$$

Proof. We make a small modification of the above argument. Assume that $k \le n$, and let A have n real eigenvalues $\lambda_1, \dots, \lambda_n$. Next we take B and C so that $Ce^{A(1-t)}B = \sum_{i=1}^{n} e^{\lambda_i t}\beta_i$, where $(\beta_1, \dots, \beta_n, \lambda_1, \dots, \lambda_n)$ are considered as system parameters.

We study the mapping

$$\begin{aligned} (\beta_1, \dots, \beta_n, \lambda_1, \dots, \lambda_n, g_1, \dots, g_k) &\mapsto \operatorname{sign}\left[\sum_{i=1}^{n} \sum_{j=1}^{k} g_j h_j(\lambda_i)\beta_i\right] \\ = \operatorname{sign}\left[\sum_{j=1}^{k} g_j \underbrace{\sum_{i=1}^{n} h_j(\lambda_i)\beta_i}_{\gamma_j}\right] &= \operatorname{sign}\left[\sum_{j=1}^{k} g_j \gamma_j\right]. \end{aligned}$$

Given $(\gamma_1, \dots, \gamma_k)$, by linear independence of $h_1, \dots, h_k$, we can find $\lambda_1, \dots, \lambda_n, \beta_1, \dots, \beta_n$ such that $\sum_{i=1}^{n} h_j(\lambda_i)\beta_i = \gamma_j$, $j = 1, \dots, k$. But $(\gamma_1, \dots, \gamma_k)$ can be viewed as a normal vector for a hyperplane through the origin in $\mathbb{R}^k$, and the concept class associated to the mapping $(g_1, \dots, g_k) \mapsto \operatorname{sign}[\sum_{j=1}^{k} g_j \gamma_j]$ as $(\gamma_1, \dots, \gamma_k)$ varies has VC dimension k. Hence $\mathrm{VC}(\mathcal{C}_{1,1}) \ge k$. □

LEMMA 5.7 (lower bound 3). *If $n \le k$, then*

$$VC(\mathcal{C}_{1,1}) \ge n.$$

Proof. Our construction for the control system is as in the previous proof, but now we assume that $n \leq k$, and we study

$$(\beta_1, \ldots, \beta_n, \lambda_1, \ldots, \lambda_n, g_1, \ldots, g_k) \mapsto \operatorname{sign}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} g_j h_j(\lambda_i)\beta_i\right]$$
$$= \operatorname{sign}\left[\sum_{i=1}^{n} \underbrace{\sum_{j=1}^{k} g_j h_j(\lambda_i)}_{\tilde{g}_i} \beta_i\right] = \operatorname{sign}\left[\sum_{i=1}^{n} \tilde{g}_i \beta_i\right],$$

and again by linear independence and the above hyperplane argument (now via first transforming $(g_1, \ldots, g_k)$) we can conclude that the above mapping has VC dimension n. Thus $\mathrm{VC}(\mathcal{C}_{1,1}) \geq n$. ∎

5.3. VC dimension upper bounds for p-dimensional outputs. We begin by proving Theorem 3.3.

Proof of the VC dimension upper bound. We develop an upper bound based on the bound for a scalar sign-observation. We have seen that under the rationality assumption (RAT) the scalar output is a piecewise rational expression P/Q. In general, the control system maps G to $(\operatorname{sign}(P_1/Q_1), \ldots, \operatorname{sign}(P_p/Q_p))^T$, which is understood as a binary representation of a number in $\{0, 1, \ldots, 2^p - 1\}$. Let $f : \mathbb{R}^{mk} \to \{0, \ldots, 2^p - 1\}$ be the mapping given by the control system, and denote the class of all such mappings by $\mathcal{F}$. For each $f \in \mathcal{F}$ introduce a loss function $L_{0\text{-}1,f}(z, a) = L_{0\text{-}1}(f(z), a) = 1$, when $f(z) \neq a$, and 0 otherwise. Define the class $L_{0\text{-}1,\mathcal{F}} = \{L_{0\text{-}1,f};\ f \in \mathcal{F}\}$.

In order to calculate the value of the output, after determining an appropriate piece, one needs to know the truth values of the expressions $P_1 > 0, Q_1 > 0, \ldots, P_p > 0$, and $Q_p > 0$, where P's and Q's are polynomials on inputs and parameters of the control system. To evaluate the value of the loss function $L_{0\text{-}1,f}(z, a)$, one needs the truth values of $y = 0, y = 1, \ldots, y = 2^p - 2$.

In the general case one needs $2nk + 2p(2k + 1)^n + 2^p - 1$ truth values. As this procedure evaluates only polynomials, we can use the Goldberg–Jerrum bound again. The maximum degree of the polynomials is $m2n^2k4(n + \ell_{\max}) + 1$, and the total number of parameters is $2pn^2m + 4n + p$, where the last term comes from the initial condition. ∎

6. A fat-shattering bound. We begin this section by proving Theorems 3.6 and 3.7. As a corollary of Theorem 3.7 we prove the fat-shattering bound appearing in Theorem 2.7 bounding the sample complexity for proper agnostic learning.

Proof of Theorem 3.6. For the first part of the proof we use a generic set B for the parameters. Assume that we can γ-shatter a set of inputs $\{u_1, \ldots, u_d\}$ and there exists $\{r_1, \ldots, r_d\}$ such that, for each assignment $b \in \{0, 1\}^d$, there exists a $\lambda \in B$ such that

$$\begin{aligned} F(\lambda, u_i) &\geq r_i + \gamma \quad \text{if } b_i = 1, \text{ and} \\ F(\lambda, u_i) &\leq r_i - \gamma \quad \text{otherwise.} \end{aligned}$$

We write $\lambda \sim \mu$ if and only if the parameters λ and μ give the same assignment for all $\{u_1, \ldots, u_d\}$. Further, let $\Lambda = \{\lambda_1, \ldots, \lambda_{2^d}\}$ be a collection of parameters that shatter $\{u_1, \ldots, u_d\}$, and let $\lambda_i, \lambda_j \in \Lambda$. Now $\lambda_i \not\sim \lambda_j$ implies that there exist $u^* \in \{u_1, \ldots, u_d\}$ and $r^* \in \{r_1, \ldots, r_d\}$ such that $F(\lambda_i, u^*) \geq \gamma + r^*$ and $F(\lambda_j, u^*) \leq$

$\gamma - r^*$, or vice versa. Hence $2\gamma \le |F(\lambda_i, u^*) - F(\lambda_j, u^*)| \le L\|\lambda_i - \lambda_j\|$ and so $\|\lambda_i - \lambda_j\| \ge 2\gamma/L$. That is, the set Λ of cardinality 2^d is a $2\gamma/L$-separated set in B. Now the fat-shattering bounds follow by calculating $2\gamma/L$-packing numbers for different sets B.

If $B = B^k_\infty(C)$, the maximum possible cardinality for an ϵ-separated set is $\lfloor 2C/\epsilon \rfloor^k$, and thus

$$2^d \le \left\lfloor \frac{2C}{2\gamma/L} \right\rfloor^k = \left\lfloor \frac{CL}{\gamma} \right\rfloor^k,$$

and solving for d yields $d \le k \log_2 \lfloor CL/\gamma \rfloor$.

Similarly, if $B = \bar{B}^k_\infty(C)$, the maximum possible cardinality for an ϵ-separated set is $(1 + \lfloor 2C/\epsilon \rfloor)^k$ and by a similar argument we arrive at the bound $d \le k \log_2(1 + \lfloor LC/\gamma \rfloor)$.

For $B = \bar{B}^k_2(C)$, let $P(\epsilon)$ be a collection of ϵ-separated sets in $\bar{B}^k_2(C)$, and let $|P(\epsilon)|$ denote its cardinality. As all open balls with radius $\epsilon/2$ with centers at ϵ-separated points have to be disjoint and their union has to be inside a ball of radius $C+\epsilon/2$, we get that $|P(\epsilon)|\alpha(k)(\epsilon/2)^k \le \alpha(k)(C+\epsilon/2)^k$, where $\alpha(k) = \pi^{k/2}/\Gamma(k/2+1)$ is the volume of a unit ball in $\mathbb{R}^k$. Hence $|P(\epsilon)| \le (C + 2/\epsilon)^k$ and $2^d \le (C + L/\gamma)^k$; i.e., $d \le k \log_2(C + L/\gamma)$. □

Next we prove Theorem 3.7 by applying the Lipschitz bound to a control system.

Proof of Theorem 3.7. Our aim is to compute the Lipschitz constant associated to the control system in Definition 2.6, and then we apply Theorem 3.6.

Denote the system parameters $(\alpha_{11}, \dots, \alpha_{nm}, a_1, b_1, \dots, a_r, b_r)$ by λ and assume $\|\lambda\|_\infty < 1$. Let

$$F(\lambda, u) = y(\tau) = \int_0^\tau \sum_{i=1}^m \sum_{\ell=1}^n \alpha_{i\ell} \xi_\ell(t) u_i(\tau - t) dt.$$

Functions $\xi_1(t), \dots, \xi_n(t)$ are of the form $\xi(t) = t^c e^{at} \sin(bt)$ or $\xi(t) = t^c e^{at} \cos(bt)$, where $a + ib$ is an eigenvalue of A and $c \in \{0, \dots, n-1\}$. Thus taking a partial derivative with respect to a or b will increase the power of t by one and change the trigonometric functions. Therefore,

$$\left| \frac{\partial F(\lambda, u)}{\partial \alpha_{\kappa\rho}} \right| = \left| \frac{\partial y(\tau)}{\partial \alpha_{\kappa\rho}} \right| = \left| \sum_{i=1}^m \sum_{\ell=1}^n d(i, \ell) \int_0^\tau \xi_\ell(t) u_i(\tau - t) dt \right|$$
$$\le nm \int_0^\tau |\xi_\ell(t) u_i(\tau - t)|\, dt \le nm\tau^{n-1} e^\tau M,$$

because $\sup_{t\in[0,\tau]} |\xi_\ell(t)| \le e^\tau \tau^n$ and $d(i, \ell) = \partial\alpha_{ij}/\partial\alpha_{\kappa\rho} = 1$ if $(i, \ell) = (\kappa, \rho)$ and zero otherwise. Similarly we calculate

$$\left| \frac{\partial F(\lambda, u)}{\partial a_\kappa} \right| = \left| \frac{\partial y(\tau)}{\partial a_\kappa} \right| \le nm\tau^n e^\tau M \text{ and}$$
$$\left| \frac{\partial F(\lambda, u)}{\partial b_\kappa} \right| = \left| \frac{\partial y(\tau)}{\partial b_\kappa} \right| \le nm\tau^n e^\tau M$$

as $\sup_{t\in[0,\tau]} |\frac{\partial \xi_\ell(t)}{\partial_\kappa}| \le e^\tau \tau^n$ and $\sup_{t\in[0,\tau]} |\frac{\partial \xi_\ell(t)}{\partial b_\kappa}| \le e^\tau \tau^n$.

Now the Lipschitz constant can be taken to be $L = n^2 m e^\tau \tau^n M$ as

$$|F(\lambda, u) - F(\lambda^*, u)| = |\nabla F \cdot (\lambda - \lambda^*)| \le L\|\lambda - \lambda^*\|_\infty.$$

The number of system parameters is at most $nm+n=(m+1)n$ and we get the level fat-shattering bound by applying Theorem 3.6 with space dimension $n(m+1)$ and $L=n^2me^\tau\tau^nM$. ∎

As a corollary, we combine the above result together with a pseudodimension bound to prove the fat-shattering bound given in Theorem 2.7.

COROLLARY 6.1 (fat-shattering bound in Theorem 2.7). *Assume that the system* $\dot{x}=Ax+Bu$, $y=Cx$, $x(0)=0$, *can be parameterized by* $\lambda\in\mathbb{R}^{n(m+1)}$ *as in Definition* 2.6 *with* $\|\lambda\|_\infty<1$, *and assume in addition that the control is given by* $u=G\omega$, *where the input basis functions* ω_j *are in* Ω *given by* (3). *We denote the corresponding control system class by*

$$\mathcal{F}_B=\{F(\lambda,\cdot):U\to\mathbb{R};\ \lambda\in B\}.$$

Then

$$\operatorname{fat}_\gamma(\mathcal{F}_B)\le\min\begin{cases}(m+1)n\log_2\left\lfloor\frac{n^2m\tau^ne^\tau kM}{\gamma}\right\rfloor,\\ 2(m+4)n\log_2\big(8e\big(nmk4(n+\ell_{\max})+1\big)\big(2nk+2(2k+1)^n\big)\big),\end{cases}$$

where $\ell_{\max}$ *is given by* (3) *and* (4) *and* M *is a constant satisfying*

$$\int_0^\tau|u_i(\tau-t)|dt\le kM$$

for all $i=1,\dots,m$.

Proof. The first part of the bound follows from Theorem 3.7 with kM in place of M.

The remaining part of the bound comes from the pseudodimension bound. First we derive the associated VC dimension bound. As we assumed that A has a fixed Jordan block structure, every entry of $e^{A(1-t)}$ is a linear combination of n functions $\xi_1(t),\dots,\xi_n(t)$. (That is, we do not need to consider all possible functions over different Jordan block structures.) This implies that in the Goldberg–Jerrum argument of section 5.1 we can take $\ell=mn+4n$, $d=nmk4(n+\ell_{\max})+1$, and $s=2nk+2(2k+1)^n$. Moreover, in that section the VC dimension bounds were derived for the time interval $[0,1]$. However, the upper bound depends on the number of system parameters and the degrees of polynomials to be evaluated. Changing the time interval to $[0,\tau]$ means just that we replace the eigenvalue parameters (referring to the proof of Proposition 5.1) $a,b,e^a\cos b,e^a\sin b$ by $a\tau,b\tau,e^{a\tau}\cos b\tau,e^{a\tau}\sin b\tau$.

The above bound is also a bound for the pseudodimension. Observe that for $\mathcal{G}=\{g:X\to\mathbb{R}\}$, the pseudodimension can be defined as $\operatorname{PD}(\mathcal{G})=\operatorname{VC}\{\operatorname{Ind}(x,y)=\operatorname{sign}(g(x)-y);\ g\in\mathcal{G}\}$. Hence we want to study the VC dimension associated to $\operatorname{sign}(y(\tau)-z)=\operatorname{sign}(P/Q-z)=\operatorname{sign}(\hat{P}/Q)$, where $\hat{P}=P-zQ$ has the same degree as P with respect to the parameters. Here z is a new input, but the bound utilizing Goldberg–Jerrum technique does not depend on the dimension of the inputs, and hence the above VC dimension bound is also a bound for the pseudodimension. (Note that here in the scale sensitive setting we do not apply the pseudodimension results of section 5.3 using loss functions, as those rescaled the outputs.) ∎

7. A class of systems with VC dimension *k*. For the control system (2) with scalar control $u(t)=\sum_{i=1}^k g_i\omega_i(t)$ and unrestricted $\omega_1,\dots,\omega_k$, the standard half-space argument gives an upper bound k. This bound is tight. We will give an example of a single-input, single-output one-parameter family of control systems in dimension two

that has VC dimension k, when the controls are of the form $u(t) = \sum_{i=1}^{k} g_i \omega_i(t)$ and $\omega_i(1-t) = 1_{[2^{-i}, 2^{-i}+2^{\alpha}]}$, where $\alpha = -2(k+1)$.

Consider a control system

$$\begin{aligned} \dot{x_1} &= x_2, \\ \dot{x_2} &= -\lambda^2 x_1 + u, \\ y &= -x_1. \end{aligned} \tag{13}$$

For time interval [0,1] and initial condition $(x_1, x_2) = (0,0)$, the output is given by $y(1) = \int_0^1 \sin(\lambda t) u(1-t)dt$.

LEMMA 7.1. *Controls* $\{\omega_1, \ldots, \omega_k\}$ *such that* $\omega_i(1-t) = 1_{[2^{-i}, 2^{-i}+2^{\alpha}]}$, *where* $\alpha = -2(k+1)$, *are shattered by the control system* (13) *with sign-observations.*

Proof. Let $T = \{2^{-i},\ i = 1, \ldots, k\}$ and $J \subseteq T$. Define $\lambda_J = \pi \sum_{i=1}^{k} a_i 2^i$, where $a_i = 1$ if $2^{-i} \notin J$ and $a_i = 0$ otherwise. Now if $t = 2^{-\ell}$, then

$$\lambda_J t = \pi \sum_{i=1}^{k} a_i 2^{i-\ell} = \pi \left(\underbrace{\sum_{i=1}^{\ell-1} a_i 2^{i-\ell}}_{1/2\, c_2} + a_\ell + \underbrace{\sum_{i=\ell+1}^{k} a_i 2^{i-\ell}}_{2c_1} \right),$$

where $c_1 \in \mathbb{N}$ and $0 \le c_2 < 2$.

Hence $\sin(\lambda_J t) = \sin(\pi(1/2\, c_2 + a_\ell))$. Note that if $a_\ell = 0$, then $1/2\, c_2 + a_\ell \in [0, 1 - 2^{-\ell}]$, and if $a_\ell = 1$, then $1/2\, c_2 + a_\ell \in [1, 2 - 2^{-\ell}]$. Thus $\sin(\pi(1/2\, c_2 + a_\ell)) \ge 0$ if $a_\ell = 0$, and $\sin(\pi(1/2\, c_2 + a_\ell)) \le 0$ if $a_\ell = 1$. Therefore, $\sin(\lambda_J t) \ge 0 \Leftrightarrow a_\ell = 0$ $\sin(\lambda_J t) \ge 0 \Leftrightarrow t \in J$. Further,

$$\int_{2^{-\ell}}^{2^{-\ell}+2^{\alpha}} \sin(\lambda_J t)dt \ge 0 \Leftrightarrow a_\ell = 0,$$

where α is taken so that

$$\sum_{j=1}^{k} 2^j 2^{\alpha} \le 2^{-(k+1)}. \tag{14}$$

This ensures that when $\ell \le k$ and $t \in [2^{-\ell}, 2^{-\ell} + 2^{\alpha}]$, $\lambda_J t \in [0, \pi(1 - 2^{-\ell} + \sum_{j=1}^{k} 2^j 2^{\alpha})] \subset [0, \pi)$ if $a_\ell = 0$ or similarly $\lambda_J t \in [\pi, 2\pi)$ if $a_\ell = 1$. In (14) we can take $\alpha = -2(k+1)$ as $\sum_{j=1}^{k} 2^j = 2^{k+1} - 2$.

In this way the integrand in $\int_{2^{-\ell}}^{2^{-\ell}+2^{\alpha}} \sin(\lambda_J t)dt$ is either positive or negative.

For $S \subseteq \{1, \ldots, k\}$, let $J = \{2^{-i}, i \in S\}$. For each ω_i,

$$\int_0^1 \sin(\lambda_J t)\omega_i(1-t)dt = \int_{2^{-i}}^{2^{-i}+2^{\alpha}} \sin(\lambda_J t)dt > 0 \Leftrightarrow i \in S;$$

i.e., the set of controls $\{\omega_1, \ldots, \omega_k\}$ is shattered by the mapping

$$\omega_i \mapsto \operatorname{sign}\left[\int_0^1 \sin(\lambda_J t)\omega_i(1-t)dt \right]. \qquad \square$$

Appendix. An example of the Goldberg–Jerrum bound.

We begin this appendix with an informal discussion on the Goldberg–Jerrum technique used to prove the VC dimension upper bounds in this paper.

We want to write $y(1) = P/Q$, where P and Q are polynomials. Unfortunately, the value of $\operatorname{sign} y(1)$ cannot be obtained by just evaluating P and Q since Q may have zeros. Therefore, we need to write

$$y(1) = \begin{cases} P_1/Q_1 & \text{if } f_1 \neq 0, \dots, f_\mu \neq 0, \\ \quad \vdots & \\ P_\gamma/Q_\gamma & \text{if } f_1 = 0, \dots, f_\mu \neq 0 \end{cases}$$

so that after evaluating μ polynomials $f_1, \dots, f_\mu$ we can pick a definition P_i/Q_i without poles in a region defined by the μ polynomials. When $y(1)$ is defined in this way, $\operatorname{sign} y(1)$ can be easily expressed by a Boolean formula evaluating $2\gamma + \mu$ polynomial inequalities and equalities.

For simplicity we assume that $p = 1$ and the initial condition $x(0) = 0$. Then using the notation of Proposition 5.1 we write

$$y(1) = \sum_{i=1}^{m} \sum_{r=1}^{n} \sum_{\ell=1}^{2n} \alpha_{ir\ell} \sum_{j=1}^{k} g_{ij} \int_0^1 \xi_\ell(x_{r1}, x_{r2}, t) \omega_j(t) dt,$$

and by the proof of Lemma 4.4

$$\int_0^1 \xi_\ell(x_{r1}, x_{r2}, t) \omega_j(t) dt = \frac{P_{\ell j}}{\left((x_{r1} + \alpha_j)^2 + (x_{r2} + \beta_j)^2\right)^{z_{\ell j}} \left((x_{r1} + \alpha_j)^2 + (x_{r2} - \beta_j)^2\right)^{z_{\ell j}}},$$

where $P_{\ell j}$ is some polynomial, $z_{\ell j} \in \mathbb{N}$, and $\omega_j(t) = t^{\ell_j} e^{\alpha_j t} \sin(\beta_j t)$ or $\omega_j(t) = t^{\ell_j} e^{\alpha_j t} \cos(\beta_j t)$. Hence the denominator of $\sum_{j=1}^{k} g_{ij} \int_0^1 \xi_\ell(x_{r1}, x_{r2}, t) \omega_j(t) dt$ is

$$\begin{aligned} &\left((x_{r1} + \alpha_1)^2 + (x_{r2} + \beta_1)^2\right)^{z_{\ell 1}} \left((x_{r1} + \alpha_1)^2 + (x_{r2} - \beta_1)^2\right)^{z_{\ell 1}} \\ &\quad \times \cdots \times \left((x_{r1} + \alpha_k)^2 + (x_{r2} + \beta_k)^2\right)^{z_{\ell k}} \left((x_{r1} + \alpha_k)^2 + (x_{r2} - \beta_k)^2\right)^{z_{\ell k}}. \end{aligned}$$

By carrying out all summations $y(1) = P/Q$. The denominator Q consists of the product

$$\begin{aligned} \prod_{r=1}^{n} \Big(&\left((x_{r1} + \alpha_1)^2 + (x_{r2} + \beta_1)^2\right)^* \left((x_{r1} + \alpha_1)^2 + (x_{r2} - \beta_1)^2\right)^* \\ &\times \cdots \times \left((x_{r1} + \alpha_k)^2 + (x_{r2} + \beta_k)^2\right)^* \left((x_{r1} + \alpha_k)^2 + (x_{r2} - \beta_k)^2\right)^* \Big), \end{aligned}$$

where $*$'s stand for some unspecified powers. Hence the zeros of Q are determined by $2nk$ polynomials $f_{ij1} = (x_{i1} + \alpha_j)^2 + (x_{i2} + \beta_j)^2$, $f_{ij2} = (x_{i1} + \alpha_j)^2 + (x_{i2} - \beta_j)^2$, and $i = 1, \dots, n$, $j = 1, \dots, k$. The number of different sign assignments determining γ is calculated as in the proof of Lemma 5.2.

Example. The purpose of the following example is to illustrate the function $y = P/Q$ used in the Goldberg–Jerrum technique together with the sequence of polynomial evaluations involved and a table for the final output depending on the outcomes of the polynomial evaluations.

Take $m = 1$, $n = 2$, $k = 2$, and assume that A has complex eigenvalues $a \pm ib$. Take basis input functions to be $\omega_1(t) = e^t$ and $\omega_2(t) = e^{2t}$. Then $y(1) = \sum_{l=1}^{2} \alpha_l \sum_{j=1}^{2} g_j \int_0^1 \xi_l(t) \omega_j(t)\, dt$, where $\xi_1(t) = e^{at} \sin(bt)$, $\xi_2(t) = e^{at} \cos(bt)$, α_1, α_2, a, b, $e^a \sin b$, and $e^a \cos b$ are system parameters and g_1, g_2 are input parameters.

By using formulae

$$\int_0^1 e^{\tilde{a}t}\sin(\tilde{b}t)\,dt = \frac{e^{\tilde{a}}(\tilde{a}\sin\tilde{b} - \tilde{b}\cos\tilde{b}) + \tilde{b}}{\tilde{a}^2+\tilde{b}^2} \quad \text{and}$$
$$\int_0^1 e^{\tilde{a}t}\cos(\tilde{b}t)\,dt = \frac{e^{\tilde{a}}(\tilde{a}\cos\tilde{b} + \tilde{b}\sin\tilde{b}) - \tilde{a}}{\tilde{a}^2+\tilde{b}^2},$$

we calculate the integrals appearing in the rationality condition, and we call them r_{11}, r_{12}, r_{21}, and r_{22}:

$$\int_0^1 \xi_1(t)\omega_1(t)\,dt = \int_0^1 e^{(a+1)t}\sin(bt)\,dt = \begin{cases} r_{11} & \text{if } (a+1)^2+b^2 \neq 0, \\ 0 & \text{if } (a+1)^2+b^2 = 0, \end{cases}$$
$$\int_0^1 \xi_1(t)\omega_2(t)\,dt = \begin{cases} r_{12} & \text{if } (a+2)^2+b^2 \neq 0, \\ 0 & \text{if } (a+2)^2+b^2 = 0, \end{cases}$$
$$\int_0^1 \xi_2(t)\omega_1(t)\,dt = \begin{cases} r_{21} & \text{if } (a+1)^2+b^2 \neq 0, \\ 1 & \text{if } (a+1)^2+b^2 = 0, \end{cases}$$
$$\int_0^1 \xi_2(t)\omega_2(t)\,dt = \begin{cases} r_{22} & \text{if } (a+2)^2+b^2 \neq 0, \\ 1 & \text{if } (a+2)^2+b^2 = 0. \end{cases}$$

The computation of $\operatorname{sign} y(1) = \operatorname{sign}(\sum_{l=1}^2 \alpha_l \sum_{j=1}^2 g_j \int_0^1 \xi_l(t)\omega_j(t)\,dt)$ is divided into three cases:

- Case $(a+1)^2+b^2 \neq 0$, $(a+2)^2+b^2 \neq 0$:

$$\operatorname{sign} y(1) = \operatorname{sign}(\alpha_1 g_1 r_{11} + \alpha_1 g_2 r_{12} + \alpha_2 g_1 r_{21} + \alpha_2 g_2 r_{22}) = \operatorname{sign}\left(\frac{P_1}{Q_1}\right).$$

- Case $(a+1)^2+b^2 = 0$, $(a+2)^2+b^2 \neq 0$:

$$\operatorname{sign} y(1) = \operatorname{sign}(\alpha_1 g_2 r_{12} + \alpha_2 g_1 + \alpha_2 g_2 r_{22}) = \operatorname{sign}\left(\frac{P_2}{Q_2}\right).$$

- Case $(a+1)^2+b^2 \neq 0$, $(a+2)^2+b^2 = 0$:

$$\operatorname{sign} y(1) = \operatorname{sign}(\alpha_1 g_1 r_{11} + \alpha_2 g_2 r_{21} + \alpha_2 g_2) = \operatorname{sign}\left(\frac{P_3}{Q_3}\right).$$

Thus we have three different expressions of the form $\frac{P}{Q}$.

Next we form the Boolean formula, $F = \operatorname{sign} y(1)$, evaluating polynomials $f_1 = (a+1)^2+b^2 = 0$, $f_2 = (a+2)^2+b^2 = 0$, $P_i > 0$, $Q_i > 0$ for $i \in \{1,2,3\}$. In the following table 1 means true and 0 means false for the above polynomial evaluation ($** = 1$ or 0, i.e., extend the table).

$f_1=0$	$f_2=0$	$P_1>0$	$Q_1>0$	$P_2>0$	$Q_2>0$	$P_3>0$	$Q_3>0$	F
0	0	1	1	**	**	**	**	1
0	0	1	0	**	**	**	**	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	**	**	1	1	**	**	1
1	0	**	**	1	0	**	**	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	1	**	**	**	**	0	0	1

In this case (see the statement of Goldberg–Jerrum bounds), $\Lambda = \{\alpha_1, \alpha_2, a, b, e^a \cos b, e^a \sin b\}$, $X = \{g_1, g_2\}$, $s = 8$, $d = 12$, and $l = 6$.

REFERENCES

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, *Scale-sensitive dimensions, uniform convergence and learnability*, J. ACM, 44 (1997), pp. 615–631.

[2] M. Anthony and B. Bartlett, *Learning in neural networks: Theoretical foundations*, Cambridge University Press, Cambridge, UK, 1999.

[3] P. Bartlett, P. Long, and R. Williamson, *Fat-shattering and learnability of real-valued functions*, in Proceedings of the 7th Annual Conference on Computational Learning Theory, ACM, New York, 1994, p. 299.

[4] P. L. Bartlett and P. M. Long, *More theorems about scale-sensitive dimensions and learning*, in Proceedings of the 8th Annual Conference on Computational Learning Theory, Santa Cruz, CA, 1995, ACM, New York, 1995, pp. 392–401.

[5] P. L. Bartlett and P. M. Long, *Prediction, learning, uniform convergence and scale-sensitive dimensions*, J. Comput. System Sci., 56 (1998), pp. 174–190.

[6] B. DasGupta and E. D. Sontag, *Sample complexity for learning recurrent perceptron mappings*, IEEE Trans. Inform. Theory, 42 (1996), pp. 1479–1487.

[7] B. DasGupta and E. D. Sontag, *Sample complexity for learning recurrent perceptron mappings*, in Advances in Neural Information Processing Systems (NIPS'95), MIT Press, Cambridge, MA, 1996, pp. 204–210.

[8] B. Eisenberg and R. Rivest, *On the sample complexity of pac-learning using random and chosen examples*, in Proceedings of the 3rd Annual Workshop on Computational Learning Theory, Morgan Kaufmann, San Francisco, 1990, pp. 154–162.

[9] C.-N. Fiechter, *PAC adaptive control of linear systems*, in Proceedings of the 10th Annual Conference on Computational Learning Theory, ACM, New York, 1997, pp. 72–80.

[10] P. Goldberg and M. Jerrum, *Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers*, Machine Learning, 18 (1995), pp. 131–148.

[11] T. Johansen and E. Weyer, *On convergence proofs in system identification—a general principle using ideas from learning theory*, Systems Control Lett., 34 (1998), pp. 85–92.

[12] P. Koiran and E. D. Sontag, *Vapnik-Chervonenkis dimension of recurrent neural networks*, Discrete Appl. Math., 86 (1998), pp. 63–80.

[13] R. Koplon and E. D. Sontag, *Linear systems with sign-observations*, SIAM J. Control Optim., 31 (1993), pp. 1245–1266.

[14] P. Kuusela and D. Ocone, *Learning with side information: An example*, 2002, submitted.

[15] P. Kuusela and D. Ocone, *Learning with side information: PAC learning bounds*, J. Comput. System Sci., 68 (2004), pp. 521–545.

[16] L. Ljung, *PAC-learning and symptotic system identification theory*, in Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 2303–2307.

[17] D. Pollard, *Convergence of Stochastic Processes*, Springer Ser. Statist., Springer-Verlag, New York, 1984.

[18] J. Shawe-Taylor, M. Anthony, and N. L. Biggs, *Bounding sample size with the Vapnik-Chervonenkis dimension*, Discrete Appl. Math., 42 (1993), pp. 65–73.

[19] E. D. Sontag, *A learning result for continuous-time recurrent neural networks*, Systems Control Lett., 34 (1998), pp. 151–158.

[20] V. Vapnik and A. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory Probab. Appl., 16 (1971), pp. 264–280.

[21] M. Vidyasagar, *A Theory of Learning and Generalization; with Applications to Neural Networks and Control Systems*, Springer-Verlag, New York, 1997.

[22] R. Wenocur and R. Dudley, *Some special Vapnik–Chervonenkis classes*, Discrete Math., 33 (1981), pp. 313–318.

[23] A. Zador and B. Pearlmutter, *VC dimension of an integrate-and-fire neuron model*, Neural Computation, 8 (1996), pp. 611–624.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 899–921

THE TOPOLOGICAL ASYMPTOTIC FOR THE HELMHOLTZ EQUATION WITH DIRICHLET CONDITION ON THE BOUNDARY OF AN ARBITRARILY SHAPED HOLE*

JULIEN POMMIER† AND BESSEM SAMET‡

Abstract. The aim of the topological sensitivity analysis is to obtain an asymptotic expansion of a design functional with respect to the creation of a small hole in the domain. In this paper, such an expansion is obtained for the Helmholtz equation, in two and three space dimensions, with a Dirichlet condition on the boundary of an arbitrarily shaped hole. In this case, the main difficulty is related to the nonhomogeneous symbol of the Helmholtz operator. In the numerical part of this work, we will show that the topological sensitivity method is very promising for solving shape inverse problems in electromagnetic applications.

Key words. topological optimization, topological asymptotic, topological gradient, nonhomogeneous problem, Helmholtz equation, shape inversion, electromagnetic applications, inverse scattering

AMS subject classifications. 49Q10, 49Q12, 78A25, 78A40, 78A45, 78A50, 35J05

DOI. 10.1137/S036301290241616X

1. Introduction.

The same numerical methods are generally used in shape inversion and optimal shape design. There are mainly two categories of shape inversion or shape optimization methods. In the first category we deform continuously the boundary of the object to be optimized in order to decrease a given cost function [5, 20, 25, 28, 31]. The final shape has the same topology as the initial shape given by the designer. Therefore, to reach the optimal geometry, we need a priori knowledge of its topology. However, the topology of the optimal shape is often the main unknown in object detection problems. For example, the knowledge of the number and the locations of buried mines is more important than their accurate shapes. The second category of algorithms allows topology modifications. Many important contributions in this field are concerned with structural mechanics and, in particular, the optimization of the compliance (external work) subject to a volume constraint [4, 16]. In view of the fact that the optimal structure has generally a large number of small holes, most authors [1, 3, 14] have considered composite material optimization. Using the homogenization theory, Allaire and Kohn [1] exhibit a class of laminated materials with an explicit expression for the optimal material at any point of the structure. In this case, the optimal solution is not a classical design—it is a distribution of composite materials. Then penalization methods must be applied in order to retrieve a realistic shape. For all these reasons, global optimization methods are used to solve more general problems [15, 26]. Unfortunately these methods are quite slow.

More recently, Eschenauer and Olhoff [7], Schumacher [27], Céa et al. [6], Garreau, Guillaume, and Masmoudi [8], Sokolowski and Zochowski [29, 30], and Nazarov and Sokolowski [21] presented a method to obtain the optimal topology by calculating the so-called topological gradient (or topological derivative). This gradient is a function

*Received by the editors October 16, 2002; accepted for publication (in revised form) December 11, 2003; published electronically September 18, 2004.
http://www.siam.org/journals/sicon/43-3/41616.html
†Département de Génie Mathématique, INSA Toulouse and CNRS UMR 5640 MIP, Complex Scientifique de Rangueil, F-31077 Toulouse cedex 4, France (pommier@gmm.insa-tlse.fr).
‡UMR MIG, Université Paul Sabatier and CNRS UMR 5640 MIP, 118 route de Narbonne, F-31062 Toulouse cedex, France (samet@mip.ups-tlse.fr).

defined in the domain of interest where, at each point, it gives the sensitivity of the cost function when a small hole is created at that point. This approach seems to be general and efficient. To present the basic idea, we consider Ω a domain of $\mathbb{R}^n$, where n equals 2 or 3, and $j(\Omega) = J(u_\Omega)$ a cost function to be minimized, where u_Ω is the solution to a given PDE problem defined in Ω. For $\varepsilon > 0$, let $\Omega_\varepsilon = \Omega \backslash \overline{x_0 + \varepsilon\omega}$ be the subset obtained by removing a small part $\overline{x_0 + \varepsilon\omega}$ from Ω, where $x_0 \in \Omega$ and $\omega \subset \mathbb{R}^n$ is a fixed open and bounded subset containing the origin. We can generally prove that the variation of the criterion is given by the asymptotic expansion

$$j(\Omega_\varepsilon) = j(\Omega) + f(\varepsilon)g(x_0) + o(f(\varepsilon)), \tag{1.1}$$

$$\lim_{\varepsilon \to 0} f(\varepsilon) = 0, \quad f(\varepsilon) > 0. \tag{1.2}$$

This expansion is called the topological asymptotic. To minimize the criterion, we have to create holes where g (called the topological gradient) is negative.

In this paper, using the adjoint method and the domain truncation technique introduced in [17], we compute the topological asymptotic expansion for the Helmholtz equation in two and three space dimensions with a Dirichlet condition on the boundary of an arbitrarily shaped hole. The originality of this work is that the symbol of the Helmholtz operator is nonhomogeneous. The basic idea is to say that the leading term of the topological asymptotic expansion is given by the principal part of the operator in the case of a Dirichlet condition on the boundary of the hole. Our work generalizes the contribution of Guillaume and Sid Idris [9] for the Poisson equation and is easily applicable to other problems for which the symbol of the operator is nonhomogeneous, as, for example, the quasi-Stokes problem and the elastic waves problem. In the numerical part, we present some applications that illustrate the ability of the topological sensitivity approach to solve inverse scattering problems.

As a background to our work, we cite the contributions of Il'in [11, 12, 13] for the construction of asymptotic expansions of solutions to boundary value problems in domains with small holes, as in the case of second order scalar equations, by the use of the method of matched asymptotic expansions. Various spectral problems in domains with small holes are investigated by Maz'ya et al. [23, 24, 18, 22]. In [32], Vogelius and Volkov provided a rigorous derivation for solutions to the time-harmonic Maxwell's equations of a transverse electric (TE) nature, in the presence of a finite number of diametrically small inhomogeneities. Based on layer potential techniques, Ammari and Kang [2] provided a rigorous derivation of complete asymptotic expansions for solutions to the Helmholtz equation in two and three dimensions, in the presence of small inhomogeneities in the domain. In our work, we derive asymptotic expansions not for solutions, but for a given cost function.

The generalized adjoint method is recalled in section 2. Next, the formulation of the Helmholtz problem is presented in section 3 and its truncated version is described in section 4. Section 5 presents the main results whose proofs are given in section 6. Finally, numerical examples illustrate in section 7 the abilities of the topological sensitivity to solve inverse scattering problems.

2. A generalized adjoint method. In this section, the generalized adjoint method introduced in [17, 8] is slightly modified. The first modification is due to the fact that the cost function is defined in a $\mathbb{C}$-Hilbert space and takes values in $\mathbb{R}$; then it is not differentiable. For this reason, the differentiability property is replaced by the formulation (2.5). The second modification is due to the fact that the sesquilinear form associated with our problem is not coercive. For this reason, the coercivity property is replaced by the inf-sup condition (see Hypothesis 2).

Let $\mathcal{V}$ be a fixed complex Hilbert space. For $\varepsilon \geq 0$, let $a_\varepsilon(.,.)$ be a sesquilinear and continuous form on $\mathcal{V}$ and let l_ε be a semilinear and continuous form on $\mathcal{V}$. We consider the following assumptions.

Hypothesis 1. There exists a sesquilinear and continuous form δa, a semilinear and continuous form δ_l, and a real function $f(\varepsilon) > 0$ defined on $\mathbb{R}^*{}_+$ such that

$$\lim_{\varepsilon \to 0} f(\varepsilon) = 0, \tag{2.1}$$

$$\|a_\varepsilon - a_0 - f(\varepsilon)\delta a\|_{\mathcal{L}_2(\mathcal{V})} = o(f(\varepsilon)), \tag{2.2}$$

$$\|l_\varepsilon - l_0 - f(\varepsilon)\delta_l\|_{\mathcal{L}(\mathcal{V})} = o(f(\varepsilon)), \tag{2.3}$$

where $\mathcal{L}(\mathcal{V})$ (respectively, $\mathcal{L}_2(\mathcal{V})$) denotes the space of continuous and semilinear (respectively, sesquilinear) forms on $\mathcal{V}$.

Hypothesis 2. There exists a constant $\alpha > 0$ such that

$$\inf_{u \neq 0} \sup_{v \neq 0} \frac{|a_0(u,v)|}{\|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}}} \geq \alpha.$$

We say that a_0 satisfies the inf-sup condition.

According to (2.2), there exists a constant $\beta > 0$ independent of ε such that

$$\inf_{u \neq 0} \sup_{v \neq 0} \frac{|a_\varepsilon(u,v)|}{\|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}}} \geq \beta.$$

For $\varepsilon \geq 0$, let u_ε be the solution to the following problem: Find $u_\varepsilon \in \mathcal{V}$ such that

$$a_\varepsilon(u_\varepsilon, v) = l_\varepsilon(v) \qquad \forall v \in \mathcal{V}. \tag{2.4}$$

We have the following lemma.

LEMMA 2.1. *If Hypotheses* 1 *and* 2 *are satisfied, then*

$$\|u_\varepsilon - u_0\|_{\mathcal{V}} = O(f(\varepsilon)).$$

Proof. It follows from Hypothesis 2 that there exists $v_\varepsilon \in \mathcal{V}, v_\varepsilon \neq 0$, such that

$$\beta \|u_\varepsilon - u_0\|_{\mathcal{V}} \|v_\varepsilon\|_{\mathcal{V}} \leq |a_\varepsilon(u_\varepsilon - u_0, v_\varepsilon)|,$$

which implies

$$\begin{aligned}
\beta \|u_\varepsilon - u_0\|_{\mathcal{V}} \|v_\varepsilon\|_{\mathcal{V}} &\leq |a_\varepsilon(u_0, v_\varepsilon) - l_\varepsilon(v_\varepsilon)| \\
&= |a_\varepsilon(u_0, v_\varepsilon) - (l_\varepsilon - l_0 - f(\varepsilon)\delta_l)(v_\varepsilon) - l_0(v_\varepsilon) - f(\varepsilon)\delta_l(v_\varepsilon)| \\
&= |(a_\varepsilon(u_0, v_\varepsilon) - a_0(u_0, v_\varepsilon)) - (l_\varepsilon - l_0 - f(\varepsilon)\delta_l)(v_\varepsilon) - f(\varepsilon)\delta_l(v_\varepsilon)| \\
&\leq |a_\varepsilon(u_0, v_\varepsilon) - a_0(u_0, v_\varepsilon) - f(\varepsilon)\delta a(u_0, v_\varepsilon)| \\
&\quad + |l_\varepsilon(v_\varepsilon) - l_0(v_\varepsilon) - f(\varepsilon)\delta_l(v_\varepsilon)| + f(\varepsilon)(|\delta a(u_0, v_\varepsilon)| + |\delta_l(v_\varepsilon)|).
\end{aligned}$$

Using Hypothesis 1 we obtain

$$\beta \parallel u_\varepsilon - u_0 \parallel_{\mathcal{V}} \|v_\varepsilon\|_{\mathcal{V}} \leq \big(o(f(\varepsilon)) + f(\varepsilon)(\|\delta a\|_{\mathcal{L}_2(\mathcal{V})} \|u_0\|_{\mathcal{V}} + \|\delta_l\|_{\mathcal{L}(\mathcal{V})})\big) \|v_\varepsilon\|_{\mathcal{V}}. \qquad \square$$

Consider now a cost function $j(\varepsilon) = J(u_\varepsilon)$, where the functional J satisfies

$$J(u+h) = J(u) + \Re(L_u(h)) + o(\|h\|) \qquad \forall u, h \in \mathcal{V}, \tag{2.5}$$

where L_u is a linear and continuous form on $\mathcal{V}$.

For $\varepsilon \geq 0$, we define the Lagrangian operator $\mathcal{L}_\varepsilon$ by

$$\mathcal{L}_\varepsilon(u,v) = J(u) + a_\varepsilon(u,v) - l_\varepsilon(v) \qquad \forall u,v \in \mathcal{V}.$$

The next theorem gives the asymptotic expansion of $j(\varepsilon)$.

THEOREM 2.2. *If Hypotheses* 1 *and* 2 *are satisfied, then*

$$j(\varepsilon) - j(0) = f(\varepsilon)\Re(\delta_{\mathcal{L}}(u_0,p_0)) + o(f(\varepsilon)), \tag{2.6}$$

where u_0 is the solution to (2.4) *with $\varepsilon = 0$, and p_0 is the solution to the following adjoint problem: Find $p_0 \in \mathcal{V}$ such that*

$$a_0(v,p_0) = -L_{u_0}(v) \qquad \forall v \in \mathcal{V} \tag{2.7}$$

and

$$\delta_{\mathcal{L}}(u,v) = \delta a(u,v) - \delta_l(v) \qquad \forall u,v \in \mathcal{V}.$$

Proof. We have that

$$j(\varepsilon) = \mathcal{L}_\varepsilon(u_\varepsilon, v) \qquad \forall \varepsilon \geq 0 \qquad \forall v \in \mathcal{V}.$$

Next, choosing $v = p_0$, we obtain

$$\begin{aligned} j(\varepsilon) - j(0) &= \mathcal{L}_\varepsilon(u_\varepsilon,p_0) - \mathcal{L}_0(u_0,p_0) \\ &= J(u_\varepsilon) - J(u_0) + a_\varepsilon(u_\varepsilon,p_0) - a_0(u_0,p_0) + l_0(p_0) - l_\varepsilon(p_0) \\ &= J(u_\varepsilon) - J(u_0) + \Re(a_\varepsilon(u_\varepsilon,p_0) - a_0(u_0,p_0)) - \Re(l_\varepsilon(p_0) - l_0(p_0)) \\ &= J(u_\varepsilon) - J(u_0) + \Re(a_\varepsilon(u_\varepsilon,p_0) - a_0(u_\varepsilon,p_0) + a_0(u_\varepsilon - u_0,p_0)) \\ &\quad -\Re(l_\varepsilon(p_0) - l_0(p_0) - f(\varepsilon)\delta_l(p_0)) - f(\varepsilon)\Re(\delta_l(p_0)). \end{aligned}$$

Using (2.5), we have that

$$J(u_\varepsilon) - J(u_0) = \Re(L_{u_0}(u_\varepsilon - u_0)) + o(\|u_\varepsilon - u_0\|).$$

Hence,

$$\begin{aligned} j(\varepsilon) - j(0) = \Re(a_\varepsilon(u_\varepsilon,p_0) - a_0(u_\varepsilon,p_0)) + \Re(a_0(u_\varepsilon - u_0,p_0) + L_{u_0}(u_\varepsilon - u_0)) \\ +o(\|u_\varepsilon - u_0\|) - \Re(l_\varepsilon(p_0) - l_0(p_0) - f(\varepsilon)\delta_l(p_0)) - f(\varepsilon)\Re(\delta_l(p_0)). \end{aligned}$$

Using that p_0 is the adjoint solution, we obtain

$$\begin{aligned} j(\varepsilon) - j(0) &= \Re(a_\varepsilon(u_\varepsilon,p_0) - a_0(u_\varepsilon,p_0)) + o(\|u_\varepsilon - u_0\|) \\ &\quad -\Re(l_\varepsilon(p_0) - l_0(p_0) - f(\varepsilon)\delta_l(p_0)) - f(\varepsilon)\Re(\delta_l(p_0)) \\ &= \Re((a_\varepsilon - a_0)(u_0,p_0)) + \Re((a_\varepsilon - a_0)(u_\varepsilon - u_0,p_0)) + o(\|u_\varepsilon - u_0\|) \\ &\quad -\Re(l_\varepsilon(p_0) - l_0(p_0) - f(\varepsilon)\delta_l(p_0)) - f(\varepsilon)\Re(\delta_l(p_0)). \end{aligned}$$

It follows from Hypothesis 1 that

$$\begin{aligned} j(\varepsilon) - j(0) = f(\varepsilon)\Re(\delta a(u_0,p_0)) + o(f(\varepsilon)) + f(\varepsilon)\Re(\delta a(u_\varepsilon - u_0,p_0)) + o(f(\varepsilon))\|u_\varepsilon - u_0\| \\ +o(\|u_\varepsilon - u_0\|) - f(\varepsilon)\Re(\delta_l(p_0)). \end{aligned}$$

Finally, from Lemma 2.1 and the hypothesis $\lim_{\varepsilon\to 0} f(\varepsilon) = 0$, we have

$$j(\varepsilon) = j(0) + f(\varepsilon)\Re(\delta a(u_0,p_0) - \delta_l(p_0)) + o(f(\varepsilon)),$$

since δ_a is continuous by assumption. ∎

3. The Helmholtz problem in a domain with a small hole. Let Ω be an open and bounded subset of $\mathbb{R}^n$ with boundary $\Gamma = \Gamma_0 \cup \Gamma_1$, $n = 2$ or 3. The Helmholtz problem is

$$\left\{\begin{array}{lll} \Delta u_\Omega + k^2 u_\Omega &= 0 & \text{in } \Omega, \\ u_\Omega &= 0 & \text{on } \Gamma_0, \\ \dfrac{\partial u_\Omega}{\partial n} &= \Lambda u_\Omega + \Theta & \text{on } \Gamma_1, \end{array}\right. \tag{3.1}$$

where $k \in \mathbb{R}^*$, $\Theta \in H_{00}^{\frac{1}{2}}(\Gamma_1)'$, and $\Lambda \in \mathcal{L}(H_{00}^{\frac{1}{2}}(\Gamma_1), H_{00}^{\frac{1}{2}}(\Gamma_1)')$.

We define

$$\left\{\begin{array}{lll} \mathcal{V}(\Omega) &=& \{v \in H^1(\Omega), v = 0 \text{ on } \Gamma_0\}, \\ a(\Omega, u, v) &=& \displaystyle\int_\Omega \nabla u.\nabla \overline{v}\, dx - k^2 \int_\Omega u\overline{v}\, dx - \langle \Lambda u, \overline{v}\rangle, \\ \ell(v) &=& \langle \Theta, \overline{v}\rangle, \end{array}\right. \tag{3.2}$$

where $\langle , \rangle$ is the duality product between $H_{00}^{\frac{1}{2}}(\Gamma_1)'$ and $H_{00}^{\frac{1}{2}}(\Gamma_1)$. The variational formulation associated with (3.1) is the following: Find $u_\Omega \in \mathcal{V}(\Omega)$ such that

$$a(\Omega, u_\Omega, v) = \ell(v) \qquad \forall v \in \mathcal{V}(\Omega). \tag{3.3}$$

We consider the following assumption.

Hypothesis 3. The operator Λ is split into $\Lambda_0 + \Lambda_1$, with $\Lambda_1 \in \mathcal{L}(H_{00}^{\frac{1}{2}}(\Gamma_1), H_{00}^{\frac{1}{2}}(\Gamma_1)')$, and satisfies

$$\Re\langle \Lambda_1 \psi, \overline{\psi}\rangle \le 0 \qquad \forall \psi \in H_{00}^{\frac{1}{2}}(\Gamma_1), \tag{3.4}$$

and $\Lambda_2 \in \mathcal{L}(H_{00}^{\frac{1}{2}}(\Gamma_1), H_{00}^{\frac{1}{2}}(\Gamma_1))$. We assume the following property of uniqueness.

Hypothesis 4. We have

$$a(\Omega, u, v) = 0 \qquad \forall v \in \mathcal{V}(\Omega) \Rightarrow u = 0, \tag{3.5}$$

$$a(\Omega, u, v) = 0 \qquad \forall u \in \mathcal{V}(\Omega) \Rightarrow v = 0. \tag{3.6}$$

From the Lax–Milgram theorem and the fact that the imbeddings $\mathcal{V}_\Omega \to L^2(\Omega)$ and $H_{00}^{\frac{1}{2}}(\Gamma_1) \to L^2(\Gamma_1)$ are compact, and due to the Fredholm alternative, we obtain the following result (see, e.g., [10] for a detailed argument).

PROPOSITION 3.1. *If Hypotheses* 3 *and* 4 *are satisfied, we have the following:*

1. *Problem* (3.3) *has one and only one solution.*
2. *The sesquilinear form* $a(\Omega, ., .)$ *satisfies the inf-sup condition: There exists a constant* $a > 0$ *such that*

$$\inf_{u\neq 0} \sup_{v\neq 0} \frac{|a_\Omega(u, v)|}{\|u\|_{\mathcal{V}(\Omega)} \|v\|_{\mathcal{V}(\Omega)}} \ge a. \tag{3.7}$$

For a given $x_0 \in \Omega$, consider the modified open subset $\Omega_\varepsilon = \Omega \backslash \overline{\omega_\varepsilon}$, $\omega_\varepsilon = x_0 + \varepsilon\omega$, where ω is a fixed open and bounded subset of $\mathbb{R}^n$ containing the origin ($\omega_\varepsilon = \emptyset$ if $\varepsilon = 0$), whose boundary $\partial\omega$ is connected and piecewise of class $\mathcal{C}^1$. The modified solution u_{Ω_ε} satisfies

$$\left\{\begin{array}{lll} \Delta u_{\Omega_\varepsilon} + k^2 u_{\Omega_\varepsilon} &= 0 & \text{in } \Omega_\varepsilon, \\ u_{\Omega_\varepsilon} &= 0 & \text{on } \Gamma_0, \\ u_{\Omega_\varepsilon} &= 0 & \text{on } \partial\omega_\varepsilon, \\ \dfrac{\partial u_{\Omega_\varepsilon}}{\partial n} &= \Lambda u_{\Omega_\varepsilon} + \Theta & \text{on } \Gamma_1. \end{array}\right. \tag{3.8}$$

The function u_{Ω_ε} is defined on the variable open set Ω_ε, and thus belongs to a functional space which depends on ε. Hence, if we want to derive the asymptotic expansion of a function of the form

$$j(\varepsilon) = J(u_{\Omega_\varepsilon}), \tag{3.9}$$

we cannot apply directly the tools of section 2, which require a fixed functional space. For this reason, we use the domain truncation method introduced in [17] to avoid this complication.

4. The truncation method. Let $R > 0$ be such that the closed ball $\overline{B(x_0, R)}$ is included in Ω. It is supposed throughout this paper that ε remains small enough so that $\overline{\omega_\varepsilon} \subset B(x_0, R)$. The truncated open subset is defined by

$$\Omega_R = \Omega \backslash \overline{B(x_0, R)}. \tag{4.1}$$

The open subset $B(x_0, R)\backslash\overline{\omega_\varepsilon}$ is denoted by D_ε (see Figure 4.1). For $\varphi \in H^{\frac{1}{2}}(\Gamma_R)$ and $\varepsilon > 0$, let u_ε^φ be the solution to the following problem: Find u_ε^φ such that

$$\left\{ \begin{array}{lll} \Delta u_\varepsilon^\varphi + k^2 u_\varepsilon^\varphi & = 0 & \text{in } D_\varepsilon, \\ u_\varepsilon^\varphi & = 0 & \text{on } \partial\omega_\varepsilon, \\ u_\varepsilon^\varphi & = \varphi & \text{on } \Gamma_R, \end{array} \right. \tag{4.2}$$

where Γ_R is the boundary of the ball $B(x_0, R)$. For $\varepsilon = 0$, u_0^φ is the solution to

$$\left\{ \begin{array}{lll} \Delta u_0^\varphi + k^2 u_0^\varphi & = 0 & \text{in } B(x_0, R), \\ u_0^\varphi & = \varphi & \text{on } \Gamma_R. \end{array} \right. \tag{4.3}$$

Using the Poincaré inequality, it can easily be seen that for $R < \frac{1}{\sqrt{2}|k|}$, (4.2) has one and only one solution.

For $\varepsilon \geq 0$, the Dirichlet-to-Neumann operator T_ε is defined by

$$\begin{array}{rccc} T_\varepsilon : & H^{1/2}(\Gamma_R) & \longrightarrow & H^{-1/2}(\Gamma_R), \\ & \varphi & \longmapsto & T_\varepsilon\varphi = \nabla u_\varepsilon^\varphi . n_{|\Gamma_R}, \end{array}$$

where the normal $n_{|\Gamma_R}$ is chosen outward to D_ε on Γ_R and $\partial\omega_\varepsilon$.

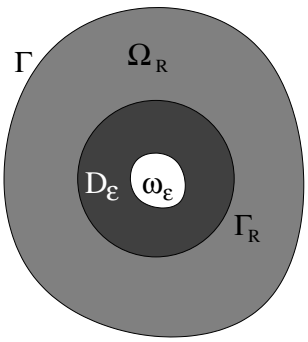


FIG. 4.1. *The truncated domain.*

Finally, we define for $\varepsilon \geq 0$ the solution u_ε to the truncated problem

$$\left\{\begin{array}{lll} \Delta u_\varepsilon + k^2 u_\varepsilon & = 0 & \text{in } \Omega_R, \\ u_\varepsilon & = 0 & \text{on } \Gamma_0, \\ \dfrac{\partial u_\varepsilon}{\partial n} & = \Lambda u_\varepsilon + \Theta & \text{on } \Gamma_1, \\ \dfrac{\partial u_\varepsilon}{\partial n} - T_\varepsilon u_{\varepsilon|\Gamma_R} & = 0 & \text{on } \Gamma_R. \end{array}\right. \tag{4.4}$$

The variational formulation associated with (4.4) is as follows: Find $u_\varepsilon \in \mathcal{V}_R$ such that

$$a_\varepsilon(u_\varepsilon, v) = \ell(v) \qquad \forall v \in \mathcal{V}_R, \tag{4.5}$$

where the functional space $\mathcal{V}_R$ and the sesquilinear form a_ε are defined by

$$\mathcal{V}_R = \{v \in H^1(\Omega_R);\ v_{|\Gamma_0} = 0\}, \tag{4.6}$$

$$a_\varepsilon(u, v) = \int_{\Omega_R} \nabla u . \nabla \overline{v}\ dx - k^2 \int_{\Omega_R} u . \overline{v}\ dx - \langle \Lambda u, \overline{v} \rangle + \int_{\Gamma_R} T_\varepsilon u_{|\Gamma_R} \overline{v}\ d\gamma(x). \tag{4.7}$$

Here, $\int_{\Gamma_R}$ denotes the duality product between $H^{1/2}(\Gamma_R)$ and $H^{-1/2}(\Gamma_R)$. The following result is standard in PDE theory.

PROPOSITION 4.1. *Problems* (3.8) *and* (4.4) *have a unique solution. Moreover, the restriction to* Ω_R *of the solution* u_{Ω_ε} *to* (3.8) *is the solution* u_ε *to* (4.4).

We now have at our disposal the fixed Hilbert space $\mathcal{V}_R$ required by section 2. We assume that the following hypothesis holds.

Hypothesis 5. The function J introduced in (3.9) is defined in a neighboring part of Γ and satisfies

$$J(u + h) = J(u) + \Re\left(L_u(h)\right) + o\left(\|h\|\right) \qquad \forall u, h \in \mathcal{V}_R,$$

where L_u is a linear and continuous form on $\mathcal{V}_R$.

Then we obtain that

$$j(\varepsilon) = J(u_{\Omega_\varepsilon}) = J(u_\varepsilon) \qquad \forall \varepsilon \geq 0. \tag{4.8}$$

Remark 1. We can also consider a more general cost function (see, e.g., [9]); the truncation method does not restrict the choice of the function. In the numerical part of this work, only measurements on the boundary of the domain are used. For this reason and to simplify the presentation, we considered the previous assumption about the cost function.

Let v_Ω be the solution to the adjoint problem

$$a(\Omega, w, v_\Omega) = -L_{u_\Omega}(w) \qquad \forall w \in \mathcal{V}(\Omega), \tag{4.9}$$

where the functional space $\mathcal{V}(\Omega)$ and the sesquilinear form $a(\Omega, ., .)$ are defined in (3.2). It has been shown in Proposition 4.1 that u_0 is the restriction to Ω_R of u_Ω. Similarly, v_0, the solution to

$$a_0(w, v_0) = -L_{u_0}(w) \qquad \forall w \in \mathcal{V}_R, \tag{4.10}$$

is the restriction to Ω_R of v_Ω.

5. The main results. This section contains the main results of this paper. All the proofs are reported in section 6. Henceforth, we have to distinguish between the cases $n = 2$ and $n = 3$. This is due to the fact that the fundamental solutions to the Laplace equation in $\mathbb{R}^2$ and $\mathbb{R}^3$ have an essentially different asymptotic expansion at infinity, and (5.1) has generally no solution if $n = 2$.

5.1. The three-dimensional case. Possibly changing the coordinate system, we can suppose for convenience that $x_0 = 0$. In order to derive the topological sensitivity of the function j, we introduce two auxiliary problems.

The first problem, called the *exterior problem*, is formulated in $\mathbb{R}^3\backslash\overline{\omega}$ and consists of finding v_ω, solution to

$$\left\{\begin{array}{lll} -\Delta v_\omega & = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega}, \\ v_\omega & = 0 & \text{at } \infty, \\ v_\omega & = u_\Omega(x_0) & \text{on } \partial\omega, \end{array}\right. \tag{5.1}$$

where u_Ω is the solution to the direct problem (3.1). Here, one can remark that just the principal part of the Helmholtz operator is used, which was described by the Laplace equation. The function v_ω can be expressed by a single layer potential on $\partial\omega$. Let

$$E(y) = \frac{1}{4\pi r} \tag{5.2}$$

with $r = ||y||$. It is a fundamental solution for the Laplace equation in $\mathbb{R}^3$. Then the function v_ω reads

$$v_\omega(y) = \int_{\partial\omega} E(y-x) p_\omega(x)\ d\gamma(x), \quad y \in \mathbb{R}^3\backslash\overline{\omega}, \tag{5.3}$$

where $p_\omega \in H^{-\frac{1}{2}}(\partial\omega)$ is the solution to boundary integral equation

$$\int_{\partial\omega} E(y-x) p_\omega(x)\ d\gamma(x) = u_\Omega(x_0) \qquad \forall y \in \partial\omega. \tag{5.4}$$

For x bounded and large $r = ||y||$, we have

$$E(y-x) = E(y) + O\left(\frac{1}{r^2}\right), \tag{5.5}$$

and the asymptotic expansion at infinity of the function v_ω is given by

$$v_\omega(y) = P_\omega(y) + W_\omega(y), \tag{5.6}$$

$$P_\omega(y) = A_\omega\left(u_\Omega(x_0)\right) E(y), \tag{5.7}$$

$$A_\omega\left(u_\Omega(x_0)\right) = \int_{\partial\omega} p_\omega(x)\ d\gamma(x), \tag{5.8}$$

$$W_\omega(y) = O\left(\frac{1}{r^2}\right). \tag{5.9}$$

Notice that $P_\omega \in L^m_{loc}$ for all $m < 3$. Clearly, the function $\alpha \longmapsto A_\omega(\alpha)$ is linear on $\mathbb{R}$, and the number $A_\omega(\alpha)$ depends on the shape of ω.

The second problem, which we call *interior problem*, is formulated in $D_0 = B(x_0, R)$ and consists to find Q^1_ω solution to

$$\left\{ \begin{array}{ll} \Delta Q^1_\omega + k^2 Q^1_\omega & = 0 \qquad \text{in } D_0, \\ Q^1_\omega & = P_{\omega|\Gamma_R} \quad \text{on } \Gamma_R. \end{array} \right. \tag{5.10}$$

Here, the idea is to consider an interior and exterior problem that gives a good "first order approximation" of $(u^\varphi_\varepsilon - u^\varphi_0)_{|D_\varepsilon}$, $\varphi = u_{\Omega|\Gamma_R}$, in the form $f(\varepsilon)(Q^1_\omega - P_\omega)$, in a way which will be stated precisely in section 6. But the given formulation (5.10) of the interior problem, which is the "natural" choice, is not sufficient to get the behavior needed by the adjoint technique described in section 2. More precisely, in this case one can construct the sesquilinear form δa but there is no positive function $f(\varepsilon)$ such that $||a_\varepsilon - a_0 - f(\varepsilon)\delta_a||_{\mathcal{L}_2(\mathcal{V}_R)} = o(f(\varepsilon))$. Indeed, one can observe through the proof of Proposition 6.7 that the behavior of $||a_\varepsilon - a_0 - f(\varepsilon)\delta_a||_{\mathcal{L}_2(\mathcal{V}_R)}$ is not of order $o(\varepsilon)$, but only of order $O(\varepsilon)$. This is due to the approximation used on the exterior problem (5.1), where just the principal part of the operator is considered. For this reason, a new term Q^2_ω is used in order to correct the error caused by this approximation. We construct Q^2_ω as the solution to

$$\left\{ \begin{array}{ll} \Delta Q^2_\omega + k^2 Q^2_\omega & = k^2 P_\omega \quad \text{in } D_0, \\ Q^2_\omega & = 0 \qquad \text{on } \Gamma_R. \end{array} \right. \tag{5.11}$$

Setting $Q_\omega = Q^1_\omega + Q^2_\omega$, then Q_ω is the solution to

$$\left\{ \begin{array}{ll} \Delta Q_\omega + k^2 Q_\omega & = k^2 P_\omega \quad \text{in } D_0, \\ Q_\omega & = P_{\omega|\Gamma_R} \quad \text{on } \Gamma_R. \end{array} \right. \tag{5.12}$$

Using the corrected interior problem (5.12), one can derive the good approximation of $(u^\varphi_\varepsilon - u^\varphi_0)_{|D_\varepsilon}$. The main result is the following, which will be proved in section 6.

THEOREM 5.1. *Let $j(\varepsilon) = J(u_{\Omega_\varepsilon})$ be a cost function satisfying Hypothesis* 5. *Then the topological asymptotic expansion is given by*

$$j(\varepsilon) - j(0) = \varepsilon \Re \left(A_\omega \left(u_\Omega(x_0) \right) \overline{v_\Omega(x_0)} \right) + o(\varepsilon), \tag{5.13}$$

where u_Ω is the direct state solution to (3.1) *and v_Ω is the adjoint state solution to* (4.9).

Then the topological gradient is given by

$$g(x) = \Re \left(A_\omega \left(u_\Omega(x) \right) \overline{v_\Omega(x)} \right) \qquad \forall x \in \Omega,$$

and only two systems must be solved in order to compute $g(x)$ for all $x \in \Omega$.

When ω is the unit ball $B(0,1)$, then $v_\omega(y)$, $P_\omega(y)$, and $W_\omega(y)$ can be computed explicitly:

$$v_\omega(y) = \frac{u_\Omega(x_0)}{r} = P_\omega(y), \quad W_\omega(y) = 0, \quad 0 \neq y \in \mathbb{R}^3. \tag{5.14}$$

Then it follows from (5.2) and (5.7) that

$$A_\omega \left(u_\Omega(x_0) \right) = 4\pi u_\Omega(x_0). \tag{5.15}$$

We have the following result.

COROLLARY 5.2. *Under the assumptions of Theorem* 5.1 *and when ω is the unit ball $B(0,1)$, the topological asymptotic expansion is given by*

$$j(\varepsilon) - j(0) = 4\pi\varepsilon \Re \left(u_\Omega(x_0) \overline{v_\Omega(x_0)} \right) + o(\varepsilon). \tag{5.16}$$

5.2. The two-dimensional case. In this section, we intend to derive the asymptotic expansion of the function j in the two-dimensional case. The technique used is similar to that of the three-dimensional case. We use the principal part of the Helmholtz operator to derive the topological sensitivity expression. Next, we briefly describe the transposition of the previous results to the two-dimensional case. As before, u_Ω and the adjoint state v_Ω are, respectively, the solutions to (3.1) and (4.9).

The exterior problem must now be defined differently than in (5.1). It consists of finding v_ω, the solution to

$$\left\{\begin{array}{lll} -\Delta v_\omega & = 0 & \text{in } \mathbb{R}^2\backslash\overline{\omega}, \\ v_\omega(y)/\log r & = u_\Omega(x_0) & \text{at } \infty, \\ v_\omega & = 0 & \text{on } \partial\omega. \end{array}\right. \tag{5.17}$$

A fundamental solution for the Laplace equation in $\mathbb{R}^2$ is given by

$$E(y) = -\frac{1}{2\pi}\log r. \tag{5.18}$$

The function v_ω has the form

$$v_\omega(y) = u_\Omega(x_0)\log\|y\| + P_\omega + W_\omega(y), \tag{5.19}$$

where P_ω is constant and $W_\omega(y) = o(1)$ at infinity [9]. In the next proposition (where ω is not supposed to be a ball), one can observe that in the two-dimensional case the topological sensitivity does not depend on the shape of the hole ω, in contrast to the three-dimensional case.

THEOREM 5.3. *The assumptions are the same as in Theorem* 5.1. *The function j has the asymptotic expansion*

$$j(\varepsilon) = j(0) - \frac{2\pi}{\log\varepsilon}\Re\left(u_\Omega(x_0)\overline{v_\Omega(x_0)}\right) + o\left(\frac{1}{\log\varepsilon}\right). \tag{5.20}$$

The proof for the two-dimensional case uses the same tools as the three-dimensional case (see section 6) and will not be repeated.

6. Proofs. This section consists of the proof of Theorem 5.1. The variation of the sesquilinear form a_ε reads

$$a_\varepsilon(u,v) - a_0(u,v) = \int_{\Gamma_R}(T_\varepsilon - T_0)u\overline{v}\; d\gamma(x). \tag{6.1}$$

Hence, the problem reduces to the analysis of $(T_\varepsilon - T_0)\varphi$ for $\varphi \in H^{\frac{1}{2}}(\Gamma_R)$. More precisely, it will be shown that there exists an operator $\delta T \in \mathcal{L}(H^{\frac{1}{2}}(\Gamma_R), H^{-\frac{1}{2}}(\Gamma_R))$ such that

$$\|T_\varepsilon - T_0 - \varepsilon\delta T\|_{\mathcal{L}(H^{\frac{1}{2}}(\Gamma_R),H^{-\frac{1}{2}}(\Gamma_R))} = O(\varepsilon^{3/2}). \tag{6.2}$$

Consequently, defining δa by

$$\delta a(u,v) = \int_{\Gamma_R}\delta T u\overline{v}\; d\gamma(x) \qquad \forall u,v \in \mathcal{V}_R \tag{6.3}$$

will yield straightforwardly

$$\|a_\varepsilon - a_0 - \varepsilon\delta a\|_{\mathcal{L}(H^{\frac{1}{2}}(\Gamma_R),H^{-\frac{1}{2}}(\Gamma_R))} = O(\varepsilon^{3/2}). \tag{6.4}$$

First we need some definitions and preliminary lemmas.

6.1. Definitions. For convenience, the following norms and seminorms are chosen for the functional spaces which will be used.

- For a bounded and open subset $\mathcal{O} \subset \mathbb{R}^3$ and $m \geq 0$, the Sobolev space $H^m(\mathcal{O})$ is equipped with the norm defined by

$$\|u\|^2_{m,\mathcal{O}} = \sum_{j=0}^{m} |u|^2_{j,\mathcal{O}},$$

where the seminorms $|u|_{j,\mathcal{O}}$ are given by

$$|u|^2_{j,\mathcal{O}} = \sum_{|\alpha|=j} \int_{\mathcal{O}} |\partial_\alpha u|^2 \, dx. \tag{6.5}$$

- For a given $\varepsilon > 0$, the space $H^{\frac{1}{2}}(\Gamma_{R/\varepsilon})$ is equipped with the norm

$$\|u\|_{\frac{1}{2},\Gamma_{R/\varepsilon}} = \inf\{\|v\|_{1,C(R/2\varepsilon,R/\varepsilon)};\ v_{|\Gamma_{R/\varepsilon}} = u\},$$

where $C(r,r') = \{x \in \mathbb{R}^3; r < ||x|| < r'\}$.

- The dual space $H^{-\frac{1}{2}}(\Gamma_{R/\varepsilon})$ is equipped with the natural norm

$$\|w\|_{-\frac{1}{2},\Gamma_{R/\varepsilon}} = \sup\{|\langle w, v\rangle_{-\frac{1}{2},\frac{1}{2}};\ v \in H^{\frac{1}{2}}(\Gamma_{R/\varepsilon}); \|v\|_{\frac{1}{2},\Gamma_{R/\varepsilon}} = 1\},$$

where $\langle , \rangle_{-\frac{1}{2},\frac{1}{2}}$ is the duality product between $H^{\frac{1}{2}}(\Gamma_{R/\varepsilon})$ and $H^{-\frac{1}{2}}(\Gamma_{R/\varepsilon})$.

6.2. Preliminary lemmas. Recall that $x_0 = 0$. We will use extensively the following change of variable: For a given function u defined on a subset $\mathcal{O}$, the function $\tilde{u}$ is defined on $\widetilde{\mathcal{O}} = \mathcal{O}/\varepsilon$ by

$$\tilde{u}(y) = u(x), \quad y = \frac{x}{\varepsilon}.$$

LEMMA 6.1. *We have that*

$$|u|_{1,\mathcal{O}} = \varepsilon^{1/2} |\tilde{u}|_{1,\widetilde{\mathcal{O}}}, \tag{6.6}$$

$$\|u\|_{0,\mathcal{O}} = \varepsilon^{3/2} \|\tilde{u}\|_{0,\widetilde{\mathcal{O}}}. \tag{6.7}$$

Proof. Due to $\nabla u(x) = \nabla \tilde{u}(y)/\varepsilon$ and to definition (6.5), we have

$$|u|^2_{1,\mathcal{O}} = \int_{\mathcal{O}} |\nabla u|^2 \, dx = \frac{1}{\varepsilon^2} \int_{\widetilde{\mathcal{O}}} |\nabla \tilde{u}|^2 \varepsilon^3 \, dy.$$

Similarly, we have

$$\|u\|_{0,\mathcal{O}} = \varepsilon^{3/2} \|\tilde{u}\|_{0,\widetilde{\mathcal{O}}}. \qquad \square$$

LEMMA 6.2 (see [9]). *For $\varphi \in H^{\frac{1}{2}}(\partial\omega)$, let v be the solution to the problem*

$$\begin{cases} -\Delta v &= 0 \quad \text{in } \mathbb{R}^3 \backslash \overline{\omega}, \\ v &= 0 \quad \text{at } \infty, \\ v &= \varphi \quad \text{on } \partial\omega. \end{cases} \tag{6.8}$$

The function v is split into

$$\begin{aligned} v(y) &= V(y) + W(y), \\ V(y) &= E(y) \int_{\partial\omega} p(x)\ d\gamma(x), \end{aligned}$$

where $E(y) = \frac{1}{4\pi\|y\|}$ *and* $p \in H^{-\frac{1}{2}}(\partial\omega)$ *is the unique solution to*

$$\int_{\partial\omega} E(y-x)p(x)\ d\gamma(x) = \varphi(y) \qquad \forall y \in \partial\omega. \tag{6.9}$$

There exists a constant $c > 0$ *(independent of* φ *and* ε*) such that*

$$\begin{aligned} \|V\|_{0,C(R/2\varepsilon,R/\varepsilon)} &\leq c\varepsilon^{-1/2}\|\varphi\|_{\frac{1}{2},\partial\omega}, \\ |V|_{1,C(R/2\varepsilon,R/\varepsilon)} &\leq c\varepsilon^{1/2}\|\varphi\|_{\frac{1}{2},\partial\omega}, \\ \|V\|_{0,D_\varepsilon/\varepsilon} &\leq c\varepsilon^{-1/2}\|\varphi\|_{\frac{1}{2},\partial\omega}, \\ |V|_{1,D_\varepsilon/\varepsilon} &\leq c\|\varphi\|_{\frac{1}{2},\partial\omega}, \\ \|W\|_{0,C(R/2\varepsilon,R/\varepsilon)} &\leq c\varepsilon^{1/2}\|\varphi\|_{\frac{1}{2},\partial\omega}, \\ |W|_{1,C(R/2\varepsilon,R/\varepsilon)} &\leq c\varepsilon^{3/2}\|\varphi\|_{\frac{1}{2},\partial\omega}, \\ \|W\|_{0,D_\varepsilon/\varepsilon} &\leq c\|\varphi\|_{\frac{1}{2},\partial\omega}. \end{aligned}$$

LEMMA 6.3. *We assume that* $R < \frac{1}{\sqrt{2}|k|}$*. For a given* $\varepsilon > 0$*,* $f_\varepsilon \in L^2(D_\varepsilon)$*, and* $\varphi \in H^{\frac{1}{2}}(\Gamma_R)$*, let* v_ε *be the solution to*

$$\begin{cases} \Delta v_\varepsilon + k^2 v_\varepsilon &= f_\varepsilon \quad \text{in } D_\varepsilon, \\ v_\varepsilon &= 0 \quad \text{on } \partial\omega_\varepsilon, \\ v_\varepsilon &= \varphi \quad \text{on } \Gamma_R. \end{cases} \tag{6.10}$$

There exists a constant $C(R,k) > 0$ *(independent of* φ *and* ε*) such that*

$$\|v_\varepsilon\|_{1,D_\varepsilon} \leq C(R,k)\left(\|\varphi\|_{\frac{1}{2},\Gamma_R} + \|f_\varepsilon\|_{0,D_\varepsilon}\right). \tag{6.11}$$

Proof. Let $\mathcal{R}\varphi$ be the lifting of φ in the space $H^1(C(R/2,R))$ such that $\mathcal{R}\varphi_{|\Gamma_{R/2}} = 0$. We extend $\mathcal{R}\varphi$ by zero to the domain D_ε. We denote this extension by $\widetilde{\mathcal{R}\varphi}$. It belongs to $H^1(D_\varepsilon)$. We introduce

$$u_\varepsilon = \widetilde{\mathcal{R}\varphi} - v_\varepsilon, \tag{6.12}$$

$$g_\varepsilon = -f_\varepsilon + \Delta\widetilde{\mathcal{R}\varphi} + k^2\widetilde{\mathcal{R}\varphi}. \tag{6.13}$$

The function g_ε belongs to the space $H^{-1}(D_\varepsilon)$ and the new unknown u_ε is the solution to

$$\begin{cases} \Delta u_\varepsilon + k^2 u_\varepsilon &= g_\varepsilon \quad \text{in } D_\varepsilon, \\ u_\varepsilon &= 0 \quad \text{on } \partial\omega_\varepsilon, \\ u_\varepsilon &= 0 \quad \text{on } \Gamma_R. \end{cases} \tag{6.14}$$

Using the Poincaré inequality and the elliptic regularity, we obtain

$$\|u_\varepsilon\|_{1,D_\varepsilon} \leq \left(\frac{1+2R^2}{1-2k^2R^2}\right)\|g_\varepsilon\|_{-1,D_\varepsilon}. \tag{6.15}$$

Finally, the result follows from (6.12), (6.13), (6.15), and the continuity of the lifting $\mathcal{R}$. ∎

Here and in what follows, we assume that $R < \frac{1}{\sqrt{2}|k|}$.

LEMMA 6.4. *For $\varepsilon > 0$ and $\psi \in H^1(D_0)$, let X_ε be the solution to the problem*

$$\left\{ \begin{array}{lll} \Delta X_\varepsilon + k^2 X_\varepsilon & = 0 & \text{in } D_\varepsilon, \\ X_\varepsilon & = \psi & \text{on } \partial\omega_\varepsilon, \\ X_\varepsilon & = 0 & \text{on } \Gamma_R. \end{array} \right. \tag{6.16}$$

There exists a constant $c > 0$ (independent of φ and ε) such that for all $\varepsilon > 0$,

$$|X_\varepsilon|_{1,C(R/2,R)} \leq c\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega}, \tag{6.17}$$

$$\|X_\varepsilon\|_{0,D_\varepsilon} \leq c\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega}, \tag{6.18}$$

$$|X_\varepsilon|_{1,D_\varepsilon} \leq c\varepsilon^{1/2} \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega}. \tag{6.19}$$

Proof. Let $\tilde{v_\varepsilon}$ be the solution to the exterior problem

$$\left\{ \begin{array}{lll} -\Delta \tilde{v_\varepsilon} & = 0 & \text{in } \mathbb{R}^3 \backslash \overline{\omega}, \\ \tilde{v_\varepsilon} & = 0 & \text{at } \infty, \\ \tilde{v_\varepsilon} & = \psi(\varepsilon y) & \text{on } \partial\omega. \end{array} \right. \tag{6.20}$$

The function X_ε can be written

$$X_\varepsilon = v_\varepsilon - w_\varepsilon,$$

where $v_\varepsilon(x) = \tilde{v_\varepsilon}\left(\frac{x}{\varepsilon}\right)$. The function w_ε itself is the solution to

$$\left\{ \begin{array}{lll} \Delta w_\varepsilon + k^2 w_\varepsilon & = k^2 v_\varepsilon & \text{in } D_\varepsilon, \\ w_\varepsilon & = 0 & \text{on } \partial\omega_\varepsilon, \\ w_\varepsilon & = v_\varepsilon & \text{on } \Gamma_R. \end{array} \right. \tag{6.21}$$

It follows from Lemma 6.3 that there exists a constant $c > 0$ such that

$$\|w_\varepsilon\|_{1,D_\varepsilon} \leq c \left(\|v_{\varepsilon|\Gamma_R}\|_{\frac{1}{2},\Gamma_R} + k^2 \|v_\varepsilon\|_{0,D_\varepsilon} \right). \tag{6.22}$$

It follows from Lemmas 6.1 and 6.2 that

$$\|v_{\varepsilon|\Gamma_R}\|_{\frac{1}{2},\Gamma_R} \leq c\|v_\varepsilon\|_{1,C(R/2,R)} \tag{6.23}$$

$$\leq c \left(\|v_\varepsilon\|_{0,C(R/2,R)} + |v_\varepsilon|_{1,C(R/2,R)} \right) \tag{6.24}$$

$$= c \left(\varepsilon^{3/2} \|\tilde{v_\varepsilon}\|_{0,C(R/2\varepsilon,R/\varepsilon)} + \varepsilon^{1/2} |\tilde{v_\varepsilon}|_{1,C(R/2\varepsilon,R/\varepsilon)} \right) \tag{6.25}$$

$$\leq c\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega}. \tag{6.26}$$

We have that

$$\|v_\varepsilon\|_{0,D_\varepsilon} = \varepsilon^{3/2} \|\tilde{v_\varepsilon}\|_{0,D_\varepsilon/\varepsilon} \tag{6.27}$$

$$\leq c\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega}. \tag{6.28}$$

From (6.22), (6.26), and (6.28), we obtain that

$$\|w_\varepsilon\|_{1,D_\varepsilon} \leq c\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega}. \tag{6.29}$$

Then we have

$$|X_\varepsilon|_{1,C(R/2,R)} = |v_\varepsilon - w_\varepsilon|_{1,C(R/2,R)} \tag{6.30}$$
$$\leq |v_\varepsilon|_{1,C(R/2,R)} + |w_\varepsilon|_{1,C(R/2,R)} \tag{6.31}$$
$$\leq c\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega} + \|w_\varepsilon\|_{1,D_\varepsilon} \tag{6.32}$$
$$\leq c\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega}, \tag{6.33}$$
$$\|X_\varepsilon\|_{0,D_\varepsilon} \leq \|v_\varepsilon\|_{0,D_\varepsilon} + \|w_\varepsilon\|_{1,D_\varepsilon} \tag{6.34}$$
$$\leq c\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega}, \tag{6.35}$$
$$|X_\varepsilon|_{1,D_\varepsilon} \leq |v_\varepsilon|_{1,D_\varepsilon} + |w_\varepsilon|_{1,D_\varepsilon} \tag{6.36}$$
$$\leq \varepsilon^{1/2} |\tilde{v_\varepsilon}|_{1,D_\varepsilon/\varepsilon} + \|w_\varepsilon\|_{1,D_\varepsilon} \tag{6.37}$$
$$\leq c\varepsilon^{1/2} \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega} + c\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega} \tag{6.38}$$
$$\leq c\varepsilon^{1/2} \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega}. \tag{6.39}$$

This completes the proof. ∎

Lemmas 6.3 and 6.4 are summarized in the following lemma.

LEMMA 6.5. *For $\varepsilon > 0$, $\varphi \in H^{\frac{1}{2}}(\Gamma_R)$, $\psi \in H^1(D_0)$, and $f_\varepsilon \in L^2(D_\varepsilon)$, let v_ε be the solution to the problem*

$$\begin{cases} \Delta v_\varepsilon + k^2 v_\varepsilon &= f_\varepsilon \quad \text{in } D_\varepsilon, \\ v_\varepsilon &= \psi \quad \text{on } \partial\omega_\varepsilon, \\ v_\varepsilon &= \varphi \quad \text{on } \Gamma_R. \end{cases} \tag{6.40}$$

There exists a constant $c > 0$ (independent of φ, ψ, f_ε, and ε) such that for all $\varepsilon > 0$,

$$|v_\varepsilon|_{1,C(R/2,R)} \leq c\left(\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega} + \|\varphi\|_{\frac{1}{2},\Gamma_R} + \|f_\varepsilon\|_{0,D_\varepsilon}\right), \tag{6.41}$$
$$\|v_\varepsilon\|_{0,D_\varepsilon} \leq c\left(\varepsilon \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega} + \|\varphi\|_{\frac{1}{2},\Gamma_R} + \|f_\varepsilon\|_{0,D_\varepsilon}\right), \tag{6.42}$$
$$|v_\varepsilon|_{1,D_\varepsilon} \leq c\left(\varepsilon^{1/2} \|\psi(\varepsilon y)\|_{\frac{1}{2},\partial\omega} + \|\varphi\|_{\frac{1}{2},\Gamma_R} + \|f_\varepsilon\|_{0,D_\varepsilon}\right). \tag{6.43}$$

LEMMA 6.6. *Let u belong to the space $H^1(C(R/2,R))$ and satisfy $\Delta u + k^2 u = 0$ in $C(R/2,R)$, $u_{|\Gamma_R} = 0$. Then there exists a constant $c > 0$ (independent of u) such that*

$$\|\nabla u.n_{|\Gamma_R}\|_{-\frac{1}{2},\Gamma_R} \leq c|u|_{1,C(R/2,R)}. \tag{6.44}$$

Proof. Let $\varphi \in H^{\frac{1}{2}}(\Gamma_R)$. We define v as the solution to the problem

$$\begin{cases} \Delta v &= 0 \quad \text{in } C(R/2,R), \\ v &= 0 \quad \text{on } \Gamma_{R/2}, \\ v &= \varphi \quad \text{on } \Gamma_R. \end{cases}$$

Using the Green formula, we obtain

$$\int_{\Gamma_R} \nabla u.n_{|\Gamma_R} \overline{\varphi}\, d\gamma(x) = \int_{C(R/2,R)} \nabla u.\nabla \overline{v}\, dx - k^2 \int_{C(R/2,R)} u\overline{v}\, dx.$$

Then we have

$$\begin{aligned} \left| \int_{\Gamma_R} \nabla u.n_{|\Gamma_R} \overline{\varphi}\, d\gamma(x) \right| &\leq |u|_{1,C(R/2,R)} \|v\|_{1,C(R/2,R)} + k^2 \|u\|_{0,C(R/2,R)} ||v||_{1,C(R/2,R)} \\ &\leq |u|_{1,C(R/2,R)} \|\varphi\|_{\frac{1}{2},\Gamma_R} + ck^2 |u|_{1,C(R/2,R)} \|\varphi\|_{\frac{1}{2},\Gamma_R} \\ &\leq c|u|_{1,C(R/2,R)} \|\varphi\|_{\frac{1}{2},\Gamma_R}. \end{aligned}$$

This completes the proof. ∎

6.3. Variation of the sesquilinear form. The variation of the sesquilinear form a_ε reads

$$a_\varepsilon(u,v) - a_0(u,v) = \int_{\Gamma_R} (T_\varepsilon - T_0) u\overline{v}\; d\gamma(x).$$

For $\varphi \in H^{\frac{1}{2}}(\Gamma_R)$, recall that u_ε^φ is the solution to (4.2), or to (4.3) if $\varepsilon = 0$. Let v_ω^φ be the solution to the problem

$$\left\{ \begin{array}{lll} \Delta v_\omega^\varphi &= 0 & \text{in } \mathbb{R}^3 \backslash \overline{\omega}, \\ v_\omega^\varphi &= 0 & \text{at } \infty, \\ v_\omega^\varphi &= u_0^\varphi(x_0) & \text{on } \partial\omega. \end{array} \right. \tag{6.45}$$

As in (5.6) and (5.7), let $P_\omega^\varphi(y) = A_\omega\left(u_0^\varphi(x_0)\right) E(y)$ be the dominant part of v_ω^φ, and let Q_ω^φ be the solution to the associated interior problem

$$\left\{ \begin{array}{lll} \Delta Q_\omega^\varphi + k^2 Q_\omega^\varphi &= k^2 P_\omega^\varphi & \text{in } D_0, \\ Q_\omega^\varphi &= P_{\omega\,|\Gamma_R}^\varphi & \text{on } \Gamma_R. \end{array} \right. \tag{6.46}$$

The linear operator δT (independent of ε) is defined as follows:

$$\begin{array}{rccc} \delta T: & H^{1/2}(\Gamma_R) & \longrightarrow & H^{-1/2}(\Gamma_R), \\ & \varphi & \longmapsto & \delta T\varphi = \nabla(Q_\omega^\varphi - P_\omega^\varphi).n_{|\Gamma_R}. \end{array} \tag{6.47}$$

PROPOSITION 6.7. *The operator T_ε admits the following asymptotic expansion:*

$$\left\| T_\varepsilon - T_0 - \varepsilon\delta T \right\|_{\mathcal{L}(H^{\frac{1}{2}}(\Gamma_R), H^{-\frac{1}{2}}(\Gamma_R))} = O(\varepsilon^{3/2}).$$

Proof. Let $\varphi \in H^{\frac{1}{2}}(\Gamma_R)$. For simplicity we drop the superscript $(.)^\varphi$. For $y = x/\varepsilon$, we have

$$v_\omega(y) = P_\omega(y) + W_\omega(y),$$

with $P_\omega(\frac{x}{\varepsilon}) = \varepsilon P_\omega(x)$ and $W_\omega(y) = O(\frac{1}{||y||^2})$. Let

$$\psi_\varepsilon(x) = \left(T_\varepsilon - T_0 - \varepsilon\delta T\right)\varphi(x).$$

We have

$$\begin{aligned} \psi_\varepsilon(x) &= \left(\nabla u_\varepsilon - \nabla u_0 - \varepsilon(\nabla Q_\omega - \nabla P_\omega)\right).n_{|\Gamma_R} \\ &= \nabla\left(w_\varepsilon(x) - W_\omega\left(\frac{x}{\varepsilon}\right)\right).n_{|\Gamma_R}, \end{aligned}$$

where w_ε is defined by

$$w_\varepsilon(x) = u_\varepsilon(x) - u_0(x) - \varepsilon Q_\omega(x) + v_\omega\left(\frac{x}{\varepsilon}\right).$$

The function w_ε is the solution to

$$\left\{ \begin{array}{lll} \Delta w_\varepsilon + k^2 w_\varepsilon &= k^2 W_\omega(x/\varepsilon) & \text{in } D_\varepsilon, \\ w_\varepsilon &= W_\omega(x/\varepsilon) & \text{on } \Gamma_R, \\ w_\varepsilon &= -u_0(x) + u_0(0) - \varepsilon Q_\omega(x) & \text{on } \partial\omega_\varepsilon. \end{array} \right. \tag{6.48}$$

In order to apply Lemma 6.5, we have to estimate the right-hand side terms, as follows.

- In D_ε, we have

$$\|W_\omega(x/\varepsilon)\|_{0,D_\varepsilon} = \varepsilon^{3/2}\|W_\omega(y)\|_{0,D_\varepsilon/\varepsilon}.$$

Using Lemma 6.2, we obtain

$$\begin{aligned}\|W_\omega(y)\|_{0,D_\varepsilon/\varepsilon} &\leq c\|u_0(x_0)\|_{\frac{1}{2},\partial\omega}\\ &\leq c|u_0(x_0)|\\ &\leq c\|\varphi\|_{\frac{1}{2},\Gamma_R}.\end{aligned}$$

Then we have

$$\|W_\omega(x/\varepsilon)\|_{0,D_\varepsilon} \leq c\varepsilon^{3/2}\|\varphi\|_{\frac{1}{2},\Gamma_R}.$$

- On Γ_R, using Lemmas 6.1 and 6.2 and the elliptic regularity, we obtain

$$\begin{aligned}\|W_\omega(x/\varepsilon)\|_{\frac{1}{2},\Gamma_R} &\leq c\|W_\omega(x/\varepsilon)\|_{1,C(R/2,R)}\\ &\leq c\left(\|W_\omega(x/\varepsilon)\|_{0,C(R/2,R)} + |W_\omega(x/\varepsilon)|_{1,C(R/2,R)}\right)\\ &= c\left(\varepsilon^{3/2}\|W_\omega(y)\|_{0,C(R/2\varepsilon,R/\varepsilon)} + \varepsilon^{1/2}|W_\omega(y)|_{1,C(R/2\varepsilon,R/\varepsilon)}\right)\\ &\leq c\varepsilon^2\|u_0(x_0)\|_{\frac{1}{2},\partial\omega}\\ &\leq c\varepsilon^2|u_0(x_0)|\\ &\leq c\varepsilon^2\|\varphi\|_{\frac{1}{2},\Gamma_R}.\end{aligned}$$

- On $\partial\omega_\varepsilon$, putting

$$\theta_\varepsilon(x) = \frac{-u_0(x) + u_0(x_0) - \varepsilon Q_\omega(x)}{\varepsilon},$$

we have for small ε

$$\begin{aligned}\|\theta_\varepsilon(\varepsilon y)\|_{\frac{1}{2},\partial\omega} &\leq c\|\theta_\varepsilon(\varepsilon y)\|_{1,\omega}\\ &= c\left\|\frac{u_0(\varepsilon y) - u_0(x_0)}{\varepsilon} + Q_\omega(\varepsilon y)\right\|_{1,\omega}\\ &\leq c\left(\|u_0\|_{\mathcal{C}^2(B(0,R/2))} + \|Q_\omega\|_{\mathcal{C}^1(B(0,R/2))}\right)\\ &\leq c\|\varphi\|_{\frac{1}{2},\Gamma_R}.\end{aligned}$$

We can now apply Lemma 6.5, which gives

$$\begin{aligned}|w_\varepsilon|_{1,C(R/2,R)} &\leq c\left(\varepsilon^{3/2}\|\varphi\|_{\frac{1}{2},\Gamma_R} + \varepsilon^2\|\varphi\|_{\frac{1}{2},\Gamma_R} + \varepsilon\|\varepsilon\theta_\varepsilon(\varepsilon y)\|_{\frac{1}{2},\partial\omega}\right)\\ &\leq c\varepsilon^{3/2}\|\varphi\|_{\frac{1}{2},\Gamma_R}.\end{aligned}$$

Finally, it follows from Lemmas 6.1 and 6.6 that

$$\begin{aligned}\|\psi\|_{-\frac{1}{2},\Gamma_R} &= \|\nabla(w_\varepsilon - W_\omega(x/\varepsilon)).n_{|\Gamma_R}\|_{-\frac{1}{2},\Gamma_R}\\ &\leq c\left(|w_\varepsilon|_{1,C(R/2,R)} + |W_\omega(x/\varepsilon)|_{1,C(R/2,R)}\right)\\ &= c\left(|w_\varepsilon|_{1,C(R/2,R)} + \varepsilon^{1/2}|W_\omega(y)|_{1,C(R/2\varepsilon,R/\varepsilon)}\right)\\ &\leq c\left(\varepsilon^{3/2}\|\varphi\|_{\frac{1}{2},\Gamma_R} + \varepsilon^2\|\varphi\|_{\frac{1}{2},\Gamma_R}\right)\\ &\leq c\varepsilon^{3/2}\|\varphi\|_{\frac{1}{2},\Gamma_R}.\end{aligned}$$

Hence,

$$\|T_\varepsilon - T_0 - \varepsilon\delta T\|_{\mathcal{L}(H^{\frac{1}{2}}(\Gamma_R),H^{-\frac{1}{2}}(\Gamma_R))} = O(\varepsilon^{3/2}). \qquad \square$$

The asymptotic expansion of the sesquilinear form a_ε follows now straightforwardly.

PROPOSITION 6.8. *Let*

$$\delta a(u,v) = \int_{\Gamma_R} \delta T u\overline{v}\; d\gamma(x), \quad u,v \in \mathcal{V}_R.$$

Then the asymptotic expansion of the sesquilinear form a_ε is given by

$$\|a_\varepsilon - a_0 - \varepsilon\delta a\|_{\mathcal{L}(H^{\frac{1}{2}}(\Gamma_R),H^{-\frac{1}{2}}(\Gamma_R))} = O(\varepsilon^{3/2}).$$

6.4. Proof of Theorem 5.1. The proof of this theorem is done in two steps. First, we prove that Hypothesis 2 is satisfied. More precisely, we prove that the sesquilinear form a_0 satisfies the inf-sup condition. Second, we apply Theorem 2.2 to compute the topological asymptotic expansion.

6.4.1. The first step: The inf-sup condition. For all $u \in \mathcal{V}_R$, we set

$$\tilde{u} = \begin{cases} u & \text{in } \Omega_R, \\ u_0^\varphi & \text{in } B(x_0,R), \end{cases}$$

where $\varphi = u_{|\Gamma_R}$ and u_0^φ is the solution to

$$\begin{cases} \Delta u_0^\varphi + k^2 u_0^\varphi = 0 & \text{in } B(x_0,R), \\ u_0^\varphi = \varphi & \text{on } \Gamma_R. \end{cases}$$

It can easily be proved that

$$a_0(u, v_{|\Omega_R}) = a(\Omega, \tilde{u}, v) \qquad \forall u \in \mathcal{V}_R \qquad \forall v \in \mathcal{V}(\Omega),$$

where the functional space $\mathcal{V}(\Omega)$ and the sesquilinear form $a(\Omega,.,.)$ are defined by (3.2). From Proposition 3.1, the sesquilinear form $a(\Omega,.,.)$ satisfies the inf-sup condition. As a consequence, there exists $v \in \mathcal{V}(\Omega), v \neq 0$, such that

$$\begin{aligned} a_0(u, v_{|\Omega_R}) = a(\Omega, \tilde{u}, v) &\geq a\|\tilde{u}\|_{\mathcal{V}(\Omega)}\|v\|_{\mathcal{V}(\Omega)} \\ &\geq a\|u\|_{\mathcal{V}_R}\|v_{|\Omega_R}\|_{\mathcal{V}_R}. \end{aligned}$$

Then a_0 satisfies the inf-sup condition and Hypothesis 2 is satisfied.

6.4.2. Applying Theorem 2.2. All the hypotheses of section 2 are satisfied and we can apply Theorem 2.2. We obtain the following asymptotic formula:

$$\begin{aligned} j(\varepsilon) - j(0) &= \varepsilon\Re(\delta a(u_\Omega, v_\Omega)) + o(\varepsilon) \\ &= \varepsilon\Re\left(\int_{\Gamma_R} \nabla(Q_\omega^\varphi - P_\omega^\varphi).n_{|\Gamma_R}\overline{v_\Omega}\; d\gamma(x)\right) + o(\varepsilon), \end{aligned}$$

where $\varphi = u_{\Omega|\Gamma_R} = u_{0|\Gamma_R}$. Thanks to Green's formula and (6.46), we obtain that

$$\begin{aligned} \int_{\Gamma_R} \nabla(Q_\omega^\varphi - P_\omega^\varphi).n_{|\Gamma_R}\overline{v_\Omega}\; d\gamma(x) = k^2 \int_{D_0} P_\omega\overline{v_\Omega}\; dx + \int_{\Gamma_R} \nabla\overline{v_\Omega}.n_{|\Gamma_R}P_\omega\; d\gamma(x) \\ - \int_{\Gamma_R} \nabla P_\omega.n_{|\Gamma_R}\overline{v_\Omega}\; d\gamma(x). \end{aligned} \tag{6.49}$$

It can be shown that

$$\begin{aligned}\int_{\Gamma_R} \nabla\overline{v_\Omega}.n_{|\Gamma_R} P_\omega \; d\gamma(x) - \int_{\Gamma_R} \nabla P_\omega.n_{|\Gamma_R}\overline{v_\Omega} \; d\gamma(x) &= A_\omega\left(u_\Omega(x_0)\right)\langle -\Delta E, \overline{v_\Omega}\psi\rangle_{\mathcal{D}'(D_0),\mathcal{D}(D_0)} \\ &\quad -k^2\int_{D_0} P_\omega\overline{v_\Omega}\; dx \\ &= A_\omega\left(u_\Omega(x_0)\right)\langle \delta, \overline{v_\Omega}\psi\rangle_{\mathcal{D}'(D_0),\mathcal{D}(D_0)} \\ &\quad -k^2\int_{D_0} P_\omega\overline{v_\Omega}\; dx \\ &= A_\omega\left(u_\Omega(x_0)\right)\overline{v_\Omega(x_0)} - k^2\int_{D_0} P_\omega\overline{v_\Omega}\; dx,\end{aligned}$$

where $\psi \in \mathcal{D}(D_0)$ satisfies $\psi(x_0) = 1$. We insert this expression into (6.49) and obtain the desired result.

7. Numerical results: Buried objects detection. We consider a simple problem of detection of metallic objects buried in soil. The aim is to find the number and the positions of metallic objects (supposedly infinite in the $\vec{e}_z$ direction) using scattered field measurements from a monostatic antenna horizontally translated above the soil. This is a rough model of the facilities described in [19]. The two-dimensional Helmholtz equation is solved with time-domain finite differences (FDTD), the frequency-domain solution obtained with a Fourier transform. The antenna is roughly approximated by a single source point, which will be translated at various locations above the soil. At each point of the mesh, the topological sensitivity will be computed.

Let $\mathcal{X} = \{x_i\}_{i=1,\dots,n_x}$ be the set of the successive locations of the source (and sensors, since the antenna is supposed to be monostatic), and let $\mathcal{F} = \{f_i\}_{i=1,\dots,n_f}$ be the set of measurement frequencies. Let ε_s be the soil permittivity. The set of metallic objects buried in the soil is denoted by Ω.

We associate with Ω a set of "measurements" $\mathcal{M}(\Omega)$. At each couple $(x_i, f_j) \in \mathcal{X} \times \mathcal{F}$, we first define the field $u^\Omega_{x_i,f_j}$, the solution of

$$\left\{\begin{array}{lll} \Delta u + k_j^2 u & = s_{x_i} & \text{in } \mathbb{R}^2 \setminus \overline{\Omega}, \\ u & = 0 & \text{on } \partial\Omega, \\ \lim\limits_{r\to\infty} \sqrt{r}(\partial_r u - iku) & = 0, & \end{array}\right. \tag{7.1}$$

where s_{x_i} represents a source point centered at x_i, and where

$$\begin{aligned} k_j^2(x) &= \varepsilon(x)\mu\omega_j^2, \\ w_j &= 2\pi f_j, \\ \varepsilon(x) &= \left\{\begin{array}{ll} \varepsilon_0 & \text{if } x \geq 0, \\ \varepsilon_s & \text{if } x < 0. \end{array}\right. \end{aligned}$$

Then the "measurements" are $\mathcal{M}(\Omega) = \{m_{x_i,f_j}(\Omega)\}$. In our numerical tests, $m_{x_i,f_j}(\Omega)$ is the value of the scattered field at point x_i.

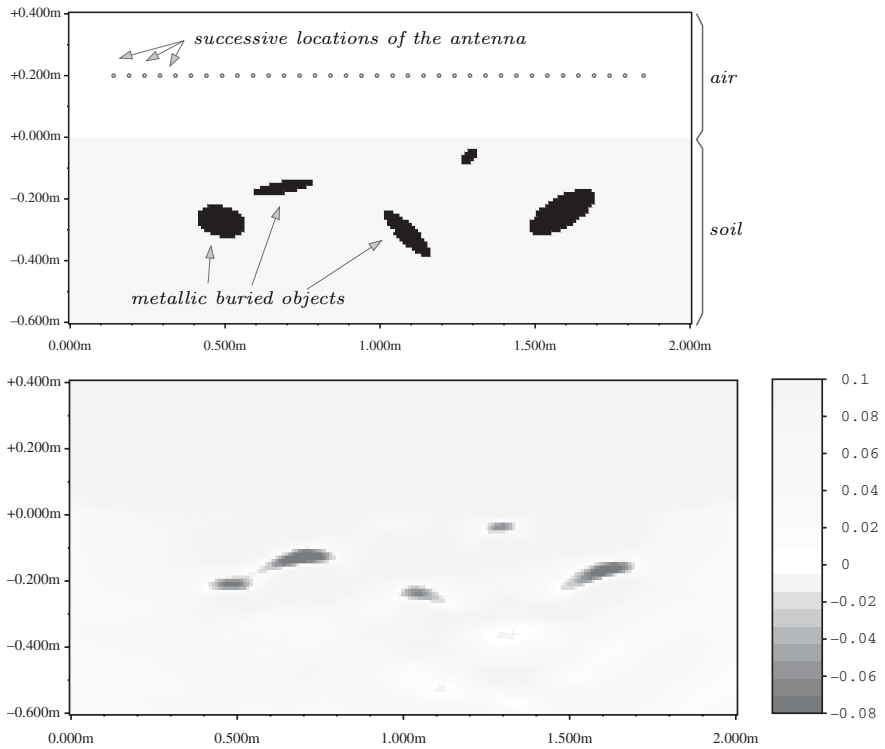


FIG. 7.1. *Repartition of metallic objects in the soil and the corresponding topological sensitivity computed on empty flat soil (dry soil, flat surface* $\varepsilon_r = 2.3$*, 20 frequencies ranging from* $400MHz$ *to* $2GHz$*).*

Reference measurements $\widetilde{\mathcal{M}} = \{\widetilde{m}_{x_i,f_j}\}$ are those values obtained from the real objects in the soil. Ideally, these would have been real measurements, but in the following numerical results, we consider only synthetical data obtained via FDTD.

The cost function, which expresses the adequacy between the measurements obtained for a distribution of metallic objects Ω and the reference data, is

$$j(\Omega) = \|\mathcal{M} - \widetilde{\mathcal{M}}\|^2 = \sum_{i,j} j_{x_i,f_j}(\Omega), \tag{7.2}$$

where

$$j_{x_i,f_j}(\Omega) = |m_{x_i,f_j}(\Omega) - \widetilde{m}_{x_i,f_j}|^2. \tag{7.3}$$

Applying the expression of the topological asymptotic (see Proposition 5.3), one has

$$j(\Omega \setminus \overline{B(x,\varepsilon)}) - j(\Omega) = \sum_{i,j} -\frac{2\pi}{\log \varepsilon} \Re\left(u^{\Omega}_{x_i,f_j}(x)\overline{v^{\Omega}_{x_i,f_j}}(x)\right) + o\left(\frac{1}{\log \varepsilon}\right), \tag{7.4}$$

where $v^{\Omega}_{x_i,f_j}$ is the adjoint state associated with the couple (x_i, f_j).

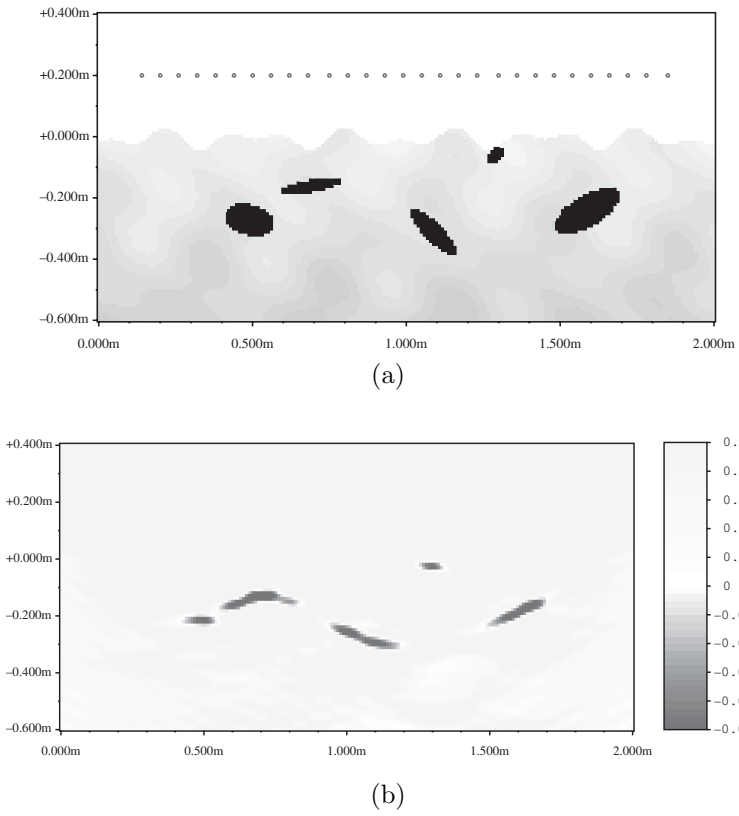
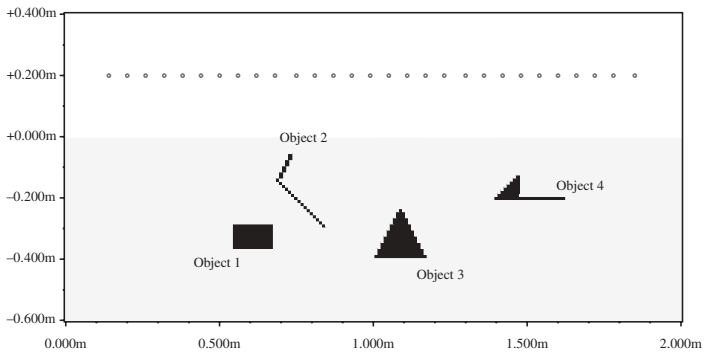


FIG. 7.2. (a) *Reference distribution of objects. The measures are computed on a dry inhomogeneous soil (ε_s ranging from* 1.6 *to* 4.15*) with a rough surface, using* 29 *measurement points and* 20 *frequencies ranging from* $260MHz$ *to* $1.86GHz$. (b) *Topological sensitivity computed on a flat empty homogeneous soil ($\varepsilon_s = 2.3$).*

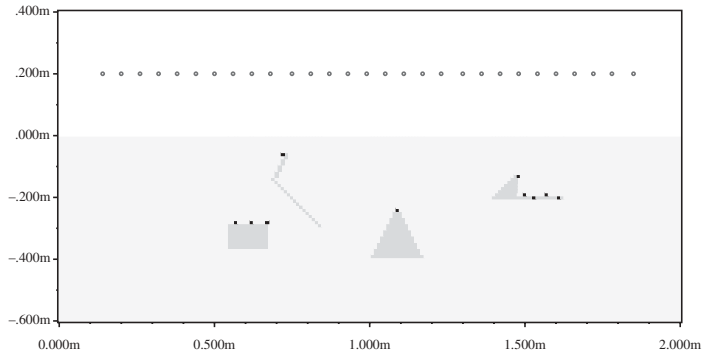
The first example (see Figure 7.1) shows the topological sensitivity computed on an "ideal" case: There is no noise on the data, and the reference soil is a flat and homogeneous dry sand soil. One can see that the top of the five objects is clearly identified by the negative values of the topological sensitivity. This topological sensitivity can be obtained very quickly since it is evaluated on an empty flat soil, which is invariant by translation: All direct states and adjoint states are just horizontal translations of a "canonical" solution. The computational cost is only 10 seconds on a 300MHz personal computer.

The second example (see Figure 7.2) is a little more realistic: The data is artificially noised since the reference data $\widetilde{\mathcal{M}}$ was obtained on a nonflat inhomogeneous soil, while the topological sensitivity was still computed on a flat homogeneous soil. One can observe that, although the objects are still located correctly, the image (see Figure 7.2(b)) is a bit distorted.

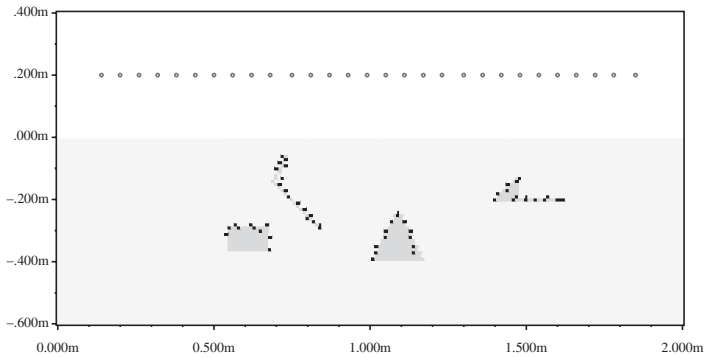
The third example shows that using an iterative process might give good results at the expense of some computational cost. In this example, the basic iterative algorithm just inserts a metal point at the point where the topological sensitivity is the most negative. Then the topological sensitivity is reevaluated, taking into account the metal points inserted at previous iterations, etc. Figure 7.3 shows the objects and the metal points that were inserted at iterations 10 and 55.



(a)



(b)



(c)

FIG. 7.3. (a) *Redistribution of objects. The measures are computed on a dry flat inhomogeneous soil* ($\varepsilon_s = 2.3$)*,* 29 *measurement points, and* 20 *frequencies ranging from* $490MHz$ *to* $3.29GHz$*.* (b) *Iteration* 10*.* (c) *Iteration* 55*.*

REFERENCES

[1] G. ALLAIRE AND R. KOHN, *Optimal design for minimum weight and compliance in plane stress using extremal microstructures*, European J. Mech. A Solids, 12 (1993), pp. 839–878.

[2] H. AMMARI AND H. KANG, *Boundary layer techniques for solving the Helmholtz equation in the presence of small inhomogeneities*, J. Math. Anal. Appl., 296 (2004), pp. 190–208.

[3] M. BENDSOE, *Optimal Topology Design of Continuum Structure: An Introduction*, Technical

report, Technical University of Denmark, Lyngby, Denmark, 1996.
[4] E. Bonnetier and C. Conca, *Approximation of Young measures by functions and application to a problem of optimal design for plates with variable thickness*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 399–422.
[5] J. Céa, *Conception optimale ou identification de forme, calcul rapide de la dérivée directionnelle de la fonction coût*, RAIRO Modél. Math. Anal. Numér., 20 (1986), pp. 371–402.
[6] J. Céa, S. Garreau, Ph. Guillaume, and M. Masmoudi, *Shape and topological optimizations connection*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 713–726.
[7] H. A. Eschenauer and N. Olhoff, *Topology optimization of continuum structures: A review*, Appl. Mech. Rev., 54 (2001), pp. 331–390.
[8] S. Garreau, Ph. Guillaume, and M. Masmoudi, *The topological asymptotic for PDE systems: The elasticity case*, SIAM J. Control Optim., 39 (2001), pp. 1756–1778.
[9] Ph. Guillaume and K. Sid Idris, *The topological asymptotic expansion for the Dirichlet problem*, SIAM J. Control Optim., 41 (2002), pp. 1042–1072.
[10] F. Ihlenburg and I. Babuska, *Finite element solution of the Helmholtz equation with high wave number* II: *The h-p version of the FEM*, SIAM J. Numer. Anal., 34 (1997), pp. 315–358.
[11] A. M. Il'in, *A boundary value problem for the elliptic equation of second order in a domain with a narrow slit.* I. *The two-dimensional case*, Math. USSR-Sb., 28 (1976), pp. 459–480 (in English).
[12] A. M. Il'in, *Study of the asymptotic behavior of the solution of an elliptic boundary value problem in a domain with a small hole*, Trudy Sem. Petrovsk., 6 (1981), pp. 57-82 (in Russian).
[13] A. M. Il'in, *Matching of Asymptotic Expansions of Solutions of Boundary Value Problems*, Transl. Math. Monogr. 102, American Mathematical Society, Providence, RI, 1992.
[14] J. Jacobsen, N. Olhoff, and E. Ronholt, *Generalized Shape Optimization of Three-Dimensional Structures Using Materials with Optimum Microstructures*, Technical report, Institute of Mechanical Engineering, Aalborg University, Aalborg, Denmark, 1996.
[15] C. Kane and M. Schoenauer, *Optimization topologique de formes par algorithmes génétiques*, Rev. Française de Mécanique, 4 (1997), pp. 237–246.
[16] R. V. Kohn and M. S. Vogelius, *Thin plates with rapidly varying thickness, and their relation to structural optimization*, in Homogenization and Effective Moduli of Materials and Media, J. L. Ericksen et al., eds., IMA Vol. Math. Appl. 1, Springer-Verlag, New York, 1986, pp. 126–149.
[17] M. Masmoudi, *The topological asymptotic expansion*, in Computational Methods for Control Applications, H. Kawarada and J. Periaux, eds., GAKUTO Internat. Ser. Math. Sci. Appl. 16, Tokyo, 2002, pp. 53–72.
[18] V. G. Maz'ya, S. A. Nazarov, and B. A. Plamenevskij, *Asymptotic expansions of the eigenvalues of boundary value problems for the Laplace operator in domains with small holes*, Math USSR-Izv., 48 (1984), pp. 347–371 (in Russian); Math. USSR-Izv., 24 (1985), pp. 321–345 (in English).
[19] P. Millot, J. C. Bureau, P. Borderies, E. Bachelier, C. Pichot, E. Lebrusq, E. Lebrusq, E. Guilianton, and J. Y. Dauvignac, *Experimental study of near surface radar imaging of buried objects with adaptive focussed synthetic aperture processing*, in Subsurface Sensing Technologies and Applications II, Proceedings of SPIE 4129, 2000, pp. 515–523.
[20] F. Murat and S. Simon, *Etudes de problèmes d'optimal design*, in Optimization Techniques: Modeling and Optimization in the Service of Man. Part 2, Lecture Notes in Comput. Sci. 41, Springer-Verlag, Berlin, 1976, pp. 54–62.
[21] S. A. Nazarov and J. Sokolowski, *Asymptotic Analysis of Shape Functionals*, Rapport de recherche de l'INRIA, RR-4633, 2002.
[22] S. A. Nazarov, *Asymptotic expansions of eigenvalues*, Leningrad University, 1987 (in Russian).
[23] S. Ozawa, *Singular Hadamard's variation of domains and eigenvalues of Laplacian*, Part 1, Proc. Japan Acad. Ser. A Math. Sci., 56 (1980), pp. 306–310.
[24] S. Ozawa, *Singular Hadamard's variation of domains and eigenvalues of Laplacian*, Part 2, Proc. Japan Acad. Ser. A Math. Sci., 57 (1981), pp. 242–246.
[25] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, New York, 1984.
[26] M. Schoenauer, L. Kallel, and F. Jouve, *Mechanics inclusions identification by evolutionary computation*, Rev. Européenne Élém. Finis, 5 (1996), pp. 619–648.
[27] A. Schumacher, *Topologieoptimierung von Bauteilstrukturen unter Verwendung von Lochpositionierungkriterien*, Doctoral Thesis, Siegen University, Siegen, Germany, 1996.
[28] J. Simon, *Differentiation with respect to the domain in boundary value problems*, Numer. Funct. Anal. Optim., 2 (1980), pp. 649–687.

[29] J. Sokolowski and A. Zochowski, *On the topological derivative in shape optimization*, SIAM J. Control Optim., 37 (1999), pp. 1251–1272.

[30] J. Sokolowski and A. Zochowski, *Topological derivatives for elliptic problems*, Inverse Problems, 15 (1999), pp. 123–134.

[31] J. Sokolowski and J. P. Zolesio, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, Springer Ser. Comput. Math. 16, Springer-Verlag, Berlin, 1992.

[32] M. S. Vogelius and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter*, Math. Model. Numer. Anal., 34 (2000), pp. 723–748.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 922–936

MEASURING DISTANCE BETWEEN SYSTEMS UNDER BOUNDED POWER EXCITATION*

P. DATE† AND G. VINNICOMBE‡

Abstract. This work suggests a way of measuring distance between two linear systems under a given bounded power excitation. The measure introduced can be used to bound from above and below the difference in closed-loop behavior of two plants with the same controller for a specified reference or disturbance spectrum. Given an unknown, single input "real" plant and its identified model, an upper bound on the distance between the plant and its model as expressed by this measure can be obtained from time domain data.

Key words. ν-gap metric, uncertainty in linear models, persistent excitation

AMS subject classifications. 93A05, 93A30

DOI. 10.1137/S0363012902410113

1. Introduction. Robust control theory is often motivated based on square summable or square integrable (bounded energy) signals. In many practical applications, the signals are better modeled as persistent disturbances. Motivating robust control design or analysis from a persistent signals viewpoint is complicated by the fact that the set of "quasi-stationary" bounded power signals which induces the infinity norm is *not* a linear vector space [5]. Nevertheless, interpreting and extending standard robust control results in a persistent signals set-up has been an active area of research for the last few years; see [13], [5] and references therein.

Given a controller which robustly stabilizes two plants (or a plant and its model), it is known that the difference in closed-loop frequency response of the two plants with the same controller can be bounded from below and above using pointwise chordal distance between the frequency responses of the two plants. A natural question to ask is whether the difference in closed-loop response of two systems for a given reference and disturbance spectrum can be bounded from above and below using an appropriate, *signal dependent* notion of distance. This work provides an affirmative answer to this question.

Specifically, the problem considered is as follows. Consider two closed-loops (P_1, C_1, C_2) and (P_2, C_1, C_2), each with the same controller $C = C_1C_2$, the same controller configuration (C_1 in the forward path, C_2 in the feedback path), and (possibly) different plants P_1, P_2. These two loops may represent the "achieved" closed-loop (i.e., with the real plant) and the "designed" closed-loop (i.e., with the model). For a given bounded power excitation (which could be a reference or a disturbance signal), the difference in the closed-loop behavior will be small if the two plants P_1 and P_2 are close in an appropriate sense. It is known that the difference in closed-loop behavior would be small (at least, for any square summable excitation) if the distance between the two systems as measured by the ν-gap metric (discussed in more details in section 3) or by the gap metric were small. However, *even if* the ν-gap is large, it is possible that the difference in the closed-loop response is small for a specific range

*Received by the editors June 24, 2002; accepted for publication (in revised form) February 10, 2004; published electronically October 8, 2004.
http://www.siam.org/journals/sicon/43-3/41011.html
†Center for Analysis of Risk and Optimisation Modelling Applications, Department of Mathematical Sciences, Brunel University, Middlesex UB8 3PH, UK (paresh.date@brunel.ac.uk).
‡Department of Engineering, University of Cambridge, CB2 1PZ, UK (gv@eng.cam.ac.uk).

of spectra of interest, provided we find a controller which stabilizes both systems with adequate stability margins. Here, a measure of distance over a subset of linear shift invariant systems is introduced which characterizes the difference in closed-loop response for a given range of signal spectra. Upper and lower bounds on this difference are established in terms of this new measure. For a plant and its candidate model, bounds on this measure are given in terms of time domain data.

The rest of the paper is organized as follows. Section 2 outlines the notation and defines the sets of signals and systems used in this paper. Section 3 introduces the ν-gap metric. A new function δ_x for measuring distance between two systems is introduced in section 4. In section 5, this function δ_x is used to bound from above and below the difference in closed-loop performance for a given range of signal spectrum. Bounds on δ_x from time domain data are introduced in section 6. Section 7 illustrates the use of this function with examples, and finally section 8 briefly summarizes the contribution of this paper.

2. Preliminaries. Let $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real and complex numbers, respectively. $\mathbb{C}^{m\times n}$ denotes the space of $m \times n$ complex matrices. $\mathbb{Z}$ denotes set of integers. $l_\infty(\mathbb{Z})$ denotes the space of bounded sequences indexed by integers. For $A \in \mathbb{C}^{m\times n}$, $\sigma_i(A)$ denotes the ith largest singular value of A. Maximum and minimum singular values of a matrix A are denoted by $\overline{\sigma}(A)$ and $\underline{\sigma}(A)$, respectively.

(i) *Signals.* Let $R_u(\tau) = \lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} u(t-\tau)u^T(t)$. Define

$$\mathcal{S}^n = \left\{ u \,|\, u \in l_\infty(\mathbb{Z}), u(t) = 0 \quad \forall\, t < 0,\ R_u(\tau) \text{ exists } \forall\, \tau, \right.$$

$$\left. \phi_u(\omega) := \sum_{\tau=-\infty}^{\infty} R_u(\tau)e^{-j\tau\omega} \text{ exists } \forall\, \omega \right\}. \tag{2.1}$$

Here, *power spectrum* $\phi_u(\omega)$ need not be bounded and may contain impulses in general. This set is called as a set of *quasi-stationary* signals in [4]. Define seminorm $\|f\|_s := \sqrt{\operatorname{trace} R_f(0)}$. For signals with continuous spectra, the equality

$$\|f\|_s = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{trace}\, \phi_f(\omega)\, d\omega}$$

also holds.

For two signals $v, w \in \mathcal{S}^n$, define

$$R_{wv}(\tau) := \lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} v(t-\tau)w^T(t)$$

$$\text{and} \quad \phi_{wv}(\omega) := \sum_{\tau=-\infty}^{\infty} R_{wv}(\tau)e^{-j\tau\omega},$$

provided the limits exist for each τ and each ω. For $v, w \in \mathcal{S}^n$, note that $v+w \in \mathcal{S}^n$ *provided* the cross-correlation function $R_{wv}(\tau)$ exists for all τ and the cross power spectrum $\phi_{wv}(\omega)$ exists for all ω. This set is obviously not a linear space and may be embedded in a linear space which includes nonstationary signals [5]. However, the use of $\mathcal{S}^n$ here to model persistent signals is motivated by two reasons. First, the correlation or spectrum based description is deemed as natural to describe persistent

disturbances, even in a nonprobabilistic setting. Second, the spectral content of a signal in $\mathcal{S}^n$ proves useful in assessing the performance of feedback systems in terms of their graph symbols restricted to the imaginary axis. This point will become apparent in sections 4 and 5.

(ii) *Systems.* Let $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$. Let $\partial\mathbb{D}$ denote the boundary of $\mathbb{D}$. $\mathcal{L}_\infty$ denotes the normed space of all functions essentially bounded on $\mathbb{D}$ and having norm $\|f\|_{\mathcal{L}_\infty} := \text{ess sup}_\omega \overline{\sigma}(f(e^{j\omega}))$, where $\overline{\sigma}(\cdot)$ represents the maximum singular value. $\mathcal{H}_\infty$ denotes the normed space of functions analytic in $\mathbb{D}$ and having norm $\|f\|_\infty := \sup_{z\in\mathbb{D}} \overline{\sigma}(f(z)) < \infty$.

Consider a linear, shift invariant discrete time system P, which can be expressed as $P = NM^{-1} = \tilde{M}^{-1}\tilde{N}$ with

1. N and M are right coprime and $G = \left[\begin{smallmatrix} N \\ M \end{smallmatrix}\right]$ inner; and
2. $\tilde{N}$ and $\tilde{M}$ are left coprime and $\tilde{G} = [-\tilde{M} \quad \tilde{N}]$ coinner.

G (resp., $\tilde{G}$) is called the normalized right (resp., normalized left) graph symbol of plant P. The set of systems of interest here are those with continuous normalized graph symbols:

$$\mathcal{P}^{m\times n} = \left\{P : G, \tilde{G} \text{ exist and are continuous on } \partial\mathbb{D}\right\}.$$

The superscript $m \times n$ is dropped when it is clear from context. The normalized graph symbols of a plant P_i will be denoted as $\tilde{G}_i$ and G_i. The set $\mathcal{P}^{m\times n}$ includes all systems whose normalized graph symbols may be uniformly approximated by real rational transfer functions; this follows from [8, Theorem 7.12]. However, it is worth mentioning that there are systems which have a continuous frequency response on the unit circle but whose normalized graph symbols are *not* continuous on the unit circle; see [9] for an example.

The set of all real rational transfer functions with n inputs and m outputs, denoted by $\mathcal{R}^{m\times n}$, is a subset of $\mathcal{P}^{m\times n}$.

The controller is denoted by C and its normalized right (left) inverse graph symbol is denoted by $K = \left[\begin{smallmatrix} V \\ U \end{smallmatrix}\right]$ ($\tilde{K} = [-\tilde{U} \quad \tilde{V}]$).

3. The ν-gap metric. The notion of measuring distance between linear systems in terms of distance between their graph spaces was introduced by Vidyasagar in [10]. In this paper, a metric called the graph metric was introduced which is characterized by the smallest distance, in a certain sense, between the coprime factors of two systems. This work was followed by a number of advances in characterization and computation of similar metrics under which feedback stability is a robust property. The pointwise gap metric [7], the gap metric [1], and the chordal metric [6] induce the same topology as the graph metric. The authors of [3] discuss the properties of gap metrics related to robustness for normalized coprime factor perturbations.

In [11], a metric called the ν-gap metric was defined. It is closely related to the gap metric but has a nicer frequency response interpretation and leads to less conservative robustness results in general. Specifically, the *ν-gap* between two plants P_1 and P_2 is defined as [11]

$$\begin{aligned} \delta_\nu(P_1, P_2) &= \inf_{Q, Q^{-1}\in\mathcal{L}_\infty} \| G_1 - G_2 Q \|_\infty \quad \text{if } I(P_1, P_2) = 0 \\ &= \quad 1 \quad \text{otherwise,} \end{aligned} \tag{3.1}$$

where $I(P_1, P_2) := \text{wno}\, \det(G_2^* G_1) = \text{wno}\, \det(\tilde{G}_1 \tilde{G}_2^*)$ and wno (g) denotes the winding number of $g(z)$ evaluated on the standard Nyquist contour indented around

any poles and zeros on $\partial\mathbb{D}$. For a real rational transfer matrix X such that $X, X^{-1} \in \mathcal{L}_\infty$, the winding number wno det $(X) = \eta(X^{-1}) - \eta(X)$, where $\eta(f)$ denotes the number of unstable poles of f. Thus, the ν-gap is seen to be the infinity norm of the smallest perturbation of the normalized coprime factorization G_1 of P_1 which yields a—not necessarily coprime—factorization G_2Q of P_2. The choice of factorization of P_2, i.e., the choice of Q, is constrained by the winding number condition. If G_2Q is constrained to be coprime instead, one gets the gap (instead of the ν-gap) between P_1 and P_2. See, e.g., section 9.3 in [12] for details.

When the winding number condition is satisfied, $\delta_\nu(P_1, P_2)$ equals the $\mathcal{L}_2$-gap,

$$\delta_{\mathcal{L}_2}(P_1, P_2) := \|\tilde{G_2}G_1\|_\infty = \sup_\omega \kappa(P_1, P_2)(e^{j\omega}). \tag{3.2}$$

$\kappa(P_1, P_2)(e^{j\omega})$ is the pointwise *chordal* distance defined by

$$\kappa(P_1, P_2)(e^{j\omega}) := \overline{\sigma}\left((I + P_2P_2^*)^{-\frac{1}{2}}(P_1 - P_2)(I + P_1^*P_1)^{-\frac{1}{2}}\right)(e^{j\omega}). \tag{3.3}$$

$\delta_\nu(P_1, P_2)$ is a measure of difference in the closed-loop performance of P_1 in feedback with a controller C and P_2 in feedback with the same controller C. Given a nominal controller C that stabilizes a (possibly frequency weighted) plant P_i, a useful closed-loop performance measure is

$$\begin{aligned} b(P_i, C) &= \|H(P_i, C)\|_\infty^{-1} = \inf_\omega \underline{\sigma}(\tilde{K}G_i)(e^{j\omega}) \\ &= \inf_\omega \underline{\sigma}(\tilde{G}_iK)(e^{j\omega}), \end{aligned} \tag{3.4}$$

where the closed-loop transfer function $H(P_i, C)$ is defined by

$$H(P_i, C) = \begin{bmatrix} P_i \\ I \end{bmatrix}(I - CP_i)^{-1}\begin{bmatrix} -C & I \end{bmatrix}.$$

It is known that any controller stabilizing a plant P_1 and achieving $b(P_1, C) > \alpha$ stabilizes the plant set $\{P_2 : \delta_\nu(P_1, P_2) \le \alpha\}$ [11]. More important, the *pointwise* difference in the closed-loop performance of nominal plant P_1 and a perturbed plant P_2 for the same controller C can be quantified in terms of $\kappa(P_1, P_2)$ as [11]:

$$\begin{aligned} \kappa(P_1, P_2)(e^{j\omega}) &\le \overline{\sigma}(H(P_1, C) - H(P_2, C))(e^{j\omega}) \\ &\le \kappa(P_1, P_2)(e^{j\omega})\,\overline{\sigma}(H(P_1, C))(e^{j\omega})\,\overline{\sigma}(H(P_2, C))(e^{j\omega}). \end{aligned} \tag{3.5}$$

The upper bound in (3.5) is useful only if C stabilizes both P_1 and P_2.

The aim here is to characterize the difference in closed-loop behavior, in a fashion similar to (3.5), for signals belonging to the set $\mathcal{S}^n$ as defined in (2.1). The next section defines a way of measuring distance between systems under a specific bounded power excitation.

4. A new measure of distance. Let Φ be a set of functions defined by

$$\begin{aligned} \Phi := \big\{ X \,|\, X : [-\pi, \pi] \to \mathbb{R},\ X(\omega) \ge 0, \\ X(\omega) \text{ is monotonic nondecreasing and bounded} \big\}. \end{aligned} \tag{4.1}$$

For $X \in \Phi$, define a seminorm

$$\|X\|_\Phi = \frac{1}{2\pi}\int_{-\pi}^{\pi} dX(\omega). \tag{4.2}$$

The definition (4.2) may be related to the set $\mathcal{S}^n$ as follows. Let $r \in \mathcal{S}^n$ be such that $\phi_r = xI^{n\times n}$, where $x(\omega) \geq 0$ is a continuous, scalar, and bounded real function over $[-\pi, \pi]$. Let

$$X(\omega) = \int_{-\pi}^{\omega} x(\tau)d\tau.$$

Then $X \in \Phi$ and $\frac{dX}{d\omega} = x$ [8, Theorem 6.20]. Also, the equality

$$\|r\|_{\mathcal{S}}^2 = n\|X\|_{\Phi} \tag{4.3}$$

follows from the definitions of seminorms $\|\cdot\|_{\mathcal{S}}$ and $\|\cdot\|_{\Phi}$. Functions belonging to the set Φ will later be used in section 5 to express bounds on the range of the spectra of interest.

Now define a function $\delta_x : \mathcal{P}^{m\times n} \times \mathcal{P}^{m\times n} \times \Phi \to \mathbb{R}_+$,

$$\delta_x(P_1, P_2, X) := \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \text{trace}\left((\tilde{G}_2G_1)^*(\tilde{G}_2G_1)(e^{j\omega})\right) dX(\omega) \right\}^{\frac{1}{2}}. \tag{4.4}$$

Thus $\delta_x^2(P_1, P_2, X)$ is seen to be a Stieltjes integral of trace $\left((\tilde{G}_2G_1)^*(\tilde{G}_2G_1)\right)$ with respect to a "weight" $X(\omega)$.[1] The choice of this weight will be determined by the shape of the spectrum of interest. For a given pair $P_1, P_2 \in \mathcal{P}^{m\times n}$, trace $\left((\tilde{G}_2G_1)^*(\tilde{G}_2G_1)\right)$ may be easily shown to be a continuous function mapping $[-\pi, \pi]$ to $\mathbb{R}$. Also, X is monotonic from the definition of Φ. From these two facts, it follows that $\delta_x(P_1, P_2, X)$ is well defined for any $P_1, P_2 \in \mathcal{P}^{m\times n}$ [8, Theorem 6.8].

The following lemma sums up properties of δ_x as a measure of distance between systems in $\mathcal{P}^{m\times n}$.

LEMMA 1.

(i) *For a given $X_0 \in \Phi$ and $P_1, P_2 \in \mathcal{P}^{m\times n}$,*

$$0 \leq \delta_x^2(P_1, P_2, X_0) \leq n\|X_0\|_{\Phi} \quad \textit{and} \tag{4.5}$$

$$\delta_x(P_1, P_2, X_0) = \delta_x(P_2, P_1, X_0). \tag{4.6}$$

(ii) *For given $X_0 \in \Phi$ and $P_1, P_2, P_3 \in \mathcal{P}^{m\times n}$,*

$$\delta_x(P_1, P_2, X_0) \leq \delta_x(P_1, P_3, X_0) + \delta_x(P_3, P_2, X_0). \tag{4.7}$$

(iii) *Suppose $X_0 \in \Phi$ is continuously differentiable, with $\frac{dX_0}{d\omega} > 0$ over an interval $[\omega_0, \omega_p] \subseteq [-\pi, \pi]$. Then*

$$\delta_x(P_1, P_2, X_0) = 0 \Leftrightarrow \kappa(P_1, P_2)(e^{j\omega}) = 0 \quad \forall\, \omega \in [\omega_0, \omega_p], \tag{4.8}$$

where $\kappa(\cdot,\cdot)$ is chordal distance as defined in (3.3).

(iv) *Suppose $X_0(\omega) = \sum_{i=1}^{\infty} a_i h(\omega - \omega_i)$, where $\{\omega_i\}$, $i = 1, 2, 3, \ldots$, is a sequence of distinct points in $[-\pi, \pi]$ and $\{a_i\}$ is such that $a_i > 0$ for all i and the sequence of partial sums $\sum a_i$ is convergent. Here $h(\cdot)$ represents the unit step function. Then*

$$\delta_x(P_1, P_2, X_0) = 0 \Leftrightarrow \kappa(P_1, P_2)(e^{j\omega_i}) = 0 \quad \forall\, \omega_i. \tag{4.9}$$

[1] Wherever square roots of positive numbers are used, it will be assumed that the positive square root is considered.

Proof. See the appendix. □

For any continuously differentiable $X \in \Phi$ such that $\frac{dX}{d\omega} > 0$ for all $\omega \in [-\pi, \pi]$, the above result shows that $\delta_x(P_1, P_2, X)$ is a metric over $\mathcal{P}^{m\times n}$. Further, for $P_1, P_2 \in \mathcal{P}^{m\times n}$ and a scalar transfer function x, $x^{-1} \in \mathcal{H}_\infty \cap \mathcal{P}^{1\times 1}$, it may be easily shown that

$$\begin{aligned} &\text{if} \quad X(\omega) = \int_{-\pi}^{\omega} |x(e^{j\tau})|^2 \, d\tau, \\ &\text{then} \quad \delta_x(P_1, P_2, X) \;=\; \|\tilde{G}_2 G_1 x\|_2. \end{aligned}$$

In general, however, X need not even be continuous for $\delta_x(P_1, P_2, X)$ to be well defined.

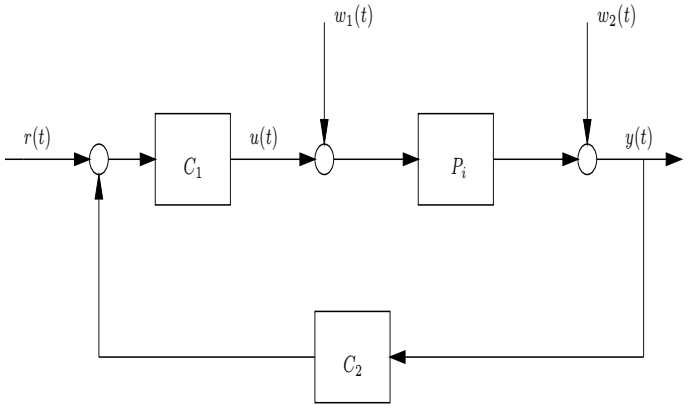


FIG. 5.1. *Closed-loop system.*

5. Closed-loop error bounds. Consider the closed-loop in Figure 5.1. Let $[-\tilde{U} \quad \tilde{V}]$ be the normalized left inverse graph symbol of controller $C = C_1 C_2$ (i.e., $C = \tilde{V}^{-1}\tilde{U}$ and $\tilde{U}$ and $\tilde{V}$ are left coprime). C_1 is a square transfer function matrix and is chosen such that $\tilde{V}C_1 \in \mathcal{H}_\infty$, $\inf_\omega \underline{\sigma}(\tilde{V}C_1)(e^{j\omega}) > 0$. The transfer function from r to $\left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right]$ in Figure 5.1 with $P = P_i$ can easily be shown to be

$$T(P_i, C_1, C_2) \;:=\; G_i(\tilde{K}G_i)^{-1}\tilde{V}C_1. \tag{5.1}$$

Define two constants dependent on controller configuration:

$$\begin{aligned} \alpha_c \;&=\; \inf_\omega \, \underline{\sigma}\,(\tilde{V}C_1)(e^{j\omega}) \\ &=\; 1 \quad \text{if } C_1 = \tilde{V}^{-1},\ C_2 = \tilde{U} \qquad (5.2) \\ &=\; \frac{1}{\sqrt{1 + \|C\|_\infty^2}} \quad \text{if } C_1 = I,\ C_2 = \tilde{V}^{-1}\tilde{U}, \qquad (5.3) \\ \text{and} \quad \beta_c \;&=\; \sup_\omega \, \overline{\sigma}\,(\tilde{V}C_1)(e^{j\omega}) \\ &=\; 1 \quad \text{if } C_1 = \tilde{V}^{-1},\ C_2 = \tilde{U} \\ &\leq\; 1 \quad \text{if } C_1 = I,\ C_2 = \tilde{V}^{-1}\tilde{U} \quad \text{or if } C_1 = \tilde{V}^{-1}\tilde{U},\ C_2 = I. \qquad (5.4) \end{aligned}$$

Note that these definitions do not exclude the "open-loop" case: $C_1 = I$, $C_2 = 0$.

Suppose that upper and lower bounds on the spectrum of disturbance or excitation r of interest are known. This information can be used to bound the difference in closed-loop response of two plants to r, as the next theorem shows.

THEOREM 1. *Suppose $T(P_i, C_1, C_2)$ as defined in (5.1) are exponentially stable, with P_1, $P_2 \in \mathcal{P}^{m\times n}$. Let $r \in \mathcal{S}^n$ be such that $\phi_r(\omega)$ is continuous, where $\mathcal{S}^n$ is the set defined in (2.1). Let x_1, x_2 be Riemann integrable functions on $[-\pi, \pi]$ such that $\exists\ \gamma > 0$ for which*

$$\gamma\, x_1 \ \le\ \sigma_i(\phi_r)(\omega)\ \le \gamma\, x_2 \quad \forall\, \omega, \quad i = 1, 2, \ldots, n. \tag{5.5}$$

Let

$$X_k(\omega)\ =\ \int_{-\pi}^{\omega} x_k(\tau) d\tau,\ k = 1, 2. \tag{5.6}$$

Then, for $b(P_i, C)$ as defined in (3.4), α_c and β_c as defined in (5.2)–(5.4), and the seminorm $\|\cdot\|_\Phi$ as defined in (4.2), the following inequalities hold:

$$\begin{aligned} \frac{\alpha_c\, \delta_x(P_1, P_2, X_1)}{\sqrt{n\, \|X_2\|_\Phi}} &\le \frac{\|(T(P_1, C_1, C_2) - T(P_2, C_1, C_2))\, r\|_{\mathcal{S}}}{\|r\|_{\mathcal{S}}} \\ &\le\ \frac{\beta_c\, \delta_x(P_1, P_2, X_2)}{\sqrt{n\, \|X_1\|_\Phi}\, b(P_1, C)\, b(P_2, C)}. \end{aligned} \tag{5.7}$$

Further, if $C_1 = \tilde{V}^{-1}(\tilde{K} G_1)$ and $C_2 = C_1^{-1} C$, then

$$\begin{aligned} \frac{\delta_x(P_1, P_2, X_1)}{\sqrt{n\, \|X_2\|_\Phi}} &\le \frac{\|(T(P_1, C_1, C_2) - T(P_2, C_1, C_2))\, r\|_{\mathcal{S}}}{\|r\|_{\mathcal{S}}} \\ &\le\ \frac{\delta_x(P_1, P_2, X_2)}{\sqrt{n\, \|X_1\|_\Phi}\, b(P_2, C)}. \end{aligned} \tag{5.8}$$

Proof. See the appendix. □

Several remarks on this result are in order.

(i) To guarantee that the filtered signal $T(P_i, C_1, C_2)\, r$ to be in $\mathcal{S}^n$, the impulse response of $T(P_i, C_1, C_2)$ should be in l_1 (i.e., should be absolutely summable) [5]. One simple way to ensure this is to impose the exponential stability condition. From a practical point of view, of course, this is a perfectly sensible requirement.

(ii) The equality (4.3) explains the presence of $\sqrt{n\|X_k\|_\Phi}$ in (5.7). To explore this further, let $\hat{r} \in \mathcal{S}^1$ be such that $\phi_{\hat{r}}$ is continuous. Let X_1, $X_2 \in \Phi$ be such that $\frac{dX_1}{d\omega} = \phi_{\hat{r}} = \frac{dX_2}{d\omega}$. Then $\|\hat{r}\|_{\mathcal{S}}^2 = \|X_1\|_\Phi = \|X_2\|_\Phi$. Define *spectral distribution function* [5]

$$F_{\hat{r}}(\omega) = \int_{-\pi}^{\omega} \phi_{\hat{r}}(\tau) d\tau.$$

Then $F_{\hat{r}}$ is continuous over $[-\pi, \pi]$ and $\frac{dF_{\hat{r}}}{d\omega} = \phi_{\hat{r}}$. Using this definition in (5.7) yields a simpler expression,

$$\alpha_c\, \delta_x(P_1, P_2, F_{\hat{r}}) \le \|(T(P_1, C_1, C_2) - T(P_2, C_1, C_2))\, \hat{r}\|_{\mathcal{S}}\ \le\ \frac{\beta_c\, \delta_x(P_1, P_2, F_{\hat{r}})}{b(P_1, C)\, b(P_2, C)}.$$

To reemphasize the main motivation of this work, note that $\delta_x(P_1, P_2, F_{\hat{r}})$ may be small for $F_{\hat{r}}$ of interest even if $\delta_\nu(P_1, P_2)$ is large.

(iii) In proving Theorem 1, we will use the equality

$$\operatorname{trace} R_r(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{trace}\, \phi_r(\omega)\, d\omega.$$

This is certainly true when ϕ_r is continuous but may not be true in general. For single input systems, ϕ_r may be allowed to have the form

$$\phi_r(\omega) = \phi_{\hat{r}}(\omega) + \sum_{i=1}^{l} \delta(\omega - \omega_i),$$

where $\phi_{\hat{r}}(\omega)$ is continuous, $\omega_i \in [-\pi, \pi]$, and $\delta(\cdot)$ is a Dirac delta function. Impulses in the spectral density of r represent periodic excitation. It is reasonable to expect that the frequencies of periodic reference (or disturbance) are known. The bounds in (5.7)–(5.8) will still make sense if we allow jump discontinuities in X_1, X_2 at the frequencies of periodic excitation. For multi-input systems, it is more difficult to account for delta functions due to the necessity of bounding singular values of the spectrum.

(iv) Note that the parameter γ in (5.5) doesn't appear in (5.7) or (5.8). This is important, since it implies that the bounds in (5.7)–(5.8) are *scale invariant* in the sense that one needs only to know the bounds on the *shape* of singular values of spectral density ϕ_r. The actual magnitude of spectral density is irrelevant.

(v) For a multi-input system, these bounds make sense for comparing performance under simultaneous excitation of all inputs. To compare behavior when $\sigma_n(\phi_r)(\omega) = 0$ for all ω (i.e., to compare the response when some but not all inputs are excited), x_1 should be zero; but that makes the lower bound in (5.7) zero, and the upper bound becomes unbounded. Even in that case, the following bound still holds:

$$\|(T(P_1, C_1, C_2) - T(P_2, C_1, C_2))\, r\|_{\mathcal{S}} \;\le\; \frac{\beta_c\, \gamma\, \delta_x(P_1, P_2, X_2)}{b(P_1, C)\, b(P_2, C)}.$$

This may be easily shown following the steps of the proof of Theorem 1.

Suppose P_2 is a model for a "true" plant P_1. For a "reasonable" controller C (i.e., which stabilizes both the true plant and the model with adequate stability margins), Theorem 1 shows that the difference in the designed and the achieved closed-loop response to a persistent excitation $r \in \mathcal{S}^n$ is small if $\delta_x(P_1, P_2, \phi_r)$ is small. The results are relevant in comparing the designed and the achieved tracking performance (when r is a reference) and in comparing noise rejection of two closed-loops (when r is a disturbance). These error bounds may also prove useful in assessing the suitability of a reduced order model to design a controller for a high order plant.

If $r_0 \in \mathcal{S}^n$ is such that the $\phi_{r_0}(\omega) = I^{n\times n}$, then

$$\|T(P_1, C_1, C_2) - T(P_2, C_1, C_2) r_0\|_{\mathcal{S}} \;=\; \|T(P_1, C_1, C_2) - T(P_2, C_1, C_2)\|_2.$$

This observation leads to the following 2-norm inequality.

COROLLARY 1. *Suppose P_1, $P_2 \in \mathcal{P}^{m\times n}$ and a controller C stabilizes both P_1 and P_2. Let $C = \tilde{V}^{-1}\tilde{U} = C_1C_2$, where $[-\tilde{U} \quad \tilde{V}]$ is a normalized left graph symbol of C and C_1 is a square matrix function such that $\tilde{V}C_1 \in \mathcal{H}_\infty$, $\inf_\omega \underline{\sigma}(\tilde{V}C_1) > 0$. Further, suppose $T(P_1, C_1, C_2)$ as defined in* (5.1) *are exponentially stable. Then*

$$\alpha_c \,\|\, \tilde{G}_2 G_1 \,\|_2 \le \|T(P_1, C_1, C_2) - T(P_2, C_1, C_2)\|_2 \le \frac{\beta_c \,\|\, \tilde{G}_2 G_1 \,\|_2}{b(P_1, C)\, b(P_2, C)}, \tag{5.9}$$

where α_c, β_c are as defined in (5.2)–(5.4).

Proof. The proof follows from (5.7), with $\phi_r = I^{n\times n}$, $\frac{dX_1}{d\omega} = \frac{dX_2}{d\omega} = 1$. ∎

Given an unknown true plant P_0, a model P_θ, and a spectral distribution F_r, it is not possible to measure $\delta_x(P_0, P_\theta, F_r)$ directly. The next section introduces bounds on $\delta_x(P_0, P_\theta, F_r)$ for a given single input "true" plant P_0 and a model P_θ in terms of data from a time domain identification experiment.

6. Bounds on $\delta_x(P_0, P_\theta, X)$. The main result in this section is restricted to single input systems. A partial generalization to the multiple input case is possible and is discussed at the end of the section.

Consider a plant $P_0 \in \mathcal{P}^{m\times 1}$. For a closed-loop system as shown in Figure 5.1 (with $P_i = P_0$), the a posteriori data is given by

$$z := \begin{bmatrix} y \\ u \end{bmatrix} = G_0(\tilde{K}G_0)^{-1}\,\tilde{V}C_1\,r + H(C, P_0)w, \tag{6.1}$$

where $H(C, P_0)$ is a closed-loop transfer function as defined in section 3, $H(C, P_0) = K(\tilde{G}_0 K)^{-1}\tilde{G}_0$. y, u (and, possibly, r) are measured signals, and $w = \left[\begin{smallmatrix} -w_1 \\ w_2 \end{smallmatrix}\right]$ is unmeasured noise or disturbance. Suppose $r, w_1, \in \mathcal{S}^1$, $w_2 \in \mathcal{S}^m$ are such that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} (r(t)w_i^T(t-\tau)) = 0 \quad \forall\, \tau, \quad i = 1, 2. \tag{6.2}$$

This deterministic assumption corresponds to the idea that the reference r and the noise signals w_i are uncorrelated. Also, suppose that the transfer function from

$$\begin{bmatrix} r & w_1 & w_2 \end{bmatrix}^T$$

to $\left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right]$ is exponentially stable. Assume that ϕ_r is continuous and define the spectral distribution function of r as before,

$$F_r(\omega) = \int_{-\pi}^{\omega} \phi_r(\tau) d\tau.$$

The data z from (6.1) can be used for identifying a model for P_0. One common method for parametric time domain identification is the prediction error method [4]. This method solves

$$\min_\theta \|\tilde{X}_\theta z\|_{\mathcal{S}} \quad \text{subject to} \quad \tilde{X}_\theta(\infty) = \begin{bmatrix} I & 0 \end{bmatrix}. \tag{6.3}$$

Here, $\tilde{X}_\theta$ is a left graph symbol of model parameterized by a (real) parameter vector θ, and z is output-input data as in (6.1). This is sometimes referred to as the "direct" prediction error in the literature [2], as opposed to methods that use a measurable reference.

The prediction error identification approach has elegant statistical properties and is widely studied in the literature; [4] offers a very comprehensive treatment. For a model P_θ obtained by the prediction error method, it is desired to obtain an estimate of $\delta_x(P_0, P_\theta, F_r)$. Let $\tilde{G}_\theta$ be the normalized left graph symbol of P_θ. The following result gives bounds on $\delta_x(P_0, P_\theta, F_r)$ in terms of time domain data.

THEOREM 2. *Let P_θ be a candidate model for P_0 with both $P_0, P_\theta \in \mathcal{P}^{m\times 1}$. Suppose $r, w_1 \in \mathcal{S}^1$, $w_2 \in \mathcal{S}^m$ are such that* (6.2) *holds. Let $C = C_1C_2$ be such that the transfer function from*

$$\begin{bmatrix} r & w_1 & w_2 \end{bmatrix}^T$$

to $\left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right]$ *in Figure* 5.1 *is exponentially stable. Let* z *be as defined in* (6.1)*, and let* α_c *and* β_c *be as defined in* (5.2)–(5.4)*. Then*

(6.4)
$$\frac{b^2(P_0,C)\,\|\tilde{G}_\theta\, z\|_{\mathcal{S}}^2 - \|\tilde{G}_0\, w\|_{\mathcal{S}}^2}{\beta_c^2} \;\le\; \delta_x^2(P_0,P_\theta,F_r) \;\le\; \frac{\|\tilde{G}_\theta\, z\|_{\mathcal{S}}^2 - b^2(P_\theta,C)\,\|\tilde{G}_0\, w\|_{\mathcal{S}}^2}{\alpha_c^2}.$$

Proof. See the appendix. ∎

From (6.4), it follows that

$$\delta_x(P_0,P_\theta,F_r) \le \frac{\|\tilde{G}_\theta\, z\|_{\mathcal{S}}}{\alpha_c}. \tag{6.5}$$

For a given model P_θ, measured data z, and a given α_c the right-hand side can be explicitly evaluated. Combining (5.7) and (6.5) gives

$$\frac{\|(T(P_0,C_1,C_2) - T(P_\theta,C_1,C_2))\, r\|_{\mathcal{S}}}{\|r\|_{\mathcal{S}}} \;\le \frac{\|\tilde{G}_\theta\, z\|_{\mathcal{S}}}{\|r\|_{\mathcal{S}}}\, \frac{\beta_c}{\alpha_c\, b(P_0,C)\, b(P_\theta,C)}.$$

If $\frac{\|\tilde{G}_\theta\, z\|_{\mathcal{S}}}{\|r\|_{\mathcal{S}}}$ is small, any "good" controller C (i.e., with sufficiently large robust stability margins $b(P_0,C)$ and $b(P_\theta,C)$) should yield a small difference in closed-loop behavior for excitation r. However, the upper bound above cannot be evaluated from data due to the presence of an unknown term $b(P_0,C)$.

If $\|\tilde{G}_0 w\|_{\mathcal{S}} \approx 0$ and if $C_1 = \tilde{V}^{-1}$, $C_2 = \tilde{U}$, then (6.4) yields an aesthetically pleasing expression,

$$b(P_0,C)\,\|\tilde{G}_\theta\, z\|_{\mathcal{S}} \;\le\; \delta_x(P_0,P_\theta,F_r) \;\le\; \|\tilde{G}_\theta\, z\|_{\mathcal{S}}.$$

Note that the model P_θ need not be obtained from the same experimental data z used in the bounds; the results hold for *any* candidate model P_θ so far as the data z is generated according to (6.1) and (6.2) holds. In particular, note that (6.5) does not use any a priori assumptions about the order or the relative stability of the plant P_0. The exponential stability of the *closed-loop* is the only assumption used here.

The result above is valid for single input systems. A generalization to multivariable plants is possible in the case when the spectrum of r is a scalar function times identity.

COROLLARY 2. *Let* P_θ *be a candidate model for* P_0 *with both* $P_0, P_\theta \in \mathcal{P}^{m\times n}$*. Suppose* $r, w_1 \in \mathcal{S}^n$, $w_2 \in \mathcal{S}^m$ *are such that* (6.2) *holds. Further, suppose* $r \in \mathcal{S}^n$ *is such that* $\phi_r = x\, I^{n\times n}$*, where* x *is a scalar, continuous spectrum. Let* $X(\omega) = \int_{-\pi}^{\omega} x(\tau)d\tau$*. Let* $C = C_1C_2$ *be such that the transfer function from*

$$\begin{bmatrix} r & w_1 & w_2 \end{bmatrix}^T$$

to $\left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right]$ *in Figure* 5.1 *is exponentially stable. Let* z *be as defined in* (6.1)*, and let* α_c *and* β_c *be as defined in* (5.2)–(5.4)*. Then*

$$\begin{aligned} \frac{b^2(P_0,C)\,\|\tilde{G}_\theta\, z\|_{\mathcal{S}}^2 - \|\tilde{G}_0\, w\|_{\mathcal{S}}^2}{\beta_c^2} &\le \delta_x^2(P_0,P_\theta,X) \\ &\le \frac{\|\tilde{G}_\theta\, z\|_{\mathcal{S}}^2 - b^2(P_\theta,C)\,\|\tilde{G}_0\, w\|_{\mathcal{S}}^2}{\alpha_c^2}. \end{aligned} \tag{6.6}$$

Proof. This may be shown following the steps of the proof of Theorem 2. Details are omitted. □

Remark 1. Under the assumption (6.2), an upper bound similar to (6.5) also holds in the case of prediction error cost in (6.3). If $\theta_\star$ is an argument which minimizes the cost in (6.3), it can be shown that

$$\|\tilde{X}_{\theta_\star}\, X_0\, r\|_S \;\le\; \|\tilde{X}_{\theta_\star}\, z\|_S,$$

where $X_0 = G_0(\tilde{K}G_0)^{-1}\,\tilde{V}C_1$. This may be easily proved from the proof of Theorem 2. The quantity $\|\tilde{X}_{\theta_\star}\, X_0\, r\|_S$, unlike $\delta_x(P_0, P_{\theta_\star}, F_r)$, depends on the choice and the configuration of controller C.

Remark 2. The bounds presented above are "distribution-free" and use only a nonprobabilistic assumption (6.2) about noise. Developing results equivalent to Theorem 2 in a probabilistic setting is an interesting and challenging area of future research.

7. Examples. Here we consider some examples to see how this new measure can be useful in comparing the closed-loop response of systems to persistent excitation. Consider a pair of plants

$$P_1(z) = \frac{2(z+1)}{z-0.6}, \qquad P_2(z) = \frac{(z+1)^2}{(z^2-0.6z+1.2)}.$$

The ν-gap error between P_1 and P_2 is significantly large ($= 0.64$). Suppose the spectrum of interest is a low frequency spectrum which satisfies

$$\gamma\, x_1(\omega) \le\; \phi_{r_0} \;\le \gamma\, x_2(\omega)$$

$$\text{for some } \gamma > 0, \text{ where} \quad x_1(\omega) = f_1(e^{j\omega})f_1(e^{-j\omega}), \quad f_1(z) = \frac{0.01z}{z-0.99},$$

$$\text{and} \quad x_2(\omega) = f_2(e^{j\omega})f_2(e^{-j\omega}), \quad f_2(z) = \frac{0.02z}{z-0.98}.$$

Note that the exact value of γ (and hence the power in the signal) is immaterial. Let $F_i(\omega) = \int_{-\pi}^{\omega} x_i(\tau)d\tau$, $i = 1, 2$. Then

$$\frac{\delta_x(P_1, P_2, F_1)}{\|F_2\|_\Phi} = 0.0294 \qquad \text{and} \qquad \frac{\delta_x(P_1, P_2, F_2)}{\|F_1\|_\Phi} = 0.0724.$$

(These may be computed as 2-norm of $\tilde{G}_2 G_1 f_i$, $i = 1, 2$.) Thus, given a controller C which stabilizes both P_1 and P_2 with "good" stability margin, the difference in closed-loop gains $T(P_1, C_1, C_2) - T(P_2, C_1, C_2)$ over this spectrum is guaranteed to be small. In this particular case, a simple integral controller $C_1 = \frac{0.025(z+1)}{(z-1)}$, $C_2 = -1$ yields stability margins $b(P_1, C) = 0.408$, $b(P_2, C) = 0.401$.

For the same plants P_1, P_2 as above, if we consider $f_1(z) = f_2(z) = 1 - 0.9z^{-1}$, then

$$\frac{\delta_x(P_1, P_2, F_1)}{\|F_2\|_\Phi} = \frac{\delta_x(P_1, P_2, F_2)}{\|F_1\|_\Phi} = 0.4469$$

so that no controller can make the difference in closed-loop gains $T(P_1, C_1, C_2) - T(P_2, C_1, C_2)$ smaller than $0.4469\,\alpha_c$ over this spectrum, with α_c as defined in (5.2).

Next, consider another pair of plants

$$P_3(z) = \frac{z-1}{z-0.99}, \qquad P_4(z) = \frac{z-1}{z-1.01}.$$

The ν-gap between P_3 and P_4 is 1. Suppose P_3 and P_4 are to be compared from the perspective of white noise rejection. Then $\|\tilde{G}_4 G_3\|_2 = 0.055$, which means any "reasonable" controller will yield a similar closed-loop white noise rejection for both P_3 and P_4 (provided such a controller exists). The controller $C_1 = 1$, $C_2 = -1$ in this case yields $b(P_3, C) = b(P_4, C) = 0.7071$.

8. Conclusion. A new measure $\delta_x(\cdot,\cdot,X)$ is introduced for measuring distance between linear shift invariant systems. It is shown that this measure can be used to characterize the difference in closed-loop behavior of two feedback systems under a given persistent excitation. For a plant P_0 and a model P_θ, bounds on this measure with respect to a given reference spectrum are obtained in terms of data from a time domain identification experiment.

Appendix. First, some technical results necessary for the proofs of Lemma 1 and Theorems 1 and 2 are given.

FACT 1. *For any complex matrix $Q \in \mathbb{C}^{m\times n}$, let the Frobenius norm be defined as usual, $\|Q\|_F^2 = \text{trace}\,(Q^*Q) = \text{trace}\,(QQ^*) = \sum_{i=1}^{m} \sigma_i^2(Q)$, where $m = \min(m,n)$. For a pair of complex matrices Q, R of compatible dimensions, with R square and invertible, the following inequalities hold:*

$$\underline{\sigma}(R)\|Q\|_F \le \|QR\|_F \le \overline{\sigma}(R)\,\|Q\|_F. \tag{A.1}$$

Proof. The upper bound is well known [13]. The lower bound follows from

$$\|Q\|_F = \|QRR^{-1}\|_F \le \overline{\sigma}(R^{-1})\|QR\|_F. \qquad \square \tag{A.2}$$

The next result proves the formula for $\inf_\omega \underline{\sigma}(\tilde{V}C_1)(e^{j\omega})$ stated in (5.3).

LEMMA 2. *Let $[-\tilde{U} \quad \tilde{V}]$ be the normalized left inverse graph symbol for $C = \tilde{V}^{-1}\tilde{U}$. Then*

$$\inf_\omega \underline{\sigma}(\tilde{V})(e^{j\omega}) = \frac{1}{\sqrt{1+\|C\|_\infty^2}}. \tag{A.3}$$

Proof. From [12, section 2.3],

$$\overline{\sigma}^2(C) = \frac{\overline{\sigma}^2(\tilde{U})}{1-\overline{\sigma}^2(\tilde{U})} \quad \text{and} \quad \underline{\sigma}^2(\tilde{V}) = 1 - \overline{\sigma}^2(\tilde{U}). \tag{A.4}$$

A rearrangement of (A.4) and taking the infimum of both sides leads to (A.3). $\square$

The proofs that follow also use some identities from [12, section 3.2] for manipulation of normalized graph symbols:

$$(\tilde{G}_1G_2)^*(\tilde{G}_1G_2) + (G_2^*G_1)(G_2^*G_1)^* = I, \tag{A.5}$$

$$(\tilde{G}_2G_1)^*(\tilde{G}_2G_1) + (G_2^*G_1)^*(G_2^*G_1) = I, \tag{A.6}$$

$$\tilde{G}_2^*\tilde{G}_2 + G_2G_2^* = I, \tag{A.7}$$

$$K(\tilde{G}_2K)^{-1}\tilde{G}_2 + G_2(\tilde{K}G_2)^{-1}\tilde{K} = I. \tag{A.8}$$

In the rest of the proofs, the argument $e^{j\omega}$ will be omitted for brevity wherever it is obvious from the context.

Proof of Lemma 1. Property (4.5) is obvious. To prove (4.6), note that

$$\begin{aligned}\sigma_i^2(\tilde{G}_2G_1)(e^{j\omega}) &= 1-\sigma_{n-i+1}^2(G_2^*G_1)(e^{j\omega}) && \text{(from (A.5))}\\ &= \sigma_i^2(\tilde{G}_1G_2)(e^{j\omega}) && \text{(from (A.6)).}\end{aligned}$$

Next, using (A.7),

$$\begin{aligned}\delta_x^2(P_1,P_2,x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{n}\sigma_i^2\left(\tilde{G}_2G_1\right)dX_0(\omega)\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{n}\sigma_i^2\left(\tilde{G}_2\left(G_3G_3^*+\tilde{G}_3^*\tilde{G}_3\right)G_1\right)dX_0(\omega)\\ &\le \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{n}\sigma_i^2\left(\tilde{G}_2\,G_3\right)dX_0(\omega)+\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{n}\sigma_i^2\left(\tilde{G}_3\,G_1\right)dX_0(\omega) \qquad \text{(A.9)}\\ &\le \left(\delta_x^2(P_2,P_3,X_0)+\delta_x^2(P_3,P_1,X_0)\right) \qquad \text{(A.10)}\end{aligned}$$

from which (4.7) follows.

To prove (4.8), let $x(\omega)=\frac{dX_0}{d\omega}$ and note that

$$\int_{-\pi}^{\pi}\left(\sum_{i=1}^{n}\sigma_i^2(\tilde{G}_2G_1)\right)dX(\omega)\ge\min_{\omega\in[\omega_0,\omega_p]}x(\omega)\int_{\omega_0}^{\omega_p}\text{trace}\,\left((\tilde{G}_2G_1)^*(\tilde{G}_2G_1)\right)d\omega.$$

On the other hand, if $X_0(\omega)=\sum_{i=1}^{\infty}a_ih(\omega-\omega_i)$, then from [8, Theorem 6.16],

$$\delta_x^2(P_1,P_2,X_0)=\frac{1}{2\pi}\sum_{i=1}^{\infty}a_i\,\text{trace}\,\left((\tilde{G}_2G_1)^*(\tilde{G}_2G_1)\right)(e^{j\omega_i})$$

from which (4.9) follows. □

Proof of Theorem 1. Note that

$$\delta_x^2(P_1,P_2,X_i)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\text{trace}\,\left((\tilde{G}_2G_1)^*(\tilde{G}_2G_1)\right)x_i\,d\omega,\quad i=1,2.$$

This follows from the definition (5.6) of $X_i(\omega)$ and from [8, Theorem 6.20].

Upper bound. We have

$$\begin{aligned}\|(T(P_1,C_1,C_2)-T(P_2,C_1,C_2))\,r\|_{\mathcal{S}}^2 &= \left\|\left(G_1(\tilde{K}G_1)^{-1}-G_2(\tilde{K}G_2)^{-1}\right)(\tilde{V}C_1)\,r\right\|_{\mathcal{S}}^2\\ &= \left\|\left(K(\tilde{G}_2K)^{-1}\tilde{G}_2G_1(\tilde{K}G_1)^{-1}\right)(\tilde{V}C_1)\,r\right\|_{\mathcal{S}}^2 \qquad \text{(A.11)}\end{aligned}$$

(premultiplying by the left-hand side of the equality in (A.7) and using the fact $\tilde{G}_2G_2=0$)

$$\le\gamma\left(\sup_{\omega}\overline{\sigma}^2\,(\tilde{G}_2K)^{-1}\right)\left(\sup_{\omega}\overline{\sigma}^2\,(\tilde{K}G_1)^{-1}\right)\left(\sup_{\omega}\overline{\sigma}^2\,(\tilde{V}C_1)\right)\delta_x^2(P_1,P_2,X_2), \qquad \text{(A.12)}$$

where the last step uses

$$\overline{\sigma}\,(\phi_r)\,\sum_{i=1}^{n}\sigma_i^2\,(\tilde{G}_2G_1)\;\leq\;\gamma\,x_2\,\sum_{i=1}^{n}\sigma_i^2\,(\tilde{G}_2G_1).$$

Also, note that

$$\frac{1}{n\,\gamma\,\|X_2\|_\Phi}\;\leq\;\frac{1}{\|r\|_{\mathcal{S}}^2}\;\leq\;\frac{1}{n\,\gamma\,\|X_1\|_\Phi}. \tag{A.13}$$

The result then follows using (A.11)–(A.13) and using the definitions of β_c and of $b(P_i,C)$ in (5.4) and (3.4), respectively.

Lower bound. Put $L=(\tilde{K}G_1)^{-1}$, $Q=(\tilde{K}G_2)^{-1}(\tilde{K}G_1)$. Then

$$\begin{aligned}
&\left\|\left(G_1(\tilde{K}G_1)^{-1}-G_2(\tilde{K}G_2)^{-1}\right)(\tilde{V}C_1)\,r\right\|_{\mathcal{S}}^2=\left\|(G_1-G_2Q)\,L(\tilde{V}C_1)\,r\right\|_{\mathcal{S}}^2\\
&\geq\alpha_c^2\,\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{n}\sigma_i^2\left((G_1-G_2Q)\phi_r^{\frac{1}{2}}\right)d\omega\quad(\text{using }\underline{\sigma}(L)\;\geq\;1)\\
&\geq\;\frac{\gamma\,\alpha_c^2}{2\pi}\int_{-\pi}^{\pi}x_1\sum_{i=1}^{n}\sigma_i^2(G_1-G_2Q)\,d\omega\\
\text{and}\quad&x_1\sum_{i=1}^{n}\sigma_i^2\,((G_1-G_2Q))\;=x_1\sum_{i=1}^{n}\sigma_i^2\,\left(\begin{bmatrix}\tilde{G}_2\\G_2^*\end{bmatrix}(G_1-G_2Q)\right)\quad(\text{using (A.6)})\\
&\geq\;x_1\sum_{i=1}^{n}\sigma_i^2\left(\tilde{G}_2G_1\right).
\end{aligned}\tag{A.14}$$

Hence

$$\left\|\left(G_1(\tilde{K}G_1)^{-1}-G_2(\tilde{K}G_2)^{-1}\right)(\tilde{V}C_1)\,r\right\|_{\mathcal{S}}^2\;\geq\;\frac{\gamma\,\alpha_c^2}{2\pi}\int_{-\pi}^{\pi}x_1\sum_{i=1}^{n}\sigma_i^2\left(\tilde{G}_2G_1\right)d\omega. \tag{A.15}$$

The result follows from (A.15) and (A.13).

The upper and lower bounds in the second part may be proved similarly by substituting $C_1=\tilde{V}^{-1}(\tilde{K}G_1)$ in (A.11) and (A.14). □

Proof of Theorem 2. From (6.1),

$$z=\begin{bmatrix}y\\u\end{bmatrix}=G_0(\tilde{K}G_0)^{-1}\tilde{V}C_1r+K(\tilde{G}_0K)^{-1}\tilde{G}_0w \tag{A.16}$$

so that

$$\|\tilde{G}_\theta z\|_{\mathcal{S}}^2=\|(\tilde{G}_\theta G_0)(\tilde{K}G_0)^{-1}(\tilde{V}C_1)\,r\|_{\mathcal{S}}^2+\|(\tilde{G}_\theta K)(\tilde{G}_0K)^{-1}\,\tilde{G}_0\,w\|_{\mathcal{S}}^2 \tag{A.17}$$

(since $\phi_{r\omega}=0$)

$$\leq\left(\sup_\omega\overline{\sigma}^2(\tilde{K}G_0)^{-1}\right)\left(\sup_\omega\overline{\sigma}^2(\tilde{V}C_1)\right)\delta_x^2(P_0,P_\theta,F_r)+\left(\sup_\omega\overline{\sigma}^2(\tilde{G}_0K)^{-1}\right)\|\,\tilde{G}_0\,w\|_{\mathcal{S}}^2. \tag{A.18}$$

The lower bound follows from (A.1) using the definitions of β_c and $b(P_0, C)$. To derive the upper bound, note that

$$\begin{aligned}\|\tilde{G}_\theta z\|_S^2 \;\geq\; & \left(\inf_\omega \underline{\sigma}^2\,(\tilde{K}G_0)^{-1}\right)\left(\inf_\omega \underline{\sigma}^2\,(\tilde{V}C_1)\right)\delta_x^2(P_0, P_\theta, F_r) \\ & +\; \left(\inf_\omega \underline{\sigma}^2\,(\tilde{G}_\theta K)\right)\left(\inf_\omega \underline{\sigma}^2\,(\tilde{G_0}K)^{-1}\right)\|\,\tilde{G}_0\,w\|_S^2.\end{aligned}$$

The result then follows using

$$\begin{aligned}&\inf_\omega \underline{\sigma}\,(\tilde{K}G_0)^{-1}(e^{j\omega}) \;\geq\; 1, \qquad \inf_\omega \underline{\sigma}\,(\tilde{G}_0K)^{-1}(e^{j\omega}) \;\geq\; 1 \\ &\text{and} \quad \inf_\omega \underline{\sigma}\,(\tilde{G}_\theta K)(e^{j\omega}) \;=\; b(P_\theta, C). \qquad \square\end{aligned}$$

REFERENCES

[1] A. K. El-Sakkary, *The gap metric: Robustness of stabilization of feedback systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 240–247.

[2] U. Forsell and L. Ljung, *Closed loop identification revisited*, Automatica J. IFAC, 35 (1999), pp. 1215–1241.

[3] T. T. Georgiou and M. C. Smith, *Optimal robustness in the gap metric*, IEEE Trans. Automat. Control, 35 (1990), pp. 673–687.

[4] L. Ljung, *System Identification: Theory for the User*, Prentice–Hall, Englewood Cliffs, NJ, 1999.

[5] P. M. Mäkilä, J. R. Partington, and T. Norlander, *Bounded power signal spaces for robust control and modeling*, SIAM J. Control Optim., 37 (1998), pp. 92–117.

[6] J. R. Partington, *Approximation of unstable infinite-dimensional systems using coprime factors*, Systems Control Lett., 16 (1992), pp. 89–96.

[7] L. Qiu and E. J. Davison, *Pointwise gap metrics on transfer matrices*, IEEE Trans. Automat. Control, 37 (1992), pp. 741–758.

[8] W. Rudin, *Principles of Mathematical Analysis*, McGraw–Hill, New York, 1976.

[9] S. Treil, *A counterexample on continuous coprime factors*, IEEE Trans. Automat. Control, 39 (1994), pp. 1262–1263.

[10] M. Vidyasagar, *The graph metric for unstable plants and robustness estimates for feedback stability*, IEEE Trans. Automat. Control, 29 (1984), pp. 403–418.

[11] G. Vinnicombe, *Frequency domain uncertainty and the graph topology*, IEEE Trans. Automat. Control, 38 (1993), pp. 1371–1383.

[12] G. Vinnicombe, *Uncertainty and Feedback: $\mathcal{H}_\infty$ Loop-Shaping and the ν-Gap Metric*, Imperial College Press, London, 2000.

[13] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice–Hall, Englewood Cliffs, NJ, 1996.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 937–954

CONTROLLABILITY OF POISSON SYSTEMS*

PETRE BIRTEA†, MIRCEA PUTA‡, AND TUDOR S. RATIU†

Abstract. Sufficient conditions for the controllability of affine nonlinear control systems on Poisson manifolds are given. The important special case when the Poisson manifold is the reduced space of a symplectic manifold by a free Lie group action is studied. The controllability of the reduced system is linked to that of the given affine nonlinear system. Several examples illustrating the theory are also presented.

Key words. controllability, symplectic manifold, Poisson manifold, reduction, weak positive Poisson stability

AMS subject classifications. 93B05, 93B29, 53D05, 53D20, 70E55

DOI. 10.1137/S0363012902401251

1. Introduction.

The phase space of classical conservative mechanical systems is usually described by a Poisson manifold $(P, \{\cdot,\cdot\})$. The dynamics on P, subject to external forces, can often be written in the form of an affine nonlinear control system as

$$\dot{x} = X(x) + \sum_{i=1}^{m} Y_i(x) u_i \,,$$

where the drift vector field X is a complete vector field tangent to the symplectic leaves of P and also preserves the symplectic volume on each one of them, $Y_1, \ldots, Y_m \in \mathfrak{X}(P)$ are smooth complete vector fields on P, the control $u := (u_1, \ldots, u_m) : (0, \infty) \longrightarrow B \subset \mathbb{R}^m$ is a measurable function, and B is a bounded subset of $\mathbb{R}^m$.

Deciding the controllability of nonlinear control systems is usually a difficult problem that has generated a large body of literature. As opposed to linear control systems, the Lie algebra rank condition is not sufficient for proving controllability of a nonlinear control system. Nevertheless, there is a link between nonlinear controllability and linear controllability given by the following well-known result: If the linearization of a nonlinear system at an equilibrium is controllable, then the nonlinear system is locally controllable. For nonlinear systems without drift, various characterizations of controllability based on Chow's theorem were obtained. These were generalized to nonlinear systems with drift in terms of the Lie algebra generated by the control vector fields. Significant results were obtained by Hermann [14], Haynes and Hermes [13], Brockett [8], Lobry [28], Sussmann and Jurdjevic [45], Krener [23], and others. Sufficient conditions for controllability of nonlinear systems satisfying the Lie algebra rank condition were obtained by Lobry [29] in the case of Poisson stable systems. This work was generalized by Jurdjevic and Quinn [18] and Bonnard [6] to the case when

*Received by the editors January 22, 2002; accepted for publication (in revised form) February 22, 2004; published electronically October 8, 2004. This research was partially supported by the European Commission and the Swiss Federal Government through funding for the Research Training Network *Mechanics and Symmetry in Europe* (MASIE) (T.S.R.) as well as the Swiss National Science Foundation through FNS grant 20-61228.00 (P.B. and T.S.R.).

http://www.siam.org/journals/sicon/43-3/40125.html

†Centre Bernoulli, École Polytechnique Fédérale de Lausanne, CH–1015 Lausanne, Switzerland (petre.birtea@epfl.ch, tudor.ratiu@epfl.ch).

‡Departamentul de Matematică, Universitatea de Vest din Timişoara, Blvd. V. Parvan 4, RO-300223, Romania (puta@geometry.uvt.ro).

only the drift vector field is required to be Poisson stable. This method was applied by Crouch [11] to the study of spacecraft attitude control problems. In order to analyze the controllability of spacecraft systems, which include attitude-orbit coupling terms and are controlled only by attitude controllers using either reaction wheels or gas jets, Lian, Wang, and Fu [26] replaced the condition on the drift vector field to be Poisson stable with the less stringent condition of weak positive Poisson stability. More precisely, they showed that if an affine nonlinear control system verifies the Lie algebra rank condition and the drift vector field is weakly positively Poisson stable, then the system is controllable.

The problem of controllability for nonlinear systems that are invariant under the action of a Lie group was studied by San Martin and Crouch [42], Jurdjevic and Kupka [17] and, in a more general setting of fiber bundles, by Nijmeijer and van der Schaft [38] (see also Grizzle and Marcus [12] and Sánchez de Alvarez [43]). Other results concerning different aspects of the relation between the given and the reduced control system can be found in Jalnapurkar and Marsden [15], [16] and Bloch, Leonard, and Marsden [5].

The aim of this paper is to give sufficient conditions for the controllability of an affine control system on a Poisson manifold. For the case when the manifold is the cotangent bundle of a Lie group, this problem was studied by Manikonda and Krishnaprasad [30]; it was this paper that has inspired the present generalization. The strategy of the proof of the main results is to give topological conditions that guarantee that the drift vector field is weakly positively Poisson stable (WPPS). In order to do this, we will use the Poincaré recursion theorem for the dynamics of the drift vector field restricted to each symplectic leaf. We will prove that if one can find a continuous function $f : P \to \mathbb{R}$ that is constant on the flow of X and is such that either f restricted to each symplectic leaf is a proper function or f is a proper function from P to $\mathbb{R}$ and all symplectic leaves are closed and embedded submanifolds of P, then X is WPPS. There is a relatively subtle technical point in the proof of this theorem: The topology of a symplectic leaf is stronger than the relative topology induced by the ambient space P; that is, every open set in the induced topology on the leaf is also open in the immersed topology of the leaf, but there exist open sets in the immersed topology of the leaf that are not open in the induced topology. This immediately implies that there are subsets in the leaf which are compact in the induced topology but are not compact in the immersed topology on the leaf.

As an important case of this first result, we will study the situation when the Poisson manifold is the reduced space of a symplectic manifold by a free proper Lie group action which also admits a momentum map. We will show that if the momentum map is proper and the reduced affine nonlinear system verifies the Lie algebra rank condition, then it is controllable. Similarly, if the momentum map is not proper but the Lie group is compact and there is a proper map $f : M/G \to \mathbb{R}$ which is constant along the trajectories of the reduced drift vector field and in addition the reduced affine nonlinear system verifies the Lie algebra rank condition, then the system is controllable. We will also give the relation between the controllability of the reduced and initial affine nonlinear systems.

The paper ends with some examples of underactuated affine nonlinear control systems; for this, some useful technical lemmas implying the properness of functions are also given.

After this paper was submitted the authors were made aware of the work of Manikonda and Krishnaprasad [31] (a preliminary version of these results can be

found in Krishnaprasad and Manikonda [24]) with which the present work has a certain amount of overlap. We shall mention in the text explicitly where this is the case and compare their results to ours.

2. Controllability and Poisson stability. In this section we shall present a controllability result for affine nonlinear control systems on a general Poisson manifold. We begin by reviewing the classical definitions and results that will be used later on by adopting the terminology in the standard textbook of Nijmeijer and van der Schaft [39].

Let M be a smooth n-dimensional connected manifold and

$$\dot{x} = X(x) + \sum_{i=1}^{m} Y_i(x)u_i \tag{2.1}$$

an affine nonlinear control system on M, where $X, Y_1, \ldots, Y_m \in \mathfrak{X}(M)$ are smooth complete vector fields on M, the control $u := (u_1, \ldots, u_m) : (0, \infty) \longrightarrow B \subset \mathbb{R}^m$ is a measurable function, and B is a bounded subset of $\mathbb{R}^m$. We will denote by $\mathcal{L}$ the Lie subalgebra of $\mathfrak{X}(M)$ generated by the vector fields $X, Y_1, \ldots, Y_m$.

DEFINITION 2.1. *The system* (2.1) *satisfies the* Lie algebra rank condition (LARC) *if* $\operatorname{span} \mathcal{L}(x) = T_x M$ *for every* $x \in M$, *where* $\mathcal{L}(x) := \{Z(x) \mid Z \in \mathcal{L}\}$.

DEFINITION 2.2. *The system* (2.1) *is* controllable *if for any two points* $x_I, x_F \in M$ *there is a control* u *which takes the system from point* $x = x_I$ *at time* $t = t_I \in \mathbb{R}$ *to the point* $x = x_F$ *at time* $t = t_F \in \mathbb{R}$, *that is, if for a certain choice of the function* u *there is an integral curve* $x(t)$ *of* (2.1) *that begins at* x_I *and ends at* x_F *in finite time.*

It is well known that for a nonlinear control system without drift (i.e., $X = 0$), the LARC implies controllability. This is Chow's theorem [9]. For the general case $X \neq 0$, the situation is more complicated and, in general, the LARC is not sufficient to guarantee controllability. A lot of work was done in this direction and we will review below only the results relevant for our purposes.

In what follows we shall need a condition, called weak positive Poisson stability, on the drift vector field X. In order to understand how this concept appeared in the literature we shall quickly relate it to standard notions in the theory of dynamical systems. The next three definitions were introduced originally in Nijmeijer and van der Schaft [39], Lobry [28], [29], and Lian, Wang, and Fu [26]. Let $X \in \mathfrak{X}(M)$ be a smooth complete vector field on M and $\{\Phi_t\}_{t\in\mathbb{R}}$ its flow.

DEFINITION 2.3. *A point* $x \in M$ *is called* positively Poisson stable *for* $X \in \mathfrak{X}(M)$ *if for any* $T > 0$ *and any neighborhood* V_x *of* x *there exists a time* $t > T$ *such that* $\Phi_t(x) \in V_x$. *The vector field* $X \in \mathfrak{X}(M)$ *is called* positively Poisson stable *if the set of positively Poisson stable points of* X *is dense in* M.

DEFINITION 2.4. *A point* $x \in M$ *is called a* nonwandering point *of* $X \in \mathfrak{X}(M)$ *if for any* $T > 0$ *and for any neighborhood* V_x *of* x *there exists a time* $t > T$ *such that* $\Phi_t(V_x) \cap V_x \neq \emptyset$.

Let Γ_X be the set of all nonwandering points of X, usually called the *nonwandering set* of X. The following result and its proof can be found in Lian, Wang, and Fu [26].

THEOREM 2.5. *The nonwandering set of a positively Poisson stable vector field* X *is the entire manifold* M, *that is,* $\Gamma_X = M$.

Proof. For a given $x \in M$ one needs to prove that for any neighborhood V_x of x and for any $T > 0$ there exists a time $t > T$ such that $\Phi_t(V_x) \cap V_x \neq \emptyset$. Let

$\underline{S_X}$ denote the set of positively Poisson stable points of $X \in \mathfrak{X}(M)$. By definition, $\overline{S_X} = M$. Thus there is a positively Poisson stable point y in V_x. This implies that for all $T > 0$ there is a time $t > T$ such that $\Phi_t(y) \in V_x$. Hence $\Phi_t(V_x) \cap V_x \neq \emptyset$. So x is nonwandering for X. Since x was arbitrary, we get $\Gamma_X = M$. □

Positive Poisson stability of a vector field is hence a sufficient condition for the nonwandering set to be the entire manifold. Since the converse is not true, one introduces a weaker definition.

DEFINITION 2.6. *A vector field is called WPPS if its nonwandering set equals M (i.e., $\Gamma_X = M$).*

A natural question that arises now is the following: When is a vector field X on a manifold WPPS? In order to answer this question, we will recall the Poincaré recursion theorem (for a proof see, e.g., Abraham and Marsden [1], Abraham, Marsden, and Ratiu [2]).

Let (M, Ω) be a manifold with a volume form Ω. Let $\mathcal{B}$ denote the collection of Borel sets on M, that is, the σ-algebra generated by the open (or closed, or compact) subsets of M. Then there exists a unique Borel measure m_Ω on $\mathcal{B}$ such that for every continuous function f with compact support

$$\int_M f dm_\Omega = \int_M f\Omega. \tag{2.2}$$

For K a compact and U an open subset of M, we have $m_\Omega(K) < \infty$ and $m_\Omega(U) > 0$. If we consider on M a vector field whose flow preserves the volume form (i.e., $\Phi_t^*\Omega = \Omega$), then $m_\Omega(\Phi_t(A)) = m_\Omega(A)$ for any measurable subset A of M.

THEOREM 2.7 (Poincaré recursion theorem). *Let (M, Ω) be a manifold with a volume form Ω and m_Ω the associated Borel measure. Let X be a time-independent, complete vector field such that its flow $\{\Phi_t\}_{t\in\mathbb{R}}$ preserves the volume. Suppose A is a measurable subset of M with $0 < m_\Omega(A) < \infty$ which is also invariant under the flow of X. Then for each measurable subset B of A with $m_\Omega(B) > 0$ and for any $T > 0$, there exists $t > T$ such that $\Phi_t(B) \cap B \neq \emptyset$.*

An immediate consequence is the following proposition.

PROPOSITION 2.8. *Let (M, Ω) be a compact manifold with a volume form Ω and X a time-independent vector field such that its flow preserves the volume form. Then X is a WPPS vector field.*

The link between the WPPS condition and controllability is given by the following theorem which is due to Lian, Wang, and Fu [26]. Earlier versions of this theorem, where the hypothesis required X to be Poisson stable, are due to Lobry [29], Bonnard [6], and Crouch [11].

THEOREM 2.9. *Suppose that X is a WPPS vector field. Then the system* (2.1) *is controllable if and only if the LARC holds.*

We now state our first result on controllability of an affine nonlinear control system on a Poisson manifold. Recall that a finite dimensional Poisson manifold is a smooth manifold P whose ring of smooth real-valued functions $C^\infty(P)$ is endowed with a Lie algebra structure $\{\cdot,\cdot\}$ satisfying the Leibniz identity in every factor. Thus, if $h \in C^\infty(P)$, the derivation $\{\cdot, h\}$ defines a vector field X_h on P, called the *Hamiltonian vector field* induced by the *Hamiltonian function* h, that is, $\langle df, X_h\rangle = \{f, h\}$ for any $f \in C^\infty(P)$. The vector fields $\{X_h \mid h \in C^\infty(P)\}$ define a singular integrable distribution whose integral manifolds are symplectic immersed submanifolds whose Poisson bracket coincides with the given one on P; these integral manifolds are called the *symplectic leaves* of P. For further information on Poisson manifolds see, for

example, Libermann and Marle [27], Marsden [34], Marsden and Ratiu [35], and Puta [41].

Note that the topology of a symplectic leaf is stronger than the topology induced by the ambient manifold P, that is, every open set in the induced topology on the leaf is also open in the immersed topology but there exist open sets in the immersed topology of the leaf that are not open in the induced topology. Therefore, there are compact subsets in the induced topology of the leaf that are not compact in the immersed topology of the leaf. In the next proof one needs to come to grips with this problem.

THEOREM 2.10. *Let $(P, \{\cdot,\cdot\})$ be a connected Poisson manifold and*

$$\dot{x} = X(x) + \sum_{i=1}^{m} Y_i(x)u_i$$

an affine nonlinear control system such that the drift vector field X is tangent to the symplectic leaves of P and also preserves the symplectic volume on each one of them. Let $f : P \to \mathbb{R}$ be a continuous function that is constant on the flow of X. Assume that one of the following hypotheses holds:

(i) *f restricted to each symplectic leaf is a proper function.*

(ii) *f is a proper function and all symplectic leaves are closed and embedded submanifolds of P.*

Then X is WPPS. If the system also verifies the LARC, then it is controllable.

Proof. Let x_0 be an arbitrary point of P and U_{x_0} an arbitrary open neighborhood of x_0 in P. Denote by L_{x_0} the symplectic leaf containing x_0 and let $c_0 = f(x_0)$. There are two possibilities: $f(L_{x_0}) = c_0$, or $f(L_{x_0}) = I$, where I is a nondegenerate connected interval in $\mathbb{R}$.

Assume first that $f(L_{x_0}) = c_0$. Under hypothesis (i), $f_{|L_{x_0}} : L_{x_0} \to \mathbb{R}$ is a proper function, so $L_{x_0} = f^{-1}_{|L_{x_0}}(c_0)$ is compact. By Proposition 2.8 it follows that X restricted to the leaf L_{x_0} is a WPPS vector field, which implies that x_0 is a nonwandering point for the flow ϕ_t of X on L_{x_0}. Thus for any $T > 0$ there exists $t > T$ such that $(L_{x_0} \cap U_{x_0}) \cap \phi_t(L_{x_0} \cap U_{x_0}) \neq \emptyset$ which, in particular, implies that $U_{x_0} \cap \phi_t(U_{x_0}) \neq \emptyset$. Since x_0 and U_{x_0} were arbitrary, it follows that X is WPPS on P. Under hypothesis (ii), $f : P \to \mathbb{R}$ is a proper function, so $f^{-1}(c_0)$ is compact in P. Since $L_{x_0} \subset f^{-1}(c_0)$ and L_{x_0} is closed and embedded in P by hypothesis, it follows that L_{x_0} is compact in P. As before, applying Proposition 2.8, we obtain that X is WPPS.

Now assume that $f(L_{x_0}) = I$, where I is a nondegenerate connected interval. Then, without loss of generality (replacing x_0 with another point in the leaf, if necessary), we can assume that c_0 lies in the interior of I and hence there is an $\varepsilon > 0$ such that $[-\varepsilon + c_0, c_0 + \varepsilon] \subset I$. The set $K := L_{x_0} \cap f^{-1}([-\varepsilon + c_0, c_0 + \varepsilon])$ is compact in L_{x_0} in hypothesis (i) because $f_{|L_{x_0}}$ is proper and in hypothesis (ii) because L_{x_0} is closed and embedded in P and $f^{-1}([-\varepsilon + c_0, c_0 + \varepsilon])$ is compact in P. This implies that $m_{L_{x_0}}(K) < \infty$, where $m_{L_{x_0}}$ is the Borel measure associated to the symplectic volume form on L_{x_0}. Also, K contains an open set of L_{x_0}, for example, $f^{-1}_{|L_{x_0}}((-\varepsilon + c_0, c_0 + \varepsilon))$, and thus $m_{L_{x_0}}(K) > 0$. By the Poincaré recursion theorem (see Theorem 2.7), for any $T > 0$ there exists a $t > T$ such that $(K \cap U_{x_0}) \cap \phi_t(K \cap U_{x_0}) \neq \emptyset$ which, in particular, implies that $U_{x_0} \cap \phi_t(U_{x_0}) \neq \emptyset$. Consequently, x_0 is a nonwandering point of the flow of X. Since x_0 was arbitrary, it follows that X is WPPS on P.

If the control system also satisfies the LARC, Theorem 2.9 implies that it is controllable. □

It should be noted that in hypothesis (ii) there are two hypotheses on the symplectic leaves of P: They need to be embedded and closed. It can happen, even in the Lie–Poisson case, that the leaves are embedded but not closed. For example, the Poisson manifold $\mathbb{R}^2$ with the bracket given by $\{f,g\}(x,y) = y(f_x g_y - f_y g_x)$ has the upper and the lower half plane as open two-dimensional symplectic leaves and the points on the x-axis as the zero-dimensional leaves.

An important case in which the drift vector field X satisfies the hypotheses of the theorem is when $X = X_h$ for some Hamiltonian function $h \in C^\infty(P)$. Indeed, X_h is always tangential to the leaves and it preserves the symplectic volume on each leaf by the Liouville theorem.

Note that if P is a Poisson manifold, in order for the above affine nonlinear control system to verify the LARC it is necessary that at least one of the vector fields $Y_1, \ldots, Y_m \in \mathfrak{X}(P)$ be non-Hamiltonian.

Theorem 2.10 immediately implies both Theorems 4.3 and 4.11 in Manikonda and Krishnaprasad [31]. Finally, it should be noted that this theorem applies in the particular, but important, case of Lie–Poisson systems. In addition, Theorem 2.10(i) can handle Poisson manifolds with nonembedded and nonclosed symplectic leaves, such as the Kirillov example of the dual of a five-dimensional semidirect product Lie group with coadjoint orbits that accumulate on themselves (see Kirillov [22] or Marsden and Ratiu [35] for a discussion of this Lie group and its coadjoint orbits).

3. Controllability of reduced systems. In this section we shall study the important case of the previous theorem when the Poisson manifold is the reduction of a symplectic manifold by a compact Lie group action. In this particular case we can give sufficient topological conditions that imply the hypotheses of Theorem 2.10. To do this, we begin with a quick review of some standard results on symplectic reduction necessary in the subsequent proofs; detailed expositions of this subject can be found in standard textbooks such as Abraham and Marsden [1], Libermann and Marle [27], Marsden [34], Marsden and Ratiu [35], Ortega and Ratiu [40], and Puta [41].

Consider a $2n$-dimensional connected symplectic manifold (M,ω) on which there is a free proper symplectic action of a Lie group G. Denote by $\{\cdot,\cdot\}_\omega$ the Poisson bracket on M defined by the symplectic form ω. Then the orbit space M/G is a smooth Poisson manifold and the projection

$$\pi : (M, \{\cdot,\cdot\}_\omega) \longrightarrow \left(M/G, \{\cdot,\cdot\}_{M/G}\right)$$

is a Poisson surjective submersion. If, in addition, the Lie group G is compact, then π is a closed proper map. (Proofs of these statements can be found in, e.g., Abraham and Marsden [1], Abraham, Marsden, and Ratiu [2], Bredon [7], Kawakubo [19], Libermann and Marle [27], and Ortega and Ratiu [40].)

Suppose that the free and proper G-action on M admits an associated momentum map $J : M \longrightarrow \mathfrak{g}^*$. If the momentum map is not equivariant with respect to the coadjoint action of G on $\mathfrak{g}^*$, then there is a $\mathfrak{g}^*$-valued group one-cocycle σ on G such that $\sigma(g) = J(g \cdot m) - \mathrm{Ad}^*_{g^{-1}} J(m)$ for every $m \in M$ and $g \in G$, where Ad^* denotes the coadjoint representation of G on $\mathfrak{g}^*$. (The connectedness of M is needed to show that the right-hand side is independent of m.) Defining the affine action of G on $\mathfrak{g}^*$ by $g \cdot \mu := \mathrm{Ad}^*_{g^{-1}} \mu + \sigma(g)$, the momentum map $J : M \to \mathfrak{g}^*$ becomes now equivariant relative to the given action on M and the just defined affine action on $\mathfrak{g}^*$. The *Marsden–Weinstein reduction theorem* states that if $\mu \in \mathfrak{g}^*$ is a value of J, then the smooth quotient manifold $M_\mu := J^{-1}(\mu)/G_\mu$ is symplectic with symplectic form

ω_μ characterized by

$$\pi_\mu^* \omega_\mu = i_\mu^* \omega,$$

where G_μ denotes the isotropy subgroup of μ under the affine action, $i_\mu : J^{-1}(\mu) \longrightarrow M$ is the inclusion, and $\pi_\mu : J^{-1}(\mu) \longrightarrow M_\mu$ is the projection. (For a proof, see the original paper Marsden and Weinstein [36], or Abraham and Marsden [1], Libermann and Marle [27], Marsden [34], and Puta [41].) The symplectic manifolds (M_μ, ω_μ) will be called *point reduced spaces*.

These point reduced spaces M_μ can be understood in a natural way as symplectic leaves of the Poisson manifold $(M/G, \{\cdot,\cdot\}_{M/G})$. Indeed, the smooth map $j_\mu : M_\mu \longrightarrow M/G$ naturally defined by the commutative diagram

$$\begin{array}{ccc} J^{-1}(\mu) & \xrightarrow{\quad i_\mu \quad} & M \\ \Big\downarrow{\scriptstyle \pi_\mu} & & \Big\downarrow{\scriptstyle \pi} \\ M_\mu & \xrightarrow{\quad j_\mu \quad} & M/G \end{array}$$

is a Poisson injective immersion. Moreover, the j_μ-images in M/G of the connected components of the symplectic manifolds (M_μ, ω_μ) are its symplectic leaves (see Manikonda and Krishnaprasad [30] or Ortega and Ratiu [40]).

Observe that, in general, j_μ is only an injective immersion. So the topology of the image of j_μ, homeomorphic to the topology of M_μ, is stronger than the subspace topology induced by the ambient space M/G. This image topology on $j_\mu(M_\mu)$ is called the *immersed topology.* As in the previous section, we draw attention to the fact that we can have a subset of $j_\mu(M_\mu)$ which is compact in the induced topology from M/G and not compact in the immersed topology. A key point in the proof of Theorem 3.4 on controllability stated in what follows is to give sufficient and easily verifiable conditions under which these two topologies coincide.

The proof of the next proposition requires compactness of G.

PROPOSITION 3.1. *Suppose that the free symplectic compact G-action on (M, ω) admits a momentum map $J : M \longrightarrow \mathfrak{g}^*$. Then the symplectic leaves of $(M/G, \{\cdot,\cdot\}_{M/G})$ are closed sets.*

Proof. Since $J^{-1}(\mu)$ is closed in M and $\pi : M \longrightarrow M/G$ is a closed map (because G is compact), the set $j_\mu(M_\mu) = \pi(J^{-1}(\mu))$ is closed in the topology of M/G. Therefore, the connected components of $j_\mu(M_\mu)$, which are the symplectic leaves of M/G, are also closed in the topology of M/G. □

We return now to the general case with G noncompact. Up to now we have regarded the symplectic leaves of $(M/G, \{\cdot,\cdot\}_{M/G})$ as the j_μ-images of the connected components of M_μ. However, as sets,

$$j_\mu(M_\mu) = J^{-1}(\mathcal{O}_\mu)/G,$$

where $\mathcal{O}_\mu \subset \mathfrak{g}^*$ is the orbit through μ relative to the affine action of G on $\mathfrak{g}^*$. The set $M_{\mathcal{O}_\mu} := J^{-1}(\mathcal{O}_\mu)/G$ is called *the orbit reduced space* associated to the orbit $\mathcal{O}_\mu$. The smooth manifold structure (and hence the topology) on $M_{\mathcal{O}_\mu}$ is the one that makes $j_\mu : M_\mu \longrightarrow M_{\mathcal{O}_\mu}$ into a diffeomorphism.

The group one-cocycle σ induces by derivation a real-valued Lie algebra two-cocycle $\Sigma : \mathfrak{g} \times \mathfrak{g} \to \mathbb{R}$ which can be shown to equal $\Sigma(\xi, \eta) = J^{[\xi,\eta]}(m) - \{J^\xi, J^\eta\}(m)$

for every $m \in M$ and $\xi, \eta \in \mathfrak{g}$; $J^\xi : M \to \mathbb{R}$ denotes the ξ-component of J, that is, $J^\xi(m) := \langle J(m), \xi \rangle$. Denote by $\xi_{\mathfrak{g}^*}(\nu) := -\mathrm{ad}^*_\xi \nu + \Sigma(\xi, \cdot)$ the infinitesimal generator of the affine action of G on $\mathfrak{g}^*$, for $\nu \in \mathfrak{g}^*$, where ad^* denotes the dual of the adjoint representation ad of $\mathfrak{g}$ on $\mathfrak{g}$ defined by $\mathrm{ad}_\xi \eta := [\xi, \eta]$, for $\xi, \eta \in \mathfrak{g}$. The affine action orbit $\mathcal{O}_\mu$ carries two symplectic forms given by

$$\omega^{\pm}_{\mathcal{O}_\mu}(\nu)(\xi_{\mathfrak{g}^*}(\nu), \eta_{\mathfrak{g}^*}(\nu)) = \pm\langle \nu, [\xi, \eta]\rangle \mp \Sigma(\xi, \eta), \tag{3.1}$$

for any $\xi, \eta \in \mathfrak{g}$. They are the natural modifications of the usual Kirillov–Kostant–Souriau symplectic forms on coadjoint orbits. For the proofs of the statements above see Abraham and Marsden [1], Libermann and Marle [27], Ortega and Ratiu [40], and Puta [41]; the formulation used above is that of Ortega and Ratiu [40].

The next theorem characterizes the symplectic form and the Hamiltonian dynamics on $M_{\mathcal{O}_\mu}$.

THEOREM 3.2 (symplectic orbit reduction). *Assume that the free proper symplectic action of the Lie group G on the symplectic manifold (M, ω) admits an associated momentum map $J : M \longrightarrow \mathfrak{g}^*$.*

(i) *On $J^{-1}(\mathcal{O}_\mu)$ there is a unique immersed smooth manifold structure such that $\pi_{\mathcal{O}_\mu} : J^{-1}(\mathcal{O}_\mu) \longrightarrow M_{\mathcal{O}_\mu}$ is a surjective submersion, where $M_{\mathcal{O}_\mu}$ is endowed with the manifold structure making j_μ into a diffeomorphism. This smooth manifold structure does not depend on the choice of μ in the orbit $\mathcal{O}_\mu$. If $J^{-1}(\mathcal{O}_\mu)$ is a submanifold of M in its own right, then the immersed topology and the induced topology on $M_{\mathcal{O}_\mu}$ coincide.*

(ii) *$M_{\mathcal{O}_\mu}$ is a symplectic manifold with the symplectic form $\omega^{\maltese}_{\mathcal{O}_\mu}$ uniquely characterized by the relation*

$$i^*_{\mathcal{O}_\mu}\omega = \pi^*_{\mathcal{O}_\mu}\omega^{\maltese}_{\mathcal{O}_\mu} + J^*_{\mathcal{O}_\mu}\omega^{+}_{\mathcal{O}_\mu},$$

where $J_{\mathcal{O}_\mu}$ is the restriction of J to $J^{-1}(\mathcal{O}_\mu)$, $i_{\mathcal{O}_\mu} : J^{-1}(\mathcal{O}_\mu) \hookrightarrow M$ is the inclusion, and $\omega^+_{\mathcal{O}_\mu}$ is the $+$orbit symplectic form on $\mathcal{O}_\mu$ given by (3.1).

(iii) *Let H be a G-invariant function on M, and define $\widetilde{H} : M/G \longrightarrow \mathbb{R}$ by $H = \widetilde{H} \circ \pi$. Then the Hamiltonian vector field X_H is also G-invariant and hence induces a vector field on M/G which coincides with the Hamiltonian vector field $X_{\widetilde{H}}$. Moreover, the flow of $X_{\widetilde{H}}$ leaves the symplectic leaves $M_{\mathcal{O}_\mu}$ of M/G invariant. This flow restricted to the symplectic leaves is again Hamiltonian relative to the symplectic form $\omega^{\maltese}_{\mathcal{O}_\mu}$ and the Hamiltonian function $\widetilde{H}_{\mathcal{O}_\mu}$ given by*

$$\widetilde{H}_{\mathcal{O}_\mu} \circ \pi_{\mathcal{O}_\mu} = H \circ i_{\mathcal{O}_\mu}.$$

The proof of this theorem in the regular case and when $\mathcal{O}_\mu$ is an embedded submanifold of $\mathfrak{g}^*$ can be found in Marle [32], Kazhdan, Kostant, and Sternberg [20], and Marsden [33]. For the general case, when $\mathcal{O}_\mu$ is not a submanifold of $\mathfrak{g}^*$, see Ortega and Ratiu [40]. Here is the main idea of the proof. Consider for each value $\mu \in \mathfrak{g}^*$ of J the G-equivariant bijection

$$s : G \times_{G_\mu} J^{-1}(\mu) \to J^{-1}(\mathcal{O}_\mu),$$

$$[g, m] \mapsto g \cdot m,$$

where $G \times_{G_\mu} J^{-1}(\mu) := (G \times J^{-1}(\mu))/G_\mu$, the G_μ-action being the diagonal action. Endow $J^{-1}(\mathcal{O}_\mu)$ with the smooth manifold structure that makes the bijection s into a diffeomorphism. Then $J^{-1}(\mathcal{O}_\mu)$ with this smooth structure is an immersed submanifold of M. This is the manifold structure on $J^{-1}(\mathcal{O}_\mu)$ used in the statement of Theorem 3.2.

In the particular case in which $J^{-1}(\mathcal{O}_\mu)$ is a smooth submanifold of M in its own right, this manifold structure coincides with the one induced by the mapping s described previously since in this situation the bijection s becomes a diffeomorphism relative to the a priori given smooth manifold structure on $J^{-1}(\mathcal{O}_\mu)$.

If μ is a regular value of J and $\mathcal{O}_\mu$ is an embedded submanifold of $\mathfrak{g}^*$, then J is transverse to $\mathcal{O}_\mu$ and hence $J^{-1}(\mathcal{O}_\mu)$ is automatically an embedded submanifold of M.

The following result is important for our work through its consequences.

PROPOSITION 3.3 (bifurcation lemma). *Let (M, ω) be a symplectic manifold and G a Lie group acting symplectically on M (not necessarily freely). Suppose also that the action has an associated momentum map $J : M \longrightarrow \mathfrak{g}^*$. For any $m \in M$,*

$$(\mathfrak{g}_m)^\circ = \text{range}(T_m J),$$

where $\mathfrak{g}_m = \{\xi \in \mathfrak{g} \mid \xi_M(m) = 0\}$ is the Lie algebra of the isotropy subgroup $G_m = \{g \in G \mid g \cdot m = m\}$ and $(\mathfrak{g}_m)^\circ = \{\mu \in \mathfrak{g}^* \mid \mu_{|\mathfrak{g}_m} = 0\}$ denotes the annihilator of $\mathfrak{g}_m$ in $\mathfrak{g}^*$.

An immediate consequence of this is the fact that when the action of G is free, then every value $\mu \in \mathfrak{g}^*$ of the momentum map J is a regular value of J.

Now we give the setting for the controllability result on M/G. Let G be a Lie group acting freely properly and symplectically on a $2n$-dimensional connected symplectic manifold (M, ω). Suppose that the action admits an associated momentum map $J : M \longrightarrow \mathfrak{g}^*$. Consider on M the affine nonlinear control system

$$\dot{x} = X_H(x) + \sum_{i=1}^{m} Y_i(x)\, u_i\,, \tag{3.2}$$

where X_H is a complete Hamiltonian vector field with G-invariant Hamiltonian H, the smooth vector fields $Y_1, \ldots, Y_m \in \mathfrak{X}(M)$ are assumed to be G-invariant and complete, and the control $u := (u_1, \ldots, u_m) : (0, \infty) \longrightarrow B \subset \mathbb{R}^m$ is a measurable function with values in a bounded subset B of $\mathbb{R}^m$. Then the system (3.2) will naturally induce the affine nonlinear control system on $(M/G, \{\cdot, \cdot\}_{M/G})$,

$$\dot{\widetilde{x}} = X_{\widetilde{H}}(\widetilde{x}) + \sum_{i=1}^{m} \widetilde{Y}_i(\widetilde{x})\, u_i\,, \tag{3.3}$$

where $X_{\widetilde{H}}$ is the Hamiltonian vector field with respect to the Poisson bracket $\{\cdot, \cdot\}_{M/G}$ and Hamiltonian function $\widetilde{H}$ given by $H = \widetilde{H} \circ \pi$, for $\pi : M \longrightarrow M/G$ the canonical projection. The following theorem generalizes Theorem 4.11 in Manikonda and Krishnaprasad [31] in the sense that it can deal with noncompact Lie group actions and nonequivariant momentum maps.

THEOREM 3.4. *Suppose that the system* (3.3) *verifies the LARC.*

(i) *If the momentum map $J : M \longrightarrow \mathfrak{g}^*$ is proper, then the system* (3.3) *is controllable.*

(ii) *If the Lie group G is compact and if there exists a continuous proper map $f : M/G \longrightarrow \mathbb{R}$ which is constant along the trajectories of $X_{\widetilde{H}}$, then the system* (3.3) *is controllable.*

Proof. The strategy to prove the controllability of (3.3) is to show that $X_{\widetilde{H}}$ is WPPS and then the conclusion follows from Theorem 2.9.

(i) As subsets of M/G, the symplectic leaves are $M_{\mathcal{O}_\mu}$ or, equivalently, $j_\mu(M_\mu)$ and the symplectic form is given by $\omega^{\maltese}_{\mathcal{O}_\mu}$. Because J is a proper map, the set $J^{-1}(\mu)$ is a compact submanifold of M. Thus M_μ is a compact manifold, which implies that the injective immersion j_μ is in fact an embedding. So the immersed topology and the induced topology on $M_{\mathcal{O}_\mu}$ coincide and, therefore, the symplectic leaves are compact embedded submanifolds of M/G.

The vector field $X_{\widetilde{H}}$ is tangent to the leaves $M_{\mathcal{O}_\mu}$ of M/G and is Hamiltonian on each of them relative to the symplectic form $\omega^{\maltese}_{\mathcal{O}_\mu}$. In particular, its flow preserves the Liouville volume on each leaf. Since the leaves are compact, the restriction of the vector field $X_{\widetilde{H}}$ to every leaf is WPPS by Proposition 2.8. Thus each point of every leaf is a nonwandering point of the flow of $X_{\widetilde{H}}$; that is, the nonwandering set of $X_{\widetilde{H}}$ equals P. Thus $X_{\widetilde{H}}$ is WPPS.

(ii) For compact G, the coadjoint orbits are submanifolds of $\mathfrak{g}^*$ and J is transverse to the coadjoint orbits that lie in its image (since by hypothesis, the action is free). So $J^{-1}(\mathcal{O}_\mu)$ is a submanifold of M in its own right and by Theorem 3.2(i) the immersed topology and the induced topology on the symplectic leaves $M_{\mathcal{O}_\mu}$ of M/G coincide. By Proposition 3.1, these leaves are also closed. So we are in the hypotheses of Theorem 2.10(ii) and the result follows. □

The relationship between the controllability of the reduced system (3.3) and the initial system (3.2) is given by the following corollary, also contained in Theorem 4.11 of Manikonda and Krishnaprasad [31].

COROLLARY 3.5. *Suppose that the initial system* (3.2) *verifies the LARC and the hypotheses in Theorem* 3.4(ii). *Then the system* (3.2) *is also controllable.*

Proof. Since the vector fields X_H and $X_{\widetilde{H}}$ are π-related, the function $f \circ \pi$ is a constant of the motion for X_H. This function is proper as a composition of two proper maps; π is proper because G is compact. We are in the hypotheses of Theorem 2.10(ii) since M is a symplectic manifold and hence its symplectic leaves, when thinking of M as a Poisson manifold, are its connected components. □

Remark 3.6. Note that for the controllability of (3.3) it is not necessary for the vector fields $\widetilde{Y}_i \in \mathfrak{X}(M/G)$ to be induced by some G-invariant vector fields on M.

4. Examples. We will illustrate the theory with several examples. In all of them we will use the following well-known lemmas to prove the properness of the integrals of motion.

LEMMA 4.1. *Let $f : \mathbb{R}^n \to \mathbb{R}^k$ be a continuous function. Then f is proper if and only if*

$$\lim_{\|x\| \to \infty} \|f(x)\| = +\infty.$$

Proof. Suppose that f is proper. If $\lim_{\|x\|\to\infty} \|f(x)\| \neq +\infty$, there exists a sequence $\{x_n\}_{n\in\mathbb{N}}$ and a constant $M > 0$ such that $\|x_n\| \to \infty$ and $\|f(x_n)\| \leq M$. Thus $\{x_n\}_{n\in\mathbb{N}}$ lies in the inverse image by f of the closed ball of radius M, which is a compact set in $\mathbb{R}^n$ because f is assumed to be proper. Hence $\{x_n\}_{n\in\mathbb{N}}$ contains a convergent subsequence. However, $\|x_n\| \to \infty$, which is a contradiction.

Conversely, assume that $\lim_{\|x\|\to\infty} \|f(x)\| = +\infty$, and let $K \subset \mathbb{R}^k$ be a compact subset. The set $f^{-1}(K)$ is closed since f is continuous. To conclude that $f^{-1}(K)$ is compact we shall show that it is also bounded. If not, there would exist a sequence $\{x_n\}_{n\in\mathbb{N}} \subset f^{-1}(K)$ such that $\|x_n\| \to \infty$. By hypothesis, $\|f(x_n)\| \to \infty$, which contradicts the fact that $f(x_n) \in K$, which is bounded. □

LEMMA 4.2. *Let M, N, and P be Hausdorff topological spaces. Let $f : M \to N$ and $g : N \to P$ be two continuous functions. If $g \circ f : M \to P$ is proper, then f is also proper.*

Proof. Let $K \subset M$ be a compact subset. Then $g(K)$ is compact in P and hence $(g\circ f)^{-1}(g(K))$ is compact in M. Since $f^{-1}(K) \subset (g\circ f)^{-1}(g(K))$ is closed, it follows that it is also compact. □

Example 1. We will study the controllability of the Hamiltonian system describing the motion of a hollow rigid body and a particle oscillating in it; the body moves about a fixed point which is also the equilibrium position of the particle that oscillates along one of the principal axes of inertia. The description of this system and the proof of its nonintegrability by the method of Ziglin can be found in Christov [10].

The equations of motion are

$$\begin{aligned}
\dot{x}_1 &= \frac{x_2x_3}{C} - \frac{x_2x_3}{B+mx_4^2},\\
\dot{x}_2 &= \frac{x_1x_3}{A+mx_4^2} - \frac{x_1x_3}{C},\\
\dot{x}_3 &= \frac{x_1x_2}{B+mx_4^2} - \frac{x_1x_2}{A+mx_4^2},\\
\dot{x}_4 &= x_5,\\
\dot{x}_5 &= \frac{x_1^2x_4}{(A+mx_4^2)^2} + \frac{x_2^2x_4}{(B+mx_4^2)^2} - \frac{\sigma x_4}{m},
\end{aligned}$$

where $A > B > C$ are the principal moments of inertia, σ is the stiffness of the spring, and m is the mass of the particle.

This is a Hamiltonian system with phase space $so(3)^* \times \mathbb{R}^2$. The Poisson bracket is the product of the Lie–Poisson bracket on $so(3)^*$ with the Poisson bracket induced by the symplectic form $m dx_4 \wedge dx_5$ on $\mathbb{R}^2$. The Hamiltonian is given by

$$H = \frac{1}{2}\left(\frac{x_1^2}{A+mx_4^2} + \frac{x_2^2}{B+mx_4^2} + \frac{x_3^2}{C} + \sigma x_4^2 + mx_5^2\right).$$

It is easy to see that the symplectic leaves are embedded closed manifolds: Every four-dimensional leaf is the product of a sphere with $\mathbb{R}^2$, and the two-dimensional leaf is $\mathbb{R}^2$. One can easily check that the Hamiltonian H is a proper function.

Consider the underactuated control system with torques

$$\begin{aligned}
\dot{x}_1 &= \frac{x_2x_3}{C} - \frac{x_2x_3}{B+mx_4^2} + u_1,\\
\dot{x}_2 &= \frac{x_1x_3}{A+mx_4^2} - \frac{x_1x_3}{C} + u_2,\\
\dot{x}_3 &= \frac{x_1x_2}{B+mx_4^2} - \frac{x_1x_2}{A+mx_4^2} + u_3,\\
\dot{x}_4 &= x_5,\\
\dot{x}_5 &= \frac{x_1^2x_4}{(A+mx_4^2)^2} + \frac{x_2^2x_4}{(B+mx_4^2)^2} - \frac{\sigma x_4}{m} + u_4,
\end{aligned}$$

where the control $u := (u_1, u_2, u_3, u_4) : (0, \infty) \to B \subset \mathbb{R}^4$ is a measurable function with values in a bounded subset $B \subset \mathbb{R}^4$. The vector fields $\frac{\partial}{\partial x_1}$, $\frac{\partial}{\partial x_2}$, $\frac{\partial}{\partial x_3}$, $\frac{\partial}{\partial x_5}$, $[X_H, \frac{\partial}{\partial x_5}]$ verify the LARC and, as a result of Theorem 2.10(ii), we obtain that the above system is controllable.

Example 2. In this example we follow the presentation of the geometric structure in Adams and Ratiu [3]. The motion of three point vortices for an ideal inviscid incompressible fluid in the plane is given by the equations

$$\begin{cases} \dot{x}_j = -\frac{1}{2\pi} \sum\limits_{\substack{i=1 \\ i \neq j}}^{3} \Gamma_i (y_j - y_i)/r_{ij}^2, \\ \dot{y}_j = -\frac{1}{2\pi} \sum\limits_{\substack{i=1 \\ i \neq j}}^{3} \Gamma_i (x_j - x_i)/r_{ij}^2, \end{cases} \tag{4.1}$$

$j = 1, 2, 3$, where $r_{ij}^2 = (x_i - x_j)^2 + (y_i - y_j)^2$ and Γ_1, Γ_2, and Γ_3 are nonzero constants, the circulations given by the corresponding point vortices. These equations are defined on $\mathbb{R}^6$ after eliminating all the diagonals $\{(x_i, y_i) = (x_j, y_j)\}$ for $i \neq j$. Kirchhoff [21] noted that (4.1) can be written in the form

$$\begin{aligned} \Gamma_j \frac{dx_j}{dt} &= \frac{\partial H}{\partial y_j}, \\ \Gamma_j \frac{dy_j}{dt} &= -\frac{\partial H}{\partial x_j}, \end{aligned}$$

where

$$H(x_1, x_2, x_3, y_1, y_2, y_3) = -\frac{1}{4\pi} \sum_{\substack{i=1 \\ i \neq j}}^{3} \Gamma_i \Gamma_j \log r_{ij}$$

is the Hamiltonian and the symplectic form is given by

$$\Omega = \sum_{i=1}^{3} \Gamma_i dx_i \wedge dy_i. \tag{4.2}$$

In what follows it is convenient to identify $\mathbb{R}^2$ with $\mathbb{C}$ by the map $(x, y) \mapsto x + \sqrt{-1}y$. The special Euclidean group $SE(2) := \{(e^{\sqrt{-1}\theta}, w) \mid \theta \in \mathbb{R}, w \in \mathbb{C}\}$ acts on $\mathbb{C}$ by $(e^{\sqrt{-1}\theta}, w) \cdot z := e^{\sqrt{-1}\theta} z + w$. This action is not free. The diagonal action of $SE(2)$ on $\mathbb{C}^3$ is free on the open invariant subset $\mathbb{C}^3 \setminus \{(z, z, z) \mid z \in \mathbb{C}\}$ that contains the open invariant subset $S := \mathbb{C}^3 \setminus \{(z_1, z_2, z_3) \mid z_i \neq z_j \text{ for } i \neq j\}$ on which the three point vortex problem is defined. It can be easily verified that this action is proper on S. This action has an associated nonequivariant momentum map $J : \mathbb{R}^6 \equiv \mathbb{C}^3 \to \mathbb{R}^3$ relative to the symplectic form (4.2) given by

$$J(\mathbf{x}, \mathbf{y}) = \left(-\frac{1}{2} \sum_{i=1}^{3} \Gamma_i (x_i^2 + y_i^2), \sum_{i=1}^{3} \Gamma_i y_i, -\sum_{i=1}^{3} \Gamma_i x_i \right).$$

If the vortex strengths Γ_1, Γ_2, and Γ_3 have the same signs, then, by applying Lemmas 4.2 and 4.1, it follows that J is a proper map and we are in the hypotheses of Theorem 3.4(i).

In Adams and Ratiu [3] it is shown that the quotient $S/SE(2)$ is diffeomorphic to $T := \mathbb{R}^3\backslash(\{(0,0,c) \mid c \in \mathbb{R}\} \cup \{(a,0,0) \mid a \geq 0\})$, that the push forward of the quotient Poisson bracket on $S/SE(2)$ to T has the matrix

$$4\begin{bmatrix} 0 & 2a_3 & -2a_2 \\ -2a_3 & 0 & 2a_1 - \|\mathbf{a}\| \\ 2a_2 & -2a_1 + \|\mathbf{a}\| & 0 \end{bmatrix},$$

and that the reduced Hamiltonian is

$$\widetilde{H}(a_1,a_2,a_3) = -\frac{1}{4\pi}(\Gamma_1\Gamma_2 \log((a_3 + \|\mathbf{a}\|)/2) + \Gamma_1\Gamma_3 \log((-a_3 + \|\mathbf{a}\|)/2) + \Gamma_2\Gamma_3 \log(-a_1 + \|\mathbf{a}\|)).$$

Therefore, the reduced equations are

$$\begin{aligned}
\dot{a}_1 &= \frac{2}{\pi}\left(\Gamma_1\Gamma_2\frac{a_2}{(a_3+\|\mathbf{a}\|)} - \Gamma_1\Gamma_3\frac{a_2}{(-a_3+\|\mathbf{a}\|)}\right),\\
\dot{a}_2 &= \frac{1}{\pi}\left(\Gamma_1\Gamma_2\frac{(-2a_1+a_3+\|\mathbf{a}\|)}{(a_3+\|\mathbf{a}\|)} + \Gamma_1\Gamma_3\frac{(2a_1+a_3-\|\mathbf{a}\|)}{(-a_3+\|\mathbf{a}\|)} + \Gamma_2\Gamma_3\frac{a_3}{(-a_1+\|\mathbf{a}\|)}\right),\\
\dot{a}_3 &= \frac{1}{\pi}\left(\Gamma_2\Gamma_3\frac{a_2}{(-a_1+\|\mathbf{a}\|)} - \Gamma_1\Gamma_2\frac{a_2}{(a_3+\|\mathbf{a}\|)} - \Gamma_1\Gamma_3\frac{a_2}{(-a_3+\|\mathbf{a}\|)}\right).
\end{aligned}$$

Consider the reduced control system

$$\begin{aligned}
\dot{a}_1 &= \frac{2}{\pi}\left(\Gamma_1\Gamma_2\frac{a_2}{(a_3+\|\mathbf{a}\|)} - \Gamma_1\Gamma_3\frac{a_2}{(-a_3+\|\mathbf{a}\|)}\right) + u_1,\\
\dot{a}_2 &= \frac{1}{\pi}\left(\Gamma_1\Gamma_2\frac{(-2a_1+a_3+\|\mathbf{a}\|)}{(a_3+\|\mathbf{a}\|)} + \Gamma_1\Gamma_3\frac{(2a_1+a_3-\|\mathbf{a}\|)}{(-a_3+\|\mathbf{a}\|)} + \Gamma_2\Gamma_3\frac{a_3}{(-a_1+\|\mathbf{a}\|)}\right) + u_2,\\
\dot{a}_3 &= \frac{1}{\pi}\left(\Gamma_2\Gamma_3\frac{a_2}{(-a_1+\|\mathbf{a}\|)} - \Gamma_1\Gamma_2\frac{a_2}{(a_3+\|\mathbf{a}\|)} - \Gamma_1\Gamma_3\frac{a_2}{(-a_3+\|\mathbf{a}\|)}\right) + u_3,
\end{aligned}$$

where the control $u := (u_1,u_2,u_3) : (0,\infty) \to B \subset \mathbb{R}^3$ is a measurable function with values in a bounded subset $B \subset \mathbb{R}^3$. It is easy to check that the vector fields $X_{\widetilde{H}}, \frac{\partial}{\partial a_1}, \frac{\partial}{\partial a_2}, \frac{\partial}{\partial a_3}$ verify the LARC and, by Theorem 3.4(i), we conclude that this reduced system is controllable.

Example 3. The next example, whose geometric study can be found in Blaom [4], is the resonant three-wave interaction. This is a Hamiltonian system whose phase space is $\mathbb{R}^6 = \mathbb{C}^3$, equipped with the symplectic structure

$$\omega = \sum_{j=1}^{3} \frac{1}{s_j\gamma_j} dq_j \wedge dp_j,$$

where $s_1, s_2, s_3 \in \{-1,1\}$ and $\gamma_1, \gamma_2, \gamma_3 \in \mathbb{R}$ are parameters subject to the constraint $\gamma_1+\gamma_2+\gamma_3 = 0$. We will restrict our attention to the particular case when $(s_1,s_2,s_3) = (1,1,1)$ and $(\gamma_1,\gamma_2,\gamma_3) = (1,1,-2)$.

In standard coordinates on $\mathbb{C}^3$, $z_j := q_j + \sqrt{-1}p_j$, $j = 1,2,3$, the Hamiltonian is given by

$$H(z_1,z_2,z_3) = -\frac{1}{2}(\bar{z}_1 z_2 \bar{z}_3 + z_1 \bar{z}_2 z_3).$$

This Hamiltonian is invariant under the action of the compact Lie group $G \equiv S^1 \times S^1$ on P given by

$$(e^{i\theta_1}, e^{i\theta_2}) \cdot (z_1, z_2, z_3) = (e^{-i\theta_1} z_1, e^{-i(\theta_1+\theta_2)} z_2, e^{-i\theta_2} z_3), \qquad 0 \leq \theta_j < 2\pi.$$

The momentum map for this action is $J : P \to \mathfrak{g}^* \cong \mathbb{R}^2$,

$$J(z_1, z_2, z_3) = \left(\frac{1}{2} \left(|z_1|^2 + |z_2|^2 \right), \frac{1}{2} \left(|z_1|^2 - \frac{1}{2} |z_3|^2 \right) \right).$$

This action is free on the open invariant subset $\left[(\mathbb{C}\backslash\{0\}) \times \mathbb{C} \times (\mathbb{C}\backslash\{0\}) \right] \cup \left[\{0\} \times (\mathbb{C}\backslash\{0\}) \times (\mathbb{C}\backslash\{0\}) \right] \cup \left[(\mathbb{C}\backslash\{0\}) \times (\mathbb{C}\backslash\{0\}) \times \{0\} \right]$. As in Blaom [4], we shall restrict the study of the resonant three-wave interaction to $S := (\mathbb{C}\backslash\{0\}) \times \mathbb{C} \times (\mathbb{C}\backslash\{0\})$, where the action is free. The smooth map $(z_1, z_2, z_3) \in S \mapsto (z_2 \bar{z}_1 \bar{z}_3 / |z_1 z_3|, |z_1|, |z_3|) \in \mathbb{R}^2 \times (0, \infty)^2$ induces a diffeomorphism $S/G \approx \mathbb{R}^2 \times (0, \infty)^2$. The push forward by this diffeomorphism of the quotient Poisson bracket on S/G to $\{(q, p, a, b) \in \mathbb{R}^4 \mid q, p \in \mathbb{R}, a > 0, b > 0\} = \mathbb{R}^2 \times (0, \infty)^2$ has the expression

$$\begin{bmatrix} 0 & 1 & -\frac{p}{a} & 2\frac{p}{b} \\ -1 & 0 & \frac{q}{a} & -2\frac{q}{b} \\ \frac{p}{a} & -\frac{q}{a} & 0 & 0 \\ -2\frac{p}{b} & 2\frac{q}{b} & 0 & 0 \end{bmatrix}$$

and the reduced Hamiltonian is

$$\widetilde{H}(q, p, a, b) = -abq.$$

The reduced equations of motion are

$$\begin{cases} \dot{q} = \dfrac{qpb}{a} - 2\dfrac{qpa}{b}, \\ \dot{p} = ab - \dfrac{q^2 b}{a} + 2\dfrac{q^2 a}{b}, \\ \dot{a} = -pb, \\ \dot{b} = 2pa. \end{cases} \tag{4.3}$$

A constant of motion for the system (4.3) is given by the function $f : Q \to \mathbb{R}$, $f(q, p, a, b) = q^2 + p^2 + a^2 + b^2$, which is proper by Lemma 4.1.

Consider now the underactuated reduced control system

$$\begin{cases} \dot{q} = \dfrac{qpb}{a} - 2\dfrac{qpa}{b} + u_1, \\ \dot{p} = ab - \dfrac{q^2 b}{a} + 2\dfrac{q^2 a}{b} + u_2, \\ \dot{a} = -pb, \\ \dot{b} = 2pa + u_3, \end{cases} \tag{4.4}$$

where the control $u := (u_1, u_2, u_3) : (0, \infty) \to B \subset \mathbb{R}^3$ is a measurable function with values in a bounded subset B. A short computation shows that the vector

fields $\{\frac{\partial}{\partial q}, \frac{\partial}{\partial p}, \frac{\partial}{\partial b}, [\frac{\partial}{\partial b}, [\frac{\partial}{\partial p}, X_{\widetilde{H}}]]\}$ generate at every point $(q,p,a,b) \in \mathbb{R}^2 \times (0,\infty)^2$ the tangent space $T_{(q,p,a,b)}(\mathbb{R}^2 \times (0,\infty)^2)$, which proves that the system (4.4) verifies the LARC. By Theorem 3.4(ii), the system (4.4) is controllable.

Example 4. We will study the controllability of the reduced system of two coupled planar rigid bodies. We take the description of the system given in Sreenath, Oh, Krishnaprasad, and Marsden [44]. After the reduction to the center of mass frame we have the configuration space $S^1 \times S^1$ with the diagonal action of S^1. The phase space is $T^*(S^1 \times S^1)$ with the canonical symplectic form of a cotangent bundle. The momentum map for the lifted action of S^1 is given by

$$J((\theta_1, \mu_1), (\theta_2, \mu_2)) = \mu_1 + \mu_2.$$

Krishnaprasad and Marsden [25] have shown that the reduced Poisson space is

$$P := T^*(S^1 \times S^1)/S^1 \cong S^1 \times \mathbb{R}^2$$

and, if we chose coordinates (θ, μ_1, μ_2) on P, the matrix of the Poisson bracket is given by

$$\begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}.$$

The reduced Hamiltonian is given by the formula

$$H = \frac{1}{2\triangle}(\widetilde{I}_2\mu_1^2 - 2\varepsilon\lambda(\theta)\mu_1\mu_2 + \widetilde{I}_1\mu_2^2),$$

where $\triangle = \widetilde{I}_1\widetilde{I}_2 - \varepsilon^2(\lambda(\theta))^2 > 0$ and

d_i	is the distance from the hinge to the center of mass of body $i = 1, 2$,
θ	is the joint angle from body 1 to body 2,
$\lambda(\theta)$	equals $d_1 d_2 \cos\theta$,
m_i	is the mass of body $i = 1, 2$,
ε	equals $m_1 m_2/(m_1 + m_2)$ (the reduced mass),
I_i	is the moment of inertia of body i about its center of mass, and
$\widetilde{I}_i$	equals $I_i + \varepsilon d_i^2$, $i = 1, 2$ (the augmented moments of inertia).

To apply Theorem 3.4 we need to show that H is a proper function. To do this, we need the following lemma.

LEMMA 4.3. *Let $f : K \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be two continuous functions, where K is compact and g is a proper function. Then the function $h : K \times \mathbb{R}^n \to \mathbb{R}$ given by $h(x,y) := f(x)g(y)$ is a proper function.*

Proof. We shall prove that $h^{-1}([a,b])$ is compact in $K \times \mathbb{R}^n$. Let $z_n := (x_n, y_n)$ be an arbitrary sequence in $h^{-1}([a,b])$. Since K is compact, we can assume that $\{x_n\}_{n\in\mathbb{N}}$ is convergent. Because $\{f(x_n)g(y_n)\}_{n\in\mathbb{N}} \subset [a,b]$ and $\{f(x_n)\}_{n\in\mathbb{N}}$ is bounded, the sequence $\{g(y_n)\}_{n\in\mathbb{N}}$ is also bounded and hence there are $a', b' \in \mathbb{R}$ such that $\{g(y_n)\}_{n\in\mathbb{N}} \subset [a', b']$. Therefore, $\{y_n\}_{n\in\mathbb{N}} \subset g^{-1}([a', b'])$, which is a compact set in $\mathbb{R}^n$ because g is a proper function. Consequently, there is a convergent subsequence of $\{y_n\}_{n\in\mathbb{N}}$. The corresponding subsequence of $\{z_n\}_{n\in\mathbb{N}}$ is convergent, which proves that $h^{-1}([a,b])$ is compact. ∎

To apply this lemma we write H in the form

$$H = \frac{1}{2\triangle}\left(\left(\sqrt{\widetilde{I}_2}\,\mu_1 - \frac{\varepsilon\lambda(\theta)}{\sqrt{\widetilde{I}_2}}\mu_2\right)^2 + \left(\widetilde{I}_1 - \frac{\varepsilon^2\lambda^2(\theta)}{\widetilde{I}_2}\right)\mu_2^2\right).$$

Since

$$\widetilde{I}_1 - \frac{\varepsilon^2\lambda^2(\theta)}{\widetilde{I}_2} > 0,$$

the smooth change of variables $(\theta, \mu_1, \mu_2) \mapsto (\theta, X, Y)$, where

$$X := \sqrt{\widetilde{I}_2}\,\mu_1 - \frac{\varepsilon\lambda(\theta)}{\sqrt{\widetilde{I}_2}}\mu_2,$$

$$Y := \left(\widetilde{I}_1 - \frac{\varepsilon^2\lambda^2(\theta)}{\widetilde{I}_2}\right)^{1/2}\mu_2$$

transforms H to the function $\frac{1}{2\triangle}\left(X^2 + Y^2\right)$ with $\frac{1}{2\triangle}$ defined on S^1. This function is proper by Lemmas 4.1 and 4.3. Thus H is a proper integral of motion for the reduced system.

Now we consider the following underactuated reduced control system with torques u_1, u_2

$$\dot{\theta} = -\frac{\partial H}{\partial \mu_1} + \frac{\partial H}{\partial \mu_2},$$

$$\dot{\mu}_1 = \frac{\partial H}{\partial \theta} + u_1,$$

$$\dot{\mu}_2 = -\frac{\partial H}{\partial \theta} + u_2,$$

where the control $u := (u_1, u_2) : (0, \infty) \to B \subset \mathbb{R}^2$ is a measurable function with values in a bounded subset B. It is easy to see that the vector fields $[X_H, \frac{\partial}{\partial \mu_1}]$, $[X_H, \frac{\partial}{\partial \mu_2}]$, $\frac{\partial}{\partial \mu_1}$, $\frac{\partial}{\partial \mu_2}$ verify the LARC and, as a consequence of Theorem 3.4(ii), we obtain that the reduced control system above is controllable.

Using Corollary 3.5 one can study the controllability of the unreduced system of the two coupled planar rigid bodies by considering any system of controls that satisfies the LARC and reduces to a system of controls that also satisfies the LARC, for example, the one above.

Acknowledgments. We would like to thank V. Timofte and I. Casu for carefully reading the manuscript and for offering valuable comments which helped us to improve the exposition. We are also grateful to the referees both for their comments that have inspired the extension of our original results to arbitrary Poisson manifolds and for their editorial suggestions.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed., Addison–Wesley, Reading, MA, 1979.

[2] R. Abraham, J. E. Marsden, and T. S. Ratiu, *Manifolds, Tensor Analysis, and Applications*, Appl. Math. Sci. 75, Springer-Verlag, New York, 1988.

[3] M. Adams and T. S. Ratiu, *The three point vortex problem: Commutative and non-commutative integrability*, in Hamiltonian Dynamical Systems, Contemp. Math. 81, K. Meyer and D. Saari, eds., AMS, Providence, RI, 1988, pp. 245–257.

[4] A. D. Blaom, *Reconstruction phases via Poisson reduction*, Differential Geom. Appl., 12 (2000), pp. 231–252.

[5] A. M. Bloch, N. E. Leonard, and J. E. Marsden, *Controlled Lagrangians and the stabilization of Euler–Poincaré mechanical systems*, Internat. J. Robust Nonlinear Control, 11 (2001), pp. 191–214.

[6] B. Bonnard, *Controlabilité des systèms nonlinéaires*, C.R. Acad. Sci. Paris Sér. I Math., 292 (1981), pp. 535–537.

[7] G. E. Bredon, *Introduction to Compact Transformation Groups*, Academic Press, New York, 1972.

[8] R. W. Brockett, *System theory on group manifolds and coset spaces*, SIAM J. Control, 10 (1972), pp. 265–284.

[9] W. L. Chow, *Über systeme von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann., 117 (1939), pp. 98–105.

[10] O. Christov, *On the non-integrability of a system describing the motion of a rigid body with a fixed point and a particle oscillating in it*, Bull. Sci. Math., 118 (1994), pp. 385–401.

[11] P. E. Crouch, *Spacecraft attitude control and stabilization: Applications of geometric control theory to rigid body models*, IEEE Trans. Automat. Control, 29 (1984), pp. 321–331.

[12] J. W. Grizzle and S. I. Marcus, *The structure of nonlinear control systems possessing symmetries*, IEEE Trans. Automat. Control, 30 (1985), pp. 248–257.

[13] G. W. Haynes and H. Hermes, *Nonlinear controllability via Lie theory*, SIAM J. Control, 8 (1970), pp. 450–460.

[14] R. Hermann, *On the accessibility problem in control theory*, in Proceedings of the International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325–332.

[15] S. M. Jalnapurkar and J. E. Marsden, *Stabilization of relative equilibria*, IEEE Trans. Automat. Control, 45 (2000), pp. 1483–1491.

[16] S. M. Jalnapurkar and J. E. Marsden, *Reduction of Hamilton's variational principle*, Dyn. Stab. Syst., 15 (2000), pp. 287–318.

[17] V. Jurdjevic and I. Kupka, *Control systems on semi-simple Lie groups and their homogeneous space*, Ann. Inst. Fourier (Grenoble), 31 (1981), pp. 151–179.

[18] V. Jurdjevic and J. P. Quinn, *Controllability and stability*, J. Differential Equations, 28 (1978), pp. 381–389.

[19] K. Kawakubo, *The Theory of Transformation Groups*, Oxford University Press, Oxford, UK, 1991.

[20] D. Kazhdan, B. Kostant, and S. Sternberg, *Hamiltonian group actions and dynamical systems of Calogero type*, Comm. Pure Appl. Math., 31 (1978), pp. 481–508.

[21] G. Kirchhoff, *Vorlesungen über Mathematische Physik*, *Vol.* 1, Kapitel 20, Teubner, Leipzig, 1883.

[22] A. A. Kirillov, *Elements of the Theory of Representations*, Grundlehren Math. Wiss. 220, Springer-Verlag, New York, 1976.

[23] A. J. Krener, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control systems*, SIAM J. Control, 12 (1974), pp. 43–52.

[24] P. S. Krishnaprasad and V. Manikonda, *Control problems of the hydrodynamics type*, in Proceedings of the 4th IFAC Nonlinear Control Systems Design Symposium (NOLCOS), University of Twente, Enschede, The Netherlands, 1998, pp. 139–144.

[25] P. S. Krishnaprasad and J. E. Marsden, *Hamiltonian structures and stability for rigid bodies with flexible attachments*, Arch. Ration. Mech. Anal., 98 (1987), pp. 71–93.

[26] K.-Y. Lian, L.-S. Wang, and L.-C. Fu, *Controllability of spacecraft systems in a central gravitational field*, IEEE Trans. Automat. Control, 39 (1994), pp. 2426–2441.

[27] P. Libermann and C.-M. Marle, *Symplectic Geometry and Analytical Mechanics*, D. Reidel, Boston, 1987.

[28] C. Lobry, *Contrôlabilité des systèmes non linéaires*, SIAM J. Control, 8 (1970), pp. 573–605.

[29] C. Lobry, *Controllability of nonlinear systems on compact manifolds*, SIAM J. Control, 12 (1974), pp. 1–4.

[30] V. Manikonda and P. S. Krishnaprasad, *Controllability of Lie-Poisson reduced dynamics*, in Proceedings of the American Control Conference, Albuquerque, NM, 1997, pp. 2203–2207.
[31] V. Manikonda and P. S. Krishnaprasad, *Controllability of a class of underactuated mechanical systems with symmetry*, Automatica J. IFAC, 38 (2002) pp. 1837–1850.
[32] C.-M. Marle, *Symplectic manifolds, dynamical groups, and Hamiltonian mechanics*, in Differential Geometry and Relativity, M. Cahen and M. Flato, eds., D. Reidel, Boston, 1976, pp. 249–269.
[33] J. E. Marsden, *Lectures on Geometric Methods in Mathematical Physics*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 37, SIAM, Philadelphia, 1981.
[34] J. E. Marsden, *Lectures on Mechanics*, London Math. Soc. Lecture Note Ser. 174, Cambridge University Press, Cambridge, UK, 1992.
[35] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, 2nd ed., Texts Appl. Math. 17, Springer-Verlag, Berlin, 2003.
[36] J. E. Marsden and A. Weinstein, *Reduction of symplectic manifolds with symmetry*, Rep. Math. Phys., 5 (1974), pp. 121–130.
[37] V. V. Nemytskii and V. V. Stepanov, *Qualitative Theory of Differential Equations*, Princeton Math. Ser. 22, Princeton University Press, Princeton, NJ, 1960.
[38] H. Nijmeijer and A. J. van der Schaft, *Controlled invariance for nonlinear systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 904–914.
[39] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, Berlin, 1990.
[40] J.-P. Ortega and T. S. Ratiu, *Momentum Maps and Hamiltonian Reduction*, Progr. Math. 222, Birkhäuser Boston, Boston, 2003.
[41] M. Puta, *Hamiltonian Mechanical Systems and Geometric Quantization*, Math. Appl. 260, Kluwer, Dordrecht, The Netherlands, 1993.
[42] L. San Martin and P. E. Crouch, *Controllability on principal fibre bundles with compact structure group*, Systems Control Lett., 5 (1984), pp. 35–40.
[43] G. Sánchez De Alvarez, *Controllability of Poisson control systems with symmetry*, Contemp. Math., 97 (1989), pp. 399–412.
[44] N. Sreenath, Y. G. Oh, P. S. Krishnaprasad, and J. E. Marsden, *The dynamics of coupled rigid bodies. Part* I: *Reduction, equilibria and stability*, Dyn. Stab. Syst., 3 (1988), pp. 25–49.
[45] H. J. Sussmann and V. J. Jurdjevic, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 955–969

A LOCAL RESULT ON INSENSITIZING CONTROLS FOR A SEMILINEAR HEAT EQUATION WITH NONLINEAR BOUNDARY FOURIER CONDITIONS*

OLIVIER BODART†, MANUEL GONZÁLEZ-BURGOS‡, AND ROSARIO PÉREZ-GARCÍA‡

Abstract. In this paper we present a local result on the existence of insensitizing controls for a semilinear heat equation when nonlinear boundary conditions of the form $\partial_n y + f(y) = 0$ are considered. The problem leads to an analysis of a special type of nonlinear null controllability problem. A sharp study of the linear case and a later application of an appropriate fixed point argument constitute the scheme of the proof of the main result. The boundary conditions we are dealing with lead us to seek a fixed point, and thus also control functions, in certain Hölder spaces. The main strategy in this paper is the construction of controls with Hölderian regularity starting from L^2-controls in the linear case. Sufficient regularity in the data and appropriate assumptions on the right-hand side term ξ of the equation are required.

Key words. controllability, nonlinear PDE of parabolic type

AMS subject classifications. 93B05, 35K55, 35K05

DOI. 10.1137/S036301290343161X

1. Statement of the problem and main result. Let $\Omega \subset \mathbb{R}^N$ be, with $N \geq 1$, a bounded domain with boundary $\partial\Omega$ of at least class C^2. Let ω and $\mathcal{O}$ be nonempty open subsets of Ω. For $T > 0$, we denote by Q the cylinder $\Omega \times (0,T)$ and by Σ its lateral boundary $\partial\Omega \times (0,T)$. We consider a semilinear heat equation with nonlinear boundary conditions of Fourier type and partially known initial data:

$$\begin{cases} \partial_t y - \Delta y + F(y) = \xi + v\mathbf{1}_\omega & \text{in } Q, \\ \partial_n y + f(y) = 0 & \text{on } \Sigma, \\ y(x,0) = y_0(x) + \tau\, \hat{y}_0(x) & \text{in } \Omega, \end{cases} \tag{1.1}$$

where F and f are given C^1 functions defined on $\mathbb{R}$; ξ and y_0 are, respectively, a known heat source and a given initial datum, both regular enough; τ is an unknown small real number; and $\hat{y}_0$ is unknown in an appropriate Banach space $X \hookrightarrow L^2(\Omega)$ (the embedding being continuous and dense), with $\|\hat{y}_0\|_X = 1$. Here, $v = v(x,t)$ is a control function to be determined, $\mathbf{1}_\omega$ is the characteristic function of the set ω, ∂_t denotes the time derivative, and ∂_n represents the derivation with respect to the outward unit normal to $\partial\Omega$.

Let us define

$$\Phi(y) = \frac{1}{2} \iint_{\mathcal{O}\times(0,T)} |y(x,t;\tau,v)|^2 \, dx\, dt, \tag{1.2}$$

$y = y(\cdot,\cdot;\tau,v)$ being a solution of (1.1) (associated to τ and v) defined in $(0,T)$, if one exists. In this paper we analyze the existence of control functions that make

*Received by the editors July 15, 2003; accepted for publication (in revised form) February 16, 2004; published electronically October 8, 2004. This work has been partially financed by D.G.E.S. (Spain), grant PB98–1134.

http://www.siam.org/journals/sicon/43-3/43161.html

†Laboratoire de Mathématiques Appliquées, UMR CNRS 6620, Université Blaise-Pascal, Clermont-Ferrand 2, 63177 Aubière, France (Olivier.Bodart@math.univ-bpclermont.fr).

‡Dpto. Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, Aptdo. 1160, 41080 Sevilla, Spain (manoloburgos@us.es, rosariopg@us.es).

the functional Φ locally insensitive to small perturbations in the initial condition. A possible physical interpretation of this problem would be the following. The function $y = y(x,t)$ can be viewed as the relative temperature of a body (with respect to the exterior surrounding air). The semilinear parabolic equation in (1.1) means that there is a fixed heat source ξ acting on the body and that we can also act on a small part ω of the body by means of a heat source $v\mathbf{1}_\omega$. On the boundary, $-\frac{\partial y}{\partial n}$ can be viewed as the *normal heat flux*, directed inward, up to a positive coefficient. Thus, the equality

$$-\frac{\partial y}{\partial n} = f(y)$$

means that this flux is a (nonlinear) function of the temperature. The problem with insensitizing Φ means that we are seeking a control function acting on ω such that the energy in $\mathcal{O}$ is invariant for small perturbations in the initial data. A natural physical hypothesis would be to suppose that f is nondecreasing with $f(0) = 0$. Throughout this paper, we will assume no special behavior on the increasing of f.

By reasons that will be seen later, in this work we will slightly change the usual notion of insensitizing controls (see [1], [4], [10], [11]), which is equivalent to the usual one in the linear case.

DEFINITION 1.1. *A control function v is said to insensitize Φ if there exists $\tau_0 > 0$ such that system* (1.1) *admits a weak solution $y(\cdot,\cdot;\tau,v) \in L^2(0,T;H^1(\Omega)) \cap C^0(\overline{Q})$ for $|\tau| \le \tau_0$ and if the following insensitivity condition holds:*

$$\left.\frac{\partial \Phi(y(\cdot,\cdot;\tau,v))}{\partial \tau}\right|_{\tau=0} = 0 \quad \forall \hat{y}_0 \in X \text{ with } \|\hat{y}_0\|_X = 1, \tag{1.3}$$

where $X = C^{2+\beta}(\overline{\Omega}) \cap H_0^2(\Omega)$.

By a weak solution of (1.1) (associated to τ and v) we will define a function (if one exists) $y = y(\cdot,\cdot;\tau,v) \in L^2(0,T;H^1(\Omega)) \cap C^0(\overline{Q})$, with $\partial_t y \in L^2(0,T;H^{-1}(\Omega))$, such that

$$\begin{cases} \langle \partial_t y(t), u\rangle_{(H^1(\Omega))',H^1(\Omega)} + \displaystyle\int_\Omega \nabla y(t)\cdot\nabla u\,dx + \int_\Omega F(y(t))u\,dx + \int_{\partial\Omega} f(y(t))u\,d\sigma \\ \qquad = \displaystyle\int_\Omega (\xi(t) + v(t)\mathbf{1}_\omega)u\,dx \quad \text{in } L^2(0,T) \quad \forall u \in H^1(\Omega), \\ y(0) = y_0 + \tau\,\hat{y}_0. \end{cases}$$

Insensitivity problems were originally introduced by J.-L. Lions in [10] and were first studied for semilinear heat equations with globally Lipschitz-continuous nonlinearities $F = F(y)$ and Dirichlet boundary conditions. In [1], the existence of the so-called ε-insensitizing controls for partially known data in both the initial and boundary conditions is proved. In [11] it is shown that one cannot expect the existence of insensitizing controls for every $y_0 \in L^2(\Omega)$ when $\Omega \setminus \overline{\omega} \neq \emptyset$, even if $F \equiv 0$. In addition, for $y_0 = 0$ and suitable assumptions on the source term ξ, de Teresa proves the existence of insensitizing controls (see Theorem 1 in [11]). This last result is extended in [2] and [3] to nonlinearities with certain superlinear growth at infinity. It is also generalized in [4] to the case of a heat equation with a nonlinear term involving the state y and its gradient. In [4], the authors also present an insensitivity result for a semilinear heat equation with a nonlinear term $F(y)$ and linear boundary conditions of Fourier type. In the present paper, we prove a local result on the existence of insensitizing controls for system (1.1), which is, to our knowledge, the first insensitivity

result in the literature for a semilinear heat equation with nonlinear Fourier boundary conditions. In the framework of the controllability, both approximate and null controllability of the classical heat equation with nonlinear Fourier boundary conditions are analyzed in [5].

Before stating the main result in this paper, let us introduce the following notation. For $p \in [1,\infty]$ and any Banach space Y, $\|\cdot\|_{L^p(Y)}$ will denote the norm in the space $L^p(0,T;Y)$. For simplicity, the norm in $L^p(Q)$ will be represented by $\|\cdot\|_{L^p}$ for $p \in [1,\infty)$, $\|\cdot\|_\infty$ will stand for the norm in $L^\infty(Q)$, and $\|\cdot\|_{\infty;\Sigma}$ will denote the norm in $L^\infty(\Sigma)$. For $r \in (2,\infty)$ and any open set $\mathcal{V} \subset \mathbb{R}^N$, we introduce the Banach space

$$X^r(0,T;\mathcal{V}) = \left\{u \in L^r(0,T;W^{2,r}(\mathcal{V})) : \partial_t u \in L^r(0,T;L^r(\mathcal{V}))\right\},$$

with its natural norm

$$\|u\|_{X^r(0,T;\mathcal{V})} = \|u\|_{L^r(W^{2,r}(\mathcal{V}))} + \|\partial_t u\|_{L^r(L^r(\mathcal{V}))}.$$

On the other hand, for $\beta \in (0,1)$ and $u \in C^0(\overline{Q})$, we define the quantity

$$[u]_{\beta,\frac{\beta}{2}} = \sup_{\overline{Q}} \frac{|u(x,t)-u(x',t)|}{|x-x'|^\beta} + \sup_{\overline{Q}} \frac{|u(x,t)-u(x,t')|}{|t-t'|^{\frac{\beta}{2}}}.$$

We will consider the space $C^{\beta,\frac{\beta}{2}}(\overline{Q}) = \{u \in C^0(\overline{Q}) : [u]_{\beta,\frac{\beta}{2}} < \infty\}$, which is a Banach space with its natural norm $|u|_{\beta,\frac{\beta}{2};\overline{Q}} = \|u\|_\infty + [u]_{\beta,\frac{\beta}{2}}$. We will also consider the Banach spaces defined by

$$C^{1+\beta,\frac{1+\beta}{2}}(\overline{Q}) = \left\{u \in C^0(\overline{Q}) : \frac{\partial u}{\partial x_i} \in C^{\beta,\frac{\beta}{2}}(\overline{Q}) \quad \forall i,\ \sup_{\overline{Q}} \frac{|u(x,t)-u(x,t')|}{|t-t'|^{\frac{1+\beta}{2}}} < \infty\right\},$$

$$C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q}) = \left\{u \in C^0(\overline{Q}) : \frac{\partial u}{\partial x_i} \in C^{1+\beta,\frac{1+\beta}{2}}(\overline{Q}) \quad \forall i,\ \partial_t u \in C^{\beta,\frac{\beta}{2}}(\overline{Q})\right\},$$

and

$$C^{3+\beta,\frac{3+\beta}{2}}(\overline{Q}) = \left\{u \in C^0(\overline{Q}) : \frac{\partial u}{\partial x_i} \in C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q}) \quad \forall i,\ \partial_t u \in C^{1+\beta,\frac{1+\beta}{2}}(\overline{Q})\right\},$$

with norms denoted by $|\cdot|_{n+\beta,\frac{n+\beta}{2};\overline{Q}}$, $n = 1,2,3$. The Banach space formed by the restrictions to $\overline{\Sigma}$ of the functions in $C^{n+\beta,\frac{n+\beta}{2}}(\overline{Q})$ will be represented by $C^{n+\beta,\frac{n+\beta}{2}}(\overline{\Sigma})$ and its norm by $|\cdot|_{n+\beta,\frac{n+\beta}{2};\overline{\Sigma}}$. Finally, we shall write $|\cdot|_{2+\beta;\overline{\Omega}}$ to denote the norm in $C^{2+\beta}(\overline{\Omega})$, and the norm in the space $L^2(0,T;H^1(\Omega))\cap C([0,T];L^2(\Omega))$ will be denoted by $\|\cdot\|_{L^2(H^1)\cap C(L^2)}$.

The main goal in this paper is to prove the following local insensitivity result for system (1.1).

THEOREM 1.2. *Assume that $\partial\Omega \in C^{3+\beta}$ for some $\beta \in (0,1)$, $\omega \cap \mathcal{O} \neq \emptyset$, and $y_0 = 0$. Let $F, f \in C^3(\mathbb{R})$ verify $F(0) = f(0) = 0$. Then, there exist two positive constants $\mathcal{M}$ and η (depending on Ω, ω, $\mathcal{O}$, T, F, and f) such that, for any $\xi \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ satisfying*

$$|\xi|_{\beta,\frac{\beta}{2};\overline{Q}} + \left\|\exp\left(\frac{\mathcal{M}}{2t}\right)\xi\right\|_{L^2} \leq \eta, \tag{1.4}$$

one can find a control function $v \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ *that insensitizes the functional* Φ *defined by* (1.2).

It is of interest to notice that the explicit way the constant $\mathcal{M}$ depends on T and F can be known (see Remark 1).

As usual in insensitivity problems, the insensitivity condition (1.3) leads us to analyze a nonstandard nonlinear null controllability problem. In the case under consideration, the following holds.

PROPOSITION 1.3. *If there exists a control function* v *insensitizing the functional* Φ *given by* (1.2)*, then this control* v *solves the null controllability problem*

$$\begin{cases} \partial_t \overline{y} - \Delta \overline{y} + F(\overline{y}) = \xi + v\mathbf{1}_\omega & \text{in } Q, \\ \partial_n \overline{y} + f(\overline{y}) = 0 & \text{on } \Sigma, \\ \overline{y}(x,0) = y_0(x) & \text{in } \Omega, \end{cases} \tag{1.5}$$

$$\begin{cases} -\partial_t q - \Delta q + F'(\overline{y})q = \overline{y}\mathbf{1}_{\mathcal{O}} & \text{in } Q, \\ \partial_n q + f'(\overline{y})q = 0 & \text{on } \Sigma, \\ q(x,T) = 0 & \text{in } \Omega, \end{cases} \tag{1.6}$$

$$q(x,0) = 0 \quad \text{in } \Omega. \tag{1.7}$$

Furthermore, if a control function v *solves* (1.5)–(1.7) *and there exists* $\tau_0 > 0$ *such that* (1.1) *admits a weak solution* $y(\cdot,\cdot;\tau,v) \in L^2(0,T;H^1(\Omega)) \cap C^0(\overline{Q})$ *for* $|\tau| \le \tau_0$*, then* v *is insensitizing the functional* Φ*.*

Proof. We reason as in [10] and [1]. Assume the existence of a control v insensitizing the functional Φ given by (1.2) in the sense of Definition 1.1. Then, system (1.5) admits a weak solution $y(\cdot,\cdot;\tau,v) \in L^2(0,T;H^1(\Omega)) \cap C^0(\overline{Q})$ for all $|\tau| \le \tau_0$, for some $\tau_0 > 0$. The derivative of $\Phi(y(\cdot,\cdot;\tau,v))$ with respect to τ at $\tau = 0$ is given by

$$\left.\frac{\partial \Phi(y(\cdot,\cdot;\tau,v))}{\partial \tau}\right|_{\tau=0} = \iint_Q \overline{y}(x,t)\mathbf{1}_{\mathcal{O}}\, y_\tau(x,t)\,dx\,dt,$$

where $\overline{y} = y(\cdot,\cdot;0,v) \in C^0(\overline{Q})$ and $y_\tau = \left.\frac{\partial y(\cdot,\cdot;\tau,v)}{\partial \tau}\right|_{\tau=0}$ is the solution of the linear system

$$\begin{cases} \partial_t y_\tau - \Delta y_\tau + F'(\overline{y})y_\tau = 0 & \text{in } Q, \\ \partial_n y_\tau + f'(\overline{y})y_\tau = 0 & \text{on } \Sigma, \\ y_\tau(x,0) = \hat{y}_0(x) & \text{in } \Omega. \end{cases}$$

Let q be the solution of (1.6). Replacing $\overline{y}\mathbf{1}_{\mathcal{O}}$ with the left-hand side of the PDE satisfied by q and integrating by parts, one obtains

$$\iint_Q \overline{y}(x,t)\mathbf{1}_{\mathcal{O}}\, y_\tau(x,t)\,dx\,dt = \int_\Omega q(x,0)\hat{y}_0(x)\,dx,$$

regardless of what $\hat{y}_0 \in L^2(\Omega)$ is. Finally, from (1.3) one deduces that

$$q(0) = 0 \quad \text{in } X',$$

whence (1.7) follows, in view of the Hahn–Banach theorem. The rest of the proof follows immediately from Definition 1.1. □

Notice that a control function v solving (1.5)–(1.7), if one exists, does not necessarily insensitize the functional Φ (think, for instance, of an initial datum $y_0+\tau\,\hat{y}_0$ not lying in $C^0(\overline{\Omega})$, for which system (1.1) admits no weak solution in $C^0(\overline{Q})$). In other words, in this case the problem of seeking insensitizing controls cannot be reformulated in an equivalent way as a null controllability problem, as is usual in insensitivity problems. In order to prove Theorem 1.2, we will thus argue as follows (see section 3). Under the assumptions in the theorem, we will first prove the existence of a control v solving (1.5)–(1.7). In a second step, we will see that, for $\tau\hat{y}_0$ regular and small enough, such a control v can be chosen so that it also insensitizes the functional Φ defined by (1.2).

The existence of a control function solving (1.5)–(1.7) will be proved by linearization and a later application of an appropriate fixed point argument. This technique, introduced in [12] in the context of the controllability of the semilinear wave equation, has been used to prove several controllability results (cf., for example, [6], [7]). Analyzing a linear null controllability problem similar to (1.5)–(1.7) (see (2.1), (2.2), and (1.7)), we realize that the potentials $a, b \in L^\infty(\Sigma)$ need to have time derivatives in $L^\infty(\Sigma)$. This requirement comes from applying Lemma 1.2 of [8] to obtain an adequate observability inequality (see Proposition 2.1) for the solutions of the corresponding adjoint systems (2.5) and (2.6). To solve the nonlinear problem, we would have to search for a fixed point in a space containing the functions $z \in L^\infty(Q)$ such that the trace of $\partial_t z$ lies in $L^\infty(\Sigma)$. As was observed in Remark 15 of [5], we are not too far from imposing $\partial_t z \in L^\infty(0,T;W^{1,N+\gamma}(\Omega))$, with $\gamma > 0$. But these spaces are too small to achieve compactness and good estimates for the fixed point mapping. We will then seek a fixed point, and thus also control functions, in the Hölder spaces introduced above. In fact, one of the main points in this paper relies on the construction, in the linear case, of control functions with Hölderian regularity starting from L^2-controls.

In order to ensure the existence of a solution to system (1.1) in the above-mentioned Hölder spaces, appropriate regularity assumptions on the data and a compatibility condition on the initial datum are required (see Lemma 3.2). This is the reason why we have introduced the space $X = C^{2+\beta}(\overline{\Omega}) \cap H_0^2(\Omega)$, with $\beta \in (0,1)$, in Definition 1.1.

In the next section, we will analyze the corresponding linear null controllability problem, while section 3 will be devoted to proving our main result.

2. The linear null controllability problem. From now on, we will assume that $\omega \cap \mathcal{O} \neq \emptyset$ and $y_0 = 0$. This section is devoted to solving a linearized version of the null controllability problem (1.5)–(1.7). We consider the linear systems

$$\begin{cases} \partial_t y - \Delta y + cy = \xi + v\mathbf{1}_\omega & \text{in } Q, \\ \partial_n y + ay = 0 & \text{on } \Sigma, \\ y(x,0) = 0 & \text{in } \Omega, \end{cases} \tag{2.1}$$

$$\begin{cases} -\partial_t q - \Delta q + dq = y\mathbf{1}_{\mathcal{O}} & \text{in } Q, \\ \partial_n q + bq = 0 & \text{on } \Sigma, \\ q(x,T) = 0 & \text{in } \Omega, \end{cases} \tag{2.2}$$

where $a, b \in L^\infty(\Sigma)$, $c, d \in L^\infty(Q)$, and $\xi \in L^2(Q)$ (at least). For each $v \in L^2(Q)$, the cascade of linear systems (2.1), (2.2) admits exactly one solution (y, q) satisfying

$$y, q \in L^2(0,T; H^1(\Omega)) \cap C([0,T]; L^2(\Omega)), \quad \partial_t y, \partial_t q \in L^2(0,T; H^{-1}(\Omega)),$$

with

$$\|y\|_{L^2(H^1)\cap C(L^2)} + \|\partial_t y\|_{L^2(H^{-1})} \le C\left(\Omega, T, \|a\|_{\infty;\Sigma}, \|c\|_\infty\right)\left(\|\xi\|_{L^2} + \|v\|_{L^2}\right), \tag{2.3}$$

$$\begin{aligned} &\|q\|_{L^2(H^1)\cap C(L^2)} + \|\partial_t q\|_{L^2(H^{-1})} \\ &\quad \le C\left(\Omega, T, \|a\|_{\infty;\Sigma}, \|b\|_{\infty;\Sigma}, \|c\|_\infty, \|d\|_\infty\right)\left(\|\xi\|_{L^2} + \|v\|_{L^2}\right). \end{aligned} \tag{2.4}$$

Under additional assumptions on the potentials and on the source term ξ, we will build a regular control v, acting on a nonempty open subset of $\omega \cap \mathcal{O}$, such that the corresponding solution (y, q) of (2.1), (2.2) satisfies (1.7).

We proceed as follows. Let us fix a nonempty open set B_0 such that $B_0 \subset\subset \omega \cap \mathcal{O}$. In a first step, using an appropriate observability inequality, we obtain an L^2-control supported on $\overline{B}_0 \times [0,T]$. Then, by means of a construction similar to that made in [2] and [3] and due to the regularizing properties of the heat equation, we will be able to furnish a regular control with a slightly larger support.

Let us consider the adjoint systems

$$\begin{cases} \partial_t \varphi - \Delta \varphi + d\varphi = 0 & \text{in } Q, \\ \partial_n \varphi + b\varphi = 0 & \text{on } \Sigma, \\ \varphi(x,0) = \varphi^0(x) & \text{in } \Omega, \end{cases} \tag{2.5}$$

$$\begin{cases} -\partial_t \psi - \Delta \psi + c\psi = \varphi \mathbf{1}_{\mathcal{O}} & \text{in } Q, \\ \partial_n \psi + a\psi = 0 & \text{on } \Sigma, \\ \psi(x,T) = 0 & \text{in } \Omega, \end{cases} \tag{2.6}$$

where $\varphi^0 \in L^2(\Omega)$. For simplicity, we will denote by a_t (resp., b_t) the time derivative of a (resp., of b). Let $B_0 \subset\subset \omega \cap \mathcal{O}$ be the open set considered above. In [4], the following observability inequality for the solutions of (2.5), (2.6) is proved.

PROPOSITION 2.1. *Assume that $a, b, a_t, b_t \in L^\infty(\Sigma)$ and $c, d \in L^\infty(Q)$. Then, there exist positive constants M and C, depending on Ω, ω, $\mathcal{O}$, T, $\|a\|_{\infty;\Sigma}$, $\|b\|_{\infty;\Sigma}$, $\|a_t\|_{\infty;\Sigma}$, $\|b_t\|_{\infty;\Sigma}$, $\|c\|_\infty$, and $\|d\|_\infty$, such that*

$$\iint_Q \exp\left(-\frac{M}{t}\right)|\psi|^2 \, dx\, dt \le C \iint_{B_0\times(0,T)} |\psi|^2 \, dx\, dt,$$

for every $\varphi^0 \in L^2(\Omega)$, where ψ solves (2.6), φ being the solution of (2.5).

The proof of this result follows the scheme of demonstration of Proposition 2 in [11] and uses a global Carleman inequality for the classical heat equation with linear boundary Fourier conditions (see Lemma 1.2 of [8]).

Remark 1. In the previous proposition, and throughout this section, the dependence of the constants with respect to T and the potentials c and d could be stated precisely. This would allow one to know the precise way the constant $\mathcal{M}$ in Theorem 1.2 depends on T and F. Nevertheless, the dependence on the boundary data a and b is not explicit. This comes from the proof of Lemma 1.2 of [8].

Due to a unique continuation property for the solutions of (2.5) and (2.6) inferred from Proposition 2.1, under suitable assumptions on ξ, one obtains L^2-controls as follows.

PROPOSITION 2.2. *Assume that $a, b, a_t, b_t \in L^\infty(\Sigma)$ and $c, d \in L^\infty(Q)$. Let M and C be the positive constants (depending on Ω, ω, $\mathcal{O}$, T, $\|a\|_{\infty;\Sigma}$, $\|b\|_{\infty;\Sigma}$, $\|a_t\|_{\infty;\Sigma}$, $\|b_t\|_{\infty;\Sigma}$, $\|c\|_\infty$, and $\|d\|_\infty$) provided by Proposition* 2.1. *Then, for any $\xi \in L^2(Q)$ verifying*

$$\iint_Q \exp\left(\frac{M}{t}\right)|\xi|^2\,dx\,dt < \infty, \tag{2.7}$$

there exists a control function $\hat{v} \in L^2(Q)$, with $\operatorname{supp}\hat{v} \subset \overline{B}_0 \times [0,T]$, such that the solution $(\hat{y}, \hat{q})$ of (2.1), (2.2) *associated to $\hat{v}$ satisfies* (1.7). *Moreover, $\hat{v}$ can be chosen so that*

$$\|\hat{v}\|_{L^2} \le \sqrt{C}\left(\iint_Q \exp\left(\frac{M}{t}\right)|\xi|^2\,dx\,dt\right)^{1/2}. \tag{2.8}$$

The proof of this proposition is given in [4] and will be omitted here. The main result in this section is the following.

PROPOSITION 2.3. *Assume that $\partial\Omega \in C^{3+\beta}$ for some $\beta \in (0,1)$, $a, b, a_t, b_t \in L^\infty(\Sigma)$, $c \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$, and $d \in C^{1+\beta,\frac{1+\beta}{2}}(\overline{Q})$. Let $M > 0$ be the constant (depending on Ω, ω, $\mathcal{O}$, T, $\|a\|_{\infty;\Sigma}$, $\|b\|_{\infty;\Sigma}$, $\|a_t\|_{\infty;\Sigma}$, $\|b_t\|_{\infty;\Sigma}$, $\|c\|_\infty$, and $\|d\|_\infty$) provided by Proposition* 2.1. *Then, for any $\xi \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ satisfying* (2.7), *one can find a control $v \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ such that the associated solution (y,q) of* (2.1), (2.2) *satisfies* (1.7). *Moreover, the $C^{\beta,\frac{\beta}{2}}$-norm of v can be estimated as follows:*

$$|v|_{\beta,\frac{\beta}{2};\overline{Q}} \le C\left(|\xi|_{\beta,\frac{\beta}{2};\overline{Q}} + \left\|\exp\left(\frac{M}{2t}\right)\xi\right\|_{L^2}\right), \tag{2.9}$$

where C is a new positive constant depending on Ω, ω, $\mathcal{O}$, T, $\|a\|_{\infty;\Sigma}$, $\|b\|_{\infty;\Sigma}$, $\|a_t\|_{\infty;\Sigma}$, $\|b_t\|_{\infty;\Sigma}$, $|c|_{\beta,\frac{\beta}{2};\overline{Q}}$, and $|d|_{1+\beta,\frac{1+\beta}{2};\overline{Q}}$.

The regularity of v and, accordingly, that of (y,q), will enable us to deal with the nonlinear null controllability problem (1.5)–(1.7).

Before proving Proposition 2.3, for the convenience of the reader we repeat some relevant material from [3] and [9] without proofs, thus making our exposition self-contained. We first recall a technical result on local regularity given in [3, Proposition 2.1 and Remark 4].

LEMMA 2.4. *Let $\widetilde{a} \in L^\infty(Q)$ and $h \in L^2(Q)$ be given. Let us consider a solution $u \in L^2(0,T;H^1(\Omega)) \cap C([0,T];L^2(\Omega))$ of*

$$\begin{cases} \partial_t u - \Delta u + \widetilde{a}u = h & \text{in } Q, \\ u(x,0) = 0 & \text{in } \Omega, \end{cases} \tag{2.10}$$

and let $\mathcal{V} \subset \Omega$ be an arbitrary open set.

(a) *If $h \in L^r(0,T;L^r(\mathcal{V}))$, with $r \in (2,\infty)$, then $u \in X^r(0,T;\mathcal{V}')$ for any open set $\mathcal{V}' \subset\subset \mathcal{V}$. Moreover, there exist two positive constants $C = C(\Omega, \mathcal{V}, \mathcal{V}', T, N, r)$ and $\mathcal{K} = \mathcal{K}(N)$ such that*

$$\|u\|_{X^r(0,T;\mathcal{V}')} \le C\left(1 + \|\widetilde{a}\|_\infty\right)^{\mathcal{K}}\left[\|h\|_{L^r(L^r(\mathcal{V}))} + \|u\|_{L^2(H^1)\cap C(L^2)}\right]. \tag{2.11}$$

(b) *Assume, in addition, that* $h \in L^r(0,T;W^{1,r}(\mathcal{V}))$, $r > 2$, *and* $\nabla \widetilde{a} \in L^\gamma(Q)^N$, *with*

$$\gamma = \begin{cases} \max\left\{r, \dfrac{N}{2}+1\right\} & \text{if} \quad r \neq \dfrac{N}{2}+1, \\ \dfrac{N}{2}+1+\varepsilon & \text{if} \quad r = \dfrac{N}{2}+1, \end{cases}$$

and ε *being an arbitrarily small positive number. Then, for any open set* $\mathcal{V}' \subset\subset \mathcal{V}$, *one has* $u \in L^r(0,T;W^{3,r}(\mathcal{V}'))$, $\partial_t u \in L^r(0,T;W^{1,r}(\mathcal{V}'))$, *and for a new positive constant* $C = C(\Omega, \mathcal{V}, \mathcal{V}', T, N, r)$, *the following estimate holds:*

$$\|u\|_{L^r(W^{3,r}(\mathcal{V}'))} + \|\partial_t u\|_{L^r(W^{1,r}(\mathcal{V}'))} \leq C\,\mathcal{H}\left[\|h\|_{L^r(W^{1,r}(\mathcal{V}))} + \|u\|_{L^2(H^1)\cap C(L^2)}\right],$$

where

$$\mathcal{H} = \mathcal{H}(N, \|\widetilde{a}\|_\infty, \|\nabla \widetilde{a}\|_{L^\gamma}) = (1+\|\widetilde{a}\|_\infty)^{\mathcal{K}+1}(1+\|\nabla \widetilde{a}\|_{L^\gamma}),$$

$\mathcal{K} = \mathcal{K}(N)$ *being as in* (2.11).

We also recall the following result, which is readily obtained by rewriting Lemma 3.3 of [9, p. 80] with the notation introduced at the beginning of this paper (also see Lemma 2.3 of [3]).

LEMMA 2.5. *Let* $\mathcal{V} \subset \mathbb{R}^N$, $N \geq 1$, *be a regular open set. The following continuous embeddings hold:*

1. *If* $r < \frac{N}{2}+1$, *then* $X^r(0,T;\mathcal{V}) \hookrightarrow L^p(\mathcal{V}\times(0,T))$, *where* $\frac{1}{p} = \frac{1}{r} - \frac{2}{N+2}$.
2. *If* $r = \frac{N}{2}+1$, *then* $X^r(0,T;\mathcal{V}) \hookrightarrow L^q(\mathcal{V}\times(0,T))$ *for all* $q < \infty$.
3. *If* $\frac{N}{2}+1 < r < N+2$, *then* $X^r(0,T;\mathcal{V}) \hookrightarrow C^{\alpha,\frac{\alpha}{2}}(\overline{\mathcal{V}}\times[0,T])$, $\alpha = 2 - \frac{N+2}{r}$.
4. *If* $r = N+2$, *then* $X^r(0,T;\mathcal{V}) \hookrightarrow C^{l,\frac{l}{2}}(\overline{\mathcal{V}}\times[0,T])$ *for all* $l \in (0,1)$.
5. *If* $r > N+2$, *then* $X^r(0,T;\mathcal{V}) \hookrightarrow C^{1+\beta,\frac{1+\beta}{2}}(\overline{\mathcal{V}}\times[0,T])$, *where* $\beta = 1 - \frac{N+2}{r}$.

We are now ready to prove Proposition 2.3. From now on, we will specify only the dependence of the constants on the arguments that will be relevant in our analysis. Thus, for instance, the dependence on the dimension N, on B_0, or on the other open sets appearing later will be omitted.

Proof of Proposition 2.3. Assume that $\partial\Omega \in C^{3+\beta}$ for some $\beta \in (0,1)$. Let a, b, c, and d be as in the statement, and let $M > 0$ be provided by Proposition 2.1. Given $\xi \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ verifying (2.7), Proposition 2.2 provides a control $\hat{v} \in L^2(Q)$ such that the associated solution $(\hat{y},\hat{q})$ of (2.1), (2.2) satisfies (1.7). Moreover, $\hat{v}$ verifies estimate (2.8) and $\operatorname{supp}\hat{v} \subset \overline{B}_0 \times [0,T]$, with B_0 being the open set considered at the beginning of this section. One has $\hat{y},\hat{q} \in L^2(0,T;H^1(\Omega)) \cap C([0,T];L^2(\Omega))$, $\partial_t\hat{y}, \partial_t\hat{q} \in L^2(0,T;H^{-1}(\Omega))$, and estimates such as (2.3) and (2.4) hold.

Let B, B_1, and B_2 be regular open sets such that $B_0 \subset\subset B_1 \subset\subset B_2 \subset\subset B \subset\subset \omega \cap \mathcal{O}$. As was anticipated above, a construction similar to the one made in [2] and [3] will allow one to construct a regular control supported on $\overline{B}\times[0,T]$. Indeed, we set

$$q = (1-\theta)\,\hat{q}, \tag{2.12}$$

$$y = (1-\theta)\,\hat{y} + 2\nabla\theta\cdot\nabla\hat{q} + (\Delta\theta)\hat{q}, \tag{2.13}$$

with $\theta \in \mathcal{D}(B)$ satisfying $\theta \equiv 1$ in B_2. We will analyze the interior regularity of $\hat{y}$ and $\hat{q}$, inferring that (y,q) solves (2.1), (2.2), and (1.7) with control term $v \in C^{\beta,\frac{\beta}{2}}(Q)$ given by

$$v = -\theta\xi + 2\nabla\theta\cdot\nabla\hat{y} + (\Delta\theta)\hat{y} + (\partial_t - \Delta + c)\left[2\nabla\theta\cdot\nabla\hat{q} + (\Delta\theta)\hat{q}\right], \tag{2.14}$$

which is, in fact, supported on $B \times [0,T]$.

First, as $c \in L^\infty(Q)$ and $\xi + \hat{v}\mathbf{1}_{B_0} \in L^2(Q) \cap L^\infty(0,T;L^\infty(\Omega \setminus \overline{B}_0))$, one can apply Lemma 2.4 with $r = (N+2)/(1-\beta)$, $\beta \in (0,1)$ given in the statement, to deduce that $\hat{y}$ lies in $X^r(0,T;(\omega \cap \mathcal{O}) \setminus \overline{B}_1)$ (notice that, without loss of generality, we can assume that $\omega \cap \mathcal{O} \subset\subset \Omega$ and that $\omega \cap \mathcal{O}$ is regular enough). Since $r > N+2$, this space is continuously embedded in $C^{1+\beta,\frac{1+\beta}{2}}(\overline{(\omega \cap \mathcal{O}) \setminus B_1} \times [0,T])$, by Lemma 2.5. Thus,

$$\hat{y} \in C^{1+\beta,\frac{1+\beta}{2}}(\overline{(\omega \cap \mathcal{O}) \setminus B_1} \times [0,T]), \tag{2.15}$$

and estimates (2.11) and (2.3) give

$$|\hat{y}|_{1+\beta,\frac{1+\beta}{2};\overline{(\omega\cap\mathcal{O})\setminus B_1}\times[0,T]} \leq C\left(\Omega,\omega,\mathcal{O},T,\|a\|_{\infty;\Sigma},\|c\|_\infty\right)\left(\|\xi\|_\infty + \|\hat{v}\|_{L^2}\right). \tag{2.16}$$

By the choice of θ, the term $v_1 = 2\nabla\theta\cdot\nabla\hat{y} + (\Delta\theta)\hat{y}$ in (2.14) then lies in $C^{\beta,\frac{\beta}{2}}(\overline{Q})$ and one can estimate

$$|v_1|_{\beta,\frac{\beta}{2};\overline{Q}} \leq C\left(\Omega,\omega,\mathcal{O},T,\|a\|_{\infty;\Sigma},\|c\|_\infty\right)\left(\|\xi\|_\infty + \|\hat{v}\|_{L^2}\right).$$

According to the interior regularity of $\hat{y}$, an argument such as the one above implies that $\hat{q} \in C^{1+\beta,\frac{1+\beta}{2}}(\overline{B \setminus B_2} \times [0,T])$, and estimates (2.11), (2.16), and (2.4) give

$$|\hat{q}|_{1+\beta,\frac{1+\beta}{2};\overline{B\setminus B_2}\times[0,T]} \leq C\left(\|\xi\|_\infty + \|\hat{v}\|_{L^2}\right), \tag{2.17}$$

with $C = C\left(\Omega,\omega,\mathcal{O},T,\|a\|_{\infty;\Sigma},\|b\|_{\infty;\Sigma},\|c\|_\infty,\|d\|_\infty\right)$. By the $C^{\beta,\frac{\beta}{2}}$-regularity of ξ and c, it is then clear that $v_2 = -\theta\xi + c[2\nabla\theta\cdot\nabla\hat{q} + (\Delta\theta)\hat{q}] \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ and

$$|v_2|_{\beta,\frac{\beta}{2};\overline{Q}} \leq C(\Omega,\omega,\mathcal{O})\left(|\xi|_{\beta,\frac{\beta}{2};\overline{Q}} + |c|_{\beta,\frac{\beta}{2};\overline{Q}}|\hat{q}|_{1+\beta,\frac{1+\beta}{2};\overline{B\setminus B_2}\times[0,T]}\right),$$

which combined with (2.17), yields

$$|v_2|_{\beta,\frac{\beta}{2};\overline{Q}} \leq C(\Omega,\omega,\mathcal{O},T,\|a\|_{\infty;\Sigma},\|b\|_{\infty;\Sigma},|c|_{\beta,\frac{\beta}{2};\overline{Q}},\|d\|_\infty)\left(|\xi|_{\beta,\frac{\beta}{2};\overline{Q}} + \|\hat{v}\|_{L^2}\right).$$

We now analyze the term $v_3 = (\partial_t - \Delta)[2\nabla\theta\cdot\nabla\hat{q} + (\Delta\theta)\hat{q}]$. To this end, we use the following result on interior Hölderian regularity, whose proof is given at the end of this section.

LEMMA 2.6. *Assume that $\partial\Omega \in C^{3+\beta}$ for some $\beta \in (0,1)$. Let us consider a solution $u \in L^2(0,T;H^1(\Omega)) \cap C([0,T];L^2(\Omega))$ of (2.10), with $\widetilde{a} \in C^{1+\beta,\frac{1+\beta}{2}}(\overline{Q})$ and $h \in L^2(Q) \cap C^{1+\beta,\frac{1+\beta}{2}}(\overline{\mathcal{V}} \times [0,T])$, $\mathcal{V}$ being a nonempty open subset of Ω. Then, for any open set $\mathcal{V}' \subset\subset \mathcal{V}$, one has $u \in C^{3+\beta,\frac{3+\beta}{2}}(\overline{\mathcal{V}}' \times [0,T])$ and*

$$|u|_{3+\beta,\frac{3+\beta}{2};\overline{\mathcal{V}}'\times[0,T]} \leq C\left(|h|_{1+\beta,\frac{1+\beta}{2};\overline{\mathcal{V}}\times[0,T]} + \|u\|_{L^2(H^1)\cap C(L^2)}\right), \tag{2.18}$$

where C is a positive constant depending on Ω, $\mathcal{V}$, $\mathcal{V}'$, T, and $|\widetilde{a}|_{1+\beta,\frac{1+\beta}{2};\overline{Q}}$.

On account of (2.15) and the regularity of the potential d, Lemma 2.6 can be applied to $u = \hat{q}$, with $\mathcal{V} = (\omega \cap \mathcal{O}) \setminus \overline{B}_1$, $\mathcal{V}' = B \setminus \overline{B}_2$, $h = \hat{y}\mathbf{1}_{\mathcal{O}}$, and $\widetilde{a} = d$, to deduce that $\hat{q}$ lies in $C^{3+\beta,\frac{3+\beta}{2}}(\overline{B \setminus B_2} \times [0,T])$. Moreover, estimates (2.18), (2.16), and (2.4) give

$$|\hat{q}|_{3+\beta,\frac{3+\beta}{2};\overline{B\setminus B_2}\times[0,T]} \leq C\left(\|\xi\|_\infty + \|\hat{v}\|_{L^2}\right),$$

with $C = C(\Omega, \omega, \mathcal{O}, T, \|a\|_{\infty;\Sigma}, \|b\|_{\infty;\Sigma}, \|c\|_\infty, |d|_{1+\beta,\frac{1+\beta}{2};\overline{Q}})$. We infer from the choice of θ that $2\nabla\theta \cdot \nabla\hat{q} + (\Delta\theta)\hat{q} \in C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q})$, and hence that v_3 lies in $C^{\beta,\frac{\beta}{2}}(\overline{Q})$, and one can estimate

$$|v_3|_{\beta,\frac{\beta}{2};\overline{Q}} \leq C(\Omega, \omega, \mathcal{O}, T, \|a\|_{\infty;\Sigma}, \|b\|_{\infty;\Sigma}, \|c\|_\infty, |d|_{1+\beta,\frac{1+\beta}{2};\overline{Q}})\left(\|\xi\|_\infty + \|\hat{v}\|_{L^2}\right).$$

In view of the previous considerations on each term v_i, $1 \leq i \leq 3$, and using estimate (2.8), one concludes that v given by (2.14) lies in $C^{\beta,\frac{\beta}{2}}(\overline{Q})$ and that (2.9) holds. Finally, it is an easy exercise to see that (y, q) defined by (2.13) and (2.12), together with this control function v, solves (2.1), (2.2), and (1.7). The only delicate point could be to check that $y(x, 0) = 0$ in Ω. But this follows immediately from the interior regularity of $\hat{q}$ (which, in particular, gives $\hat{q} \in C([0,T]; H^1(B \setminus \overline{B}_2))$), the choice of θ, and the fact that $\hat{q}(x, 0) = 0$ in Ω. This ends the proof of Proposition 2.3. □

We end this section by giving the proof of Lemma 2.6, which relies on a localization argument.

Proof of Lemma 2.6. Assume the hypothesis in the statement, with $\beta \in (0, 1)$ and $\mathcal{V}$ being fixed. Given an open set $\mathcal{V}' \subset\subset \mathcal{V}$, we consider a regular open set $\mathcal{V}_1$, with $\mathcal{V}' \subset\subset \mathcal{V}_1 \subset\subset \mathcal{V}$. According to the regularity of h and the potential $\widetilde{a}$, we can apply the second point of Lemma 2.4, with $r = (N + 2)/(1 - \beta)$ (thus $\gamma = r$, since $r > N + 2$), and deduce that u, $\frac{\partial u}{\partial x_i} \in X^r(0, T; \mathcal{V}_1)$, $i = 1, \ldots, N$, together with the estimate

$$\|u\|_{X^r(0,T;\mathcal{V}_1)} + \|\nabla u\|_{X^r(0,T;\mathcal{V}_1)^N} \leq C\left[\|h\|_{L^r(W^{1,r}(\mathcal{V}))} + \|u\|_{L^2(H^1)\cap C(L^2)}\right],$$

with $C > 0$ depending on Ω, $\mathcal{V}$, $\mathcal{V}'$, T, $\|\widetilde{a}\|_\infty$ and $\|\nabla\widetilde{a}\|_{L^r}$. Since $r > N + 2$, by Lemma 2.5 we get

$$u,\ \frac{\partial u}{\partial x_i} \in C^{1+\beta,\frac{1+\beta}{2}}(\overline{\mathcal{V}}_1 \times [0,T]), \quad i = 1, \ldots, N, \tag{2.19}$$

with

$$\begin{aligned} &|u|_{1+\beta,\frac{1+\beta}{2};\overline{\mathcal{V}}_1\times[0,T]} + |\nabla u|_{1+\beta,\frac{1+\beta}{2};\overline{\mathcal{V}}_1\times[0,T]} \\ &\leq C\left(\Omega, \mathcal{V}, \mathcal{V}', T, \|\widetilde{a}\|_\infty, \|\nabla\widetilde{a}\|_{L^r}\right)\left(\|h\|_{L^r(W^{1,r}(\mathcal{V}))} + \|u\|_{L^2(H^1)\cap C(L^2)}\right). \end{aligned} \tag{2.20}$$

We claim that, indeed, u lies in $C^{3+\beta,\frac{3+\beta}{2}}(\overline{\mathcal{V}}' \times [0,T])$ and satisfies (2.18). To this end, let $\zeta \in \mathcal{D}(\mathcal{V}_1)$ be such that $\zeta \equiv 1$ in $\mathcal{V}'$ and set $w = \zeta u$. Then w solves

$$\begin{cases} \partial_t w - \Delta w = \widetilde{h} & \text{in } Q, \\ w = 0 \quad \text{on } \Sigma, \quad w(x, 0) = 0 & \text{in } \Omega, \end{cases} \tag{2.21}$$

with $\widetilde{h} = \zeta h - [\widetilde{a}\zeta u + 2\nabla\zeta \cdot \nabla u + (\Delta\zeta)u]$. The regularity of h and $\widetilde{a}$, together with (2.19), gives $\widetilde{h} \in C^{1+\beta,\frac{1+\beta}{2}}(\overline{Q})$, and using (2.20), one has

$$|\widetilde{h}|_{1+\beta,\frac{1+\beta}{2};\overline{Q}} \leq C\left(|h|_{1+\beta,\frac{1+\beta}{2};\overline{\mathcal{V}}\times[0,T]} + \|u\|_{L^2(H^1)\cap C(L^2)}\right), \tag{2.22}$$

with $C = C(\Omega, \mathcal{V}, \mathcal{V}', T, |\widetilde{a}|_{1+\beta,\frac{1+\beta}{2};\overline{Q}})$. Since $\partial\Omega \in C^{3+\beta}$ and the compatibility condition of order 1 for system (2.21) is trivially fulfilled, one can apply Theorem 5.2 of [9] to obtain $w \in C^{3+\beta,\frac{3+\beta}{2}}(\overline{Q})$, with

$$|w|_{3+\beta,\frac{3+\beta}{2};\overline{Q}} \leq C(\Omega, T)|\widetilde{h}|_{1+\beta,\frac{1+\beta}{2};\overline{Q}}. \tag{2.23}$$

Finally, recalling that $u \equiv w$ in $\mathcal{V}'$, one infers the desired interior regularity of u, and estimate (2.18) holds, using (2.23) and (2.22). □

3. Proof of Theorem 1.2. We begin this section by recalling the following result for linear systems of the form

$$\begin{cases} \partial_t u - \Delta u + cu = h & \text{in } Q, \\ \partial_n u + au = 0 & \text{on } \Sigma, \\ u(x,0) = u_0(x) & \text{in } \Omega, \end{cases} \tag{3.1}$$

whose proof is given in [9, Theorem 5.3, p. 320].

LEMMA 3.1. *Assume that $\partial\Omega \in C^{2+\beta}$ for some $\beta \in (0,1)$. Let $a \in C^{1+\beta,\frac{1+\beta}{2}}(\overline{\Sigma})$ and $c \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ be given. Then, for any $h \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ and $u_0 \in C^{2+\beta}(\overline{\Omega})$ satisfying the compatibility condition*

$$\partial_n u_0(x) + a(x,0)u_0(x) = 0 \quad \text{on } \partial\Omega,$$

system (3.1) *admits exactly one solution $u \in C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q})$ verifying the estimate*

$$|u|_{2+\beta,1+\frac{\beta}{2};\overline{Q}} \leq C\left(\Omega, T, |a|_{1+\beta,\frac{1+\beta}{2};\overline{\Sigma}}, |c|_{\beta,\frac{\beta}{2};\overline{Q}}\right)\left(|h|_{\beta,\frac{\beta}{2};\overline{Q}} + |u_0|_{2+\beta;\overline{\Omega}}\right). \tag{3.2}$$

Let us now prove Theorem 1.2. Assume the hypothesis in the statement. From the considerations in the first section, the proof falls naturally into two steps.

Step 1. *Existence of a regular control solving the nonlinear null controllability problem* (1.5)–(1.7): *The fixed point argument.* Let G and g be the C^2 functions defined by

$$G(s) = \begin{cases} \dfrac{F(s)}{s} & \text{if } s \neq 0, \\ F'(0) & \text{if } s = 0, \end{cases} \qquad g(s) = \begin{cases} \dfrac{f(s)}{s} & \text{if } s \neq 0, \\ f'(0) & \text{if } s = 0. \end{cases}$$

Let us set

$$Z = C^1(\overline{Q}) \cap C^{1+\beta,\frac{1+\beta}{2}}(\overline{Q}).$$

For fixed $z \in \overline{B}(0;1) \subset Z$, we consider the linear systems

$$\begin{cases} \partial_t y - \Delta y + G(z)y = \xi + v\mathbf{1}_\omega & \text{in } Q, \\ \partial_n y + g(z)y = 0 & \text{on } \Sigma, \\ y(x,0) = 0 & \text{in } \Omega, \end{cases} \tag{3.3}$$

$$\begin{cases} -\partial_t q - \Delta q + F'(z)q = y\mathbf{1}_{\mathcal{O}} & \text{in } Q, \\ \partial_n q + f'(z)q = 0 & \text{on } \Sigma, \\ q(x,T) = 0 & \text{in } \Omega, \end{cases} \tag{3.4}$$

with potentials $G(z), F'(z) \in Z$ and $g(z), f'(z) \in \widetilde{Z} = C^1(\overline{\Sigma}) \cap C^{1+\beta,\frac{1+\beta}{2}}(\overline{\Sigma})$. By abuse of notation, from now on we will let $g(z)$ (resp., $f'(z)$) stand for both the function $g(z)$ in Z and its restriction to $\overline{\Sigma}$ (resp., for both $f'(z) \in Z$ and its restriction to $\overline{\Sigma}$). Let us set

$$\mathcal{M}(\Omega, \omega, \mathcal{O}, T, F, f) = \sup_{z \in \overline{B}(0;1)} M_z, \tag{3.5}$$

where M_z is, for fixed $z \in \overline{B}(0;1)$, the positive constant (depending on Ω, ω, $\mathcal{O}$, T, $\|g(z)\|_Z$, $\|f'(z)\|_Z$, $\|G(z)\|_Z$, and $\|F'(z)\|_Z$) provided by Proposition 2.1. Let ξ in $C^{\beta,\frac{\beta}{2}}(\overline{Q})$ satisfy (1.4), with $\eta > 0$ to be chosen later (hence also verifying (2.7) with $M = M_z$ for all $z \in \overline{B}(0;1)$). From Proposition 2.3, there exists $v_z \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ such that the associated solution (y_z, q_z) of (3.3), (3.4) lies in $C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q}) \times C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q})$ (by Lemma 3.1) and satisfies (1.7). Moreover, one has

$$|v_z|_{\beta,\frac{\beta}{2};\overline{Q}} \leq C(\Omega, \omega, \mathcal{O}, T, z) \left(|\xi|_{\beta,\frac{\beta}{2};\overline{Q}} + \left\| \exp\left(\frac{M_z}{2t}\right) \xi \right\|_{L^2} \right),$$

with $C(\Omega, \omega, \mathcal{O}, T, z) = C(\Omega, \omega, \mathcal{O}, T, \|g(z)\|_Z, \|f'(z)\|_Z, \|G(z)\|_Z, \|F'(z)\|_Z)$, and y_z satisfies an estimate such as (3.2); hence

$$|y_z|_{2+\beta,1+\frac{\beta}{2};\overline{Q}} \leq C_1(\Omega, \omega, \mathcal{O}, T, z) \left(|\xi|_{\beta,\frac{\beta}{2};\overline{Q}} + \left\| \exp\left(\frac{M_z}{2t}\right) \xi \right\|_{L^2} \right),$$

with $C_1(\Omega, \omega, \mathcal{O}, T, z) = C_1(\Omega, \omega, \mathcal{O}, T, \|g(z)\|_Z, \|f'(z)|_Z, \|G(z)\|_Z, \|F'(z)\|_Z)$ (and a similar estimate for q_z holds). Then, for any $z \in \overline{B}(0;1)$ one has

$$|v_z|_{\beta,\frac{\beta}{2};\overline{Q}} \leq \widetilde{C}(\Omega, \omega, \mathcal{O}, T, F, f) \left(|\xi|_{\beta,\frac{\beta}{2};\overline{Q}} + \left\| \exp\left(\frac{\mathcal{M}}{2t}\right) \xi \right\|_{L^2} \right), \tag{3.6}$$

$$|y_z|_{2+\beta,1+\frac{\beta}{2};\overline{Q}} \leq C_2(\Omega, \omega, \mathcal{O}, T, F, f) \left(|\xi|_{\beta,\frac{\beta}{2};\overline{Q}} + \left\| \exp\left(\frac{\mathcal{M}}{2t}\right) \xi \right\|_{L^2} \right), \tag{3.7}$$

together with a similar estimate for q_z, with

$$\widetilde{C}(\Omega, \omega, \mathcal{O}, T, F, f) = \sup_{z \in \overline{B}(0;1)} C(\Omega, \omega, \mathcal{O}, T, z),$$
$$C_2(\Omega, \omega, \mathcal{O}, T, F, f) = \sup_{z \in \overline{B}(0;1)} C_1(\Omega, \omega, \mathcal{O}, T, z).$$

For each $z \in \overline{B}(0;1) \subset Z$, we consider the families

$$\mathcal{A}(z) = \left\{ v \in C^{\beta,\frac{\beta}{2}}(\overline{Q}) : (y, q) \text{ satisfies (3.3), (3.4), and (1.7), } v \text{ verifying (3.6)} \right\},$$
$$\Lambda(z) = \left\{ y : (y, q) \text{ solves (3.3), (3.4) with } v \in \mathcal{A}(z) \right\}.$$

One can then define the set-valued mapping $\Lambda : z \in \overline{B}(0;1) \subset Z \mapsto \Lambda(z) \subset Z$. For fixed $z \in \overline{B}(0;1)$, each $y \in \Lambda(z)$ lies in $C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q})$ and satisfies (3.7), and thus

$$\|y\|_Z \leq C_2(\Omega, \omega, \mathcal{O}, T, F, f) \left(|\xi|_{\beta,\frac{\beta}{2};\overline{Q}} + \left\| \exp\left(\frac{\mathcal{M}}{2t}\right) \xi \right\|_{L^2} \right). \tag{3.8}$$

We claim that there exists $\eta(\Omega,\omega,\mathcal{O},T,F,f) > 0$ such that if a source term $\xi \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ satisfies (1.4), with $\mathcal{M}$ given by (3.5), then the Kakutani fixed point theorem can be applied to Λ. First, for fixed $z \in \overline{B}(0;1) \subset Z$, it is easy to check that $\Lambda(z)$ is a nonempty closed convex subset of Z (here we use the linear character of systems (3.3) and (3.4)). By estimate (3.7), $\Lambda(z)$ is a bounded set in $C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q})$. Since this space is compactly embedded into Z, one infers that each $\Lambda(z)$ is a compact subset of Z. Furthermore, there exists a fixed compact set $K \subset Z$ such that $\Lambda(z) \subset K$ for all $z \in \overline{B}(0;1)$.

In the second place, Λ is proved to be an upper hemicontinuous multivalued mapping, or, equivalently, it is proved that for any bounded linear form $\mu \in Z'$, the function

$$z \in \overline{B}(0;1) \subset Z \mapsto \sup_{y\in\Lambda(z)} \langle \mu, y\rangle \in \mathbb{R}$$

is upper semicontinuous. To this end, it suffices to show that the set

$$B_{\lambda,\mu} = \left\{ z \in \overline{B}(0;1) : \sup_{y\in\Lambda(z)} \langle \mu, y\rangle \geq \lambda \right\}$$

is closed in Z for any $\lambda \in \mathbb{R}$ and any $\mu \in Z'$. Let us fix $\lambda \in \mathbb{R}$ and $\mu \in Z'$, and consider a sequence $\{z_n\}_{n\geq 1} \subset B_{\lambda,\mu}$ such that

$$z_n \to z \text{ in } Z. \tag{3.9}$$

Our aim is to see that $z \in B_{\lambda,\mu}$. As stated above, each $\Lambda(z_n)$ is a compact set in Z. Then for fixed $n \geq 1$ one has

$$\sup_{y\in\Lambda(z_n)} \langle \mu, y\rangle = \langle \mu, y_n\rangle \geq \lambda \tag{3.10}$$

for some $y_n \in \Lambda(z_n)$. By the definition of $\Lambda(z_n)$ and $\mathcal{A}(z_n)$, there exist $v_n \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ satisfying

$$|v_n|_{\beta,\frac{\beta}{2};\overline{Q}} \leq \widetilde{C}(\Omega,\omega,\mathcal{O},T,F,f)\left(|\xi|_{\beta,\frac{\beta}{2};\overline{Q}} + \left\|\exp\left(\frac{\mathcal{M}}{2t}\right)\xi\right\|_{L^2}\right) \tag{3.11}$$

($\widetilde{C}(\Omega,\omega,\mathcal{O},T,F,f)$ as in (3.6)) and $q_n \in C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q})$ such that (y_n,q_n) together with v_n solve

$$\begin{cases} \partial_t y_n - \Delta y_n + G(z_n)y_n = \xi + v_n\mathbf{1}_\omega & \text{in } Q,\\ \partial_n y_n + g(z_n)y_n = 0 & \text{on } \Sigma,\\ y_n(x,0) = 0 & \text{in } \Omega, \end{cases} \tag{3.12}$$

$$\begin{cases} -\partial_t q_n - \Delta q_n + F'(z_n)q_n = y_n\mathbf{1}_{\mathcal{O}} & \text{in } Q,\\ \partial_n q_n + f'(z_n)q_n = 0 & \text{on } \Sigma,\\ q_n(x,T) = 0, \quad q_n(x,0) = 0 & \text{in } \Omega. \end{cases} \tag{3.13}$$

From (3.11) and (3.7), $\{v_n\}$ and $\{(y_n,q_n)\}$ are uniformly bounded in $C^{\beta,\frac{\beta}{2}}(\overline{Q})$ and $C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q}) \times C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q})$, respectively. Taking into account the compact embedding of $C^{\beta,\frac{\beta}{2}}(\overline{Q})$ (resp., $C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q})$) into $C^0(\overline{Q})$ (resp., Z), there exist subsequences (still denoted by $\{v_n\}$ and $\{(y_n,q_n)\}$) such that

$$v_n \to \widetilde{v} \quad \text{in } C^0(\overline{Q}), \qquad (y_n,q_n) \to (\widetilde{y},\widetilde{q}) \quad \text{in } Z \times Z$$

for some $\widetilde{v} \in C^0(\overline{Q})$, $(\widetilde{y}, \widetilde{q}) \in Z \times Z$. It is easily seen that, in fact, $\widetilde{v} \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$. On account of the regularity of F and f, from (3.9) one also has

$$\begin{aligned} G(z_n) \to G(z) \quad \text{and} \quad F'(z_n) \to F'(z) \quad \text{in } Z, \\ g(z_n) \to g(z) \quad \text{and} \quad f'(z_n) \to f'(z) \quad \text{in } \widetilde{Z}. \end{aligned}$$

We can then pass to the limit in (3.11)–(3.13) and deduce that $(\widetilde{y}, \widetilde{q})$ solves (3.3), (3.4), and (1.7) with control term $\widetilde{v} \in \mathcal{A}(z)$. Thus, $\widetilde{y} \in \Lambda(z)$, and taking limits in (3.10), one infers that

$$\sup_{y \in \Lambda(z)} \langle \mu, y \rangle \geq \langle \mu, \widetilde{y} \rangle \geq \lambda.$$

We conclude that $z \in B_{\lambda,\mu}$; hence, Λ is an upper hemicontinuous mapping.

Now let $\eta = \eta(\Omega, \omega, \mathcal{O}, T, F, f) > 0$ be such that $\eta \leq C_2(\Omega, \omega, \mathcal{O}, T, F, f)^{-1}$. Then, for a given source term $\xi \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ satisfying (1.4), with $\mathcal{M}$ given by (3.5), we infer from (3.8) that any $y \in \Lambda\left(\overline{B}(0;1)\right)$ verifies $\|y\|_Z \leq 1$; that is, Λ maps the nonempty closed convex set $\overline{B}(0;1)$ into itself. We can then apply the Kakutani fixed point theorem and conclude that there exists $\overline{y} \in Z$ such that $\overline{y} \in \Lambda(\overline{y})$. Hence, there exists $v \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$, solving the nonlinear null controllability problem (1.5)–(1.7) (for $y_0 = 0$). Moreover, by (3.6) one can estimate

$$|v|_{\beta,\frac{\beta}{2};\overline{Q}} \leq \widetilde{C}(\Omega, \omega, \mathcal{O}, T, F, f)\eta. \tag{3.14}$$

Step 2. *Existence of a control insensitizing the functional* Φ. Let us see that there exists $\eta(\Omega, \omega, \mathcal{O}, T, F, f) > 0$ such that for any $\xi \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ satisfying (1.4), with $\mathcal{M}$ given by (3.5), the control v in the previous step can be chosen so that, for τ small enough, the existence of a solution of (1.1) (with $y_0 = 0$) in $C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q})$ is ensured. This will conclude the proof of the theorem, since such a control v will then insensitize the functional Φ given by (1.2), in view of Proposition 1.3.

We use the following result, which can be proved by linearizing and applying an appropriate fixed point argument.

LEMMA 3.2. *Assume that* $\partial\Omega \in C^{2+\beta}$ *for some* $\beta \in (0,1)$. *Let* $F \in C^2(\mathbb{R})$ *and* $f \in C^3(\mathbb{R})$ *be given. Then, there exists* $\delta > 0$ *(depending on* Ω, T, F, *and* f*) with the property that for any* $h \in C^{\beta,\frac{\beta}{2}}(\overline{Q})$ *and* $u_0 \in C^{2+\beta}(\overline{\Omega})$ *satisfying*

$$|h - F(0)|_{\beta,\frac{\beta}{2};\overline{Q}} + |f(0)| + |u_0|_{2+\beta;\overline{\Omega}} \leq \delta$$

and the compatibility condition

$$\partial_n u_0 + f(u_0) = 0 \quad \text{on } \partial\Omega, \tag{3.15}$$

the nonlinear system

$$\begin{cases} \partial_t u - \Delta u + F(u) = h & \text{in } Q, \\ \partial_n u + f(u) = 0 & \text{on } \Sigma, \\ u(x,0) = u_0(x) & \text{in } \Omega, \end{cases}$$

admits a unique solution $u \in C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q})$.

Let us consider $X = C^{2+\beta}(\overline{\Omega}) \cap H_0^2(\Omega)$, with $\beta \in (0,1)$ as in the statement. Let $\delta > 0$ be provided by Lemma 3.2 and let $\mathcal{M}(\Omega, \omega, \mathcal{O}, T, F, f)$ be given by (3.5).

Recalling (3.14), one can choose $\eta = \eta(\Omega, \omega, \mathcal{O}, T, F, f) > 0$ small enough so that for any $\xi \in C^{\beta, \frac{\beta}{2}}(\overline{Q})$ verifying (1.4), $\hat{y}_0 \in X$ with $\|\hat{y}_0\|_X = 1$, and $\tau \in \mathbb{R}$ small enough, one has

$$|\xi + v\mathbf{1}_\omega|_{\beta, \frac{\beta}{2}; \overline{Q}} + |\tau \hat{y}_0|_{2+\beta; \overline{\Omega}} \leq \delta.$$

Since the initial datum $\tau \hat{y}_0$ satisfies (3.15) (by choice of X), one infers from Lemma 3.2 that system (1.1) possesses a solution $y(\cdot, \cdot; \tau, v) \in C^{2+\beta, 1+\frac{\beta}{2}}(\overline{Q})$, which ends the proof of Theorem 1.2. □

REFERENCES

[1] O. Bodart and C. Fabre, *Controls insensitizing the norm of the solution of a semilinear heat equation*, J. Math. Anal. Appl., 195 (1995), pp. 658–683.

[2] O. Bodart, M. González-Burgos, and R. Pérez-García, *Insensitizing controls for a semilinear heat equation with a superlinear nonlinearity*, C. R. Acad. Sci. Paris, Sér. I Math., 335 (2002), pp. 677–682.

[3] O. Bodart, M. González-Burgos, and R. Pérez-García, *Existence of insensitizing controls for a semilinear heat equation with a superlinear nonlinearity*, Comm. Partial Differential Equations, to appear.

[4] O. Bodart, M. González-Burgos, and R. Pérez-García, *Insensitizing controls for a heat equation with a nonlinear term involving the state and the gradient*, Nonlinear Anal., 57 (2004), pp. 687–711.

[5] A. Doubova, E. Fernández-Cara, and M. González-Burgos, *On the controllability of the heat equation with nonlinear boundary Fourier conditions*, J. Differential Equations, 196 (2004), pp. 385–417.

[6] C. Fabre, J.-P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 36–61.

[7] E. Fernández-Cara and E. Zuazua, *Null and approximate controllability for weakly blowing up semilinear heat equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 583–616.

[8] A. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Seoul, Korea, 1996.

[9] O. A. Ladyzenskaya, V. A. Solonnikov, and N. N. Uraltzeva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1967.

[10] J.-L. Lions, *Quelques notions dans l'analyse et le contrôle de systèmes à données incomplètes*, in Proceedings of the XIth Congress on Differential Equations and Applications/First Congress on Applied Mathematics, University of Málaga, Málaga, Spain, 1990, pp. 43–54.

[11] L. De Teresa, *Insensitizing controls for a semilinear heat equation*, Comm. Partial Differential Equations, 25 (2000), pp. 39–72.

[12] E. Zuazua, *Exact boundary controllability for the semilinear wave equation*, in Nonlinear Partial Differential Equations and Their Applications, Vol. X, Pitman Res. Notes Math. Ser. 220, H. Brezis and J.-L. Lions, eds., Longman Sci. Tech., Harlow, UK, 1991, pp. 357–391.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 970–985

SUPERCONVERGENCE PROPERTIES OF OPTIMAL CONTROL PROBLEMS*

C. MEYER† AND A. RÖSCH‡

Abstract. An optimal control problem for a two-dimensional (2-d) elliptic equation is investigated with pointwise control constraints. This paper is concerned with discretization of the control by piecewise constant functions. The state and the adjoint state are discretized by linear finite elements. Approximations of the optimal solution of the continuous optimal control problem will be constructed by a projection of the discrete adjoint state. It is proved that these approximations have convergence order h^2.

Key words. linear-quadratic optimal control problems, error estimates, elliptic equations, numerical approximation, control constraints, superconvergence

AMS subject classifications. 49K20, 49M25, 65N30

DOI. 10.1137/S0363012903431608

1. Introduction.

The paper is concerned with the discretization of the two-dimensional (2-d) elliptic optimal control problem

$$J(u) = F(y,u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u\|^2_{L^2(\Omega)} \tag{1.1}$$

subject to the state equations

$$\begin{aligned} Ay + a_0 y &= u \qquad \text{in } \Omega, \\ y &= 0 \qquad \text{on } \Gamma \end{aligned} \tag{1.2}$$

and subject to the control constraints

$$a \le u(t,x) \le b \qquad \text{for a.a. } x \in \Omega, \tag{1.3}$$

where Ω is a bounded domain and Γ is the boundary of Ω; A denotes a second order elliptic operator of the form

$$Ay(x) = -\sum_{i,j=1}^{2} D_i(a_{ij}(x)D_j y(x)),$$

where D_i denotes the partial derivative with respect to x_i and a and b are real numbers. Moreover, $\nu > 0$ is a fixed positive number. We denote the set of admissible controls by U_{ad}:

$$U_{ad} = \{u \in L^2(\Omega) : \ a \le u \le b \ \text{ a.e. in } \Omega\}.$$

*Received by the editors July 16, 2003; accepted for publication (in revised form) December 24, 2003; published electronically October 8, 2004. This research was supported by the DFG Research Center "Mathematics for Key Technologies" (FZT 86) in Berlin.
http://www.siam.org/journals/sicon/43-3/43160.html

†Technische Universität Berlin, Fakultät II Mathematik und Naturwissenschaften, Straße des 17. Juni 136, D-10623 Berlin, Germany (cmeyer@math.tu-berlin.de).

‡Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria (arnd.roesch@oeaw.ac.at).

We discuss here the full discretization of the control and the state equations by a finite element method. The asymptotic behavior of the discretized problem is studied, and superconvergence results are established.

The approximation of the discretization for semilinear elliptic optimal control problems is discussed in Arada, Casas, and Tröltzsch [1]. The optimal control problem (1.1)–(1.3) is a linear-quadratic counterpart of such a semilinear problem. Our aim is to construct controls $\tilde{u}$ which have an approximation order of h^2. This higher convergence order is the difference between our work and [1].

The discretization of optimal control problems by piecewise constant functions is well investigated; we refer to Falk [7] and Geveci [8]. Piecewise constant and piecewise linear discretization in space are discussed for a parabolic problem in Malanowski [12]. Theory and numerical results for elliptic boundary control problems are contained in Casas and Tröltzsch [5] and Casas, Mateos, and Tröltzsch [4].

Piecewise linear control discretizations for elliptic optimal control problems are studied by Casas and Tröltzsch; see [5]. In an abstract optimization problem, piecewise linear approximations are investigated in Rösch [14]. In all papers, the convergence order is h or $h^{3/2}$.

A quadratic convergence result is proved by Hinze [10]. In that approach only the state equation is discretized. The control is obtained by a projection of the adjoint state to the set of admissible controls.

In this paper, we combine the advantages of the different approaches. After solving a fully discretized optimal control problem, a control $\tilde{u}$ is calculated by the projection of the adjoint state p_h in a postprocessing step. Although the approximation of the discretized solution is only of order h, we will show that this postprocessing step improves the convergence order to h^2.

The paper is organized as follows: In section 2 the discretizations are introduced and the main results are stated. Section 3 contains auxiliary results. The proofs of the superconvergence results are placed in section 4. The paper ends with numerical experiments shown in section 5.

2. Discretization and superconvergence results. Throughout this paper, Ω denotes a convex bounded open subset in $\mathbb{R}^2$ of class $C^{1,1}$. The coefficients a_{ij} of the operator A belong to $C^{0,1}(\bar{\Omega})$ and satisfy the ellipticity condition

$$m_0|\xi|^2 \le \sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \qquad \forall (\xi, x) \in \mathbb{R}^2 \times \bar{\Omega}, \quad m_0 > 0.$$

Moreover, we require $a_{ij}(x) = a_{ji}(x)$ and $y_d \in L^p(\Omega)$ for some $p > 2$. For the function $a_0 \in L^\infty(\Omega)$, we assume $a_0 \ge 0$. Next, we recall some results from Bonnans and Casas [2].

LEMMA 2.1 (see [2]). *For every $p > 2$ and every function $g \in L^p(\Omega)$, the solution y of*

$$Ay + a_0 y = g \qquad \text{in } \Omega, \quad y|_\Gamma = 0,$$

belongs to $H_0^1(\Omega) \cap W^{2,p}(\Omega)$. Moreover, there exists a positive constant c independent of a_0 such that

$$\|y\|_{W^{2,p}(\Omega)} \le c\|g\|_{L^p(\Omega)}.$$

Next, we introduce the adjoint equation

$$\begin{aligned} Ap + a_0 p &= y - y_d && \text{in } \Omega, \\ p &= 0 && \text{on } \Gamma. \end{aligned} \tag{2.1}$$

Due to Lemma 2.1, the state equation and the adjoint equation admit unique solutions in $H_0^1(\Omega) \cap W^{2,p}(\Omega)$ if $y_d \in L^p(\Omega)$ for $p > 2$. This space is embedded in $C^{0,1}(\bar{\Omega})$.

We call the solution y of (1.2) for a control u an associated state to u and write $y(u)$. In the same way, we call the solution p of (2.1) corresponding to $y(u)$ an associated adjoint state to u and write $p(u)$.

Introducing the projection

$$\Pi_{[a,b]}(f(x)) = \max(a, \min(b, f(x))),$$

we can formulate the necessary and sufficient first order optimality condition for (1.1)–(1.3).

LEMMA 2.2. *A necessary and sufficient condition for the optimality of a control $\bar{u}$ with corresponding state $\bar{y} = y(\bar{u})$ and adjoint state $\bar{p} = p(\bar{u})$, respectively, is that the equation*

$$\bar{u}(x) = \Pi_{[a,b]}\left(-\frac{1}{\nu}\bar{p}\right) \tag{2.2}$$

holds.

Since the optimal control problem is strictly convex, we obtain the existence of a unique optimal solution. The optimality condition can be formulated as a variational inequality (3.11). A standard pointwise a.e. discussion of this variational inequality leads to the above formulated projection formula; see [12].

We are now able to introduce the discretized problem. We define a finite element based approximation of the optimal control (1.1)–(1.3). To this aim, we consider a family of triangulations $(T_h)_{h>0}$ of $\bar{\Omega}$. With each element $T \in T_h$, we associate two parameters $\rho(T)$ and $\sigma(T)$, where $\rho(T)$ denotes the diameter of the set T and $\sigma(T)$ is the diameter of the largest ball contained in T. The mesh size of the grid is defined by $h = \max_{T \in T_h} \rho(T)$. We suppose that the following regularity assumptions are satisfied.

(A1) There exist two positive constants ρ and σ such that

$$\frac{\rho(T)}{\sigma(T)} \le \sigma, \qquad \frac{h}{\rho(T)} \le \rho$$

hold for all $T \in T_h$ and all $h > 0$.

(A2) Let us define $\bar{\Omega}_h = \cup_{T \in T_h} T$, and let Ω_h and Γ_h denote its interior and its boundary, respectively. We assume that $\bar{\Omega}_h$ is convex and that the vertices of T_h placed on the boundary of Γ_h are points of Γ. From [13, estimate (5.2.19)], it is known that

$$|\Omega \setminus \Omega_h| \le Ch^2,$$

where $|.|$ denotes the measure of the set. Next, to every boundary triangle T of T_h we associate another triangle $\hat{T}$ with curved boundary as follows: The edge between the two boundary nodes of T is substituted by the corresponding curved part of Γ. We

denote by $\hat{T}_h$ the union of these curved boundary triangles with the interior triangles to Ω of T_h, such that $\bar{\Omega} = \cup_{\hat{T}\in\hat{T}_h}\hat{T}$. Moreover, we set

$$U_h = \{u \in L^\infty(\Omega) :\ u|_{\hat{T}} \text{ is constant on all } \hat{T} \in \hat{T}_h\}, \quad U_h^{ad} = U_h \cap U_{ad},$$
$$V_h = \{y_h \in C(\bar{\Omega}) :\ y_h \in \mathcal{P}_1\, \forall\, T \in T_h, \text{ and } y_h = 0 \text{ on } \bar{\Omega} \setminus \Omega_h\},$$

where $\mathcal{P}_1$ is the space of polynomials of degree less than or equal to 1. For each $u_h \in U_h$, we denote by $y_h(u_h)$ the unique element of V_h that satisfies

$$a(y_h(u_h), v_h) = \int_\Omega u_h v_h\, dx \qquad \forall v_h \in V_h, \tag{2.3}$$

where $a :\ V_h \times V_h \to \mathbb{R}$ is the bilinear form defined by

$$a(y_h, v_h) = \int_\Omega \left(a_0(x) y_h(x) v_h(x) + \sum_{i,j=1}^{2} a_{ij}(x) D_i y_h(x) D_j v_h(x) \right) dx.$$

In other words, $y_h(u_h)$ is the approximated state associated with u_h. Because of $y_h = v_h = 0$ on $\bar{\Omega}\setminus\Omega_h$, the integrals over Ω can be replaced by integrals over Ω_h. The finite dimensional approximation of the optimal control problem is defined by

$$\inf J(u_h) = \frac{1}{2}\|y_h(u_h) - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u_h\|^2_{L^2(\Omega)}, \qquad u_h \in U_h^{ad}. \tag{2.4}$$

The adjoint equation is discretized in the same way:

$$a(p_h(u_h), v_h) = \int_\Omega (y_h(u_h) - y_d) v_h\, dx \qquad \forall v_h \in V_h. \tag{2.5}$$

The approximation order of the solutions of (2.4) in the L^2-sense is investigated in [1].

We will go a different way. For our superconvergence result we need an additional assumption for $\bar{u}$. We know already that the associated adjoint state $\bar{p}$ belongs to a space $W^{2,p}(\Omega)$ for a certain $p > 2$. The optimal control $\bar{u}$ is obtained by the projection formula (2.2). Therefore, we can classify the triangles T_i in two sets K_1 and K_2:

$$K_1 = \{T_i :\ \bar{u} \text{ is only Lipschitz continuous on } T_i\}, \qquad K_2 = \{T_i :\ \bar{u} \in W^{2,p}(T_i)\}.$$

This classification is correct: $W^{2,p}(\Omega)$ is embedded in $C^{0,1}(\bar{\Omega})$. Moreover, the projection operator is continuous from $C^{0,1}(\bar{\Omega})$ to $C^{0,1}(\bar{\Omega})$. Clearly, the number of triangles in K_1 grows for decreasing h. Nevertheless, the following additional assumption is fulfilled in many practical cases.

(A3) $|K_1| \le c \cdot h$.

Let $\bar{u}$ be the optimal solution of (1.1)–(1.3). Next, we denote by S_i the centroid of the triangle T_i. We define a piecewise constant function by the values of $\bar{u}(S_i)$:

$$w_h(x) = \bar{u}(S_i) \quad \text{ if } x \in T_i. \tag{2.6}$$

It is easy to verify that $w_h \in U_h^{ad}$.

Now we are able to formulate our first superconvergence result.

THEOREM 2.3. *Assume that the assumptions* (A1)–(A3) *hold. Let* u_h *be the solutions of* (2.4). *Then the estimate*

$$\|u_h - w_h\|_{L^2(\Omega)} \le ch^2 \tag{2.7}$$

holds true.

The proof of Theorem 2.3 is contained in section 4.

Moreover, we can construct controls in a postprocessing step. We start with the solution u_h of (2.4). The control $\tilde{u}$ is calculated by a projection of the discrete adjoint state $p_h(u_h)$ to the admissible set

$$\tilde{u}(x) = \Pi_{[a,b]}\left(-\frac{1}{\nu}(p_h(u_h))(x)\right).$$

THEOREM 2.4. *Assume that the assumptions* (A1)–(A3) *hold. Let* $\tilde{u}$ *be the control constructed above. Then the estimate*

$$\|\bar{u} - \tilde{u}\|_{L^2(\Omega)} \le ch^2 \tag{2.8}$$

holds true.

The proof of Theorem 2.4 is also derived in section 4.

3. Auxiliary results. First, we recall some well-known results for finite element method approximations [6]. We start with the so-called Aubin–Nitsche lemma.

LEMMA 3.1. *Let* (A1) *and* (A2) *be fulfilled and* $u \in L^2(\Omega)$. *Then we have*

$$\|y(u) - y_h(u)\|_{L^2(\Omega)} \le ch^2\|u\|_{L^2(\Omega)}, \tag{3.1}$$

$$\|p(u) - p_h(u)\|_{L^2(\Omega)} \le ch^2(\|u\|_{L^2(\Omega)} + \|y_d\|_{L^2(\Omega)}). \tag{3.2}$$

Next, we prove an estimate for the numerical integration.

LEMMA 3.2. *Let* f *be a function belonging to* $H^2(T_i)$ *for all* i *in a certain index set* I. *Then the estimates*

$$\left|\int_{T_i} (f(x) - f(S_i))\ dx\right| \le ch^2\sqrt{|T_i|}|f|_{H^2(T_i)}$$

and

$$\sum_{i\in I}\left|\int_{T_i} (f(x) - f(S_i))\ dx\right| \le ch^2\left(\sum_{i\in I}|f|^2_{H^2(T_i)}\right)^{1/2}$$

are valid.

Proof. The proof is almost standard. First, we remark that $|.|_{H^2(T_i)}$ denotes the H^2-seminorm. Next, we transform the integral to an integral over the reference element by $E\hat{x} = x$ and apply the Bramble–Hilbert lemma:

$$\begin{aligned}
\left|\int_{T_i} (f(x) - f(S_i))\ dx\right| &= \frac{|T_i|}{|\hat{T}|}\left|\int_{\hat{T}} (f(E\hat{x}) - f(S_i))\ d\hat{x}\right| \\
&\le c|T_i|\left(\int_{\hat{T}}\sum_{|\alpha|=2}|D^\alpha_{\hat{x}} f(E\hat{x})|^2\ d\hat{x}\right)^{1/2} \\
&\le ch^2|T_i|\left(\frac{|\hat{T}|}{|T_i|}\int_{T_i}\sum_{|\alpha|=2}|D^\alpha_x f(x)|^2\ dx\right)^{1/2} \\
&\le ch^2\sqrt{|T_i|}|f|_{H^2(T_i)}.
\end{aligned}$$

This implies

$$\sum_{i\in I}\left|\int_{T_i}(f(x)-f(S_i))\;dx\right|\le ch^2\left(\sum_{i\in I}|f|^2_{H^2(T_i)}\right)^{1/2}$$

by the Cauchy–Schwarz inequality. ∎

LEMMA 3.3. *Let w_h be the functions defined by* (2.6). *In addition, we assume that the assumptions* (A1)–(A3) *are fulfilled. Then the estimate*

$$\|y_h(\bar u)-y_h(w_h)\|_{L^2(\Omega)}\le ch^2\|\bar p\|_{W^{2,p}(\Omega)} \tag{3.3}$$

holds true.

Proof. We start with the transformation

$$\begin{aligned}\|y_h(\bar u)-y_h(w_h)\|^2_{L^2(\Omega)}&=(y_h(\bar u)-y_h(w_h),y_h(\bar u)-y_h(w_h))_{L^2(\Omega)}\\&=(p_h(\bar u)-p_h(w_h),\bar u-w_h)_{L^2(\Omega)}\\&=\int_{K_1}(p_h(\bar u)-p_h(w_h))(\bar u-w_h)\;dx\\&\quad+\int_{K_2}(p_h(\bar u)-p_h(w_h))(\bar u-w_h)\;dx.\end{aligned}\tag{3.4}$$

It remains to estimate these two integrals. The K_1-part can be estimated by the following arguments: The function $\bar p$ belongs to $W^{2,p}(\Omega)$ with $p>2$. Hence, we have

$$\|\bar u\|_{C^{0,1}(\bar\Omega)}\le\frac{1}{\nu}\|\bar p\|_{C^{0,1}(\bar\Omega)}\le c\|\bar p\|_{W^{2,p}(\Omega)}.$$

Because of $\bar u(S_i)=w_h(S_i)$ and the fact that w_h is constant on T_i, this implies $|\bar u(x)-w_h(x)|\le c\|\bar p\|_{W^{2,p}(\Omega)}\cdot|x-S_i|\le ch\|\bar p\|_{W^{2,p}(\Omega)}$. Consequently, we obtain

$$\begin{aligned}\left|\int_{K_1}(p_h(\bar u)-p_h(w_h))(\bar u-w_h)\;dx\right|&\le\sum_{T_i\in K_1}\int_{T_i}|(p_h(\bar u)-p_h(w_h))(\bar u-w_h)|\;dx\\&\le\sum_{T_i\in K_1}ch\|\bar p\|_{W^{2,p}(\Omega)}\|p_h(\bar u)-p_h(w_h)\|_{C(\bar\Omega)}\int_{T_i}dx\\&\le ch\|\bar p\|_{W^{2,p}(\Omega)}\|p_h(\bar u)-p_h(w_h)\|_{C(\bar\Omega)}\int_{K_1}dx\\&\le ch^2\|\bar p\|_{W^{2,p}(\Omega)}\|p_h(\bar u)-p_h(w_h)\|_{C(\bar\Omega)}\end{aligned}\tag{3.5}$$

by means of assumption (A3). For a triangle T_i of the K_2-part we have for an arbitrary function $v_h\in V_h$

$$\int_{T_i}w_hv_h\;dx=\int_{T_i}\bar u(S_i)v_h\;dx=\int_{T_i}\bar u(S_i)v_h(S_i)\;dx.$$

This is a formula for the numerical integration of $\bar uv_h$. Consequently, we obtain by Lemma 3.2

$$\begin{aligned}\left|\int_{K_2}(\bar u-w_h)v_h\;dx\right|&\le\sum_{T_i\in K_2}\left|\int_{T_i}(\bar u-\bar u(S_i))v_h\;dx\right|\\&\le ch^2\left(\sum_{T_i\in K_2}|\bar uv_h|^2_{H^2(T_i)}\right)^{1/2}.\end{aligned}\tag{3.6}$$

Next, we divide each triangle T_i of K_2 in an "active" part (A_i) and an "inactive" part (I_i) with $A_i \cup I_i = T_i$. The optimal control $\bar{u}$ is constant on the active component A_i ($\bar{u} = a$ or $\bar{u} = b$). Therefore, the seminorm is 0 on these parts. On the inactive parts I_i, we have

$$\bar{u} = -\frac{1}{\nu}\bar{p}.$$

Therefore, we can estimate

$$|\bar{u}v_h|_{H^2(T_i)} = |\bar{u}v_h|_{H^2(I_i)} = \frac{1}{\nu}|\bar{p}v_h|_{H^2(I_i)} \le c|\bar{p}v_h|_{H^2(T_i)}.$$

Hence, we can continue by

$$\begin{aligned}
\left|\int_{K_2} (\bar{u} - w_h)v_h \, dx\right| &\le ch^2 \left(\sum_{T_i \in K_2} |\bar{u}v_h|^2_{H^2(T_i)}\right)^{1/2} \\
&\le ch^2 \left(\sum_{T_i \in K_2} |\bar{p}v_h|^2_{H^2(T_i)}\right)^{1/2} \\
&\le ch^2 \left(\sum_{T_i \in K_2} \sum_{|\alpha|,|\beta|=1} \|D^{\alpha+\beta}\bar{p}v_h\|^2_{L^2(T_i)} + \|D^{\alpha}\bar{p}D^{\beta}v_h\|^2_{L^2(T_i)}\right)^{1/2} \\
&\le ch^2 \|\bar{p}\|_{W^{2,p}(\Omega)} \|v_h\|_{H^1_0(\Omega)} \qquad (3.7)
\end{aligned}$$

by means of Hölder's inequality in the last step. Next, we set $v_h = p_h(\bar{u}) - p_h(w_h)$ and obtain

$$\left|\int_{K_2} (\bar{u} - w_h)(p_h(\bar{u}) - p_h(w_h)) \, dx\right| \le ch^2 \|\bar{p}\|_{W^{2,p}(\Omega)} \|p_h(\bar{u}) - p_h(w_h)\|_{H^1_0(\Omega)}. \qquad (3.8)$$

Inserting (3.5) and (3.8) in (3.4), we get

$$\|y_h(\bar{u}) - y_h(w_h)\|^2_{L^2(\Omega)} \le ch^2 \|\bar{p}\|_{W^{2,p}(\Omega)} (\|p_h(\bar{u}) - p_h(w_h)\|_{C(\bar{\Omega})} + \|p_h(\bar{u}) - p_h(w_h)\|_{H^1_0(\Omega)}).$$

We benefit now from the fact that $p_h(\bar{u})$ and $p_h(w_h)$ are the solutions of the discretized adjoint equation (2.5). Hence, we have

$$\|p_h(\bar{u}) - p_h(w_h)\|_{C(\bar{\Omega})} + \|p_h(\bar{u}) - p_h(w_h)\|_{H^1_0(\Omega)} \le c\|y_h(\bar{u}) - y_h(w_h)\|_{L^2(\Omega)}$$

with a positive constant c independent of h. The C-estimate can be obtained as follows: Take the adjoint equation (2.1) and the discretized adjoint equation (2.5), but with a right-hand side $f_h \in V_h$ instead of $y - y_d$. Then we find for the corresponding solutions z and z_h of the continuous and discretized adjoint equation

$$\begin{aligned}
\|z_h\|_{C(\bar{\Omega})} &\le \|z_h - z\|_{C(\bar{\Omega})} + \|z\|_{C(\bar{\Omega})} \\
&\le ch\|f_h\|_{L^2(\Omega)} + c\|z\|_{H^2(\Omega)} \\
&\le ch\|f_h\|_{L^2(\Omega)} + c\|f_h\|_{L^2(\Omega)}.
\end{aligned}$$

Substituting $f_h = y_h(\bar{u}) - y_h(w_h)$ and $z_h = p_h(\bar{u}) - p_h(w_h)$ delivers the desired estimate. For the estimate of the first expression, we refer to Braess [3].

Finally, we get

$$\|y_h(\bar{u}) - y_h(w_h)\|_{L^2(\Omega)} \le ch^2 \|\bar{p}\|_{W^{2,p}(\Omega)},$$

which is exactly inequality (3.3). □

COROLLARY 3.4. *Assume that the assumptions of Lemma* 3.3 *are fulfilled. Then, we have*

$$\|p_h(\bar{u}) - p_h(w_h)\|_{L^2(\Omega)} \le ch^2 \|\bar{p}\|_{W^{2,p}(\Omega)}. \tag{3.9}$$

By means of Lemma 3.1*, we obtain*

$$\|\bar{p} - p_h(w_h)\|_{L^2(\Omega)} \le ch^2 (\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}). \tag{3.10}$$

LEMMA 3.5. *The following variational inequalities are necessary and sufficient for the optimality of the unique solutions of* (1.1)–(1.3) *and* (2.4)*, respectively.*

$$(\bar{p} + \nu\bar{u}, u - \bar{u})_{L^2(\Omega)} \ge 0 \qquad \forall u \in U_{ad}, \tag{3.11}$$

$$(p_h(u_h) + \nu u_h, \zeta_h - u_h)_{L^2(\Omega)} \ge 0 \qquad \forall \zeta_h \in U_h^{ad}. \tag{3.12}$$

The variational inequality (3.11) is an equivalent formulation for the projection formula (2.2).

Next, we derive a variational inequality for the function w_h. First, formula (3.11) is true for all $u \in U_{ad}$. Therefore, we have pointwise a.e.

$$(\bar{p}(x) + \nu\bar{u}(x)) \cdot (u - \bar{u}(x)) \ge 0 \qquad \forall u \in [a, b].$$

We apply this formula for $x = S_i$ and $u = u_h(S_i)$. This is correct because of the continuity of $\bar{u}$, $\bar{p}$, and u_h in these points. We arrive at

$$(\bar{p}(S_i) + \nu\bar{u}(S_i)) \cdot (u_h(S_i) - \bar{u}(S_i)) \ge 0 \qquad \forall S_i.$$

Due to (2.6), this is equivalent to

$$(\bar{p}(S_i) + \nu w_h(S_i)) \cdot (u_h(S_i) - w_h(S_i)) \ge 0 \qquad \forall S_i.$$

We integrate this formula over T_i and add up over all i

$$(\hat{p} + \nu w_h, u_h - w_h)_{L^2(\Omega)} \ge 0 \tag{3.13}$$

with

$$\hat{p}(x) = \bar{p}(S_i) \qquad \text{if } x \in T_i.$$

Moreover, we can test inequality (3.12) with the function w_h and get

$$(p_h(u_h) + \nu u_h, w_h - u_h)_{L^2(\Omega)} \ge 0. \tag{3.14}$$

We add these two inequalities and obtain

$$(\hat{p} - p_h(u_h) + \nu(w_h - u_h), u_h - w_h)_{L^2(\Omega)} \ge 0.$$

This is equivalent to

$$\nu \|w_h - u_h\|^2_{L^2(\Omega)} \le (\hat{p} - p_h(u_h), u_h - w_h)_{L^2(\Omega)}. \tag{3.15}$$

4. Superconvergence properties. Inequality (3.15) is the starting point for the proofs of the superconvergence results. Now, we are ready to prove Theorem 2.3.

Proof. For the right-hand side of (3.15), we find

$$\begin{aligned}(\hat{p} - p_h(u_h), u_h - w_h)_{L^2(\Omega)} &= (p_h(w_h) - p_h(u_h), u_h - w_h)_{L^2(\Omega)} \\ &\quad + (\bar{p} - p_h(w_h), u_h - w_h)_{L^2(\Omega)} \\ &\quad + (\hat{p} - \bar{p}, u_h - w_h)_{L^2(\Omega)}. \end{aligned} \tag{4.1}$$

Next we estimate these three terms. We start with

$$(p_h(w_h) - p_h(u_h), u_h - w_h)_{L^2(\Omega)} = (y_h(w_h) - y_h(u_h), y_h(u_h) - y_h(w_h))_{L^2(\Omega)} \le 0. \tag{4.2}$$

The second term can be estimated by formula (3.10):

$$(\bar{p} - p_h(w_h), u_h - w_h)_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}) \cdot \|w_h - u_h\|_{L^2(\Omega)}. \tag{4.3}$$

The third term represents again a formula for the numerical integration. Using that u_h and w_h are constant on each triangle T_i, we obtain by Lemma 3.2

$$\begin{aligned}(\hat{p} - \bar{p}, u_h - w_h)_{L^2(\Omega)} &= \sum_i \int_{T_i} (\hat{p}(x) - \bar{p}(x))(u_h(x) - w_h(x))\, dx \\ &= \sum_i (u_h(S_i) - w_h(S_i)) \int_{T_i} (\bar{p}(S_i) - \bar{p}(x))\, dx \\ &\le \sum_i ch^2 |u_h(S_i) - w_h(S_i)| \sqrt{|T_i|} \cdot |\bar{p}|_{H^2(T_i)} \\ &\le ch^2 \cdot \|w_h - u_h\|_{L^2(\Omega)} \cdot \|\bar{p}\|_{W^{2,p}(\Omega)}. \end{aligned} \tag{4.4}$$

Inserting (4.2)–(4.4) in (4.1), we get

$$(\hat{p} - p_h(u_h), u_h - w_h)_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}) \cdot \|w_h - u_h\|_{L^2(\Omega)}.$$

Next, we combine this inequality with (3.15)

$$\nu \|w_h - u_h\|^2_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}) \cdot \|w_h - u_h\|_{L^2(\Omega)}.$$

This formula is equivalent to

$$\|w_h - u_h\|_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}),$$

which was the assertion of Theorem 2.3. ∎

Theorem 2.3 means that the values of the numerical solution u_h in the centroids already have a quadratic convergence rate. By the projection of the associated adjoint state $p_h(u_h)$, we obtain an admissible control $\tilde{u}$ that has a quadratic convergence order with respect to the L^2-norm. This was the assertion of Theorem 2.4.

Proof. We start with the result of Theorem 2.3

$$\|w_h - u_h\|_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}).$$

This inequality implies

$$\|p_h(w_h) - p_h(u_h)\|_{L^2(\Omega)} \le c\|w_h - u_h\|_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}).$$

From Corollary 3.4, we know formula (3.10):

$$\|\bar{p} - p_h(w_h)\|_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}).$$

Therefore, we obtain by the triangle inequality

$$\|\bar{p} - p_h(u_h)\|_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}).$$

The projection operator $\Pi_{[a,b]}$ is Lipschitz continuous with constant 1 from $L^2(\Omega)$ to $L^2(\Omega)$. Finally, we get

$$\|\bar{u} - \tilde{u}\|_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}).$$

The superconvergence result is proved. □

COROLLARY 4.1. *By the arguments of the proof of Theorem* 2.4*, we get another result. We find for the* L^∞*-error*

$$\|\bar{u} - \tilde{u}\|_{L^\infty(\Omega)} \le ch(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}).$$

Proof. From formula (3.3)

$$\|y_h(w_h) - y_h(\bar{u})\|_{L^2(\Omega)} \le ch^2\|\bar{p}\|_{W^{2,p}(\Omega)}$$

and the inequality

$$\|y_h(w_h) - y_h(u_h)\|_{L^2(\Omega)} \le c\|w_h - u_h\|_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}),$$

we get by the triangle inequality

$$\|y_h(\bar{u}) - y_h(u_h)\|_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}).$$

This inequality implies

$$\|p_h(\bar{u}) - p_h(u_h)\|_{L^\infty(\Omega)} \le c\|y_h(\bar{u}) - y_h(u_h)\|_{L^2(\Omega)} \le ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}).$$

Denoting the solution of (2.5) with $\bar{y}$ instead of $y_h(u_h)$, by $\bar{p}_h$, we continue with

$$\begin{aligned}\|\bar{p} - p_h(u_h)\|_{L^\infty(\Omega)} \le{}& \|\bar{p} - \bar{p}_h\|_{L^\infty(\Omega)} + \|\bar{p}_h - p_h(\bar{u})\|_{L^\infty(\Omega)} \\ &+ \|p_h(\bar{u}) - p_h(u_h)\|_{L^\infty(\Omega)}.\end{aligned} \tag{4.5}$$

The first term can be estimated by $ch\|\bar{p}\|_{H^2(\Omega)}$ (see [3]). For the second term, we use the argumentation of Lemma 3.1 with $z_h = p_h(\bar{u}) - p_h(u_h)$ and $f_h = y_h(\bar{u}) - \bar{y}$:

$$\begin{aligned}\|\bar{p} - p_h(u_h)\|_{L^\infty(\Omega)} &\le ch\|\bar{p}\|_{H^2(\Omega)} + c\|y_h(\bar{u}) - \bar{y}\|_{L^2(\Omega)} + ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}) \\ &\le ch\|\bar{p}\|_{H^2(\Omega)} + ch^2(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}) \\ &\le ch(\|\bar{p}\|_{W^{2,p}(\Omega)} + \|y_d\|_{L^2(\Omega)}).\end{aligned}$$

The properties of the projection operator imply the assertion. □

COROLLARY 4.2. *The first estimate can be improved if all data are sufficiently smooth:*

$$\|\bar{p} - \bar{p}_h\|_{L^\infty(\Omega)} \le ch^2|\ln h|^{3/2}\|\bar{p}\|_{W^{2,\infty}(\Omega)}$$

(see Braess [3]*). In this case, formula* (4.5) *implies*

$$\|\bar{u} - \tilde{u}\|_{L^\infty(\Omega)} \le ch^2|\ln h|^{3/2}(\|\bar{p}\|_{W^{2,\infty}(\Omega)} + \|y_d\|_{L^2(\Omega)}).$$

5. Numerical tests. We have tested the convergence theory by two examples. In both cases, the Laplace operator $-\Delta$ was chosen for the elliptic operator A. The first example exactly fits the presented theory.

Both optimization problems were solved numerically by a primal-dual active set strategy; see, for instance, [11]. The discretization was already described in section 2: The control function u is discretized by piecewise constant functions, whereas the state y and the adjoint state p were approximated by piecewise linear functions. We used uniform meshes. The additional numerical effort for the calculation of $\tilde{u}$ is very small. We have only to evaluate the pointwise projection of the function $-\frac{1}{\nu}p_h$ to the interval $[a, b]$.

In contrast to this, the numerical evaluation of the L^2-norm $\|\bar{u}-\tilde{u}\|_{L^2(\Omega)}$ and the graphical representation are not so simple. Therefore, we briefly sketch these aspects. We want to point out that this additional effort is only needed to confirm the theoretical results. This effort is not necessary for the computation of the approximated optimal control.

For the computation of the L^2-norms $\|\bar{u}-u_h\|_{L^2(\Omega)}$ and $\|\bar{u}-\tilde{u}\|_{L^2(\Omega)}$, respectively, we introduce sets $\tilde{K}_1$, $\tilde{K}_2$, analogous to the sets K_1 and K_2:

$$\tilde{K}_1 = \{T_i : \tilde{u} \text{ is only Lipschitz continuous on } T_i\}, \quad \tilde{K}_2 = \{T_i : \tilde{u} \in C^\infty(T_i)\}.$$

Moreover, we set $M_1 = K_1 \cup \tilde{K}_1$, $M_2 = K_2 \cap \tilde{K}_2$. The numerical evaluation of $\|\bar{u}-\tilde{u}\|_{L^2(\Omega)}$ differs on the sets M_1 and M_2. Therefore, we split the L^2-norm up into

$$\|\bar{u}-\tilde{u}\|^2_{L^2(\Omega)} = \|\bar{u}-\tilde{u}\|^2_{L^2(M_1)} + \|\bar{u}-\tilde{u}\|^2_{L^2(M_2)}.$$

In our examples, the part $\bar{u}|_{K_2} \in C^\infty(K_2)$ is smooth. Thus, $\|\bar{u}-\tilde{u}\|_{L^2(M_2)}$ can be evaluated with sufficient accuracy applying an appropriate quadrature formula. In contrast to this, the difference $\bar{u}-\tilde{u}$ belongs only to $C^{0,1}(T_i)$ for all triangles $T_i \in M_1$. Hence, an arbitrary accurate quadrature formula would only admit an error of order h. Therefore, we introduce a subgrid of a significantly smaller mesh size in each triangle $T_i \in M_1$ and evaluate the norm $\|\bar{u}-\tilde{u}\|_{L^2(M_1)}$ on this subgrid to ensure sufficient accuracy. We want to point out that this subgrid is only used for the evaluation of the norm $\|\bar{u}-\tilde{u}\|_{L^2(M_1)}$ with a sufficient high accuracy.

Example 1. In this example, we investigate the Laplace equation with homogeneous Dirichlet boundary conditions. Therefore, we choose $a_0 \equiv 0$ in (1.2). Thus, the state equation is given by

$$\begin{aligned} -\Delta y &= u \text{ in } \Omega, \\ y &= 0 \text{ on } \Gamma. \end{aligned} \tag{5.1}$$

Now, we define the optimal state by

$$\bar{y} = y_a - y_g$$

with an analytical part $y_a = \sin(\pi\, x_1)\, \sin(\pi\, x_2)$ and a less smooth function y_g. The function y_g is defined as the solution of

$$\begin{aligned} -\Delta y_g &= g \text{ in } \Omega, \\ y_g &= 0 \text{ on } \Gamma. \end{aligned}$$

The function g is given by

$$g(x_1,x_2)=\begin{cases} u_f(x_1,x_2)-a & \text{if } u_f(x_1,x_2)<a,\\ 0 & \text{if } u_f(x_1,x_2)\in[a,b],\\ u_f(x_1,x_2)-b & \text{if } u_f(x_1,x_2)>b \end{cases}$$

with $u_f(x_1,x_2)=2\,\pi^2\,\sin(\pi\,x_1)\,\sin(\pi\,x_2)$. Due to the state equation (5.1), we obtain for the exact optimal control $\bar{u}$

$$\bar{u}(x_1,x_2)=\begin{cases} a & \text{if } u_f(x_1,x_2)<a,\\ u_f(x_1,x_2) & \text{if } u_f(x_1,x_2)\in[a,b],\\ b & \text{if } u_f(x_1,x_2)>b. \end{cases}$$

For the optimal adjoint state $\bar{p}$, we find

$$\bar{p}(x_1,x_2)=-2\,\pi^2\,\nu\,\sin(\pi\,x_1)\,\sin(\pi\,x_2).$$

The desired state is given by

$$y_d(x_1,x_2)=\bar{y}+\Delta\bar{p}=y_a-y_g+4\,\pi^4\,\nu\,\sin(\pi\,x_1)\,\sin(\pi\,x_2).$$

It is easy to see that these functions fulfill the necessary and sufficient first order optimality conditions. Moreover, the sets with

$$-\frac{1}{\nu}\bar{p}=a \quad \text{or} \quad -\frac{1}{\nu}\bar{p}=b$$

are a finite number (here two) of curves γ_i. Hence, the measure of the set K_1 is bounded by the total length of these curves

$$|K_1|\le 2h\sum|\gamma_i|,$$

and assumption (A3) is fulfilled. We chose $\nu=1$ for the numerical calculations.

Figures 5.1 and 5.2 show the numerical solutions $\tilde{u}$ for $h=0.04$ and $h=0.02$. As described above, the function $\tilde{u}$ is obtained by the pointwise projection $\Pi_{[a,b]}(-\frac{1}{\nu}p_h)$ in a postprocessing step. Therefore, $\tilde{u}$ contains sharp breaks on the subset $\tilde{K}_1$. To visualize these breaks, we introduce new mesh points in all triangles $T_i\in\tilde{K}_1$. Notice that these new grid points are only used for the graphical presentation of the projection.

Figures 5.3 and 5.4 show the convergence behavior of $\|\bar{u}-u_h\|_{L^2(\Omega)}$ and $\|\bar{u}-\tilde{u}\|_{L^2(\Omega)}$, respectively, for $h=0.04$, 0.02, 0.01, and 0.005. In the figures, $\bar{u}$ is denoted by u_{opt}.

As one can see, the theoretical predictions are fulfilled and one obtains quadratic convergence for $\|\bar{u}-\tilde{u}\|_{L^2(\Omega)}$. Furthermore, the absolute value of the error is significantly reduced by the projection, as Table 5.1 shows.

Theoretical results with respect to the $L^\infty(\Omega)$-norm were addressed in Corollaries 4.1 and 4.2. Again, we used finer subgrids for the numerical evaluation of the norms (see Figures 5.5 and 5.6).

Example 2. A Neumann boundary problem is studied in this example. In this case, the theoretical results do not exactly fit the problem. The Lax–Milgram theorem implies the existence of weak solutions ($\bar{y},\bar{p}\in H^1(\Omega)$) for the state equation and the adjoint equation. The $W^{2,p}$-regularity of the solution of the adjoint equation

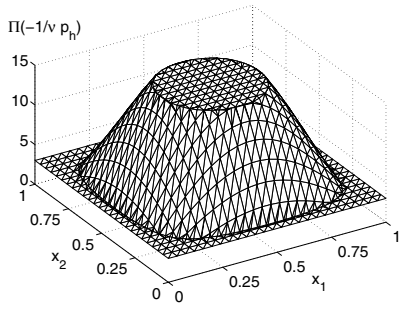


FIG. 5.1. $\tilde{u}$ *at* $h = 0.04$.

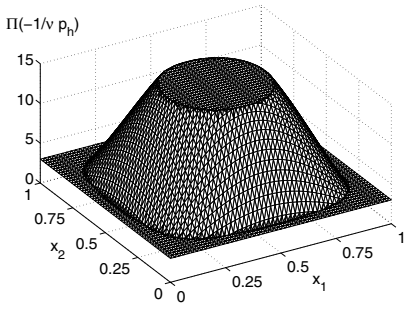


FIG. 5.2. $\tilde{u}$ *at* $h = 0.02$.

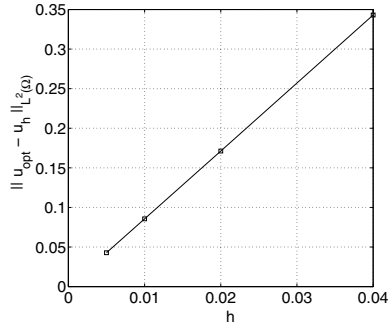


FIG. 5.3. $\|\bar{u} - u_h\|_{L^2(\Omega)}$.

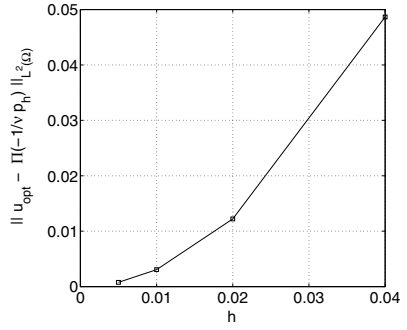


FIG. 5.4. $\|\bar{u} - \tilde{u}\|_{L^2(\Omega)}$.

TABLE 5.1

$h/\sqrt{2}$	$\|\bar{u} - u_h\|_{L^2(\Omega)}$	$\|\bar{u} - \tilde{u}\|_{L^2(\Omega)}$	$\|u_h - w_h\|_{L^2(\Omega)}$	$\|\bar{u} - u_h\|_{L^\infty(\Omega)}$	$\|\bar{u} - \tilde{u}\|_{L^\infty(\Omega)}$
0.04	0.34312	0.04856	0.05335	1.61552	0.13760
0.02	0.17155	0.01221	0.01342	0.81633	0.03513
0.01	0.08556	0.00306	0.00335	0.40975	0.00884
0.005	0.04281	0.00077	0.00084	0.20485	0.00221

is obtained by a result of Grisvard [9, Theorem 4.4.1.2]. For this special structure (Laplace operator, homogeneous Neumann data, $\Omega = \Omega_h$), the proofs of our main results can be adapted. A general discussion of Neumann boundary conditions requires more detailed investigations. We discuss here the problem

$$\begin{aligned} -\Delta y + c\, y &= u \text{ in } \Omega, \\ \partial_n y &= 0 \text{ on } \Gamma, \end{aligned} \tag{5.2}$$

where ∂_n denotes the normal derivative with respect to the outward normal vector. Again, we construct the optimal state $\bar{y}$ by $\bar{y} = y_a - y_g$, with an analytical part $y_a(x_1, x_2) = \cos(\pi\, x_1)\, \cos(\pi\, x_2)$. The function y_g is now determined by the equation

$$\begin{aligned} -\Delta y_g + c\, y_g &= g \text{ in } \Omega, \\ \partial_n y_g &= 0 \text{ on } \Gamma, \end{aligned}$$

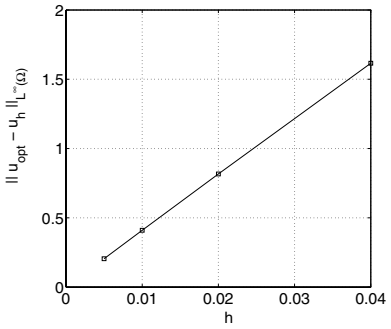


FIG. 5.5. $\|\bar{u} - u_h\|_{L^\infty(\Omega)}$.

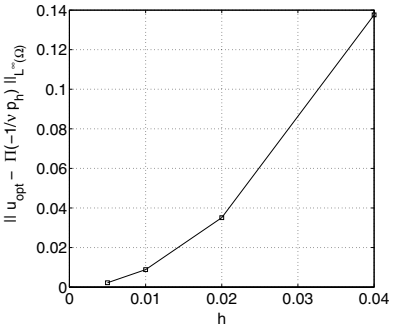


FIG. 5.6. $\|\bar{u} - \tilde{u}\|_{L^\infty(\Omega)}$.

with the inhomogenity

$$g(x_1, x_2) = \begin{cases} u_f(x_1, x_2) - a & \text{if } u_f(x_1, x_2) < a, \\ 0 & \text{if } u_f(x_1, x_2) \in [a, b], \\ u_f(x_1, x_2) - b & \text{if } u_f(x_1, x_2) > b \end{cases}$$

and $u_f(x_1, x_2) = (2\,\pi^2 + c)\,\cos(\pi\,x_1)\,\cos(\pi\,x_2)$. The optimal control $\bar{u}$ is given by (5.2):

$$\bar{u}(x_1, x_2) = \begin{cases} a & \text{if } u_f(x_1, x_2) < a, \\ u_f(x_1, x_2) & \text{if } u_f(x_1, x_2) \in [a, b], \\ b & \text{if } u_f(x_1, x_2) > b. \end{cases}$$

The optimal adjoint state is defined by

$$\bar{p}(x_1, x_2) = -(2\,\pi^2 + c)\,\nu\,\sin(\pi\,x_1)\,\sin(\pi\,x_2).$$

Moreover, the desired state y_d is chosen as

$$\begin{aligned} y_d(x_1, x_2) &= \bar{y} + \Delta\bar{p} - c\,\bar{p} \\ &= y_a - y_g + (4\,\pi^4\,\nu + 4\,\pi^2\,\nu\,c + \nu\,c^2)\,\sin(\pi\,x_1)\,\sin(\pi\,x_2). \end{aligned}$$

Again, it is easy to see that these functions fulfill the necessary and sufficient first order optimality conditions. Assumption (A3) can be verified with the same arguments as in Example 1. In the numerical test, we chose $\nu = c = 1$. The projected function $\tilde{u}$ for $h = 0.04$ and $h = 0.02$ is shown in Figures 5.7 and 5.8. For the visualization of this projection we introduced again new grid points.

Figures 5.9 and 5.10 illustrate that we obtain the same convergence results for this Neumann boundary example.

Comparable with the first example, the absolute error is considerable reduced by the projection, as Table 5.2 shows.

The convergence behavior of the $L^\infty(\Omega)$-errors is illustrated in Figures 5.11 and 5.12.

Figures 5.13 and 5.14 show the convergence behavior of $\|u_h - w_h\|_{L^2(\Omega)}$ (analyzed in Theorem 2.3) for the two examples.

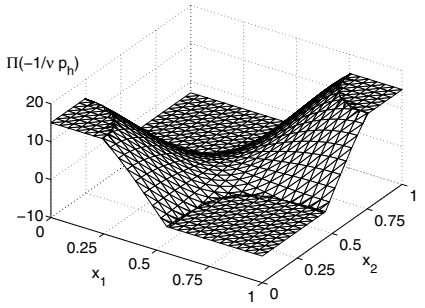


FIG. 5.7. $\tilde{u}$ *at* $h = 0.04$.

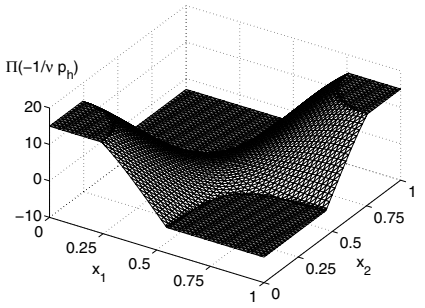


FIG. 5.8. $\tilde{u}$ *at* $h = 0.02$.

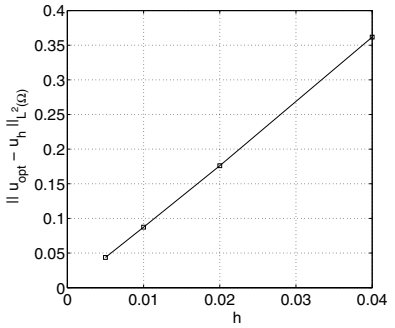


FIG. 5.9. $\|\bar{u} - u_h\|_{L^2(\Omega)}$.

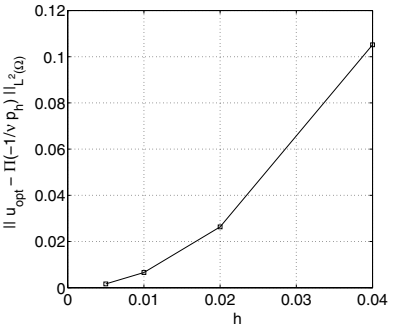


FIG. 5.10. $\|\bar{u} - \tilde{u}\|_{L^2(\Omega)}$.

TABLE 5.2

$h/\sqrt{2}$	$\|\bar{u} - u_h\|_{L^2(\Omega)}$	$\|\bar{u} - \tilde{u}\|_{L^2(\Omega)}$	$\|u_h - w_h\|_{L^2(\Omega)}$	$\|\bar{u} - u_h\|_{L^\infty(\Omega)}$	$\|\bar{u} - \tilde{u}\|_{L^\infty(\Omega)}$
0.04	0.36168	0.10517	0.10659	1.84963	0.21285
0.02	0.17610	0.02632	0.02657	0.89765	0.05352
0.01	0.08744	0.00656	0.00663	0.44174	0.01340
0.005	0.04366	0.00164	0.00166	0.21905	0.00336

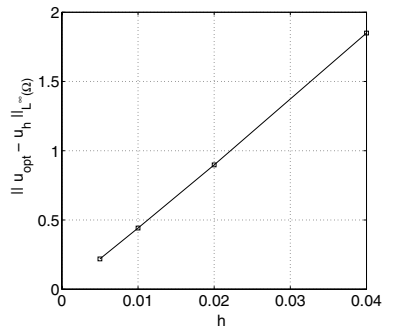


FIG. 5.11. $\|\bar{u} - u_h\|_{L^\infty(\Omega)}$.

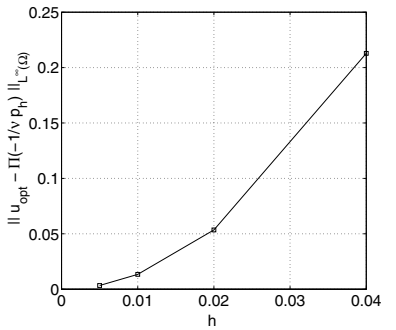


FIG. 5.12. $\|\bar{u} - \tilde{u}\|_{L^\infty(\Omega)}$.

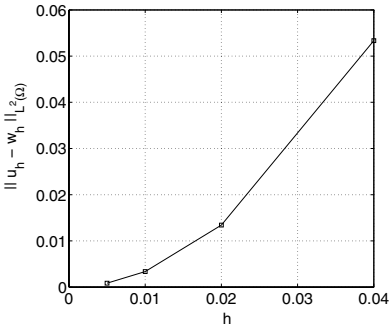


FIG. 5.13. $\|u_h - w_h\|_{L^2(\Omega)}$ *for Example* 1.

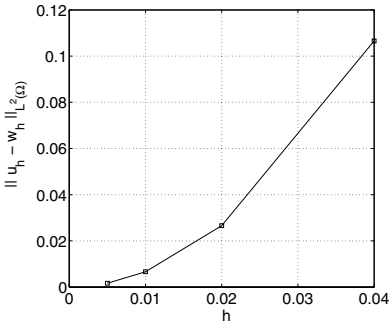


FIG. 5.14. $\|u_h - w_h\|_{L^2(\Omega)}$ *for Example* 2.

Acknowledgment. The authors are very grateful to T. Apel for several helpful discussions.

REFERENCES

[1] N. ARADA, E. CASAS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Comput. Optim. Appl., 23 (2002), pp. 201–229.

[2] F. BONNANS AND E. CASAS, *An extension of Pontryagin's principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities*, SIAM J. Control Optim., 33 (1995), pp. 274–298.

[3] D. BRAESS, *Finite Elements*, Springer-Verlag, Berlin, Heidelberg, 1992.

[4] E. CASAS, M. MATEOS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl., submitted.

[5] E. CASAS AND F. TRÖLTZSCH, *Error estimates for linear-quadratic elliptic control problems*, in Analysis and Optimization of Differential Systems, V. Barbu et al., eds., Kluwer Academic Publishers, Boston, 2003, pp. 89–100.

[6] P. CIARLET, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis Vol. II, Handb. Numer. Anal. II, North–Holland, Amsterdam, 1991, pp. 17–352.

[7] R. FALK, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.

[8] T. GEVECI, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, RAIRO Anal. Numér., 13 (1979), pp. 313–328.

[9] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[10] M. HINZE, *A Generalized Discretization Concept for Optimal Control Problems with Control Constraints*, Tech. rep. MATH-NM-02-2003, Technische Universität Dresden, Dresden, Germany, 2003.

[11] K. KUNISCH AND A. RÖSCH, *Primal-dual active set strategy for a general class of constrained optimal control problems*, SIAM J. Optim., 13 (2002), pp. 321–334.

[12] K. MALANOWSKI, *Convergence of approximations vs. regularity of solutions for convex, control–constrained optimal control problems*, Appl. Math. Optim., 8 (1981), pp. 69–95.

[13] P. RAVIART AND J. THOMAS, *Introduction à l' Analyse Numérique des Équations aux Dérivées Partielles*, Masson, Paris, 1992.

[14] A. RÖSCH, *Error estimates for linear-quadratic control problems with control constraints*, Optim. Methods and Softw., submitted.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 986–1010

A CLASS OF NONLINEAR DEGENERATE INTEGRODIFFERENTIAL CONTROL SYSTEMS*

HANG GAO†, PEIDONG LEI†, AND BO ZHANG‡

Abstract. In this paper, we study the approximate controllability and noncontrollability for a class of nonlinear degenerate integrodifferential control systems. Based on the property of finite propagation for the disturbances, we prove noncontrollability. We also establish sufficient conditions for approximate controllability of the system in $L^2(\Omega)$. Moreover, we prove the global existence and uniqueness of generalized solutions.

Key words. nonlinear degenerate integrodifferential equation, global existence, uniqueness, finite propagation, noncontrollability, approximate controllability

AMS subject classifications. 93B05, 35B37

DOI. 10.1137/S036301290241767X

1. Introduction.

In this paper, we study the control system

$$\begin{cases} \dfrac{\partial}{\partial t}y(x,t) + Ly + \displaystyle\int_0^t f(y(x,s))ds = m(x)u(x,t) \quad \text{in } Q_T, \\ y(x,t) = 0 \ \text{ on } \partial\Omega \times (0,T), \quad y(x,0) = y_0(x) \ \text{ in } \Omega, \end{cases} \tag{1.1}$$

where

$$Ly =: -\mathrm{div}(|\nabla y(x,t)|^{p-2}\nabla y(x,t)), \quad p > 2,$$

$Q_T = \Omega \times (0,T)$, $\Omega \subset R^3$, is a bounded domain with smooth boundary $\partial\Omega$, u is a control input, m is the characteristic function of an open set $\omega \subset \Omega$, while $f : R \to R$ is a given function.

Integrodifferential control systems of the type considered here arise in a variety of applications ranging from heat flow in material with memory to wave propagation. For $p = 2$, (1.1) becomes a standard integrodifferential equation which has a vast literature (see Fitzgibbon [4], Heard [6], Hussain [7], Webb [13]). In this case ($p = 2$), solutions have the property of infinite propagation of disturbances; i.e., a solution with nontrivial nonnegative initial data becomes positive after the initial time. For $p > 2$, (1.1) degenerates if $\nabla y = 0$. This degenerate equation appears in some nonlinear models with concentration dependent mobility. Just as the porous medium equation, solutions of (1.1) have the property of finite propagation. Such a property and the degenerative nature of (1.1) present a significant challenge to many investigators and have been the subject of intensive study in the last two decades. So far, a series of fine theories has been established for this p-Laplacian equation, namely, $f \equiv 0$ in (1.1); see Dibenedetto [3], Gao and Yin [5], Kalashnikov [8], Wu et al. [14], and Yin [15, 16]

*Received by the editors November 12, 2002; accepted for publication (in revised form) January 26, 2004; published electronically October 8, 2004. This research was supported in part by NSF of China (10071012).

http://www.siam.org/journals/sicon/43-3/41767.html

†Department of Mathemathcs, Northeast Normal University, Changchun, Jilin 130024, People's Republic of China (hangg@nenu.edu.cn, leipd168@nenu.edu.cn).

‡Department of Mathemathics and Computer Science, Fayetteville State University, Fayetteville, NC 28301-4298 (bzhang@uncfsu.edu).

for reference. However, to our knowledge, little is known for the nonlinear degenerate integrodifferential equation (1.1).

The goal of this paper is to study the noncontrollability and approximate controllability of the system (1.1). The problem of controllability in systems of partial differential equations, including integrodifferential equations for the case $p = 2$, has been studied by many researchers; we refer to the work of Balachandran, Balasubramaniam, and Dauer [1], Balachandran and Dauer [2], Teresa [10], Teresa and Zuazua [11], Wang and Wang [12], and Zuazua [17] for reference, fundamental theory, and recent development. In this paper, we are concerned with the degenerate case, i.e., $p > 2$.

To give readers an overview of results in the subsequent sections, we outline our main theorems below, beginning with the following definition.

DEFINITION 1.1. *A function y is called a generalized solution of* (1.1) *in Q_T if the following conditions are fulfilled:*

(1) $y \in L^\infty(0,T;W_0^{1,p}(\Omega))$ *and* $\frac{\partial y}{\partial t} \in L^2(Q_T)$.

(2) *For any $\varphi \in C^1(\bar{Q}_T)$ with $\varphi(x,t) = 0$ on $\partial\Omega \times (0,T)$, the following equality holds:*

$$\int_{Q_T} \frac{\partial y}{\partial t}\varphi dxdt + \int_{Q_T} |\nabla y|^{p-2}\nabla y \cdot \nabla\varphi dxdt + \int_{Q_T}\int_0^t f(y(x,s))ds\varphi dxdt = \int_{Q_T} mu\varphi dxdt.$$

(3) $\operatorname{ess\,lim}_{t\to 0^+} \int_\Omega |\nabla(y(x,t) - y_0(x))|^p dx = 0$.

If for any $T > 0$ the control system (1.1) *admits a generalized solution in Q_T, we say that* (1.1) *has a global solution.*

In order to obtain the global existence of a generalized solution, the following assumption is needed.

(H_1) *$f(\cdot)$ is Lipschitz continuous in R with a Lipschitz constant K.*

THEOREM 1.1. *Suppose condition* (H_1) *holds. Then for $y_0 \in W_0^{1,p}(\Omega)$ and $u \in L^q(Q_T)$ with $q > 3$, there exists a unique global solution $y(x,t) = y(x,t;u)$ of problem* (1.1).

Now we define the controllability and noncontrollability of the system (1.1).

We say that system (1.1) is approximately controllable in $L^2(\Omega)$ at time $T > 0$ if the following holds: "For any $y_0 \in W^{1,p}(\Omega)$, the set of reachable states at time T

$$E(T) = \{y(x,T); y \text{ is the solution of (1.1) with } u \in L^2(Q_T)\}$$

is dense in $L^2(\Omega)$." In other words, we say that system (1.1) is approximately controllable in $L^2(\Omega)$ at time $T > 0$ if and only if for any $y_1 \in L^2(\Omega)$ and for all $\varepsilon > 0$, there exists a control function $u \in L^2(Q_T)$ such that

$$\|y(\cdot,T;u) - y_1\|_{L^2(\Omega)} < \varepsilon. \tag{1.2}$$

We say that system (1.1) is noncontrollable at time $T > 0$ if for every $y_0 \in W^{1,p}(\Omega)$ there exists a function $y_1(x) \in L^\infty(\Omega)$ and $\varepsilon_0 > 0$ such that

$$\|y(x,T) - y_1(x)\|_{L^\delta(\Omega)} > \varepsilon_0$$

for any $\delta \geq 1$, where $y(x,t)$ is a generalized solution of (1.1) with $y(x,0) = y_0(x)$.

The following two theorems give the noncontrollability and controllability results.

(H_2) supp $m(x) \subset$ supp $y_0 \subset \Omega$, $f(0) \leq 0$.

THEOREM 1.2. *Suppose conditions* (H_1) *and* (H_2) *hold. If* $y_0 \geq 0$, *then the system* (1.1) *is noncontrollable.*

Remark 1.1. The condition $f(0) \leq 0$ is necessary; that is, if $f(0) > 0$, then Theorem 1.2 fails.

THEOREM 1.3. *Suppose condition* (H_1) *is fulfilled. If* $m(x) \equiv 1$ *in* Ω, *then for any* $T > 0$, (1.1) *is approximately controllable in* $L^2(\Omega)$ *at time* T.

Remark 1.2. It is easy to see that $m(x) \equiv 1$ means $\omega = \Omega$.

We shall prove these theorems in sections 2, 3, and 4, respectively.

2. Global existence and uniqueness. In this section, we discuss the global existence and uniqueness of solutions of (1.1). First, for any $T > 1$, consider the perturbed system

$$\begin{cases} \dfrac{\partial y}{\partial t} + L_n y + \displaystyle\int_0^t f(y(x,s))ds = m(x)u(x,t) \quad \text{in } Q_T, \\ y(x,t) = 0 \ \text{ on } \partial\Omega \times (0,T), \quad y(x,0) = y_0(x) \ \text{ in } \Omega, \end{cases} \tag{2.1}$$

where

$$L_n y = -\operatorname{div}\left(\left(|\nabla y|^2 + \frac{1}{n}\right)^{(p-2)/2} \nabla y\right). \tag{2.2}$$

For f, u, m, y_0 given above and $z \in L^2(0,T;W^{1,p}(\Omega))$, we can find $f_k \in C^2(R)$, $u_k \in C^2(Q_T)$, $m_k \in C^2(\Omega)$, and $y_{0k} \in C_0^2(\Omega)$ such that

$$\int_{Q_T} \left[\int_0^t f_k(z(x,s))ds - \int_0^t f(z(x,s))ds\right]^2 dxdt \to 0,$$
$$\|y_{0k} - y_0\|_{W^{1,p}(\Omega)} \to 0, \quad \|m_k u_k - mu\|_{q,Q_T} \to 0,$$

where q is defined in Theorem 1.1. By the classical theory of parabolic equations (see [9, p. 452]), we know that for any $z \in L^2(0,T;W^{1,p}(\Omega))$, the problem

$$\begin{cases} \dfrac{\partial y}{\partial t} + L_n y + \displaystyle\int_0^t f_k(z(x,s))ds = m_k(x)u_k(x,t) \quad \text{in } Q_T, \\ y(x,t) = 0 \ \text{ on } \partial\Omega \times (0,T), \quad y(x,0) = y_{0k}(x) \ \text{ in } \Omega \end{cases} \tag{2.3}$$

has one and only one classical solution y_k.

LEMMA 2.1. *Let* $z \in L^2(0,T;W^{1,p}(\Omega))$ *and* $y_k(x,t)$ *be a solution of* (2.3). *Then*

$$\|y_k\|_{L^p(0,T;W^{1,p}(\Omega))} \leq C, \tag{2.4}$$

$$\left\|\frac{\partial y_k}{\partial t}\right\|_{L^2(Q_T)} \leq C, \tag{2.5}$$

$$\|L_n y_k\|_{L^\infty(0,T;W^{-1,p'}(\Omega))} \leq C, \tag{2.6}$$

where $\frac{1}{p} + \frac{1}{p'}$, *and* C *is a constant depending on* $T, p, |\Omega|$, $\|u\|_{2,Q_T}$, $\|z\|_{L^2(0,T;W^{1,p}(\Omega))}$, *and* $\|y_0\|_{W^{1,p}(\Omega)}$.

Moreover, if $\|z\|_{L^2(0,T;W^{1,p}(\Omega))} \le C^*$ *(C^* is to be determined later), then there are positive constants* $t_1 \le T$ *such that*

$$\|y_k\|_{L^\infty(0,t_1;W^{1,p}(\Omega))} \le C^*, \tag{2.7}$$

$$\left\| \frac{\partial y_k}{\partial t} \right\|_{L^2(Q_{t_1})} \le C^*, \tag{2.8}$$

$$\|L_n y_k\|_{L^\infty(0,t_1;W^{-1,p'}(\Omega))} \le C^*. \tag{2.9}$$

Proof. Multiplying (2.3) by $\frac{\partial y_k}{\partial t}$ and integrating the resulting relation over $Q_t = \Omega \times (0,t)$, we have

$$\begin{aligned} &\int_{Q_t} \left(\frac{\partial y_k}{\partial t} \right)^2 dx d\tau + \int_{Q_t} \left(|\nabla y_k|^2 + \frac{1}{n} \right)^{(p-2)/2} \nabla y_k \cdot \nabla \frac{\partial y_k}{\partial t} dx d\tau \\ &\le \left\{ \int_{Q_t} \left[\int_0^\tau f_k(z(x,s)) ds \right]^2 dx d\tau + \int_{Q_t} u_k^2 dx d\tau \right\} + \frac{1}{2} \int_{Q_t} \left(\frac{\partial y_k}{\partial t} \right)^2 dx d\tau. \end{aligned} \tag{2.10}$$

Letting $\varepsilon > 0$ $(0 < \varepsilon < 1/2)$, we choose $\|u_k\|^2_{2,Q_t} \le \|u\|^2_{2,Q_t} + \varepsilon$ and

$$\int_{Q_t} \left[\int_0^\tau f_k(z(x,s)) ds \right]^2 dx d\tau \le \int_{Q_t} \left[\int_0^\tau f(z(x,s)) ds \right]^2 dx d\tau + \varepsilon.$$

Use condition (H_1) and apply Hölder's and Poincaré's inequalities to get

$$\begin{aligned} \int_{Q_t} \left[\int_0^\tau f(z(x,s)) ds \right]^2 dx d\tau &\le 2t^2 \int_{Q_t} [K^2 z^2(x,\tau) + f^2(0)] dx d\tau \\ &\le 2t^{\frac{3p-2}{p}} |\Omega|^{\frac{p-2}{p}} K^2 \|z\|^2_{L^2(0,t;W^{1,p}(\Omega))} + 2t^3 |\Omega| f^2(0). \end{aligned}$$

It follows from (2.10) that

$$\begin{aligned} &\frac{1}{2} \int_{Q_{t_1}} \left(\frac{\partial y_k}{\partial t} \right)^2 dx d\tau + \int_{Q_{t_1}} \left(|\nabla y_k|^2 + \frac{1}{n} \right)^{(p-2)/2} \nabla y_k \cdot \nabla \frac{\partial y_k}{\partial t} dx d\tau \\ &\le \|u\|^2_{2,Q_t} + 2t^{\frac{3p-2}{p}} |\Omega|^{\frac{p-2}{p}} K^2 \|z\|^2_{L^2(0,t;W^{1,p}(\Omega))} + 2t^3 |\Omega| f^2(0) + 1. \end{aligned} \tag{2.11}$$

Since

$$\int_{Q_t} \left(|\nabla y_k|^2 + \frac{1}{n} \right)^{(p-2)/2} \nabla y_k \cdot \nabla \frac{\partial y_k}{\partial t} dx d\tau = \frac{1}{p} \int_{Q_t} \frac{d}{d\tau} \left[\left(|\nabla y_k|^2 + \frac{1}{n} \right)^{p/2} \right] dx d\tau,$$

we get from (2.11) that

$$\begin{aligned} &\int_{Q_t} \left(\frac{\partial y_k}{\partial t} \right)^2 dx d\tau + \int_\Omega \left(|\nabla y_k(x,t)|^2 + \frac{1}{n} \right)^{p/2} dx \\ &\le p[\|u\|^2_{2,Q_t} + 2|\Omega|(K^2 t^{\frac{3p-2}{p}} \|z\|^2_{L^2(0,T;W^{1,p}(\Omega))} + f^2(0)t^3) + \|\nabla y_0\|^p_{p,\Omega} + 1]. \end{aligned} \tag{2.12}$$

This yields (2.4) and (2.5). To prove (2.6), we need only to use (2.4) and notice that

$$\left|\left(|\nabla y_k|^2+\frac{1}{n}\right)^{(p-2)/2}\nabla y_k\right|^{p/(p-1)} \le \left(|\nabla y_k|^2+\frac{1}{n}\right)^{p/2}.$$

From (2.12), we can select $t_1>0$ so small that

$$\begin{aligned}\int_{Q_t}\left(\frac{\partial y_k}{\partial t}\right)^2 dxd\tau+\int_\Omega\left(|\nabla y_k(x,t)|^2+\frac{1}{n}\right)^{p/2}dx\\ \le p(\|u\|^2_{2,Q_T}+\|\nabla y_0\|^p_{p,\Omega}+2)=:C^*\end{aligned}\tag{2.13}$$

for all $t\in[0,t_1]$. This completes the proof. □

LEMMA 2.2. *Assume that all conditions of Theorem* 1.1 *hold. Then there exists a unique generalized solution* $y(x,t)$ *of* (2.1).

Proof. For convenience, set $Y=L^2(0,T;W^{1,p}(\Omega))$. First, for any $z\in Y$, we prove the existence of a solution for the following problem:

$$\begin{cases}\dfrac{\partial}{\partial t}y(x,t)+L_ny+\displaystyle\int_0^t f(z(x,s))ds=m(x)u(x,t) & \text{in } Q_T,\\ y(x,t)=0 \text{ on } \partial\Omega\times(0,T), \quad y(x,0)=y_0(x) & \text{in } \Omega.\end{cases}\tag{2.14}$$

In fact, there exist sufficient smooth functions f_k, m_k, u_k, and y_{0k}, which satisfy

$$\int_{Q_T}\left[\int_0^t f_k(z(x,s))ds-\int_0^t f(z(x,s))ds\right]^2 dxdt\to 0,$$
$$\|y_{0k}-y_0\|_{W^{1,p}(\Omega)}\to 0, \quad \|m_ku_k-mu\|_{q,Q_T}\to 0.$$

Let y_k be the solution of (2.3). According to Lemma 2.1 we see that there exist a function $y\in L^\infty(0,T;W^{1,p}(\Omega))$ and a subsequence $\{y_{k_i}\}$ of $\{y_k\}$ such that as $k_i\to\infty$,

$$y_{k_i}(x,t)\to y(x,t) \quad \text{a.e. for } (x,t)\in Q_T;$$

$$y_{k_i}\to y, \quad \text{weakly-* in } \quad L^\infty(0,t_1;W^{1,p}(\Omega));$$

$$\frac{\partial y_{k_i}}{\partial t}\to\frac{\partial y}{\partial t}, \quad \text{weakly in } \quad L^2(Q_{t_1});$$

$$L_ny_{k_i}\to\eta, \quad \text{weakly-* in } \quad L^\infty(0,t_1;W^{-1,p'}(\Omega)).$$

It remains to show that $\eta=L_ny$. For this purpose, we define a functional

$$F_n(y)=\frac{1}{p}\int_\Omega\left(|\nabla y|^2+\frac{1}{n}\right)^{p/2}dx.$$

Then, $F_n(\cdot)$ is a convex functional on $W^{1,p}(\Omega)$, and $F_n'(y)=L_ny$ is a monotone operator. For any $g\in C_0^\infty(Q_T)$, we have

$$\int_0^{t_1}[F_n(g)-F_n(y_{k_i})]dt\ge\int_0^{t_1}\langle L_ny_{k_i},g-y_{k_i}\rangle dt,$$

where $\langle \cdot, \cdot \rangle$ denotes the dual product of $W^{1,p}(\Omega)$ and $W^{-1,p'}(\Omega)$. After a limiting process in the inequality above, we obtain

$$\int_0^{t_1} [F_n(g) - F_n(y)]dt \geq \int_0^{t_1} \langle \eta, g - y \rangle dt,$$

which implies $\eta = F_n'(y)$ immediately. Thus, $y = y(x,t)$ is a generalized solution of (2.14) in Q_{t_1}.

We can use a similar argument to obtain a solution of (2.14) on $(0,T)$. In fact, for the initial function $y(\cdot, t_1) \in W^{1,p}(\Omega)$, we can get a solution $y \in L^\infty(t_1, t_2; W^{1,p}(\Omega))$; that is, we have a solution $y \in L^\infty(0, t_2; W^{1,p}(\Omega))$. Continue this procedure to obtain a function $y \in L^\infty(0,T; W^{1,p}(\Omega))$, which is the solution of (2.14).

For $t_1 (0 < t_1 < 1)$, since $\|y\|_{L^\infty(0,t_1;W^{1,p}(\Omega))} \leq C^*$, we have $\|y\|_{L^2(0,t_1;W^{1,p}(\Omega))} \leq C^*$. We now introduce a set

$$Y_0(0,t_1) = \{y \in Y \ \mid \ \|y\|_{L^2(0,t_1;W^{1,p}(\Omega))} \leq C^*\},$$

where C^* is as given in Lemma 2.1. It is easy to see that $Y_0(0,t_1)$ is a bounded convex closed set in the Banach space $L^2(0,t_1;W^{1,p}(\Omega))$.

From the discussion above, we know that for any $z \in Y_0(0,t_1)$, there is a $y \in Y_0(0,t_1)$, which is the solution of problem (2.14). Now define a map $F : Y_0(0,t_1) \to L^2(0,t_1;W^{1,p}(\Omega))$ by $y = F(z)$. Obviously, $F(Y_0(0,t_1)) \subset Y_0(0,t_1)$. We will show that F has a fixed point y, which is a solution of (2.1).

We need only to prove that F is a continuous compact map. To this end, let $z_{k_1}, z_{k_2} \in Y_0(0,t_1)$, y_{k_1}, y_{k_2} be solutions of (2.14) corresponding to z_{k_1} and z_{k_2}, respectively. Then

$$(2.15)\quad \begin{cases} \dfrac{\partial}{\partial t}(y_{k_1} - y_{k_2}) + (L_n y_{k_1} - L_n y_{k_2}) \\ \qquad = \displaystyle\int_0^t [f(z_{k_2}(x,s)) - f(z_{k_1}(x,s))]ds \quad \text{in } Q_{t_1}, \\ (y_{k_1} - y_{k_2})(x,t) = 0 \quad \text{on } \partial\Omega \times (0,t_1), \quad (y_{k_1} - y_{k_2})(x,0) = 0 \quad \text{in } \Omega. \end{cases}$$

Observe that

$$\int_{Q_{t_1}} \left[\left(|\nabla y_{k_1}|^2 + \frac{1}{n} \right)^{\frac{p-2}{2}} \nabla y_{k_1} - \left(|\nabla y_{k_2}|^2 + \frac{1}{n} \right)^{\frac{p-2}{2}} \nabla y_{k_2} \right] \cdot \nabla (y_{k_1} - y_{k_2}) dxdt$$
$$\geq \left(\frac{1}{n} \right)^{\frac{p-2}{2}} \int_{Q_{t_1}} |(\nabla y_{k_1} - \nabla y_{k_2})|^2 dxdt.$$

Multiply (2.15) by $y_{k_1} - y_{k_2}$ and integrate in Q_{t_1} to derive

$$\int_\Omega (y_{k_1}(x,t_1) - y_{k_2}(x,t_1))^2 dx + \left(\frac{1}{n} \right)^{\frac{p-2}{2}} \int_{Q_{t_1}} |\nabla (y_{k_1} - y_{k_2})|^2 dxdt$$
$$\leq KC(\varepsilon) \int_{Q_{t_1}} |z_{k_1} - z_{k_2}|^2 dxdt + \varepsilon \int_{Q_{t_1}} (y_{k_1} - y_{k_2})^2 dxdt.$$

Here we have used condition (H_1). By Poincaré's inequality, we obtain

$$\int_\Omega (y_{k_1}(x,t_1)-y_{k_2}(x,t_1))^2dx+\frac{1}{2}\left(\frac{1}{n}\right)^{\frac{p-2}{2}}\int_{Q_{t_1}}|\nabla(y_{k_1}-y_{k_2})|^2dxdt$$
$$\le KC(\varepsilon)\int_{Q_{t_1}}|z_{k_1}(x,t)-z_{k_2}(x,t)|^2dxdt.$$

When $\|z_{k_1}-z_{k_2}\|_{L^2(0,t_1;W^{1,p}(\Omega))}\to 0$, we have

$$\|y_{k_1}-y_{k_2}\|_{2,Q_{t_1}}+\|\nabla(y_{k_1}-y_{k_2})\|_{2,Q_{t_1}}\to 0.$$

Since $\|y_{k_i}\|_{L^\infty(0,t_1;W^{1,p}(\Omega))}\le C^*$, we have

$$\|y_{k_1}-y_{k_2}\|_{L^2(0,t_1;W^{1,p}(\Omega))}\to 0.$$

This implies that F is a continuous map.

To prove that F is a compact map, we need to introduce a new space $W(q)=W_q^{2,1}(Q_{t_1})$ (see [9, p. 5]), endowed with the norm

$$\|y\|_W=\|D_{xx}y\|_{q,Q_{t_1}}+\|D_xy\|_{q,Q_{t_1}}+\|D_ty\|_{q,Q_{t_1}}+\|y\|_{q,Q_{t_1}}.$$

Suppose $\{y_k\}\subset F(Y_0(0,t_1))$. Then, there exists $\{z_k\}\subset Y_0(0,t_1)$ such that $y_k=F(z_k)$. At the same time, for any natural number k, there exists $a_k\in C^1(\bar{Q}_T)$ such that $\|(|\nabla y_k|^2+\frac{1}{n})^{(p-2)/2}-a_k\|_{p/(p-2),Q_{t_1}}<\frac{1}{k}$. We consider the following problem:

$$\begin{cases}\dfrac{\partial}{\partial t}y(x,t)-\operatorname{div}\big[a_k(x,t)\nabla y(x,t))\big]vspace.5pc\\ \quad =m(x)u(x,t)-\displaystyle\int_0^t f(z_k(x,s))ds=:\tilde{u}_k(x,t)\quad\text{in } Q_{t_1},\\ y=0\quad\text{on }\partial\Omega\times(0,t_1),\quad y(x,0)=y_0(x)\quad\text{in }\Omega.\end{cases}$$

Since $u\in L^q(Q_T)$ and $z_k\in L^2(0,T;W^{1,p}(\Omega))$, we have $\tilde{u}_k\in L^{q^*}(Q_T)$, where $q^*<3$. From [9, Chapter 4], we know that there exists a solution $\tilde{y}_k\in W(q^*)$ for the problem above, and $\|\tilde{y}_k\|_{W(q^*)}\le C\|\tilde{u}_k\|_{q^*,Q_T}\le C$. According to Sobolev's embedding theorem, we have $|D_x\tilde{y}_k|\le C$, and there exists a subsequence $\{\tilde{y}_{k_i}\}\subset\{\tilde{y}_k\}$ such that $\{\tilde{y}_{k_i}\}$ is a convergent sequence in $L^2(0,t_1;W^{1,p}(\Omega))$. Since

$$\begin{cases}\dfrac{\partial}{\partial t}(y_k-\tilde{y}_k)-\operatorname{div}\left[\left(|\nabla y_k|^2+\dfrac{1}{n}\right)^{\frac{p-2}{2}}\nabla(y_k-\tilde{y}_k)\right]\\ \qquad=\operatorname{div}\left[\left(\left(|\nabla y_k|^2+\dfrac{1}{n}\right)^{\frac{p-2}{2}}-a_k\right)\nabla\tilde{y}_k\right],\\ (y_k-\tilde{y}_k)|_{\partial\Omega\times(0,t_1)}=0,\quad (y_k-\tilde{y}_k)(x,0)=0,\end{cases}$$

we can obtain

$$\|\tilde{y}_k-y_k\|_{L^2(0,t_1;W^{1,p}(\Omega))}\le\frac{C}{k}.$$

Thus, $\{y_{k_i}\}$ is a convergent sequence in $Y_0(0,t_1)$, and F is a compact map.

Applying Schauder's fixed point theorem, we know that F has a fixed point y, which is a solution of (2.1) in Q_{t_1}. Similarly, for the initial function $y(\cdot, t_1) \in W^{1,p}(\Omega)$, we get a solution $y \in Y_0(t_1, t_2)$, that is, a solution $y \in Y_0(0, t_2)$. By repeating the discussion above, we obtain a function $y \in Y_0(0, T)$, which is a solution of (2.1). □

Now, we begin to prove Theorem 1.1.

Proof of Theorem 1.1. Suppose that $y_n \in L^2(0, T; W^{1,p}(\Omega))$ is the solution of problem (2.1). From Lemma 2.1, we see that there exist a function $y \in Y$ and a subsequence $\{y_{n_k}\}$ of $\{y_n\}$ such that as $n_k \to \infty$,

$$y_{n_k}(x,t) \to y(x,t) \quad \text{a.e. } (x,t) \in Q_T;$$

$$y_{n_k} \to y, \quad \text{weakly-* in} \quad L^\infty(0, T; W^{1,p}(\Omega));$$

$$\frac{\partial y_{n_k}}{\partial t} \to \frac{\partial y}{\partial t}, \quad \text{weakly in} \quad L^2(Q_T);$$

$$L_{n_k} y_{n_k} \to \eta, \quad \text{weakly-* in} \quad L^\infty(0, T; W^{-1,p'}(\Omega)).$$

Next, we show that $\eta = Ly$. To this end, we introduce two functionals

$$F_n(y) = 1/p \int_\Omega \left(|\nabla y|^2 + \frac{1}{n} \right)^{p/2} dx, \quad F(y) = 1/p \int_\Omega (|\nabla y|^2)^{p/2} dx.$$

It is clear that $F_n(\cdot)$ is a convex functional on $W^{1,p}(\Omega)$, and $F_n'(y) = L_n y$ is a monotone operator. For any $g \in C_0^\infty(Q_T)$, we have

$$\int_0^T [F_{n_k}(g) - F_{n_k}(y_{n_k})] dt \geq \int_0^T \langle L_{n_k} y_{n_k}, g - y_{n_k} \rangle dt,$$

where $\langle \cdot, \cdot \rangle$ denotes the dual product of $W^{1,p}(\Omega)$ and $W^{-1,p'}(\Omega)$. After a limiting process in the inequality above, we obtain

$$\int_0^T [F(g) - F(y)] dt \geq \int_0^T \langle \eta, g - y \rangle dt,$$

which implies $\eta = F'(y)$ immediately. Thus, $y \in L^2(0, T; W_0^{1,p}(\Omega))$ is a solution of (1.1). Furthermore, similar to Lemma 2.1, $y \in L^\infty(0, T; W_0^{1,p}(\Omega))$ is a solution of (1.1).

Finally, we prove the uniqueness of the generalized solution of (1.1). Set $z(x,t) = e^{-\beta t} y(x,t)$, where $y(x,t)$ is the solution of problem (1.1) and $\beta > \frac{K^2 T^2 + 1}{2}$ is a constant. Then,

$$\begin{cases} \dfrac{\partial}{\partial t} z(x,t) - \operatorname{div}(|\nabla y(x,t)|^{p-2} \nabla z(x,t)) + \beta z(x,t) \\ \qquad + \ e^{-\beta t} \displaystyle\int_0^t f(y(x,s)) ds = e^{-\beta t} m(x) u(x,t) \quad \text{in } Q_T, \\ z = 0 \quad \text{on } \partial\Omega \times (0,T), \quad z(x,0) = y_0(x) \quad \text{in } \Omega. \end{cases}$$

Let $y_1(x,t)$ and $y_2(x,t)$ be two generalized solutions of (1.1). Then

$$\begin{cases} \dfrac{\partial}{\partial t}(z_1-z_2) - \operatorname{div}\{|\nabla y_1|^{p-2}\nabla z_1 - |\nabla y_2|^{p-2}\nabla z_2\} + \beta(z_1-z_2) \\ \qquad + \ e^{-\beta t}\displaystyle\int_0^t [f(y_1(x,s)) - f(y_2(x,s))]ds = 0 \quad \text{in } \ Q_T, \\ (z_1-z_2) = 0 \quad \text{on } \partial\Omega\times(0,T), \quad (z_1-z_2)(x,0) = 0 \quad \text{in } \Omega. \end{cases} \tag{2.16}$$

We proceed to multiply (2.16) by $z_1 - z_2$ and integrate in Q_T. Taking into account the inequality

$$\int_Q \{|\nabla y_1|^{p-2}\nabla z_1 - |\nabla y_2|^{p-2}\nabla z_2\}\cdot\nabla(z_1-z_2)dxdt \geq 0, \tag{2.17}$$

we obtain

$$\begin{aligned} &\int_\Omega [z_1\ (x,T) - z_2(x,T)]^2dx + \beta\int_{Q_T}[z_1(x,t)-z_2(x,t)]^2dxdt \\ &\qquad \leq \int_{Q_T} e^{-\beta t}\left(\int_0^t Ke^{\beta s}|z_1(x,s)-z_2(x,s)|ds\right)|z_1(x,t)-z_2(x,t)|dxdt \\ &\qquad \leq \frac{1}{2}\int_{Q_T} e^{-2\beta t}\left(\int_0^t e^{\beta s}K|z_1(x,s)-z_2(x,s)|ds\right)^2 dxdt \\ &\qquad + \ \frac{1}{2}\int_{Q_T}|z_1(x,t)-z_2(x,t)|^2dxdt \\ &\qquad \leq \frac{K^2T^2+1}{2}\int_{Q_T}|z_1(x,t)-z_2(x,t)|^2dxdt. \end{aligned}$$

Since $\beta > \frac{K^2T^2+1}{2}$, we have $\int_{Q_T}|z_1(x,t)-z_2(x,t)|^2dxdt \leq 0$. Therefore, $y_1(x,t) = y_2(x,t)$, and the proof is complete. □

3. Noncontrollability. In this section, we discuss the noncontrollability of system (1.1). In order to obtain the noncontrollability, we need to prove that the speed of propagation for disturbances is finite. First, we prove the following lemma.

LEMMA 3.1. *Suppose* $f(0) \leq 0$ *and* (H$_1$) *holds. Let* $y(x,t)$ *be the solution of* (1.1) *with* $y_0(x) \geq 0$ *and* $u(x,t) \geq 0$. *Then* $y(x,t) \geq 0$ *in* Q.

Proof. Denote $y_- = \max\{-y, 0\}$. Multiplying (1.1) by y_- and integrating over Q_t, we obtain

$$\begin{aligned} &\frac{1}{2}\int_\Omega y_-^2(x,t)dx - \frac{1}{2}\int_\Omega y_-^2(x,0)dx + \int_{Q_t}|\nabla y_-(x,s)|^p dxds \\ &= \int_{Q_t}\int_0^t f(y(x,\tau))d\tau y_-(x,s)dxds - \int_{Q_t} m(x)u(x,s)y_-(x,s)dxds. \end{aligned}$$

Since $y_0(x) \geq 0$, $u(x,t) \geq 0$, we have $y_-^2(x,0) = 0$ a.e. in Ω and $u(x,s)y_-(x,s) \geq 0$ a.e. in Q_t. Thus

$$\frac{1}{2}\int_\Omega y_-^2(x,t)dx + \int_{Q_t}|\nabla y_-(x,s)|^p dxds \leq \int_{Q_t}\int_0^t f(y(x,\tau))d\tau y_-(x,s)dxds. \tag{3.1}$$

Using $f(0) \le 0$, condition (H_1), and Hölder's inequality, we obtain

$$\begin{aligned}
&\int_{Q_t} \int_0^t f(y(x,\tau))d\tau y_-(x,s)dxds \\
&\le K \int_{Q_t} \int_0^t |y(x,\tau)|d\tau y_-(x,s)dxds + tf(0) \int_{Q_t} y_-(x,s)dxds \\
&\le K \left(\int_{Q_t} \left(\int_0^t |y_-(x,\tau)|d\tau \right)^2 dxds \right)^{1/2} \left(\int_{Q_t} y_-^2(x,s)dxds \right)^{1/2} \\
&\le KT \int_{Q_t} y_-^2(x,s)dxds.
\end{aligned} \tag{3.2}$$

Substituting (3.2) into (3.1), we have

$$\int_\Omega y_-^2(x,t)dx \le KT \int_{Q_t} y_-^2(x,s)dxds.$$

By Gronwall's inequality, we get $\int_\Omega y_-^2(x,t)dx = 0$ immediately, which implies that $y(x,t) \ge 0$.

Now, we turn to prove the property of finite propagation of disturbances. Denote

$$\begin{aligned}
&G = \{(x,t) : y(x,t) > 0, t > 0\}, \quad G(t) = \{x : y(x,t) > 0\}, \\
&P = \partial G \cap \{(x,t) : t > 0\}, \quad P(t) = \partial G(t).
\end{aligned}$$

If $y(x,t)$ is continuous, then G and $G(t)$ are open sets of Q_T and Ω, respectively. We will call P the interface of the generalized solution $y(x,t)$.

PROPOSITION 3.1. *Suppose conditions* (H_1) *and* (H_2) *hold. If* $y_0 \ge 0$ *and* $u \ge 0$, *then there exists a time* $t^* > 0$ *such that* supp $y(\cdot,t) \subset \Omega$ *for* $0 < t < t^*$.

Proof. First, we consider a simple case, namely, supp $y_0 \subset \Pi_{i=1}^3 [\alpha_i, \beta_i] \subset \Omega$, where $[\alpha_i, \beta_i]$ for $i = 1,2,3$ is a closed bounded interval in R.

Set $z(x,t) = e^{-\beta t} y(x,t)$, $\beta > KT + KT^2$, where K is the Lipschitz constant given in (H_1). Then, supp$y(\cdot,t) =$ supp$z(\cdot,t)$. For each $r_1 \ge 0$, define

$$\Omega(r_1) = \{x | x = (x_1, x') \in \Omega, x_1 \le r_1\}.$$

We now prove supp$y(\cdot,t) \subset \Omega(\alpha_1(t), \beta_1(t))$, where

$$\Omega(\alpha_1(t), \beta_1(t)) = (\Omega \setminus \Omega(\alpha_1(t))) \cap \Omega(\beta_1(t))$$

for some functions $\alpha_1(t)$ and $\beta_1(t)$ to be determined later. From (1.1), we have

$$\begin{cases}
\dfrac{\partial}{\partial t} z(x,t) - \operatorname{div}(|\nabla y(x,t)|^{p-2} \nabla z(x,t)) + \beta z(x,t) \\
\qquad = e^{-\beta t} m(x) u(x,t) - e^{-\beta t} \displaystyle\int_0^t f(y(x,s))ds \quad \text{in } Q_T, \\
z(x,t) = 0 \text{ on } \partial\Omega \times (0,T), \quad z(x,0) = y_0(x).
\end{cases} \tag{3.3}$$

Let $\varphi(x, r_1) = (x_1 - r_1)^+$, $r_1 \ge \beta_1$ (given in (H_2)), where $(\cdot)^+ = \max\{\cdot, 0\}$, and

$x = (x_1, x_2, x_3)$. Taking $z\varphi^k$ as a test function, we get from Definition 1.1 that

$$\begin{aligned}
&\frac{1}{2}\int_\Omega z^2(x,t)\varphi^k dx + \int_{Q_t}(e^{\beta\tau}|\nabla z|)^{p-2}\nabla z\cdot\nabla(z\varphi^k)dxd\tau + \beta\int_{Q_t} z^2\varphi^k dxd\tau \\
&\quad = \int_{Q_t}\left[m(x)u(x,\tau) - \int_0^\tau f(e^{\beta s}z(x,s))ds\right]e^{-\beta\tau}z\varphi^k dxd\tau + \frac{1}{2}\int_\Omega y_0^2(x)\varphi^k dx \\
&\quad \le \int_{Q_t}\int_0^\tau [e^{-\beta(\tau-s)}Kz(x,s) + e^{-\beta\tau}f(0)]ds(z\varphi^k)dxd\tau \qquad (3.4)\\
&\quad \le K\int_{Q_t}\int_0^\tau z(x,s)ds(z\varphi^k)dxd\tau = \frac{1}{2}K\int_{Q_t}\frac{d}{d\tau}\left[\int_0^\tau z(x,s)ds\right]^2\varphi^k dxd\tau \\
&\quad = \frac{1}{2}K\int_\Omega\left[\int_0^t z(x,s)ds\right]^2\varphi^k dx \le \frac{1}{2}KT\int_{Q_t} z^2\varphi^k dxds.
\end{aligned}$$

Since $\beta > KT$, we obtain

$$\frac{1}{2}\int_\Omega z^2(x,t)\varphi^k dx + \int_{Q_t}(e^{\beta\tau}|\nabla z|)^{p-2}\nabla z\cdot\nabla(z\varphi^k)dxd\tau \le 0. \tag{3.5}$$

Next, observe

$$\begin{aligned}
&\int_{Q_t}(e^{\beta\tau}|\nabla z|)^{p-2}\nabla z\cdot\nabla(z\varphi^k)dxd\tau \\
&= \int_{Q_t}|e^{\beta\tau}\nabla z|^{p-2}|\nabla z|^2\varphi^k dxd\tau + k\int_{Q_t}|e^{\beta\tau}\nabla z|^{p-2}\nabla z\cdot\nabla\varphi\cdot z\varphi^{k-1}dxd\tau \\
&\ge \int_{Q_t} e^{\beta(p-2)\tau}|\nabla z|^p\varphi^k dxd\tau - k\int_{Q_t} e^{\beta(p-2)\tau}|\nabla z|^{p-1}\cdot z\varphi^{k-1}dxd\tau.
\end{aligned}$$

From this inequality and (3.5), we have

$$\int_{Q_t}|\nabla z(x,\tau)|^p\varphi^k dxd\tau \le C_1\int_{Q_t}|\nabla z(x,\tau)|^{p-1}\cdot z\varphi^{k-1}dxd\tau. \tag{3.6}$$

Choose a sufficiently large k such that $k - p - p^2 > 0$. Using Young's inequality on the right-hand side of (3.6) with $\mu = \frac{p}{p-1}$ and $\mu' = p$, we obtain

$$\int_{Q_t}|\nabla z|^p\varphi^k dxd\tau \le \frac{1}{2}\int_{Q_t}|\nabla z|^p\varphi^k dxd\tau + C_2\int_{Q_t} z^p\varphi^{k-p}dxd\tau.$$

This yields

$$\int_{Q_t}|\nabla z(x,\tau)|^p\varphi^k dxd\tau \le 2C_2\int_{Q_t} z^p(x,\tau)\varphi^{k-p}dxd\tau. \tag{3.7}$$

In order to estimate the right-hand side of (3.7), we take $z^{p-1}\varphi^{k-p}$ as a test function

and use an argument similar to that of (3.4) to obtain

$$\begin{aligned}
&\frac{1}{p}\int_\Omega z^p(x,t)\varphi^{k-p}dx + \int_{Q_t} |e^{\beta t}\nabla z|^{p-2}\nabla z\cdot\nabla(z^{p-1}\varphi^{k-p})dxd\tau \\
&\quad + \ \beta\int_{Q_t} z^p\varphi^{k-p}dxd\tau \le K\int_{Q_t}\left[\int_0^\tau z(x,s)ds\right] z^{p-1}(x,\tau)\varphi^{k-p}dxd\tau \\
&\quad \le K\int_{Q_t}\left\{\int_0^t z^p(x,s)ds\right\}^{\frac{1}{p}} t^{(p-1)/p} z^{p-1}(x,\tau)\varphi^{k-p}dxd\tau \\
&\quad \le Kt^{(2p-1)/p}\int_\Omega\left\{\int_0^t z^p(x,s)ds\right\}^{\frac{1}{p}}\left\{\int_0^t z^p(x,\tau)d\tau\right\}^{(p-1)/p} t^{1/p}\varphi^{k-p}dx \\
&\quad \le KT^2\int_{Q_t} z^p(x,\tau)\varphi^{k-p}dxd\tau.
\end{aligned}$$

Taking into account $\beta > KT^2$, we have

$$\frac{1}{p}\int_\Omega z^p(x,t)\varphi^{k-p}dx + \int_{Q_t} |e^{\beta t}\nabla z)|^{p-2}\nabla z\cdot\nabla(z^{p-1}\varphi^{k-p})dxd\tau \le 0. \tag{3.8}$$

Since

$$\begin{aligned}
&\int_{Q_t} |e^{\beta\tau}\nabla z|^{p-2}\nabla z\cdot\nabla(z^{p-1}\varphi^{k-p})dxd\tau \\
&\ge (p-1)\int_{Q_t} e^{(p-2)\beta\tau}|\nabla z|^p z^{p-2}\varphi^{k-p}dxd\tau \\
&\quad - \ (k-p)\int_{Q_t} e^{(p-2)\beta\tau}|\nabla z|^{p-1}z^{p-1}\varphi^{k-p-1}dxd\tau,
\end{aligned}$$

we obtain

$$\frac{1}{p}\int_\Omega z^p(x,t)\varphi^{k-p}dx \le C_3\int_{Q_t}|\nabla z|^{p-1}z^{p-1}\varphi^{k-p-1}dxd\tau$$

and thus

$$\sup_{\tau\in[0,t]}\int_\Omega z^p(x,\tau)\varphi^{k-p}dx \le C_3 p\int_{Q_t}|\nabla z|^{p-1}z^{p-1}\varphi^{k-p-1}dxd\tau.$$

Substituting this inequality into (3.7), we find

$$\int_{Q_t}|\nabla z|^p\varphi^k dxd\tau \le C_4 t\int_{Q_t}|\nabla z|^{p-1}z^{p-1}\varphi^{k-p-1}dxd\tau. \tag{3.9}$$

Apply Young's inequality to obtain

$$\begin{aligned}
&C_4 t\int_{Q_t}|\nabla z|^{p-1}z^{p-1}\varphi^{k-p-1}dxd\tau \\
&\le \frac{1}{2}\int_{Q_t}|\nabla z|^p\varphi^k dxd\tau + C_5 t^p\int_{Q_t} z^{(p-1)p}\varphi^{k-p-p^2}dxd\tau.
\end{aligned}$$

Combine this inequality and (3.9) to get

$$\int_{Q_t} |\nabla z|^p \varphi^k dx d\tau \le 2C_5 t^p \int_{Q_t} z^{(p-1)p} \varphi^{k-p-p^2} dx d\tau. \tag{3.10}$$

Again, using Sobolev's embedding theorem, we obtain

$$\int_{Q_t} z^{(p-1)p} \varphi^{k-p-p^2} dx d\tau \le C_6 \|z\|^{(p-1)p}_{1,p,Q(r_1)_t}, \tag{3.11}$$

where $Q(r_1)_t = \{(x,\tau)|\ \ x \in \Omega \setminus \Omega(r_1), \tau \in (0,t)\}$. Finally, we get from Poincaré's inequality that

$$\int_{Q_t} |\nabla z(x,\tau)|^p \varphi^k dx d\tau \le C_7 t^p \left(\int_{Q(r_1)_t} |\nabla z(x,\tau)|^p dx d\tau \right)^{p-1}. \tag{3.12}$$

Next, define

$$g_k(r_1) = \int_{Q_t} |\nabla z(x,\tau)|^p \varphi^k dx d\tau, \quad g_0(r_1) = \int_{Q(r_1)_t} |\nabla z(x,\tau)|^p dx d\tau.$$

Then

$$g_k(r_1) \le C t^p (g_0(r_1))^{p-1}.$$

Using Hölder's inequality, we obtain

$$\begin{aligned} g_1(r_1) &\le \left(\int_{Q_t} |\nabla z|^p \varphi^k dx d\tau \right)^{\frac{1}{k}} \left(\int_{Q(r_1)_t} |\nabla z|^p dx d\tau \right)^{\frac{k-1}{k}} \\ &\le C t^{\frac{p}{k}} (g_0(r_1))^{\frac{p-2+k}{k}}. \end{aligned}$$

Note that $\frac{p-2+k}{k} > 1$. Letting $\frac{p-2+k}{k} = 1+\theta$ ($\theta = \frac{p-2}{k} > 0$) and taking into account $g_1'(r_1) = -g_0(r_1)$, we have

$$g_1(r_1) \le C t^{\frac{p}{k}} (-g_1'(r_1))^{1+\theta}.$$

This implies

$$g_1'(r_1) \le -C t^{-\frac{p}{k(1+\theta)}} (g_1(r_1))^{\frac{1}{1+\theta}}. \tag{3.13}$$

If $g_1(\beta_1) = 0$, then $\nabla z(x,t) = 0$ for $(x,t) \in Q \setminus \Omega(\beta_1)_T$. Since $z(x,t) = 0$ on $\partial\Omega \times (0,T)$, we see that $z(x,t) = 0$ for $(x,t) \in Q \setminus \Omega(\beta_1)_T$. Thus we have supp $y(\cdot,t) \subset \Omega(\beta_1)$. If $g_1(\beta_1) > 0$, then there exists a maximal interval (β_1, β_1^*) on which $g_1(r_1) > 0$ and

$$(g_1(r_1)^{\frac{\theta}{1+\theta}})' = \frac{\theta}{1+\theta} \left(\frac{g_1'(r_1)}{[g_1(r_1)]^{\frac{1}{1+\theta}}} \right) \le -C t^{\frac{-p}{k(1+\theta)}}.$$

Integrating the above inequality over (β_1, β_1^*), we obtain

$$-(g_1(\beta_1)^{\frac{\theta}{1+\theta}}) \le -C(\beta_1^* - \beta_1) t^{-\frac{p}{k(1+\theta)}},$$

which implies that

$$\beta_1^* \le \beta_1 + C(p)t^{\frac{p}{k(1+\theta)}}[g_1(\beta_1)]^{\frac{\theta}{1+\theta}} \le \beta_1 + C(p)t^{\frac{p}{p-2+k}} =: \beta_1(t), \tag{3.14}$$

where $C(p)$ is a constant depending only upon p. This shows supp $y(\cdot,t) \subset \Omega(\beta_1(t))$.

Similarly, supp $y(\cdot,t) \subset \Omega \setminus \Omega(\alpha_1(t))$. Hence, supp $y(\cdot,t) \subset \Omega(\alpha_1(t),\beta_1(t))$. The discussion in the direction of x_i is similar to that for x_1. We finally obtain supp $y(\cdot,t) \subset \Omega(\alpha_i(t),\beta_i(t))$, $i = 1,2,3$.

Now we treat the general case, that is, supp$y_0 \subset \Omega$. Obviously, we can select a finite closed covering D of suppy_0 with $D = \cup_{j=1}^n D^{(j)} \subset \Omega$ and $D^{(j)} = \Pi_{i=1}^3[\alpha_i^{(j)},\beta_i^{(j)}]$, where $[\alpha_i^{(j)},\beta_i^{(j)}]$ for $i = 1,2,3,\ j = 1,2,\dots,n$, is a closed bounded interval in R. Moreover, we can require that the profile of $D^{(j)}$ is contained in suppy_0 (we call it the inner profile) or the cap set of the profile and the inner of suppy_0 is empty (we call it the outer profile). We are interested in the cuboids which are not contained in suppy_0 only. As a matter of convenience, we denote these cuboids by $D^{(1)},\dots,D^{(n_0)}$, where $n_0 \le n$.

For the cuboid $D^{(j)} = \Pi_{i=1}^3[\alpha_i^{(j)},\beta_i^{(j)}]$ $(j = 1,2,\dots,n_0)$, we can also obtain the finite propagation of the outer profile of $D^{(1)}$. Using the same technique for deriving (3.14), we can establish the estimate of the speed of finite propagation, and hence the proof is complete. □

Proof of Theorem 1.2. For simplicity, we assume that supp$y_0 \subset \Pi_{i=1}^3[\alpha_i,\beta_i] \subset \Omega$. We say that the interface P of the generalized solution y reaches $\partial\Omega$ "fully" at the time t if $P(t) = \partial\Omega$. Set

$$t^* = \inf\{t > 0 : P \text{ reaches } \partial\Omega \text{ fully}\}.$$

Now we estimate the time t^*. For any $z = (z_1,z_2,z_3) \in \partial\Pi_{i=1}^3[\alpha_i,\beta_i]$, if $z_i = \alpha_i$ or $z_i = \beta_i$, denote the line

$$l_i = \{x = z + te_i : \quad e_i \text{ is the unit normal of the hyperplane } x_i = 0\}$$

and the opposite point of z

$$w(z) = \begin{cases} w_\alpha(z) = \partial\Omega \cap \{x \in \Omega : x_i \le \alpha_i\} \cap l_i & \text{if } z_i = \alpha_i, \\ w_\beta(z) = \partial\Omega \cap \{x \in \Omega : x_i \ge \beta_i\} \cap l_i & \text{if } z_i = \beta_i. \end{cases}$$

Set

$$d_z = dist(z,\partial\Omega) = |z - w(z)|, \quad d = \sup_{z \in \partial\Pi_{i=1}^3[\alpha_i,\beta_i]} d_z.$$

In order to ensure that the interface of y reaches $\partial\Omega$, we get from the estimate of the speed of propagation (3.14) that

$$C(p)t^{\frac{p}{p-2+k}} \ge d$$

must be fulfilled. This implies that

$$t^* \ge C(p)d^{\frac{p-2+k}{p}}. \tag{3.15}$$

Using an argument similar to that in the proof of Proposition 3.1, we can obtain inequality (3.15) for the general case.

Thus, for any $0 < t < C(p)d^{\frac{p-2+k}{p}}$ and any input control u,

$$\sup\{|x-y| : \quad x \in \text{supp } y(\cdot,t;u), y \in \partial\Omega\} > 0.$$

This means that we can select a measurable subset $D \subset \Omega \backslash \text{supp} y(\cdot,t;u)$ with positive measure, such that $y = 0$ in D, and hence the system (1.1) is noncontrollable. The proof is complete. □

4. Approximate controllability. Under the condition $m(x) \equiv 1$ in Ω, the speed of propagation of the disturbances is infinite provided that $u(x,t) > 0$. In this section, we will discuss the approximate controllability of (1.1); that is, for all $t_1 > 0$, for all $\varepsilon > 0$, for all $y_1 \in L^2(\Omega)$, there exists a control function u such that $\|y(\cdot,t_1;u) - y_1\| \le \varepsilon$. For this purpose, we set $z(x,t) = e^{-\beta t}y(x,t)$ (β is to be determined later). Consider the control system

$$\begin{cases} \dfrac{\partial}{\partial t}z(x,t) - \text{div}(|\nabla e^{\beta t}z(x,s)|^{p-2}\nabla z(x,t)) + \beta z(x,t) \\ \quad + e^{-\beta t}\displaystyle\int_0^t f(e^{\beta t}z(x,s))ds = e^{-\beta t}u(x,t) \quad \text{in } \ Q_T, \\ z(x,t) = 0 \ \text{ on } \ \partial\Omega \times (0,T), \quad z(x,0) = y_0(x) \ \text{ in } \ \Omega. \end{cases} \tag{4.1}$$

Obviously, system (1.1) is approximately controllable if and only if system (4.1) is approximately controllable at time t_1. So, we need only to discuss the approximate controllability of (4.1) at t_1.

Remark 4.1. The choice of β depends not only on K but also on t_1.

We now introduce a Hilbert space $V_2(Q_T)$, endowed with the norm

$$\|y\|_V = \max_{t\in(0,T)} \|y(\cdot,t)\|_{2,\Omega} + \|D_x y\|_{2,Q_T}.$$

Define $V_2^{1,0}(Q_T) = V_2(Q_T) \cap C(0,T;L^2(\Omega))$, and set

$$V =: V_2^{1,1/2}(Q_T) = \left\{ z \in V_2^{1,0}(Q_T) \ \middle| \ \lim_{h\to 0}\int_0^{T-h}\int_\Omega h^{-1}[z(x,t+h) - z(x,t)]^2 = 0 \right\}.$$

First, we discuss the approximate controllability of the system

$$\begin{cases} \dfrac{\partial}{\partial t}z(x,t) + \tilde{L}_n z + \beta z(x,t) \\ \quad + e^{-\beta t}\displaystyle\int_0^t f(e^{\beta t}z(x,s))ds = e^{-\beta t}u(x,t) \quad \text{in } \ Q_T, \\ z(x,t) = 0 \ \text{ on } \ \partial\Omega \times (0,T), \quad z(x,0) = y_0(x) \ \text{ in } \ \Omega, \end{cases} \tag{4.2}$$

where

$$\tilde{L}_n z = -\text{div}\left(\left(|e^{\beta t}\nabla z(x,t)|^2 + \frac{1}{n}\right)^{(p-2)/2}\nabla z(x,t)\right);$$

that is, for any $z_1 \in L^2(\Omega)$ and $\varepsilon > 0$, we can find a $u \in L^p(Q_T)$ such that $\|z(\cdot,t_1) - z_1\|_{2,\Omega} \le \varepsilon$.

We endow a new norm to $z \in V$:

$$|z|_V^2 = \left(\frac{1}{n}\right)^{(p-2)/2} \|D_x z\|_{2,Q_T}^2 + \beta \|z\|_{2,Q_T}^2$$

for fixed n. Then $(V, |\cdot|_V)$ is a Banach space. Let

$$Z = \{z \ \mid \ |z|_V \le C^*\},$$

where C^* is a constant to be determined later. It is obvious that Z is a bounded closed convex set in V. For any $w \in Z$, we discuss the approximate controllability of the system

$$\text{(4.3)} \quad \begin{cases} \dfrac{\partial}{\partial t} z(x,t) + \tilde{L}_n(w) z + \beta z + e^{-\beta t} \displaystyle\int_0^t f(e^{\beta s} w(x,s)) ds = u(x,t) \quad \text{in } \ Q_T, \\ z(x,t) = 0 \ \text{ on } \ \partial\Omega \times (0,T), \quad z(x,0) = 0 \ \text{ in } \ \Omega, \end{cases}$$

where

$$\tilde{L}_n(w) z =: -\text{div}\left(\left(|e^{\beta t} \nabla w(x,t)|^2 + \frac{1}{n} \right)^{\frac{p-2}{2}} \nabla z(x,t) \right).$$

From [9, pp. 153–157], we know that for any $u \in L^p(Q_T)$, there exists a unique solution $z \in V$ with

$$\text{(4.4)} \qquad |z|_V \le \|u\|_{2,Q_T} + \left\| e^{-\beta t} \int_0^t f(e^{\beta s} w(x,s)) ds \right\|_{2,Q_T}.$$

Remark 4.2. If $\int_0^t f(e^{\beta s} w(x,s)) ds = 0$ and $z_1 = 0$, then we can choose $u = 0$ such that $z(x,t_1) = 0 = z_1$ for system (4.3).

For any $\varphi^0 \in L^2(\Omega)$, we discuss the following problem:

$$\text{(4.5)} \quad \begin{cases} -\dfrac{\partial}{\partial t} \varphi(x,t) + \tilde{L}_n(w) \varphi(x,t) + \beta \varphi(x,t) = 0 \quad \text{in } \ Q_T, \\ \varphi(x,t) = 0 \ \text{ on } \ \partial\Omega \times (0,T), \quad \varphi(x,t_1) = \varphi^0(x) \ \text{ in } \ \Omega. \end{cases}$$

We also know from [9, p. 153] that there exists a unique solution $\varphi \in V \cap L^\infty(\bar{Q})$ with

$$\text{(4.6)} \qquad |\ \varphi\ |_V \le \|\ \varphi^0\|_{2,\Omega}.$$

Remark 4.3. For any $\theta^0 \in L^2(\Omega)$, if θ is the solution of problem (4.5) with $\theta(x,t_1) = \theta^0(x)$, then

$$\int_\Omega z(x,t_1)\theta^0(x) dx = -\int_{Q_{t_1}} e^{-\beta t} \int_0^t f(e^{\beta s} w(x,s)) ds \theta(x,t) dx dt$$

if and only if $u = 0$ in (4.3), where $z(x,t)$ and $w(x,t)$ are given in (4.3).

For any given $z_1 \in L^2(\Omega)$ and $\varepsilon > 0$, we define a functional on $L^2(\Omega)$ as follows:

$$\text{(4.7)} \quad \begin{aligned} J(\varphi^0) = {} & \frac{1}{2} \int_{Q_{t_1}} \varphi^2(x,t) dx dt - \int_\Omega z_1(x) \varphi^0(x) dx + \varepsilon \|\varphi^0\|_{2,\Omega} \\ & - \int_{Q_{t_1}} e^{-\beta t} \int_0^t f(e^{\beta s} w(x,s)) ds \varphi(x,t) dx dt, \end{aligned}$$

where $\varphi(x,t)$ is the solution of (4.5) with $\varphi(x,t_1)=\varphi^0(x)$.

LEMMA 4.1. *There exists a unique* $\bar{\varphi}^0 \in L^2(\Omega)$ *such that*

$$J(\bar{\varphi}^0) \le J(\varphi^0) \quad \forall \varphi^0 \in L^2(\Omega).$$

Proof. First, we show that there exists a constant $M>0$ such that

$$J(\varphi^0) \ge -M \quad \forall \varphi^0 \in L^2(\Omega). \tag{4.8}$$

In fact, we need only to consider the case of $\|\varphi_k^0\|_{2,\Omega} \to +\infty$. Set $\tilde{\varphi}_k^0 = \varphi_k^0/\|\varphi_k^0\|_{2,\Omega}$ so that $\|\tilde{\varphi}_k^0\|_{2,\Omega}=1$. Let φ_k and $\tilde{\varphi}_k$ be solutions of (4.5) with $\varphi_k(x,t_1)=\varphi_k^0(x)$ and $\tilde{\varphi}_k(x,t_1)=\tilde{\varphi}_k^0(x)$, respectively. Then, we have $\varphi_k = \|\varphi_k^0\|_{2,\Omega}\tilde{\varphi}_k$. So, this implies

$$\begin{aligned}\frac{J(\varphi_k^0)}{\|\varphi_k^0\|_{2,\Omega}} &= \frac{1}{2}\int_{Q_{t_1}} (\tilde{\varphi}_k(x,t))^2 dxdt \|\varphi_k^0\|_{2,\Omega} - \int_\Omega z_1(x)\tilde{\varphi}_k^0(x)dx + \varepsilon\|\tilde{\varphi}_k^0\|_{2,\Omega} \\ &\quad - \int_{Q_{t_1}} e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\tilde{\varphi}_k(x,t)dxdt.\end{aligned} \tag{4.9}$$

From this equality, we see that $\frac{J(\varphi_k^0)}{\|\varphi_k^0\|_{2,\Omega}} \to +\infty$ as $\|\varphi_k^0\|_{2,\Omega} \to +\infty$. Thus (4.8) holds.

According to (4.8), we can choose a minimizing sequence $\{\varphi_k^0\}$. From the discussion above, we know $\|\varphi_k^0\|_{2,\Omega} \le C$. So, there exist $\bar{\varphi}^0 \in L^2(\Omega)$ and a subsequence $\varphi_{k_i}^0$ such that $\varphi_{k_i}^0 \to \bar{\varphi}^0$ weakly in $L^2(\Omega)$. Suppose that φ_k and $\bar{\varphi}$ are solutions of (4.5) with $\varphi_k(x,t_1)=\varphi_k^0(x)$ and $\bar{\varphi}(x,t_1)=\bar{\varphi}^0(x)$, respectively. Since $\varphi_{k_i}^0 \to \bar{\varphi}^0$ weakly in $L^2(\Omega)$, then $\varphi_{k_i} \to \bar{\varphi}$ weakly in V. It follows from Sobolev's embedding theorem that $\varphi_{k_i} \to \bar{\varphi}$ in $L^2(Q_{t_1})$. Thus, we have

$$J(\bar{\varphi}^0) \le \lim_{k_i\to\infty} J(\varphi_{k_i}^0) = \inf\left\{J(\varphi^0) \mid \ \varphi^0 \in L^2(\Omega)\right\}.$$

Since the functional $J(\cdot)$ is strictly convex, the minimum function is unique. Lemma 4.1 is proved. ∎

LEMMA 4.2. *Suppose* $\bar{\varphi}^0$ *is the minimum function of functional* (4.7), $\bar{\varphi}$ *is the solution of* (4.5) *with* $\bar{\varphi}(x,t_1)=\bar{\varphi}^0(x)$, *and there exists* $\varphi_1^0 \in L^2(\Omega)$ *such that*

$$\int_\Omega z_1(x)\varphi_1^0(x)dx \ne -\int_{Q_{t_1}} e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\varphi_1(x,t)dxdt,$$

where φ_1 *is the solution of* (4.5) *with* $\varphi_1(x,t_1)=\varphi_1^0(x)$. *Then,* $\bar{\varphi}^0 \ne 0$ *and for any* $\theta^0 \in L^2(\Omega)$ *and* θ, *which is the solution of* (4.5) *with* $\theta(x,t_1)=\theta^0(x)$, *we have*

$$\begin{aligned}\int_{Q_{t_1}} \bar{\varphi}(x,t)\,\theta(x,t)dxdt &= \int_\Omega z_1(x)\theta^0(x)dx - \int_\Omega \varepsilon\frac{\bar{\varphi}^0(x)}{\|\bar{\varphi}^0\|_{2,\Omega}}\theta^0(x)dx \\ &\quad + \int_{Q_{t_1}} e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\theta(x,t)dxdt.\end{aligned} \tag{4.10}$$

Proof. First, we define $\varphi_m^0 = \frac{1}{m}\varphi_1^0$ so that

$$\begin{aligned}J(\varphi_m^0) &= \frac{1}{2m^2}\int_{Q_{t_1}} \varphi_1^2(x,t)dxdt - \frac{1}{m}\int_\Omega z_1(x)\varphi_1^0(x)dx + \varepsilon\frac{1}{m}\|\varphi_1^0\|_{2,\Omega} \\ &\quad - \frac{1}{m}\int_{Q_{t_1}} e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\varphi_1(x,t)dxdt.\end{aligned}$$

We may assume

$$\int_\Omega z_1(x)\varphi_1^0(x)dx + \int_{Q_{t_1}} e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\varphi_1(x,t)dxdt > 0.$$

If it is not, we choose $\hat{\varphi}_1^0(x) = -\varphi_1^0(x)$ so that

$$\int_\Omega z_1(x)\hat{\varphi}_1^0(x)dx + \int_{Q_{t_1}} e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\hat{\varphi}_1(x,t)dxdt > 0.$$

We can also choose sufficiently large m and sufficiently small ε such that $J(\varphi_m^0) < 0$. So, $J(\bar{\varphi}^0) < 0$. Since $J(0) = 0$, we have $\bar{\varphi}^0 \neq 0$.

For any $\theta^0 \in L^2(\Omega)$, set $\varphi_\rho^0 = \bar{\varphi}^0 + \rho\theta^0$, $\rho \in (-1,1)$. Let φ_ρ satisfy (4.5) with $\varphi_\rho(x,t_1) = \varphi_\rho^0(x)$. Then $\varphi_\rho(x,t) = \bar{\varphi}(x,t) + \rho\theta(x,t)$. We carry out the following calculation:

$$\begin{aligned}
J(\varphi_\rho^0) - J(\bar{\varphi}^0) &= \frac{1}{2}\int_{Q_{t_1}} [\varphi_\rho^2(x,t) - \bar{\varphi}^2(x,t)]dxdt \\
&+ \varepsilon\frac{1}{\|\varphi_\rho^0\| + \|\bar{\varphi}^0\|}\int_\Omega \Big[(\varphi_\rho^0(x))^2 - (\bar{\varphi}^0(x))^2\Big]dx \\
&- \int_\Omega z_1(x)\rho\theta^0(x)dx - \rho\int_{Q_{t_1}} e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\theta(x,t)dxdt \\
&= \rho\int_{Q_{t_1}} \bar{\varphi}(x,t)\theta(x,t)dxdt + \frac{1}{2}\rho^2\int_{Q_{t_1}} \theta^2(x,t)dxdt \\
&+ \varepsilon\frac{1}{\|\varphi_\rho^0\| + \|\bar{\varphi}^0\|}\int_\Omega \Big[2\rho\bar{\varphi}^0(x)\theta^0(x) + \rho^2(\theta^0(x))^2\Big]dx \\
&- \rho\int_\Omega z_1(x)\theta^0(x)dx - \rho\int_{Q_{t_1}} e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\theta(x,t)dxdt.
\end{aligned}$$

Dividing the above equality by ρ and letting $\rho \to 0$, since $0 = \delta J(\bar{\varphi}^0)$, we obtain (4.10). □

LEMMA 4.3. *System* (4.3) *is approximately controllable.*

Proof. For any $t_1 > 0$, $\varepsilon > 0$, there exists a unique $\bar{\varphi}^0$, which is the minimum function of functional (4.7). So there exists a unique solution $\bar{\varphi}(x,t)$ of (4.5) with $\bar{\varphi}(x,t_1) = \bar{\varphi}^0(x)$. Now, let $u(x,t) = \bar{\varphi}(x,t)$; that is,

$$\text{(4.11)} \quad \begin{cases} \dfrac{\partial}{\partial t}z(x,t) + \tilde{L}_n(w)z + \beta z + e^{-\beta t}\displaystyle\int_0^t f(e^{\beta s}w(x,s))ds = \bar{\varphi}(x,t) \quad \text{in } Q_T, \\ z(x,t) = 0 \ \text{ on } \ \partial\Omega\times(0,T), \quad z(x,0) = 0 \ \text{ in } \ \Omega. \end{cases}$$

Multiplying (4.11) by $\theta(x,t)$, where $\theta(x,t)$ is the solution of problem (4.5) with $\theta(x,t_1) = \theta^0(x)$, and integrating by parts the resulting relation over Q_{t_1}, we obtain

$$\begin{aligned}
&\int_\Omega z(x,t_1)\theta^0(x)\,dx + \int_{Q_{t_1}} e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\theta(x,t)dxdt \\
&\qquad = \int_{Q_{t_1}} \bar{\varphi}(x,t)\theta(x,t)dxdt.
\end{aligned}$$

Using (4.10), we have

$$\int_\Omega \left[z(x,t_1) - z_1(x) + \varepsilon \frac{\bar{\varphi}^0(x)}{\|\bar{\varphi}^0\|_{2,\Omega}} \right] \theta^0(x)dx = 0 \quad \forall\ \theta^0 \in L^2(\Omega).$$

From this equality, we see that $\|z(\cdot,t_1) - z_1\|_{2,\Omega} \le \varepsilon$. □

COROLLARY 4.1. *For $y_0 \in W^{1,p}(\Omega)$ and $w \in Z$, the system*

$$\text{(4.12)} \quad \begin{cases} \dfrac{\partial}{\partial t} z(x,t) + \tilde{L}_n(w)z + \beta z + e^{-\beta t} \displaystyle\int_0^t f(e^{\beta s} w(x,s))ds = u(x,t) & \text{in } Q_T, \\ z(x,t) = 0 \ \text{ on } \partial\Omega \times (0,T), \quad z(x,0) = y_0(x) \ \text{ in } \Omega \end{cases}$$

is approximately controllable.

Proof. Let z^1 and z^2 be solutions of the following problems, respectively:

$$\begin{cases} \dfrac{\partial}{\partial t} z^1(x,t) + \tilde{L}_n(w)z^1 + \beta z^1 + e^{-\beta t} \displaystyle\int_0^t f(e^{\beta s} w(x,s))ds = u(x,t) & \text{in } Q_T, \\ z^1(x,t) = 0 \ \text{ on } \partial\Omega \times (0,T), \quad z^1(x,0) = 0 \ \text{ in } \Omega, \end{cases}$$

$$\begin{cases} \dfrac{\partial}{\partial t} z^2(x,t) + \tilde{L}_n(w)z^2(x,t) + \beta z^2 = 0 & \text{in } Q_T, \\ z(x,t) = 0 \ \text{ on } \partial\Omega \times (0,T), \quad z^2(x,0) = y_0(x) \ \text{ in } \Omega. \end{cases}$$

It is easy to see that $z(x,t) = z^1(x,t) + z^2(x,t)$ is a solution of (4.12). Let

$$\text{(4.13)} \quad \begin{aligned} J(\varphi^0) = {} & \frac{1}{2} \int_{Q_{t_1}} \varphi^2(x,t)dxdt - \int_\Omega [z_1(x) - z^2(x,t_1)]\varphi^0(x)dx + \varepsilon \|\varphi^0\|_{2,\Omega} \\ & - \int_{Q_{t_1}} e^{-\beta t} \int_0^t f(e^{\beta s} w(x,s))ds \varphi(x,t)dxdt. \end{aligned}$$

From Lemma 4.3, we know that there exists $\bar{\varphi} \in W_2^{1,1}(Q_T)$ (see [9, p. 6]); when $u = \bar{\varphi}$, we have

$$\|z(\cdot,t_1) - z_1\|_{2,\Omega} = \|z^1(\cdot,t_1) - (z_1 - z^2(\cdot,t_1))\|_{2,\Omega} \le \varepsilon. \qquad \square$$

THEOREM 4.1. *Suppose condition* (H_1) *is fulfilled. Then system* (4.2) *is approximately controllable at t_1.*

Proof. From Corollary 4.1, we know that system (4.12) is approximately controllable. We define a map $F : Z = \{z \in V \ \mid \ |z|_V \le C^*\} \to Z$ by $F(w) = z$ for any $w \in Z$, with $z = z(x,t)$ being the solution of the equation

$$\text{(4.14)} \quad \begin{cases} \frac{\partial}{\partial t} z(x,t) + \tilde{L}_n(w)z(x,t) + \beta z + e^{-\beta t} \int_0^t f(e^{\beta s} w(x,s))ds = \bar{\varphi}(x,t) & \text{in } Q_T, \\ z(x,t) = 0 \ \text{ on } \partial\Omega \times (0,T), \quad z(x,0) = y_0(x) \ \text{ in } \Omega, \end{cases}$$

where $\bar{\varphi}$ is the solution of problem (4.5) with $\bar{\varphi}(x,t_1) = \bar{\varphi}^0(x)$, and $\bar{\varphi}^0$ is the minimum function of the functional (4.13).

First, we should show that for sufficient large C^*, we have $F(Z) \subset Z$. In fact, multiplying (4.14) by z and integrating over Q_{t_1}, we have

$$
\begin{aligned}
|z|_V^2 &\le \left[\|\bar\varphi\|_{2,Q_T}^2 + \|y_0\|_{2,\Omega}^2 + \frac{1}{2}\int_{Q_{t_1}} \left(e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\right)^2 dxdt\right] \\
&\le \left[\frac{1}{\beta}\|\bar\varphi^0\|_{2,\Omega}^2 + \|y_0\|_{2,\Omega}^2 + t_1^2\int_{Q_{t_1}} (K^2w^2(x,t) + f^2(0))dxdt\right] \\
&\le \left[\frac{1}{\beta}\|\bar\varphi^0\|_{2,\Omega}^2 + \|y_0\|_{2,\Omega}^2 + T^2\|f(0)\|_{2,Q_T}^2 + \frac{T^2K^2}{\beta}|w|_V^2\right] \\
&\le \left[\frac{1}{\beta}\|\bar\varphi^0\|_{2,\Omega}^2 + \|y_0\|_{2,\Omega}^2 + T^2\|f(0)\|_{2,Q_T}^2 + \frac{T^2K^2}{\beta}(C^*)^2\right].
\end{aligned} \tag{4.15}
$$

From (4.13), we have

$$
\begin{aligned}
0 > J(\bar\varphi^0) > \frac{1}{2}\int_{Q_{t_1}} (\tilde\varphi(x,t))^2 dxdt\|\bar\varphi^0\|_{2,\Omega}^2 - \int_\Omega [z_1(x) - z^2(x,t_1)]\bar\varphi^0(x)dx \\
- \int_{Q_{t_1}} e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\bar\varphi(x,t)dxdt,
\end{aligned} \tag{4.16}
$$

where $\tilde\varphi(x,t)$ is the solution of (4.5) with $\tilde\varphi(x,t_1) = \frac{\bar\varphi^0(x)}{\|\bar\varphi^0\|_{2,\Omega}}$. Since $\tilde\varphi \in V$ and $\int_\Omega \tilde\varphi^2(x,t_1)dx = 1$, there exists a positive constant c_0, which is independent of $\tilde\varphi$, such that

$$
c_0 \le \frac{1}{2}\int_{Q_{t_1}} \tilde\varphi^2(x,t)dxdt.
$$

From (4.16), we obtain

$$
\begin{aligned}
c_0\|\bar\varphi^0\|_{2,\Omega}^2 &< C(\eta)\|z_1 - z^2(\cdot,t_1)\|_{2,\Omega}^2 + \eta\|\bar\varphi^0\|_{2,\Omega}^2 \\
&+ \frac{1}{2}\int_{Q_{t_1}} \left[e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\right]^2 dxdt + \frac{1}{2}\|\bar\varphi\|_{2,Q_{t_1}}^2 \\
&\le C(\eta)\|z_1 - z^2(\cdot,t_1)\|_{2,\Omega}^2 + \eta\|\bar\varphi^0\|_{2,\Omega}^2 \\
&+ \frac{T^2}{2}\int_{Q_{t_1}} [2K^2w^2(x,t) + 2f^2(x,t,0)]dxdt + \frac{1}{2\beta}\|\bar\varphi^0\|_{2,\Omega}^2,
\end{aligned} \tag{4.17}
$$

where we have used the inequality $\|\bar\varphi\|_{2,Q_{t_1}}^2 \le \frac{1}{\beta}\|\bar\varphi^0\|_{2,\Omega}^2$. Furthermore, we choose $\eta = \frac{c_0}{4}$ and $\frac{1}{2\beta} \le \frac{c_0}{4}$ so that

$$
\begin{aligned}
\|\bar\varphi^0\|_{2,\Omega}^2 &\le \frac{2}{c_0}\left[C(\eta)\|z_1 - z^2(\cdot,t_1)\|_{2,\Omega}^2 + \frac{T^2K^2}{\beta}|w|_V^2 + T^2\|f(0)\|_{2,Q_T}^2\right] \\
&\le \frac{2C(\eta)}{c_0}\|z_1 - z^2(\cdot,t_1)\|_{2,\Omega}^2 + \frac{2T^2}{c_0}\|f(0)\|_{2,Q_T}^2 + \frac{2T^2K^2}{c_0\beta}(C^*)^2.
\end{aligned}
$$

Substituting this inequality into (4.15), we have

$$
\begin{aligned}
|z|_V^2 &\le \frac{2C(\eta)}{c_0\beta}\|z_1 - z^2(\cdot,t_1)\|_{2,\Omega}^2 + \|y_0\|_{2,\Omega}^2 \\
&+ \left(T^2 + \frac{2T^2}{c_0\beta}\right)\|f(0)\|_{2,Q_T}^2 + \frac{T^2K^2}{\beta}\left(1 + \frac{2}{c_0\beta}\right)(C^*)^2.
\end{aligned} \tag{4.18}
$$

We can choose sufficiently large β such that $\frac{TK^2}{\beta}(1+\frac{2}{c_0\beta})=\frac{1}{2}$. Set

$$(C^*)^2=2\left[\frac{2C(\eta)}{c_0\beta}\|z_1-z^2(\cdot,t_1)\|^2_{2,\Omega}+\|y_0\|^2_{2,\Omega}+\left(T^2+\frac{2T^2}{c_0\beta}\right)\|f(0)\|^2_{2,Q_T}\right].$$

Then, from (4.18), we have $|z|^2_V\le(C^*)^2$. This implies that $F(Z)\subset Z$.

Next, we shall prove that F is a continuous map. Set $F=F_3\circ F_2\circ F_1$, where $F_1:Z\to L^2(\Omega)$, $F_2:L^2(\Omega)\to V\cap L^\infty(\bar{Q}_T)$, $F_3:V\to Z$; that is,

$$F_1(w)=\bar{\varphi}^0,\quad F_2(\bar{\varphi}^0)=\bar{\varphi},\quad F_3(\bar{\varphi})=z.$$

From (4.4) and (4.6), it is easy to see that F_2 and F_3 are continuous maps. So, we need only to prove that F_1 is a continuous map.

Suppose $w_k\in Z$, $w\in Z$, and $w_k\to w$ in Z; we prove that $F_1(w_k)\to F_1(w)$ in $L^2(\Omega)$. Set $\bar{\varphi}^0_k=F_1(w_k)$, $\bar{\varphi}^0=F_1(w)$; $\bar{\varphi}_k$ and $\bar{\varphi}$ are solutions of (4.5) with $\bar{\varphi}_k(x,t_1)=\bar{\varphi}^0_k(x)$, $\bar{\varphi}(x,t_1)=\bar{\varphi}^0(x)$, respectively. We have $\|\bar{\varphi}^0_k\|_{2,\Omega}\le C$ and $\|\bar{\varphi}_k\|_V\le C$. So there exist subsequences $\bar{\varphi}^0_{k_i}$ of $\bar{\varphi}^0_k$ and $\tilde{\varphi}^0\in L^2(\Omega)$ such that $\bar{\varphi}^0_{k_i}\to\tilde{\varphi}^0$ weakly in $L^2(\Omega)$, and $\bar{\varphi}_{k_i}\to\bar{\varphi}$ weakly in V.

For any $\varphi^0\in L^2(\Omega)$, we have $J(\bar{\varphi}^0_k,w_k)\le J(\varphi^0,w_k)$ and

$$J(\tilde{\varphi}^0,w)\le\lim_{k_i\to\infty}J(\bar{\varphi}^0_{k_i},w_{k_i}).$$

This implies

$$J(\tilde{\varphi}^0,w)\le\lim_{k_i\to\infty}J(\bar{\varphi}^0_{k_i},w_{k_i})\le\lim_{k_i\to\infty}J(\varphi^0,w_{k_i})=J(\varphi^0,w).$$

Thus, $\tilde{\varphi}^0$ is the minimal function of $J(\cdot,w)$. According to the uniqueness of the minimal function for $J(\cdot,w)$, we have $\tilde{\varphi}^0=\bar{\varphi}^0$.

Since $\bar{\varphi}_{k_i}\to\bar{\varphi}$ weakly in V, so $\bar{\varphi}_{k_i}\to\bar{\varphi}$ strongly in $L^2(\Omega)$. Thus,

$$\begin{aligned}\lim_{k_i\to\infty}&\left[\int_{Q_{t_1}}\bar{\varphi}^2_{k_i}(x,t)dxdt-\int_\Omega[z_1(x)-z^2(x,t_1)]\bar{\varphi}^0_{k_i}(x)dx\right.\\&\left.-\int_{Q_{t_1}}e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\bar{\varphi}_{k_i}(x,t)dxdt\right]\\&=\int_{Q_{t_1}}\bar{\varphi}^2(x,t)dxdt-\int_\Omega[z_1(x)-z^2(x,t_1)]\bar{\varphi}^0(x)dx\\&\quad-\int_{Q_{t_1}}e^{-\beta t}\int_0^t f(e^{\beta s}w(x,s))ds\bar{\varphi}(x,t)dxdt.\end{aligned}$$

Thus, from (4.7), we have $\lim_{k_i\to\infty}\|\bar{\varphi}^0_{k_i}\|_{2,\Omega}=\|\bar{\varphi}^0\|_{2,\Omega}$. Furthermore, we have $\lim_{k_i\to\infty}\|\bar{\varphi}^0_{k_i}-\bar{\varphi}^0\|_{2,\Omega}=0$.

We need to show that $\lim_{k\to\infty}\|\bar{\varphi}^0_k-\bar{\varphi}^0\|_{2,\Omega}=0$. If it is not true, then there exists a subsequence $\{\bar{\varphi}^0_{k_i}\}$ such that $\bar{\varphi}^0_{k_i}$ does not converge to $\bar{\varphi}^0$; that is, there exists a number $\delta>0$ such that $\|\bar{\varphi}^0_{k_i}-\bar{\varphi}^0\|_{2,\Omega}\ge\delta$ for all k_i. On the other hand, $\|\bar{\varphi}^0_{k_i}\|_{2,\Omega}\le C$. Using a similar argument as above, we obtain a subsequence $\{\bar{\varphi}^0_{k_i'}\}$ of $\{\bar{\varphi}^0_{k_i}\}$ with $\lim_{k_i'\to\infty}\|\bar{\varphi}^0_{k_i'}-\bar{\varphi}^0\|_{2,\Omega}=0$. This contradicts $\|\bar{\varphi}^0_{k_i}-\bar{\varphi}^0\|_{2,\Omega}\ge\delta>0$. So, $\lim_{k\to\infty}\|\bar{\varphi}^0_k-\bar{\varphi}^0\|_{2,\Omega}=0$. Thus, we have proved the continuity of F_1. Therefore, F is continuous.

Finally, we prove that F is a compact map. In fact, if $\{z_k\} \subset F(Z)$, then there exists $\{w_k\} \subset Z$ such that $z_k = F(w_k)$. Again, we choose $a_k \in C^\infty(Q_T)$ such that $\|(|e^{\beta t}\nabla w_k|^2 + \frac{1}{n})^{(p-2)/2} - a_k\|_{p,Q_T} < \frac{1}{k}$. Now consider the following problem:

$$\begin{cases} \dfrac{\partial}{\partial t}z(x,t) - \operatorname{div}\big[a_k(x,t)\nabla z(x,t))\big] + \beta z(x,t) \\ \qquad\qquad = u(x,t) - e^{-\beta t}\displaystyle\int_0^t f(e^{\beta s}w_k(x,s))ds =: \tilde{u}_k(x,t) \quad \text{in } Q_T, \\ z(x,t) = 0 \ \text{ on } \partial\Omega \times (0,T), \quad z(x,0) = y_0(x) \ \text{ in } \Omega. \end{cases}$$

From [9, Chapter 4], we know that there exists a solution $\tilde{z}_k \in W$ to the above problem and $\|\tilde{z}_k\|_W \le C\|\tilde{u}_k\|_{p,Q_T} \le C$. According to Sobolev's embedding theorem, there exists a subsequence $\{\tilde{z}_{k_i}\} \subset \{\tilde{z}_k\}$ such that $\{\tilde{z}_{k_i}\}$ is a convergent sequence in $(V, |\cdot|_V)$. Since $\|\tilde{z}_k - z_k\|_V \le \frac{C}{k}$, we conclude that $\{z_{k_i}\}$ is a convergent sequence in V. Thus $F : Z \to Z$ is a continuous compact map. By Schauder's fixed point theorem, F has a fixed point z, and thus, system (4.14) becomes

$$\begin{cases} \dfrac{\partial}{\partial t}z(x,t) + \tilde{L}_n z(x,t) + \beta z + e^{-\beta t}\displaystyle\int_0^t f(e^{\beta s}z(x,s))ds = \bar{\varphi}(x,t) \quad \text{in } Q_T, \\ z(x,t) = 0 \ \text{ on } \partial\Omega \times (0,T), \quad z(x,0) = y_0(x) \ \text{ in } \Omega, \end{cases}$$

and

$$\|z(\cdot,t_1) - z_1\|_{2,\Omega} \le \varepsilon.$$

This proves that (2.1) is approximately controllable.

Now, we give the proof of Theorem 1.3.

Proof. To complete the proof of Theorem 1.3, we need only to prove the following conclusion: For any $\varepsilon > 0$, there is a N such that $n > N$ implies

$$\| \ y_n(\cdot,t_1) - y(\cdot,t_1) \ \|_{2,\Omega} < \varepsilon,$$

where y and y_n are solutions of problems (1.1) and (2.1), respectively.

In fact, we repeat the argument in section 2 by setting $z_n = e^{-\beta s}y_n$ and $z = e^{-\beta s}y$, $\beta > K$. From (1.1) and (2.1), we find

$$(4.19)\quad \begin{cases} \dfrac{\partial}{\partial t}(z_n - z) - \operatorname{div}\Bigg[\left(|e^{\beta t}\nabla z_n(x,t)|^2 + \dfrac{1}{n}\right)^{(p-2)/2}\nabla z_n(x,t) \\ \qquad - \left(|e^{\beta t}\nabla z(x,t)|^2 + \dfrac{1}{n}\right)^{(p-2)/2}\nabla z(x,t)\Bigg] + \beta(z_n - z) \\ \qquad = \operatorname{div}\left\{\left[\left(|e^{\beta t}\nabla z(x,t)|^2 + \dfrac{1}{n}\right)^{(p-2)/2} - |e^{\beta t}\nabla z(x,t)|^{p-2}\right]\nabla z(x,t)\right\} \\ \qquad + \ e^{-\beta t}\displaystyle\int_0^t [f(e^{\beta s}z(x,s)) - f(e^{\beta s}z_n(x,s))]ds \quad \text{in } Q_T, \\ (z_n - z) = 0 \ \text{ on } \partial\Omega \times (0,T), \quad (z_n - z)(x,0) = 0 \ \text{ in } \Omega. \end{cases}$$

Multiplying (4.19) by $z_n - z$ and integrating in Q_{t_1}, we obtain

$$\begin{aligned}
&\frac{1}{2} \parallel z_n(\cdot,t_1) - z(\cdot,t_1) \parallel^2_{2,\Omega} \\
&\quad \le \int_{Q_{t_1}} \Bigg[\left(|e^{\beta t}\nabla z(x,t)|^2 + \frac{1}{n} \right)^{(p-2)/2} \\
&\qquad - \ |e^{\beta t}\nabla z(x,t)|^{p-2} \Bigg] \nabla z(x,t) \nabla (z(x,t) - z_n(x,t)) dxdt = I.
\end{aligned} \tag{4.20}$$

If $2 < p \le 4$, we set

$$Q_\eta = \{(x,t) \in Q_{t_1} \,|\, |e^{\beta t}\nabla z(x,t)| \ge \eta\}.$$

For any $\varepsilon > 0$, we can find an $\eta > 0$ such that

$$\begin{aligned}
I_1 = \int_{Q_{t_1} \setminus Q_\eta} &\Bigg[\left(|e^{\beta t}\nabla z(x,t)|^2 + \frac{1}{n} \right)^{(p-2)/2} \\
&- \ |e^{\beta t}\nabla z(x,t)|^{p-2} \Bigg] \nabla z(x,t) \nabla (z(x,t) - z_n(x,t)) dxdt < \frac{\varepsilon^2}{4},
\end{aligned}$$

$$\begin{aligned}
I_2 = \int_{Q_\eta} &\Bigg[\left(|e^{\beta t}\nabla z(x,t)|^2 + \frac{1}{n} \right)^{(p-2)/2} \\
&- \ |e^{\beta t}\nabla z(x,t)|^{p-2} \Bigg] \nabla z(x,t) \nabla (z(x,t) - z_n(x,t)) dxdt \\
&= \frac{p-2}{2} \frac{1}{n} \int_{Q_\eta} \left(|e^{\beta t}\nabla z(x,t)|^2 + \theta(x,t)\frac{1}{n} \right)^{(p-4)/2} \nabla z(x,t) \nabla (z(x,t) - z_n(x,t)) dxdt \\
&\le \frac{p-2}{2} \frac{1}{n} \left(\frac{1}{\eta^2} \right)^{(4-p)/2} \int_{Q_\eta} \nabla z(x,t) \nabla (z(x,t) - z_n(x,t)) dxdt \\
&< \frac{C}{n} \|\nabla z\|_{2,Q_\eta} \|\nabla (z - z_n)\|_{2,Q_\eta} = \frac{C_1}{n}.
\end{aligned}$$

Thus, there exists a number N, such that $n > N$ implies $I_2 < \frac{\varepsilon^2}{4}$. So we have

$$\frac{1}{2} \parallel z_n(\cdot,t_1) - z(\cdot,t_1) \parallel^2_{2,\Omega} \le I = I_1 + I_2 < \frac{\varepsilon^2}{2},$$

that is, $\|z_n(\cdot,t_1) - z(\cdot,t_1)\|_{2,\Omega} \le \varepsilon$.

If $p > 4$, from (4.20), we have

$$
\begin{aligned}
I &= \frac{p-2}{2}\frac{1}{n}\int_{Q_{t_1}} \left(|e^{\beta t}\nabla z(x,t)|^2 + \theta(x,t)\frac{1}{n}\right)^{(p-4)/2} \nabla z(x,t)\nabla(z(x,t) - z_n(x,t))dxdt \\
&\le \frac{p-2}{2}\frac{1}{n}2^{(p-4)/2}\int_{Q_{t_1}\setminus Q_1} |\nabla z(x,t)||\nabla(z(x,t) - z_n(x,t))|dxdt \\
&+ \frac{p-2}{2}\frac{1}{n}(2e^{\beta T})^{(p-4)/2}\int_{Q_1} |\nabla z(x,t)|^{p-3}|\nabla(z(x,t) - z_n(x,t))|dxdt \\
&\le \frac{C_{21}}{n}\|\nabla z\|_{2,Q_{t_1}}\|\nabla(z - z_n)\|_{2,Q_{t_1}} \\
&+ \frac{C_{22}}{n}\int_0^{t_1}\left(\int_\Omega |\nabla z(x,t)|^p dx\right)^{(p-3)/p}\left(\int_\Omega |\nabla(z(x,t) - z_n(x,t))|^{p/3}dx\right)^{3/p} dt \\
&\le \frac{C_{21}}{n}\|\nabla z\|_{2,Q_{t_1}}\|\nabla(z - z_n)\|_{2,Q_{t_1}} \\
&+ \frac{C_{22}^*}{n}\|z\|^{p-3}_{L^\infty(0,T;W^{1,p}(\Omega))}\|z - z_n\|_{L^\infty(0,T;W^{1,p}(\Omega))} \\
&= \frac{C_2}{n}.
\end{aligned}
$$

Thus, there exists a number N such that $n > N$ implies

$$\| z_n(\cdot,t_1) - z(\cdot,t_1) \|^2_{2,\Omega} \le C\frac{1}{n} \le \varepsilon^2.$$

The proof of Theorem 1.3 is complete. □

Acknowledgment. The authors would like to thank the referees for their valuable comments and suggestions on this paper.

REFERENCES

[1] K. Balachandran, P. Balasubramaniam, and J. P. Dauer, *Controllability of nonlinear integrodifferential systems in Banach space,* J. Optim. Theory Appl., 84 (1995), pp. 83–91.

[2] K. Balachandran and J. P. Dauer, *Controllability of Sobolev-type integrodifferential systems in Banach spaces,* J. Math. Anal. Appl., 217 (1995), pp. 335–348.

[3] E. Dibenedetto, *Degenerate Parabolic Equations,* Springer-Verlag, New York, 1993.

[4] W. E. Fitzgibbon, *Semilinear integrodifferential equations in a Banach space,* Nonlinear Anal., 4 (1980), pp. 745–760.

[5] H. Gao and J. Yin, *On a class of anisotropic diffusion equations,* Systems Sci. Math. Sci., 8 (1995), pp. 311–318.

[6] M. L. Heard, *An abstract semilinear hyperbolic Volterra integrodifferential equation,* J. Math. Anal. Appl., 80 (1981), pp. 175–202.

[7] M. A. Hussain, *On a nonlinear integrodifferential equation in Banach space,* Indian J. Pure Appl. Math., 19 (1988), pp. 516–529.

[8] A. S. Kalashnikov, *Propagation of perturbation in the first boundary-value problem for a double nonlinear degenerate parabolic equation,* J. Soviet Math., 32 (1986), pp. 315–320.

[9] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'tseva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[10] L. De Teresa, *Approximate controllability of semilinear heat equation in* $\mathbb{R}^N$*,* SIAM J. Control Optim., 36 (1998), pp. 2128–2147.

[11] L. De Teresa and E. Zuazua, *Approximate controllability of semilinear heat equation in unbounded domains,* Nonlinear Anal., 37 (1999), pp. 1059–1090.

[12] G. Wang and L. Wang, *The Carleman inequality and its application to periodic optimal control governed by semilinear parabolic differential equations,* J. Optim. Theory Appl., 118 (2003), pp. 429–461.
[13] G. F. Webb, *An abstract semilinear Volterra integrodifferential equation,* Proc. Amer. Math. Soc., 69 (1978), pp. 255–260.
[14] Z. Wu, J. Zhao, J. Yin, and H. Li, *Nonlinear Diffusion Equations,* World Scientific, River Edge, NJ, 2001.
[15] J. Yin, *The property of finite speed of propagation of generalized solutions to nonlinear diffusion equations,* Acta Math. Sinica, 34 (1991), pp. 360–364.
[16] J. Yin, *Solutions with compact support for nonlinear diffusion equations,* Nonlinear Anal., 19 (1992), pp. 309–321.
[17] E. Zuazua, *Approximate controllability of semilinear heat equation with globally Lipschitz nonlinearities,* Control Cybernet., 28 (1999), pp. 665–683.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 1011–1036

A CONVEX OPTIMIZATION APPROACH TO ARMA(n,m) MODEL DESIGN FROM COVARIANCE AND CEPSTRAL DATA*

P. ENQVIST†

Abstract. Methods for determining ARMA(n,m) filters from covariance and cepstral estimates are proposed. In [C. I. Byrnes, P. Enqvist, and A. Lindquist, *SIAM J. Control Optim.*, 41 (2002), pp. 23–59], we have shown that an ARMA(n,n) model determines and is uniquely determined by a window $r_0, r_1, \dots, r_n$ of covariance lags and $c_1, c_2, \dots, c_n$ of cepstral lags. This unique model can be determined from a convex optimization problem which was shown to be the dual of a maximum entropy problem. In this paper, generalizations of this problem are analyzed. Problems with covariance lags $r_0, r_1, \dots, r_n$ and cepstral lags $c_1, c_2, \dots, c_m$ of different lengths are considered, and by considering different combinations of covariances, cepstral parameters, poles, and zeros, it is shown that only zeros and covariances give a parameterization that is consistent with generic data.

However, the main contribution of this paper is a regularization of the optimization problems that is proposed in order to handle generic data. For the covariance and cepstral problem, if the data does not correspond to a system of desired order, solutions with zeros on the boundary occur and the cepstral coefficients are not interpolated exactly. In order to achieve strictly minimum phase filters for estimated covariance and cepstral data, a barrier-like term is introduced to the optimization problem. This term is chosen so that convexity is maintained and so that the unique solution will still interpolate the covariances but only approximate the cepstral lags. Furthermore, the solution will depend analytically on the covariance and cepstral data, which provides robustness, and the barrier term increases the entropy of the solution.

Key words. cepstrum, covariance, ARMA, entropy, convex optimization

AMS subject classifications. 94A17, 93E12, 93B15, 90C25

DOI. 10.1137/S0363012901399751

1. Introduction.

In this paper some new methods for determining autoregressive (AR), moving average (MA), and ARMA filters from a finite sample of data are analyzed. These methods are based on an optimization approach to interpolation first introduced in [6] and then extended in various directions in [5, 8, 3, 4]. In particular, the data is matched using two different sets of characterizing parameters, that is, covariance lags and cepstral lags. Before going into details, the basic system identification problem considered here is described.

Let

$$w(z) = \frac{\sigma(z)}{a(z)}$$

be a real rational function, where z denotes a forward shift operator and

$$a(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_n \qquad (a_0 > 0), \tag{1.1}$$

$$\sigma(z) = z^m + \sigma_1 z^{m-1} + \cdots + \sigma_m \tag{1.2}$$

are stable polynomials, i.e., real polynomials having all their roots in the open unit disc. Consider the zero-mean stationary process $\{Y_t\}_{t=-\infty}^{\infty}$ obtained by passing a

*Received by the editors December 17, 2001; accepted for publication (in revised form) February 10, 2004; published electronically November 9, 2004. This work was supported in part by grants from the Swedish Research Council for Engineering Sciences and the Göran Gustafsson Foundation.
http://www.siam.org/journals/sicon/43-3/39975.html

†Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden (pere@math.kth.se).

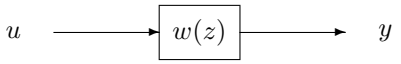


FIG. 1.1. *Shaping filter.*

white noise $\{U_t\}_{t=-\infty}^{\infty}$ through a filter with transfer function $w(z)$, as depicted in Figure 1.1, and let the system come to steady state. Such a process is called an *ARMA*(n, m) *process*, and the corresponding model

$$a_0 y_t + a_1 y_{t-1} + \cdots + a_n y_{t-n} = u_t + \sigma_1 u_{t-1} + \cdots + \sigma_m u_{t-m} \tag{1.3}$$

is called an ARMA(n, m) model. The stable polynomial $\sigma(z)$ defines the MA part, or the zeros of the filter, while the stable polynomial $a(z)$ defines the AR part, or the poles of the filter. Assume that the stochastic process $\{Y_t\}_{t=-\infty}^{\infty}$ is observed for $t = 1, \ldots, N$ to generate a string $\{y_t\}_{t=1}^{N}$ of data. In system identification the filter $w(z)$ that "best" matches the data $\{y_t\}_{t=1}^{N}$ is chosen in some prescribed class of models. The class of models that will be considered in this paper is the ARMA(n,m) filters. It will be assumed that the ARMA filters in the model class will be of fixed order, but how to choose the model class parameters m and n is not considered in this paper.

Design of AR, MA, and ARMA filters have a long history, and a number of different approaches have been used. Most methods minimize some error criterion over a set of parameterized models determining a model class. These approaches include, e.g., the maximum likelihood (ML) method [1], prediction error (PE) methods [18, 23], and others [14, 21, 16]. Even subspace methods [25] are based on optimization, but there the solution is given analytically by projections. This is appreciated in applications, but the subspace methods may yield a solution that is outside of the model class (e.g., nonstable models [10]), which is the main drawback of this approach. A common problem for the other methods mentioned above is that they may fail to deliver the global optima in the model class. For the ML and PE methods the error criterion function is not convex in the model parameters. This may lead to the minimization procedure getting stuck in a local optima which may be very different from the optimal solution. Consequently, although these methods are used with great success, they are not guaranteed to converge to a global optima.

The methods proposed in this paper can be used for system identification performed in two steps. The first step is the estimation of covariances and cepstral parameters from real data, and the second step determines the model that (approximatively) interpolates the estimated parameters. This is sometimes called estimation by the method of moments. In this paper we analyze only the second step. The proposed interpolation-based identification methods can be seen as generalizations of the widely used linear predictive coding (LPC) method [22], also called the maximum entropy method [2], which interpolates a set of covariances with an AR filter. The LPC method applies exact interpolation of estimated covariances, but as Burg describes it [2, p. xi], "Maximum entropy spectral analysis is based on choosing the spectrum which corresponds to the most random or the most unpredictable time series whose autocorrelation function agrees with the known values." The LPC filter parameters can be determined from a linear equation system. In the methods proposed in this paper the solution is derived from a well-behaved optimization problem. In fact, the information about the process is used in such a way that full ARMA models can be

designed using a convex error criterion. Since the optimization problems are convex there are no problems with local optima and there are optimization procedures that can be guaranteed to converge to the global optima for all initial points.

Next, the interpolation parameters are considered and standard methods for estimating them from data are presented.

The *covariances* of the stochastic process $\{Y_t\}_{t=-\infty}^{\infty}$ are defined by the second order moments

$$r_k \stackrel{\triangle}{=} \mathrm{E}(Y_t Y_{t+k}), \quad k = 0, 1, \ldots.$$

Taking the Fourier transform of the covariances, the *spectral density*

$$\Phi(e^{i\theta}) \stackrel{\triangle}{=} \sum_{k=-\infty}^{\infty} r_k e^{-ik\theta}$$

of the process is obtained. We will assume the estimated spectral densities are *coercive*, i.e., positive and bounded away from zero on the unit circle. For the sake of reference, we note that

$$r_k = \langle z^k, \Phi \rangle = \langle z^{-k}, \Phi \rangle, \quad k = 0, 1, 2, \ldots, \tag{1.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in the real $L_2[-\pi, \pi]$ space, i.e.,

$$\langle f, g \rangle \stackrel{\triangle}{=} \frac{1}{2\pi i} \int_{|z|=1} f(z)\overline{g(z)}\, \frac{dz}{z} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{i\theta}) g(e^{-i\theta})\, d\theta. \tag{1.5}$$

If the process is generated by a shaping filter $w(z)$, the spectral density can be extended to a Laurent series, valid in an annulus around the unit circle, given by

$$\Phi(z) \stackrel{\triangle}{=} \sum_{k=-\infty}^{\infty} r_k z^{-k} = w(z) w(z^{-1}). \tag{1.6}$$

The last equality in (1.6) gives an important relation between the spectral density and the shaping filter. If all covariances were known, the shaping filter $w(z)$ could be determined from Φ by spectral factorization [23, 9]. However, only a finite number of covariances can be estimated, and then the spectral density is not completely determined. The covariances $r_0, r_1, \ldots, r_n$ can be estimated from data using ergodic estimates. For example, the estimate

$$\hat{r}_k = \frac{1}{N} \sum_{j=1}^{N-k} y_j y_{j+k}, \quad k \ll N, \tag{1.7}$$

can be used. This particular estimate has the nice property that the matrix of estimated covariances

$$\mathbf{R}_n \stackrel{\triangle}{=} \begin{bmatrix} \hat{r}_0 & \hat{r}_1 & \ldots & \hat{r}_n \\ \hat{r}_1 & \hat{r}_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \hat{r}_1 \\ \hat{r}_n & \ldots & \hat{r}_1 & \hat{r}_0 \end{bmatrix} \tag{1.8}$$

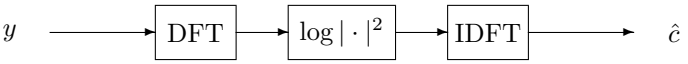


FIG. 1.2. *Cepstral estimation.*

is positive definite. The set of covariances $\hat{r}_0, \hat{r}_1, \ldots, \hat{r}_n$ such that $\mathbf{R}_n$ is positive definite is denoted by $\mathcal{R}_n$. For more about the statistical properties of the estimation of covariances, the reader is referred to [23, 17].

Since the spectral density is a positive function on the unit circle, its logarithm is well defined, and it has a Laurent series expansion in a neighborhood of the unit circle given by

$$\log \Phi(z) = \sum_{k=-\infty}^{\infty} c_k z^{-k}. \tag{1.9}$$

This defines the *cepstral coefficients*

$$c_k = \langle z^k, \log \Phi \rangle, \quad k = 0, 1, 2, \ldots. \tag{1.10}$$

A finite window of the cepstrum coefficients can also be estimated directly from data using ergodic estimates. In particular, taking the DFT of the data $\{y_j\}_{j=1}^N$, the absolute value square, the logarithm and the IDFT determines an estimate of the cepstrum [22, 13] as depicted in Figure 1.2. However, the problem of estimating cepstral coefficients directly from data is not by far as well analyzed as the problem of estimating covariances.

Two new design methods are presented in this paper. The first method is a regularization of the *cepstral-covariance matching* (CCM) method introduced in [3]. The CCM method determines an ARMA filter that simultaneously interpolates a finite window of covariances and cepstral lags, if that is possible, using a convex optimization formulation. In this paper an approximation in the interpolation of the cepstral parameters is introduced to make the global optima correspond to a stable minimum phase filter for generic covariance and cepstral estimates.

The second method is based on covariance matching. It is well known that design of MA models using covariance interpolation constraints lacks solutions for generic covariances. In this case some approximation in the covariance interpolation constraints is necessary, and an approximation similar to the one used for the CCM method is proposed.

The outline of this paper is as follows. In section 2, the optimization problem from [3] for determining the covariance and cepstrum interpolating filter is restated and analyzed. Actually, it is stated in a slightly more general form, allowing for ARMA(n,m) models, and the relation of poles and zeros to covariances and cepstral lags is discussed. In section 3, duality theory from mathematical programming is used for deriving some new optimization problems for solving interpolation problems. In section 4, a regularization of the CCM filter optimization problem is determined. By adding a barrier-like term to the objective function, stable minimum phase models are obtained at the price of approximate cepstral interpolation. It is also shown that the barrier term increases the entropy of the resulting model. In section 5, a regularization of an optimization problem derived in section 3 for designing zeros based on covariance interpolation is determined. By adding the same barrier-like term to this objective

function, stable minimum phase models are obtained at the price of approximate covariance interpolation. The barrier term is shown to increase the entropy of the resulting model. In section 6, a number of experiments are carried out in order to compare the method to other methods that are also well posed [24, 20].

2. Local coordinates for ARMA models in terms of cepstral and covariance windows. As a preliminary and to fix notation, pertinent facts on a method introduced in [3, 4] are now reviewed. As described in the introduction, the aim of this problem is to interpolate the covariances $r_0, r_1, \ldots, r_n \in \mathcal{R}_n$ and the cepstral coefficients $c_1, \ldots, c_m$. Therefore, the resulting filter is called a CCM filter [3].

The spectral density corresponding to a stable ARMA(n,m) filter can be parameterized as

$$\Phi = \frac{P}{Q},$$

where P and Q are pseudopolynomials

$$P(z) = p_0 + \frac{1}{2}p_1(z + z^{-1}) + \cdots + \frac{1}{2}p_m(z^m + z^{-m}), \tag{2.1}$$

$$Q(z) = q_0 + \frac{1}{2}q_1(z + z^{-1}) + \cdots + \frac{1}{2}q_n(z^n + z^{-n}) \tag{2.2}$$

such that $P(z) \in \hat{\mathcal{D}}_m^+$ and $Q(z) \in \mathcal{D}_n^+$, where $\hat{\mathcal{D}}_m^+$ and $\mathcal{D}_n^+$ are defined next.

Let $\mathcal{D}_n$ denote the class of real symmetric pseudopolynomials of degree at most n which are nonnegative on the unit circle, and let $\mathcal{D}_n^+$ be the subclass of all $D \in \mathcal{D}_n$ which are positive on the unit circle. Moreover, let $\boldsymbol{\mathcal{D}}_n$ and $\boldsymbol{\mathcal{D}}_n^+$ be the corresponding $(n+1)$-vectors of coefficients. In this notation, $P \in \mathcal{D}_m^+$, $Q \in \mathcal{D}_n^+$, $\mathbf{p} := (p_0, p_1, \ldots, p_m) \in \boldsymbol{\mathcal{D}}_m^+$, and $\mathbf{q} := (q_0, q_1, \ldots, q_n) \in \boldsymbol{\mathcal{D}}_n^+$. Finally, let $\hat{\mathcal{D}}_m^+$ and $\hat{\boldsymbol{\mathcal{D}}}_m^+$ be the subsets of $P \in \mathcal{D}_m^+$ and $\mathbf{p} \in \boldsymbol{\mathcal{D}}_m^+$, respectively, which have been normalized so that $p_0 = 1$. Clearly, these are all convex sets.

The following theorem is a generalization of Theorem 3.1 in [3] (see also Theorem 3.1 in [4]) in that it allows for $m \neq n$.

THEOREM 2.1. *Each* ARMA(n, m) *model* (1.3) *such that the polynomials* σ *and* a *are coprime determines and is uniquely determined by its window* $r_0, r_1, \ldots, r_n$ *of covariance lags and its window* $c_1, c_2, \ldots, c_m$ *of cepstral coefficients.*

Consequently, there is a one-to-one correspondence between the $n + m + 1$ coefficients $r_0, r_1, \ldots, r_n, c_1, c_2, \ldots, c_m$ and the $n + m + 1$ coefficients $a_0, a_1, \ldots, a_n$, $\sigma_1, \sigma_2, \ldots, \sigma_m$, provided σ and a are coprime. The proof is mutatis mutandis the same as in [3, 4]. The statement that $r_0, r_1, \ldots, r_n$, $c_1, c_2, \ldots, c_m$ are uniquely determined by $a_0, a_1, \ldots, a_n$, $\sigma_1, \sigma_2, \ldots, \sigma_m$ is trivial. The converse statement follows from the fact that the optimization problem

$$(\mathcal{P}) \quad \left[\begin{array}{ll} \min & \varphi(\mathbf{p}, \mathbf{q}), \\ \text{subject to (s.t.)} & (\mathbf{p}, \mathbf{q}) \in \hat{\boldsymbol{\mathcal{D}}}_m \times \boldsymbol{\mathcal{D}}_n, \end{array} \right] \tag{2.3}$$

has a unique solution, which is an interior point whenever the coefficients $r_0, r_1, \ldots, r_n$, $c_1, c_2, \ldots, c_m$ are the exact theoretical ones. Here $\varphi : \hat{\boldsymbol{\mathcal{D}}}_m \times \boldsymbol{\mathcal{D}}_n \to \mathbb{R}$ is the convex function

$$\varphi(\mathbf{p}, \mathbf{q}) \stackrel{\triangle}{=} -\sum_{k=1}^{m} c_k p_k + \sum_{k=0}^{n} r_k q_k + \left\langle P, \log \frac{P}{Q} \right\rangle. \tag{2.4}$$

In fact problem $(\mathcal{P})$ is the dual of maximizing the entropy

$$\mathcal{E}(\Phi) \triangleq \langle 1, \log \Phi \rangle \tag{2.5}$$

with covariance and cepstral interpolation constraints, and it has a solution in an interior point $(\tilde{\mathbf{p}}, \tilde{\mathbf{q}}) \in \hat{\boldsymbol{\mathcal{D}}}_m^+ \times \boldsymbol{\mathcal{D}}_n^+$ if and only if the gradient

$$\frac{\partial \varphi}{\partial p_k} = -c_k + \left\langle z^k, \log \frac{P}{Q} \right\rangle, \quad k = 1, \dots, m, \tag{2.6}$$

$$\frac{\partial \varphi}{\partial q_k} = r_k - \left\langle z^k, \frac{P}{Q} \right\rangle, \qquad k = 0, 1, \dots, n, \tag{2.7}$$

equals zero there. This happens precisely when

$$\left\langle z^k, \log \frac{P}{Q} \right\rangle = c_k, \quad k = 1, \dots, m, \tag{2.8}$$

$$\left\langle z^k, \frac{P}{Q} \right\rangle = r_k, \quad k = 0, 1, \dots, n, \tag{2.9}$$

which are the cepstral matching conditions (1.10) and the covariance matching conditions (1.4), respectively, when $\Phi = P/Q$.

However, for observed parameters $r_0, r_1, \dots, r_n, c_1, c_2, \dots, c_m$, which may be corrupted by measurement errors, or are otherwise generic, φ may fail to have an interior minimum. In this case, as shown in [4] using the methods of [6], the condition $\tilde{\mathbf{q}} \in \boldsymbol{\mathcal{D}}_n^+$ will nevertheless be satisfied so that the covariance matching condition (2.9) is fulfilled, whereas $\tilde{\mathbf{p}}$ will be on the boundary of $\hat{\boldsymbol{\mathcal{D}}}_m$. In section 4, problem $(\mathcal{P})$ will be regularized so that approximate cepstral matching is achieved for a $\tilde{\mathbf{p}} \in \hat{\boldsymbol{\mathcal{D}}}_m^+$.

It is clear that the number of covariance lags $r_0, r_1, \dots, r_n$ and cepstral parameters $c_1, \dots, c_m$, respectively, determines the number of poles and zeros of the ARMA(n,m) filter. This suggests that there is a deeper connection between the poles and the initial covariance lags, and between the zeros and the initial cepstral coefficients. The covariances are thus connected to the long-term effects of the filter and the cepstral parameters to the short-term effects. For example, a MA(m) filter can be determined as a special case when $n = 0$, where the zeros are determined using the cepstral parameters $c_1, \dots, c_m$ and the gain of the filter is determined by the variance r_0. The MA(m) filter has a finite impulse response—actually the impulse response is zero for lags larger than m—and thus only shapes the short-term response. In fact, for minimum phase systems the initial cepstral coefficients are equivalent to the initial impulse response parameters [4]. In case the long-term effects (i.e., the poles) of an ARMA(n,m) filter are known, the zeros can be designed by fixing the corresponding pseudopolynomial $Q \in \mathcal{D}_n^+$ in problem $(\mathcal{P})$. The amplification of the resulting filter has to be determined a posteriori, using, for example, the variance r_0. There is, however, no guarantee that there exists an ARMA(n,m) filter matching a generic cepstral sequence.

Using symmetrical arguments, an AR(n) filter can be determined as a special case when $m = 0$, where the poles are determined using the covariances $r_0, r_1, \dots, r_n \in \mathcal{R}_n$. This special case corresponds to the well-known LPC method. In case the short-term effects (i.e., the zeros) of an ARMA(n,m) filter are known, the poles can be designed by fixing the corresponding pseudopolynomial $P \in \mathcal{D}_n^+$ in problem $(\mathcal{P})$. The resulting optimization problem is thoroughly analyzed in [6, 11] and provides a solution to a

problem studied in [15, 7]. The filter designed with this method is called a *lattice-ladder notch* (LLN) filter [3], and it will always meet the covariance interpolation constraints. This problem is the starting point of the next section.

For curiosity it can be noted that the tails of the covariance and autocorrelation sequences hold similar complementary information. In [15] it was shown that the tail of the autocorrelation sequence asymptotically satisfies a recursion determined by the zeros. Similarly, it is well known that the tail of the covariances, and also the impulse response parameters, satisfies a recursion determined by the poles.

3. Optimization problems for cepstral and covariance interpolation. Next we consider four ARMA interpolation problems involving different combinations of covariances, cepstral parameters, poles and zeros as summarized in Table 3.1. By consistent we mean that there exists a matching ARMA model for each combination of valid parameters and by local coordinates we mean that the ARMA model locally depends analytically on the parameters.

TABLE 3.1
Summary of the interpolation problems.

Interpolation method	Consistent	Local coord.
Covariance matching with fixed numerator	Yes	Yes
Covariance matching with fixed denominator	No	Yes
Cepstral matching with fixed numerator	No	Yes
Cepstral matching with fixed denominator	No	Yes

Covariance matching with fixed numerator. Fixing P to be a given constant pseudopolynomial, the optimization problem $(\mathcal{P})$ is reduced to the problem $(\mathcal{P}_0)$ to minimize the strictly convex functional

$$\mathbb{J}_P(\mathbf{q}) \stackrel{\triangle}{=} \sum_{k=0}^{n} r_k q_k - \langle P, \log Q \rangle \tag{3.1}$$

over all $\mathbf{q} \in \boldsymbol{\mathcal{D}}_n^+$, i.e., the optimization problem introduced in [6]. As shown there, it has a unique solution $\hat{\mathbf{q}}$ which is an interior point, i.e., $\hat{\mathbf{q}} \in \boldsymbol{\mathcal{D}}_n^+$. This can be seen from the fact that

$$\frac{\partial \mathbb{J}_P}{\partial q_k} = r_k - \left\langle z^k, \frac{P}{Q} \right\rangle, \quad k = 0, 1, \ldots, n, \tag{3.2}$$

become infinite on the boundary of $\boldsymbol{\mathcal{D}}_n$. Setting the gradient (3.2) equal to zero, it is seen that $\hat{\Phi} = P/\hat{Q}$ satisfies the covariance matching condition

$$\langle z^k, \hat{\Phi} \rangle = r_k, \quad k = 0, 1, \ldots, n, \tag{3.3}$$

and that it is the unique such interpolant. This is the main result in [6].

In [3, 4] it was shown that this problem is the dual (in the sense of mathematical programming) of the problem of finding the unique spectral density Φ which minimizes the linear combination

$$p_0 c_0 + p_1 c_1 + \cdots + p_m c_m$$

of the cepstral coefficients

$$c_k \stackrel{\triangle}{=} \langle z^k, \log \Phi \rangle, \quad k = 0, 1, \ldots, m, \tag{3.4}$$

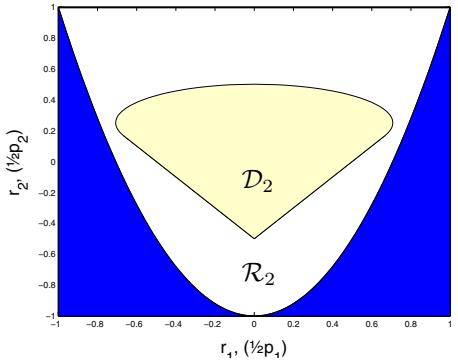


FIG. 3.1. *Comparison of the region of positive covariance sequences $\mathcal{R}_2$ and the region of positive pseudopolynomials $\mathcal{D}_2$.*

i.e., the functional

$$\langle P, \log \Phi \rangle$$

subject to the covariance matching condition (3.3). As a special case, we have the problem to maximize the entropy gain $\mathcal{E}(\Phi)$ subject to (3.3) which yields the well-known maximum entropy solution to the covariance extension problem. This duality is considered in [6].

Covariance matching with fixed denominator. Whereas, for each $P \in \mathcal{D}_m^+$, there is one and only one $Q \in \mathcal{D}_n^+$ such that

$$\Phi = \frac{P}{Q}$$

matches a prescribed window $r_0, r_1, \dots, r_n$ of covariance lags, in general there does not exist a Φ with a prescribed denominator Q matching $r_0, r_1, \dots, r_m$. This is illustrated by the following example, where $Q = 1$.

Example 3.1. Consider a MA(2) process defined by $\sigma(z) = z^2 + \sigma_1 z + \sigma_2$, and the corresponding pseudopolynomial $P = \sigma(z)\sigma(z^{-1})$ normalized so that $P \in \hat{\mathcal{D}}_2^+$. Then, in view of (1.6) and (2.1), $p_1/2$ and $p_2/2$ are the normalized covariances of the MA(2) process and they take values in the convex region $\mathcal{D}_2$ in Figure 3.1, defined by the curve $p_1 = 2\sqrt{2p_2(1-p_2)}$ and the lines $p_1 = 1 + p_2$ and $p_1 = -1 - p_2$. The covariances $r_0, r_1, r_2 \in \mathcal{R}_2$, normalized so that $r_0 = 1$, correspond to the convex region $\mathcal{R}_2$ restricted by the parabola $r_2 = 2r_1^2 - 1$ and the line $r_2 = 1$. It is clear that there does not exist a MA(2) process with covariances taken arbitrarily in $\mathcal{R}_2$. This is a special case of the fact that sample covariances do not form sufficient statistics for the MA parameter estimation problem.

Nonetheless, we may consider the optimization problem

$$(\mathcal{R}) \quad \left[\begin{array}{ll} \min\limits_{\mathbf{f} \in \boldsymbol{\mathcal{F}}^+} & \frac{1}{2}\langle \Phi^2, Q \rangle, \\ \text{s.t.} & \langle \Phi, z^k \rangle = r_k, \quad k = 0, 1, \dots, m, \end{array} \right]$$

where $\mathbf{f} = (f_0, f_1, \dots) \in \boldsymbol{\mathcal{F}}^+$ are the Fourier coefficients,

$$\Phi(z) = \sum_{k=-\infty}^{\infty} f_k z^{-k}, \quad f_{-k} = f_k,$$

to be chosen so that $\Phi(e^{i\theta}) > 0$.

Here the objective function

$$\mathbb{F}_Q(\mathbf{f}) \stackrel{\triangle}{=} \frac{1}{2}\langle \Phi^2, Q\rangle \tag{3.5}$$

puts a large cost on the peaks of Φ, at least where Q is not small, and this will thus provide a spectral density without unnecessary poles.

Taking $p_0, p_1, \ldots, p_m$ to be the Lagrange multipliers, the Lagrangian becomes

$$\begin{aligned} L_Q^{\mathcal{R}}(\mathbf{f}, \mathbf{p}) &= \frac{1}{2}\langle \Phi^2, Q\rangle - \sum_{k=0}^{m} p_k(\langle \Phi, z^k\rangle - r_k) \\ &= \sum_{k=0}^{m} p_k r_k + \frac{1}{2}\langle \Phi^2, Q\rangle - \langle \Phi, P\rangle. \end{aligned} \tag{3.6}$$

Clearly $\inf\{L_Q^{\mathcal{R}}(\mathbf{f}, \mathbf{p}) | \mathbf{f} \in \boldsymbol{\mathcal{F}}^+\} > -\infty$ if and only if $Q \in \mathcal{D}_n^+$. Then $\mathbf{f} \mapsto L_Q^{\mathcal{R}}(\mathbf{f}, \mathbf{p})$ has a unique minimum in $\hat{\mathbf{f}} \in \boldsymbol{\mathcal{F}}^+$ if and only if

$$\frac{\partial L_Q^{\mathcal{R}}}{\partial f_k} = \langle z^k, \Phi Q - P\rangle = 0, \quad k = 0, 1, 2, \ldots, \tag{3.7}$$

in this point, which in turn holds if and only if

$$\Phi = \frac{P}{Q},$$

requiring that $P \in \mathcal{D}_m^+$. Then, the dual functional $\psi : \boldsymbol{\mathcal{D}}_m^+ \to \mathbb{R}$ becomes

$$\begin{aligned} \psi(\mathbf{p}) &\stackrel{\triangle}{=} \inf_{\mathbf{f} \in \boldsymbol{\mathcal{F}}^+} L_Q^{\mathcal{R}}(\mathbf{f}, \mathbf{p}) \\ &= \sum_{k=0}^{n} r_k p_k - \frac{1}{2}\left\langle 1, \frac{P^2}{Q}\right\rangle. \end{aligned} \tag{3.8}$$

This is a strictly concave function. Consequently, we have the dual problem

$$(\mathcal{M}) \quad \left[\max_{\mathbf{p} \in \boldsymbol{\mathcal{D}}_m^+} \psi(\mathbf{p}) \right].$$

If there is a maximizing solution to problem $(\mathcal{M})$, then the gradient

$$\frac{\partial \psi}{\partial p_k} = r_k - \left\langle z^k, \frac{P}{Q}\right\rangle, \quad k = 0, 1, \ldots, m, \tag{3.9}$$

must be zero in this point, enforcing the covariance matching condition

$$\langle z^k, \Phi\rangle = r_k, \quad k = 0, 1, \ldots, m. \tag{3.10}$$

However, unlike (3.2), this gradient is finite on the boundary, and there is no reason why the optimum might not be there.

THEOREM 3.1. *If $Q \in \mathcal{D}_n^+$, then problem $(\mathcal{R})$ has a unique solution. Furthermore, if the dual function ψ has a maximum $\hat{\mathbf{p}} \in \boldsymbol{\mathcal{D}}_m^+$, which then must be unique, then*

$$\Phi = \frac{\hat{P}}{Q}$$

is the unique solution of problem $(\mathcal{R})$.

Proof. If $Q \in \mathcal{D}_n^+$, then $\mathbb{F}_Q(\mathbf{f})$ is strictly convex, and since the feasible region is the intersection of the convex cone $\boldsymbol{\mathcal{F}}^+$ and the linear constraints, problem $(\mathcal{R})$ has a unique optimum. Let $\hat{\mathbf{p}} \in \boldsymbol{\mathcal{D}}_m^+$ be the unique solution to the dual problem $(\mathcal{M})$, let $\hat{P} \in \mathcal{D}_n^+$ be the corresponding pseudopolynomial, and let

$$\hat{f}_k = \left\langle z^k, \frac{\hat{P}}{Q} \right\rangle.$$

Clearly $\hat{\mathbf{f}} \in \boldsymbol{\mathcal{F}}^+$. Since the gradient (3.9) is zero for $\mathbf{p} = \hat{\mathbf{p}}$, the covariance matching condition (3.10) is fulfilled for $\mathbf{f} = \hat{\mathbf{f}}$, and, therefore, $\mathbb{F}_Q(\hat{\mathbf{f}}) = L_Q^{\mathcal{R}}(\hat{\mathbf{f}}, \hat{\mathbf{p}})$. But by the construction above,

$$L_Q^{\mathcal{R}}(\hat{\mathbf{f}}, \hat{\mathbf{p}}) = \inf_{\mathbf{f} \in \boldsymbol{\mathcal{F}}^+} L_Q^{\mathcal{R}}(\mathbf{f}, \hat{\mathbf{p}}) \le L_Q^{\mathcal{R}}(\mathbf{f}, \hat{\mathbf{p}}) \quad \forall \mathbf{f} \in \boldsymbol{\mathcal{F}}^+.$$

Then for any $\mathbf{f} \in \boldsymbol{\mathcal{F}}^+$ which satisfies the covariance matching condition (3.10),

$$\mathbb{F}_Q(\mathbf{f}) = L_Q^{\mathcal{R}}(\mathbf{f}, \hat{\mathbf{p}}) \ge \mathbb{F}_Q(\hat{\mathbf{f}}),$$

which establishes the optimality of $\hat{\mathbf{f}}$. ∎

Cepstral matching with fixed denominator. An interesting question is whether a symmetric cepstral matching optimization problem can be formulated. That is, given a cepstral window

$$c_0, c_1, \dots, c_m,$$

consider the optimization problem to find a spectral density Φ which minimizes the linear combination

$$q_0 r_0 + q_1 r_1 + \cdots + q_n r_n$$

of theoretical covariances $r_k = \langle z^k, \Phi \rangle$ subject to the cepstral matching constraint

$$\langle z^k, \log \Phi \rangle = c_k, \quad k = 0, 1, \dots, m, \tag{3.11}$$

i.e., the optimization problem

$$(\mathcal{S}) \quad \left[\begin{array}{ll} \min\limits_{\mathbf{g} \in \boldsymbol{\mathcal{G}}} & \mathbb{I}_Q(\mathbf{g}), \\ \text{s.t.} & \langle \log \Phi, z^k \rangle = c_k, \quad k = 0, 1, \dots, m, \end{array} \right],$$

where

$$\mathbb{I}_Q(\mathbf{g}) \triangleq \langle \Phi, Q \rangle \tag{3.12}$$

and $\mathbf{g} = (g_0, g_1, \dots) \in \boldsymbol{\mathcal{G}}$ are the Fourier coefficients of the logarithm of the spectral density

$$\log \Phi(z) = \sum_{k=-\infty}^{\infty} g_k z^{-k}, \quad g_{-k} = g_k.$$

There is a dual optimization problem to problem $(\mathcal{S})$ (see [12]), namely, to find the supremum of

$$\mathbb{J}_Q(\mathbf{p}) = p_0 + \sum_{k=0}^{n} c_k p_k - \left\langle P, \log \frac{P}{Q} \right\rangle \tag{3.13}$$

over all $\mathbf{p} \in \boldsymbol{\mathcal{D}}_m$. Although this function is strictly concave, as is $-\mathbb{J}_P$, unlike $\mathbb{J}_P$ it may not have an optimum in the open region $\boldsymbol{\mathcal{D}}_m^+$. In fact, the gradient

$$\frac{\partial \mathbb{J}_Q}{\partial p_k} = c_k - \left\langle z^k, \log \frac{P}{Q} \right\rangle, \quad k = 0, 1, \ldots, n, \tag{3.14}$$

is finite on the boundary, by Szegö's theorem [9], so there is no guarantee that the stationary point is an interior point.

However, if there is a stationary point $\hat{\mathbf{p}} \in \boldsymbol{\mathcal{D}}_m^+$ at which the gradient (3.14) is zero, it is unique and

$$\hat{\Phi} = \frac{\hat{P}}{Q}$$

satisfies the cepstral matching condition (3.11).

THEOREM 3.2. *If $Q \in \mathcal{D}_n^+$, then problem $(\mathcal{S})$ has a unique solution. Furthermore, if the function $\mathbb{J}_Q$ has a maximum $\hat{\mathbf{p}} \in \boldsymbol{\mathcal{D}}_m^+$, which then must be unique, then*

$$\Phi = \frac{\hat{P}}{Q}$$

is the unique solution of problem $(\mathcal{S})$.

The proof follows the lines of Theorem 3.1 and is omitted.

Cepstral matching with fixed numerator. It is easy to see that we can instead prescribe the numerator polynomial P in cepstrum interpolation. In fact, if

$$\Phi = \frac{P}{Q}$$

is a coercive spectral density, then so is

$$\Phi^{-1} = \frac{Q}{P}.$$

Moreover, the cepstral matching condition (3.11) may be written as

$$\langle z^k, \log \Phi^{-1} \rangle = -c_k, \quad k = 0, 1, \ldots, n. \tag{3.15}$$

Consequently, there is an analogous optimization problem to find a coercive spectral density Φ minimizing

$$\langle \Phi^{-1}, P \rangle$$

subject to (3.15), which has a dual problem to maximize

$$\hat{\mathbb{J}}_P(\mathbf{q}) = q_0 - \sum_{k=0}^{n} c_k q_k - \left\langle Q, \log \frac{Q}{P} \right\rangle.$$

As before, an interpolant Φ with prescribed numerator P and prescribed cepstral window $(c_0, c_1, \dots, c_n)$ exists and is then unique if and only if $\hat{\mathbb{J}}_P$ has a maximum in $\boldsymbol{\mathcal{D}}_n^+$.

In this section some new optimization problems for cepstral and covariance matching have been proposed. However, like the simultaneous covariance and cepstral matching problem revisited in section 2, these problems do not have the consistency of problem $(\mathcal{P}_0)$ in that solutions may not exist (see Table 3.1). It is, therefore, natural to ask for good approximate solutions, based on the same principles. This will be done via regularization, a topic to be addressed in the next section.

4. Regularization of problem $(\boldsymbol{\mathcal{P}})$. Since the optimization problem $(\mathcal{P})$ defined in section 2 may have solutions on the boundary of the feasible region for $\mathbf{p}$, a regularization of that problem is proposed in this section. In order to force the solution to the interior of the feasible region, the optimization problem $(\mathcal{P})$ is modified by adding a barrier-like term to the objective function φ. In problem $(\mathcal{P}_0)$, the term $-\langle P, \log Q\rangle$ of the objective function $\mathbb{J}_P(\mathbf{q})$ pushes Q from the boundary. By adding the term $\beta = -\langle 1, \log P\rangle$ to the objective function φ of the problem $(\mathcal{P})$ the optimal P is forced into the interior of $\hat{\mathcal{D}}_m$.

It follows from [6] that $\beta = -\langle 1, \log P\rangle$ has bounded level sets, is finite for all $P \in \hat{\mathcal{D}}_m$, and has a derivative that tends to infinity as P tends to the boundary of $\hat{\mathcal{D}}_m$. It is further strictly convex in P. Since the function values of β are finite at the boundary, it is not a barrier function from a mathematical programming viewpoint [19], but since the derivatives of β with respect to P are infinite, it will constrain the optimum away from the boundary as a barrier function would. It is shown later that this term will also increase the entropy of the solution.

The regularized optimization problem is formulated as

$$(\mathcal{P}_\lambda) \quad \left[\begin{array}{ll} \min & \varphi_\lambda(\mathbf{p}, \mathbf{q}), \\ \text{s.t.} & (\mathbf{p}, \mathbf{q}) \in \hat{\boldsymbol{\mathcal{D}}}_m \times \boldsymbol{\mathcal{D}}_n \end{array} \right],$$

where the modified objective function is given by

$$\varphi_\lambda(\mathbf{p}, \mathbf{q}) \triangleq -\sum_{k=1}^{m} c_k p_k + \sum_{k=0}^{n} r_k q_k + \left\langle P, \log \frac{P}{Q} \right\rangle - \langle \lambda, \log P\rangle. \tag{4.1}$$

LEMMA 4.1. *For each $\lambda > 0$, the function φ_λ has compact sublevel sets; i.e., for all $\mu \in \mathbb{R}$, $\varphi_\lambda^{-1}(-\infty, \mu]$ is compact.*

Proof. First note that the set $\hat{\boldsymbol{\mathcal{D}}}_m$ of all $\mathbf{p}$ such that $P \in \mathcal{D}_m$ and $p_0 = 1$ is compact. In fact, let

$$\sigma(z) = \sigma_0 z^m + \sigma_1 z^{m-1} + \cdots + \sigma_m$$

be the stable spectral factor of P, i.e., the Schur polynomial σ such that $\sigma\sigma^* = P$. Then

$$\sigma_0^2 + \sigma_1^2 + \cdots + \sigma_m^2 = p_0 = 1,$$

and hence $|\sigma_k| \le 1$ for $k = 0, 1, \dots, m$. Then it is easy to check that $|p_k| \le 2m$ for each $k = 0, 1, \dots, m$ proving compactness, since $\hat{\boldsymbol{\mathcal{D}}}_m$ is closed.

Now suppose that $(\mathbf{p}^{(k)}, \mathbf{q}^{(k)})$ is a sequence in $\varphi_\lambda^{-1}(-\infty, \mu]$. To show that $\varphi_\lambda^{-1}(-\infty, \mu]$ is compact, it suffices to show that $(\mathbf{p}^{(k)}, \mathbf{q}^{(k)})$ has a subsequence which

converges to a point in $\varphi_\lambda^{-1}(-\infty, \mu]$. Clearly $\mathbf{p}^{(k)} \in \hat{\boldsymbol{\mathcal{D}}}_m$, which is compact, so it has a convergent subsequence. As for $\mathbf{q}^{(k)}$, we can write

$$Q^{(k)}(z) = \rho_k \hat{Q}^{(k)}(z),$$

where $\hat{q}_0 = 1$ so that $\hat{\mathbf{q}}^{(k)}$ belongs to the compact set $\hat{\boldsymbol{\mathcal{D}}}_n$. Therefore, the sequence $\mathbf{q}^{(k)}$ has a convergent subsequence if and only if ρ_k has. However,

$$\varphi_\lambda(\mathbf{p}^{(k)}, \mathbf{q}^{(k)}) = \alpha_k + \beta_k \rho_k - \gamma_k \log \rho_k,$$

where

$$\alpha_k \stackrel{\triangle}{=} -\sum_{j=1}^{m} c_j p_j^{(k)} + \langle P^{(k)} - \lambda, \log P^{(k)} \rangle - \langle P^{(k)}, \log \hat{Q}^{(k)} \rangle$$

is bounded from above and below,

$$\beta_k \stackrel{\triangle}{=} \sum_{j=0}^{n} r_j \hat{q}_j^{(k)} = \langle \Phi_{\mathrm{ME}}, \hat{Q}^{(k)} \rangle > 0$$

and bounded, where Φ_{ME} denotes the maximum entropy spectral density with covariances $r_0, r_1, \ldots, r_n$, and

$$\gamma_k \stackrel{\triangle}{=} \langle P^{(k)}, 1 \rangle > 0$$

and bounded. From this we see that $\varphi_\lambda(\mathbf{p}^{(k)}, \mathbf{q}^{(k)})$ would exceed μ if either ρ_k were to tend to infinity or to zero. Hence $\{\rho_k\}$ is bounded with a convergent subsequence as claimed. □

It is easy to see that φ_λ is strictly convex on its closed convex domain. Therefore, Lemma 4.1 implies that φ_λ achieves a unique minimum in $\hat{\boldsymbol{\mathcal{D}}}_m \times \boldsymbol{\mathcal{D}}_n$. If this minimum is located in the interior of $\hat{\boldsymbol{\mathcal{D}}}_m \times \boldsymbol{\mathcal{D}}_n$, the gradient

$$\frac{\partial \varphi_\lambda}{\partial p_k} = -c_k + \left\langle z^k, \log \frac{P}{Q} \right\rangle - \left\langle z^k, \frac{\lambda}{P} \right\rangle, \quad k = 1, \ldots, m, \tag{4.2}$$

$$\frac{\partial \varphi_\lambda}{\partial q_k} = r_k - \left\langle z^k, \frac{P}{Q} \right\rangle, \quad k = 0, 1, \ldots, n, \tag{4.3}$$

is zero at this point. The barrier term will make sure that the optimum will occur at an interior point. This is guaranteed by the following lemma, which is proved along the lines of Lemma 5.4 in [6].

LEMMA 4.2. *If $\lambda > 0$, the function φ_λ never attains a minimum on the boundary $\partial(\hat{\boldsymbol{\mathcal{D}}}_m \times \boldsymbol{\mathcal{D}}_n)$.*

Hence we have the following theorem.

THEOREM 4.3. *For each $\lambda > 0$, problem $(\mathcal{P}_\lambda)$ has a unique solution in $\hat{\boldsymbol{\mathcal{D}}}_m^+ \times \boldsymbol{\mathcal{D}}_n^+$. At this point*

$$\left\langle z^k, \frac{P}{Q} \right\rangle = r_k, \qquad k = 0, 1, \ldots, n, \tag{4.4}$$

$$\left\langle z^k, \log \frac{P}{Q} \right\rangle = c_k + \lambda \epsilon_k, \quad k = 1, \ldots, m, \tag{4.5}$$

where $\epsilon_k = \langle z^k, \frac{1}{P} \rangle$.

Proof. At the stationary point $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ the gradient is zero,

$$\frac{\partial \varphi_\lambda}{\partial p_k} = -c_k + \left\langle z^k, \log \frac{\hat{P}}{\hat{Q}} \right\rangle - \left\langle z^k, \frac{\lambda}{\hat{P}} \right\rangle = 0, \quad k = 1, \dots, m, \tag{4.6}$$

$$\frac{\partial \varphi_\lambda}{\partial q_k} = r_k - \left\langle z^k, \frac{\hat{P}}{\hat{Q}} \right\rangle = 0, \quad k = 0, 1, \dots, n, \tag{4.7}$$

and at that point

$$\frac{\hat{P}(z)}{\hat{Q}(z)} = \sum_{k=-\infty}^{\infty} R_k z^{-k},$$

where $R_k = r_k$ for $k = 0, 1, \dots, n$, and

$$\log \frac{\hat{P}(z)}{\hat{Q}(z)} = \sum_{k=-\infty}^{\infty} C_k z^{-k},$$

where $C_k = c_k + \lambda \epsilon_k$ for $k = 1, \dots, m$, with ϵ_k defined by the Laurent series

$$\frac{1}{\hat{P}(z)} = \sum_{k=-\infty}^{\infty} \epsilon_k z^{-k}. \tag{4.8}$$

□

If we would like to choose λ so that the error $\lambda \epsilon_k$ in the interpolation of the cepstral parameter c_k is as small as possible, there might seem to exist two ways to achieve a zero error. The first way is to choose $\lambda = 0$, in which case problem $(\mathcal{P}_\lambda)$ reduces to problem $(\mathcal{P})$. If problem $(\mathcal{P})$ has a solution in the interior of $\hat{\boldsymbol{\mathcal{D}}}_m \times \boldsymbol{\mathcal{D}}$, the coefficients ϵ_k are finite and the error will be zero. But if the solution is at the boundary, the expansion of $1/P$ will diverge, and the product $\lambda \epsilon_k$ will not tend to zero as λ tends to zero. The other way is to have $\epsilon_k = 0$ for $k = 1, \dots, m$. This is accomplished if P is constant, i.e., $P = p_0$. But since the solution then corresponds to the maximum entropy solution, and this solution does not interpolate generic cepstral parameters, λ must tend to infinity in order for P to tend to p_0. The connection to entropy is studied after the following example.

Example 4.1. An identification experiment is carried out to study the effect of the parameter λ on the filters determined by problem $(\mathcal{P}_\lambda)$. A MA(8) filter was driven by white noise to generate a data sequence of length 512, and from this data the variance $\hat{r}_0$ was estimated using (1.7) and cepstral parameters were estimated as described in Figure 1.2. Using problem $(\mathcal{P}_\lambda)$ with $m = 8$, $n = 0$, and six different values of λ, MA(8) filters were determined and the L_2 norm of the error, relative to the generating filter, was calculated. This was repeated for 100 MA(8) filters, and the mean of the L_2 norm of the errors for each value of λ was determined. As depicted in Figure 4.1, the mean L_2 norm of the error decreases as λ decreases, and it is almost constant for $\lambda < 10^{-3}$. This indicates that λ should be chosen small enough but not necessarily very close to zero. In Table 4.1 the means of the relative cepstral estimation errors and the means of the relative correction terms $\lambda \epsilon_k$, relative to the sums $\sum_{\ell=1}^{8} |c_\ell|$, are displayed in percents. It is clear that the estimated cepstral values are approximated better the smaller the value of λ is and that the correction term is about the same size as the estimation error for λ around $0.01 - 0.1$ and less than 10% of the estimation error for $\lambda \leq 0.0001$.

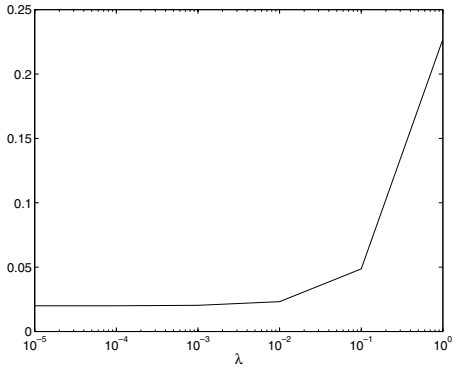


FIG. 4.1. *The L_2 norm for the errors of the method for different values of λ.*

TABLE 4.1
The means of the relative cepstral estimation errors and the correction terms in percents.

Cepstrum	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
Est. error	2.58	2.31	2.36	2.53	2.59	2.15	2.29	2.33
Corr. terms	$\lvert\lambda\epsilon_1\rvert$	$\lvert\lambda\epsilon_2\rvert$	$\lvert\lambda\epsilon_3\rvert$	$\lvert\lambda\epsilon_4\rvert$	$\lvert\lambda\epsilon_5\rvert$	$\lvert\lambda\epsilon_6\rvert$	$\lvert\lambda\epsilon_7\rvert$	$\lvert\lambda\epsilon_8\rvert$
$\lambda = 1$	9.51	7.10	5.20	5.24	4.35	5.61	3.16	4.32
$\lambda = 0.1$	3.91	3.41	2.54	2.73	2.25	2.85	1.68	2.32
$\lambda = 0.01$	1.11	1.12	0.87	1.02	0.86	1.01	0.68	0.90
$\lambda = 0.001$	0.38	0.38	0.34	0.38	0.35	0.37	0.31	0.36
$\lambda = 0.0001$	0.23	0.22	0.23	0.24	0.22	0.22	0.21	0.23
$\lambda = 0.00001$	0.21	0.19	0.20	0.21	0.20	0.20	0.19	0.21

If it is known that the estimated model is in the model class, to get consistent estimates the parameter λ should tend to zero as the number of data points tends to infinity and the estimation error tends to zero. In practice, such knowledge about the estimated model is uncommon, consistency cannot be guaranteed, and it is suggested that λ tends to some small positive value.

The entropy of a process is defined as in (2.5) (see, for example, [2]). For problem $(\mathcal{P})$ with fixed Q, the barrier term is equal to $-\mathcal{E}(\Phi)$ up to a constant. The barrier term is thus a measure of the entropy of the process. Then λ is a weight that determines how much the entropy should be maximized relative to the importance of the interpolation.

Next we will show that the entropy of the solution to problem $(\mathcal{P}_\lambda)$ is monotonically nondecreasing in λ. From this we can see that the price of a decrease in λ is a decreased entropy.

PROPOSITION 4.1. *The entropy of the filter Φ_λ, which solves problem $(\mathcal{P}_\lambda)$, is a monotonic nondecreasing function of λ.*

Proof. Denote the entropy of the filter $\Phi_\lambda = P_\lambda/Q_\lambda$ by $\mathcal{E}(\Phi_\lambda) = \langle \log(P_\lambda/Q_\lambda), 1\rangle$. This is a differentiable function of λ. Then

$$\frac{\mathrm{d}\mathcal{E}(\Phi_\lambda)}{\mathrm{d}\lambda} = \left\langle \frac{1}{P_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle - \left\langle \frac{1}{Q_\lambda}\frac{\mathrm{d}Q_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle, \tag{4.9}$$

and taking the derivative of (4.2) and (4.3) with respect to λ gives

$$\left\langle \frac{1}{P_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda} - \frac{1}{Q_\lambda}\frac{\mathrm{d}Q_\lambda}{\mathrm{d}\lambda}, z^k \right\rangle - \left\langle \frac{1}{P_\lambda} - \frac{\lambda}{P_\lambda^2}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}, z^k \right\rangle = 0, \quad k = 1, \dots, m, \tag{4.10}$$

$$-\left\langle \frac{1}{Q_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}, z^k \right\rangle + \left\langle \frac{P_\lambda}{Q_\lambda^2}\frac{\mathrm{d}Q_\lambda}{\mathrm{d}\lambda}, z^k \right\rangle = 0, \quad k = 0, 1, \dots, n. \tag{4.11}$$

Taking the linear combination of (4.11) corresponding to Q_λ gives

$$-\left\langle \frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle + \left\langle \frac{P_\lambda}{Q_\lambda}\frac{\mathrm{d}Q_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle = 0, \tag{4.12}$$

and that of (4.10) corresponding to P_λ yields

$$\left\langle \frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle - \left\langle \frac{P_\lambda}{Q_\lambda}\frac{\mathrm{d}Q_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle - 1 + \left\langle \frac{\lambda}{P_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle = p_0 \left\langle \frac{1}{P_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda} + \frac{\lambda}{P_\lambda^2}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda} - \frac{1}{Q_\lambda}\frac{\mathrm{d}Q_\lambda}{\mathrm{d}\lambda} - \frac{1}{P_\lambda}, 1 \right\rangle. \tag{4.13}$$

Setting $p_0 = 1$, (4.12) and (4.13) yield

$$-\left\langle \frac{1}{Q_\lambda}\frac{\mathrm{d}Q_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle = -1 + \lambda \left\langle \frac{1}{P_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle - \left\langle \frac{1}{P_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda} + \frac{\lambda}{P_\lambda^2}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda} - \frac{1}{P_\lambda}, 1 \right\rangle, \tag{4.14}$$

which inserted into (4.9) gives

$$\frac{\mathrm{d}\mathcal{E}(\Phi_\lambda)}{\mathrm{d}\lambda} = \left\langle \left(\frac{1}{P_\lambda} - 1\right)\left(1 - \lambda \frac{1}{P_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}\right), 1 \right\rangle.$$

For $\lambda = 0$ this reduces to

$$\frac{\mathrm{d}\mathcal{E}(\Phi_0)}{\mathrm{d}\lambda} = \left\langle \frac{1}{P_\lambda} - 1, 1 \right\rangle,$$

which is nonnegative by Lemma A.1 (see the appendix). We have thus proven that $\mathcal{E}(\Phi_\lambda)$ is a nondecreasing function in λ for $\lambda = 0$. For an arbitrary $\breve{\lambda} > 0$, let $\lambda = \breve{\lambda} + \tilde{\lambda}$ and $\breve{c}_k = c_k - \breve{\lambda}\breve{\epsilon}_k$, where

$$\frac{1}{P_{\breve{\lambda}}(z)} = \sum_{k=-\infty}^{\infty} \breve{\epsilon}_k z^{-k}.$$

Also, let $\breve{\varphi}_{\tilde{\lambda}}$ be given by (4.1) but with the parameters c_k replaced by $\breve{c}_k$. Then the equality

$$\frac{\partial \varphi_\lambda}{\partial p_k}(P_{\breve{\lambda}}, Q_{\breve{\lambda}}) = \frac{\partial \breve{\varphi}_{\tilde{\lambda}}}{\partial p_k}(P_{\breve{\lambda}}, Q_{\breve{\lambda}}) = 0$$

holds and the argument above applied on $\breve{\varphi}_{\tilde{\lambda}}$ instead of φ_λ holds for $\tilde{\lambda} = 0$; i.e., the derivative $\frac{\mathrm{d}\mathcal{E}(\Phi_\lambda)}{\mathrm{d}\lambda} \geq 0$ for $\lambda = \breve{\lambda}$. □

What happens to the solution if λ tends to infinity? If λ tends to infinity, the minimization over $\mathbf{p} \in \hat{\boldsymbol{\mathcal{D}}}$ will tend to minimize the barrier term $-\langle 1, \log P\rangle$. Taking $R = 1$ in the following lemma, which is proven in the appendix, the optimal P is seen to be $\hat{P}(z) = 1$.

LEMMA 4.4. *Let $P, R \in \hat{\mathcal{D}}_m^+$; then*

$$-\langle R, \log R\rangle \leq -\langle R, \log P\rangle.$$

Inserting $\hat{P} = 1$ into the objective function φ_λ results in a function equivalent to the maximum entropy objective function $\mathbb{J}_{\hat{P}}(\mathbf{q})$. Consequently, as λ tends to infinity, the resulting filter tends towards the maximum entropy filter.

Remark 4.1. It is clear from Lemma 4.4 that if a barrier of the form $\langle R, \log P\rangle$ is used, where R is in $\mathcal{D}_m^+$, the solution will tend to the LLN filter corresponding to $P = R$ as λ tends to infinity. Such a barrier will also guarantee a unique interior point solution, but the analogue of Proposition 4.1 does not hold. If an initial estimate R of P exists, this barrier will improve the solution by merging the information given in the interpolation parameters and in the initial estimate R.

Standard optimization algorithms can be used to solve problem $(\mathcal{P}_\lambda)$. For example, a damped Newton method, similar to the one proposed in [3] for problem $(\mathcal{P}_0)$, can be used. The best performance is probably achieved by using a homotopy method similar to the one presented in [11]. Since the function φ_λ is strictly convex, the global optimum can always be found. The strict convexity follows by considering the second derivatives of φ_λ in direction $(\delta\mathbf{p}, \delta\mathbf{q})$ at $(\mathbf{p}, \mathbf{q})$,

$$\left\langle (Q\delta P - P\delta Q)^2, \frac{1}{PQ^2} \right\rangle + \lambda \left\langle (\delta P)^2, \frac{1}{P^2} \right\rangle > 0 \quad \text{for } \delta\mathbf{p}, \delta\mathbf{q} \text{ not both zero.}$$

Since the Hessian is positive definite for $\lambda > 0$, φ_λ is strictly convex for all $\lambda > 0$.

5. Regularization of problem $(\mathcal{M})$. Next we return to problem $(\mathcal{M})$ introduced in section 3. There is no intrinsic reason why ψ should have a minimum in $\boldsymbol{\mathcal{D}}_m^+$ for generic values of the covariance lags, merely satisfying the usual positivity (Toeplitz) condition. In fact, it was demonstrated in [7, 3, 4] that covariance matching cannot be achieved with arbitrary poles, only with arbitrary zeros, and illustrated again in Example 3.1.

There is thus a need for a regularization of problem $(\mathcal{M})$. Reformulating problem $(\mathcal{M})$ as a minimization problem, the same barrier term β that was used for problem $(\mathcal{P})$ can be used here. To this end, given the window $r_0, r_1, \ldots, r_m \in \mathcal{R}_m$ of covariance lags and the pseudopolynomial $Q \in \mathcal{D}_n^+$, for each $\lambda > 0$, introduce the regularized problem

$$(\mathcal{M}_\lambda) \quad \left[\min_{\mathbf{p} \in \boldsymbol{\mathcal{D}}_m} \quad \psi_\lambda(\mathbf{p}) \right],$$

where $\psi_\lambda : \boldsymbol{\mathcal{D}}_m \to \mathbb{R}$ is given by

$$\psi_\lambda(\mathbf{p}) \triangleq \frac{1}{2} \left\langle 1, \frac{P^2}{Q} \right\rangle - \sum_{k=0}^{m} r_k p_k - \lambda \langle 1, \log P \rangle.$$

It is easy to see that ψ_λ is strictly convex for each $\lambda > 0$, and we know from before that $\boldsymbol{\mathcal{D}}_m$ is convex. Consequently, it follows from the following lemma that problem $(\mathcal{M}_\lambda)$ has a unique minimum.

LEMMA 5.1. *For each $\lambda > 0$, the function ψ_λ has compact sublevel sets; i.e., for all $\mu \in \mathbb{R}$, $\psi_\lambda^{-1}(-\infty, \mu]$ is compact.*

Proof. First note as before that the set $\hat{\boldsymbol{\mathcal{D}}}_m$ is compact. Next suppose that $\mathbf{p}^{(k)}$ is a sequence in $\psi_\lambda^{-1}(-\infty, \mu]$. To show that $\psi_\lambda^{-1}(-\infty, \mu]$ is compact, it suffices to show that $\mathbf{p}^{(k)}$ has a subsequence which converges to a point in $\psi_\lambda^{-1}(-\infty, \mu]$. We can write

$$P^{(k)}(z) = \rho_k \hat{P}^{(k)}(z),$$

where $\hat{p}_0 = 1$ so that $\hat{\mathbf{p}}^{(k)}$ belongs to the compact set $\hat{\boldsymbol{\mathcal{D}}}_m$. Therefore, the sequence $\mathbf{p}^{(k)}$ has a convergent subsequence if and only if ρ_k has. However,

$$\psi_\lambda(\mathbf{p}^{(k)}) = \alpha_k + \beta_k\rho_k + \gamma_k\rho_k^2 - \lambda\log\rho_k,$$

where

$$\begin{aligned}\alpha_k &\triangleq -\lambda\langle 1, \log\hat{P}^{(k)}\rangle,\\ \beta_k &\triangleq -\sum_{j=0}^{n} r_j\hat{p}_j^{(k)},\\ \gamma_k &\triangleq \frac{1}{2}\left\langle 1, \frac{(\hat{P}^{(k)})^2}{Q}\right\rangle > 0\end{aligned}$$

are all bounded from above and below.

From this we see that $\psi_\lambda(\mathbf{p}^{(k)})$ would exceed μ if either ρ_k were to tend to infinity or to zero. Hence $\{\rho_k\}$ is bounded with a convergent subsequence as claimed. ∎

The barrier term in ψ_λ ensures that the minimum does not occur on the boundary $\partial\mathcal{D}_m$ of $\mathcal{D}_m$ but in the interior $\mathcal{D}_m^+$. In fact, the following lemma can be proven along the same lines as Lemma 5.4 in [6].

LEMMA 5.2. *If $\lambda > 0$, the function ψ_λ never attains a minimum on the boundary $\partial\mathcal{D}_m$.*

Since thus the unique minimum of ψ_λ lies in the interior of $\mathcal{D}_m$, the gradient

$$\frac{\partial\psi_\lambda}{\partial p_k} = \left\langle z^k, \frac{P}{Q}\right\rangle - r_k - \left\langle z^k, \frac{\lambda}{P}\right\rangle, \quad k = 0, 1, \ldots, m, \tag{5.1}$$

is zero at this point. Consequently, we have the following theorem.

THEOREM 5.3. *For each $\lambda > 0$, problem $(\mathcal{M}_\lambda)$ has a unique solution in $\boldsymbol{\mathcal{D}}_m^+$. At this point*

$$\left\langle z^k, \frac{P}{Q}\right\rangle = r_k + \lambda\epsilon_k, \quad k = 0, 1, \ldots, m, \tag{5.2}$$

where $\epsilon_k = \langle z^k, \frac{1}{P}\rangle$.

Proof. This follows immediately from Lemma 5.2 and (5.1). ∎

Example 5.1. Example 4.1 is now revisited and repeated for the MA design method based on problem $(\mathcal{M}_\lambda)$. The influence of the parameter λ on the filters is studied by varying the parameter λ and plotting the error of the estimated filter. Using the same data, covariances were estimated using (1.7), the $(\mathcal{M}_\lambda)$ method was applied, and the L_2 norm for the error of the estimated filters compared to the generating filters was determined as a function of λ. As depicted in Figure 5.1, the L_2 norm of the error varies in a way very similar to Example 4.1; that is, the L_2 norm of the error decreases as λ decreases, and it is almost constant for $\lambda < 10^{-3}$. This indicates that λ should be chosen small enough but not necessarily very close to zero. It also indicates that the method corresponding to problem $(\mathcal{P}_\lambda)$ gives better results than that of problem $(\mathcal{M}_\lambda)$. This suggests that it is better to use cepstra than covariances for estimating zeros. In Table 5.1, the mean of the relative estimation errors of the covariances and the mean of the relative values of the correction terms $\lambda\epsilon_k$, relative to the sums $\sum_{\ell=0}^{8}|r_\ell|$, are displayed in percents. It is clear that the estimated covariance values are approximated better the smaller the value of λ is and that the correction term is

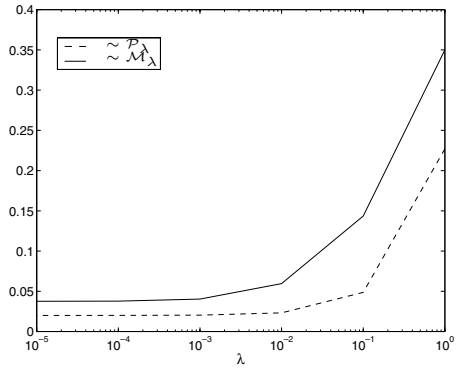


FIG. 5.1. *Comparison of the L_2 norm for the errors of the two methods for different values of λ.*

TABLE 5.1
The means of the relative covariance estimation errors and correction terms in percents.

Covariance	r_0	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
Est. error	2.75	1.80	1.65	1.50	1.57	1.70	1.68	1.89	1.75
Corr. terms	$\lvert\lambda\epsilon_0\rvert$	$\lvert\lambda\epsilon_1\rvert$	$\lvert\lambda\epsilon_2\rvert$	$\lvert\lambda\epsilon_3\rvert$	$\lvert\lambda\epsilon_4\rvert$	$\lvert\lambda\epsilon_5\rvert$	$\lvert\lambda\epsilon_6\rvert$	$\lvert\lambda\epsilon_7\rvert$	$\lvert\lambda\epsilon_8\rvert$
$\lambda = 1$	24.88	3.16	2.24	2.10	1.67	1.47	1.59	1.18	1.18
$\lambda = 0.1$	5.76	1.17	0.84	0.67	0.64	0.52	0.66	0.43	0.54
$\lambda = 0.01$	1.29	0.37	0.30	0.23	0.25	0.19	0.27	0.20	0.28
$\lambda = 0.001$	0.45	0.17	0.16	0.15	0.16	0.13	0.18	0.16	0.22
$\lambda = 0.0001$	0.33	0.14	0.14	0.14	0.14	0.12	0.17	0.16	0.22
$\lambda = 0.00001$	0.32	0.14	0.14	0.14	0.14	0.12	0.17	0.16	0.22

about the same size as the estimation error for λ around $0.1 - 1$ and about 10% of the estimation error for $\lambda \leq 0.001$.

A similar result as in Proposition 4.1 holds here.

PROPOSITION 5.1. *The entropy of the filter Φ_λ, which solves problem $(\mathcal{M}_\lambda)$, is a monotonic nondecreasing function of λ.*

Proof. First note that $\lambda \mapsto \mathcal{E}(\Phi_\lambda)$ is a differentiable function. Then

$$\frac{\mathrm{d}\mathcal{E}(\Phi_\lambda)}{\mathrm{d}\lambda} = \left\langle \frac{1}{P_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle.$$

Moreover, taking the derivative of (5.1) with respect to λ gives

$$\left\langle \frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}\frac{1}{Q}, z^k \right\rangle - \left\langle \frac{1}{P_\lambda}, z^k \right\rangle + \left\langle \lambda \frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}\frac{1}{P_\lambda^2}, z^k \right\rangle = 0, \quad k = 0, 1, \ldots, m. \tag{5.3}$$

Now, forming the appropriate linear combination of (5.3) (corresponding to $\mathrm{d}P_\lambda/\mathrm{d}\lambda$) shows that

$$\frac{\mathrm{d}\mathcal{E}(\Phi_\lambda)}{\mathrm{d}\lambda} = \left\langle \frac{1}{P_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}, 1 \right\rangle = \left\langle \frac{1}{Q} + \frac{\lambda}{P_\lambda^2}, \left(\frac{\mathrm{d}P_\lambda}{\mathrm{d}\lambda}\right)^2 \right\rangle,$$

which is positive for nonnegative λ. This proves the lemma. □

The optimization methods that were suggested for problem $(\mathcal{P}_\lambda)$ should work for problem $(\mathcal{M}_\lambda)$ as well. Since problem $(\mathcal{M}_\lambda)$ is strictly convex the global optimum

can always be obtained. The strict convexity can be shown by considering the second order Fréchet derivative in the direction $\delta\mathbf{p}$ at $\mathbf{p}$

$$(\delta\mathbf{p})^\top \nabla^2_{\mathbf{pp}} \psi_\lambda(\delta\mathbf{p}) = \left\langle (\delta P)^2, \frac{1}{Q} \right\rangle + \lambda \left\langle (\delta P)^2, \frac{1}{P^2} \right\rangle > 0$$

for $\lambda > 0$ and nonzero $\delta\mathbf{p}$. Therefore, the Hessian is positive definite and the objective function ψ_λ is strictly convex.

6. Simulations. To illustrate our methods, we conclude with a few simulations where we fix $\lambda = 0.0005$ throughout.

6.1. MA filter design.

Experiment 1. For comparison, Example 2 in [24] is considered. The two methods were applied to 20 sets of data samples, each of length 400, which were generated from a MA(6) filter with its zeros located at $0.99e^{\pm i(\pi/4)}$, $0.99e^{\pm i(3\pi/4)}$, and $0.999e^{\pm i(\pi/2)}$. The power spectral densities of the estimated filters using problem $(\mathcal{P}_\lambda)$ and problem $(\mathcal{M}_\lambda)$ to estimate MA(6) filters are depicted in Figures 6.1 and 6.2, respectively. For comparison, the power spectral density of the generating filter is included in the graphs as a solid line. It is clear from the figures that the error is smaller for problem $(\mathcal{P}_\lambda)$ than for problem $(\mathcal{M}_\lambda)$. This suggests again that the cepstral lags are better than covariance lags for estimating zeros.

To allow comparison to the examples in [24], the experiments are repeated using enhanced estimates of the interpolation parameters. The covariance estimates are enhanced using the algorithm in [24] with $m = 40$. These estimates are determined by taking the least-squares error prediction of the covariances $r_0, r_1, \ldots, r_n$ using the estimates $\hat{r}_0, \hat{r}_1, \ldots, \hat{r}_m$, given that the true covariances $r_{n+1}, r_{n+2}, \ldots, r_m$ are zero. Due to the lack of an analogous enhanced cepstral estimate, the so-called LPC cepstrum is used. The cepstrum is estimated by first determining a 40th order maximum entropy AR filter (also called LPC an filter) and using the cepstra that this filter generates. The results are depicted in Figures 6.3 and 6.4, and they are of a similar quality to the results obtained in [24].

The results depicted in Figure 6.4 using the enhanced covariance estimates are of considerably better quality than the results depicted in Figure 6.2. In Table 6.1 it

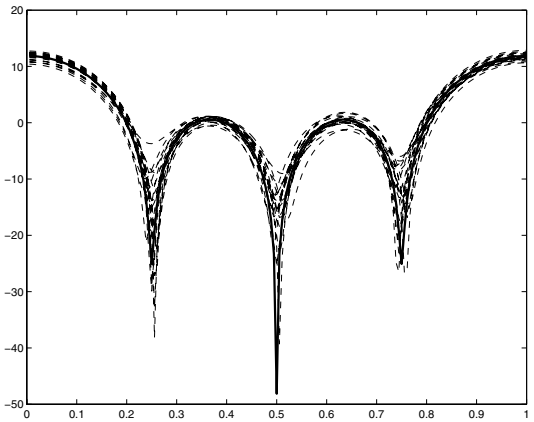


FIG. 6.1. *Comparison of estimation of MA*(6) *model using problem* $(\mathcal{P}_\lambda)$ *(dashed) relative to the true system (solid),* $\lambda = 5e - 4$.

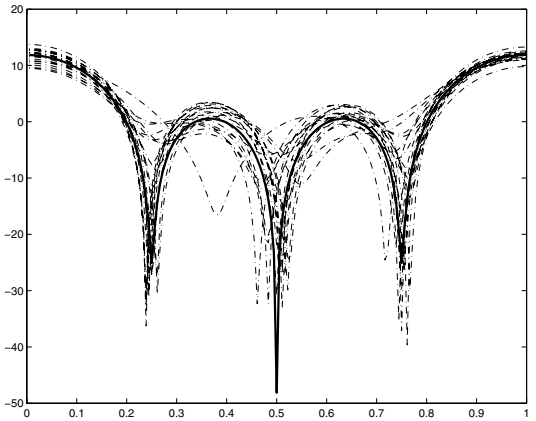


FIG. 6.2. *Comparison of estimation of MA*(6) *model using problem* $(\mathcal{M}_\lambda)$ *(dashed) relative to the true system (solid),* $\lambda = 5e-4$.

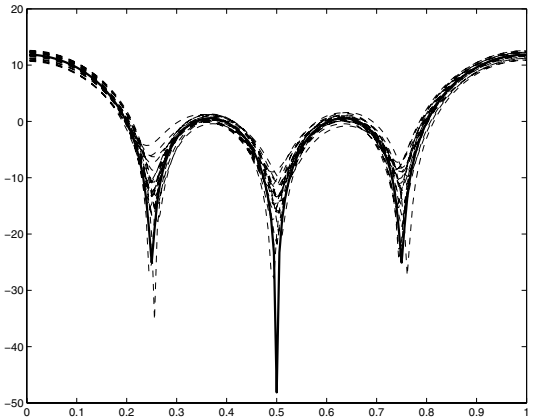


FIG. 6.3. *Comparison of MA*(6) *filters estimated using problem* $(\mathcal{P}_\lambda)$ *with the LPC cepstrum (dashed) relative to the true system (solid),* $\lambda = 5e-4$.

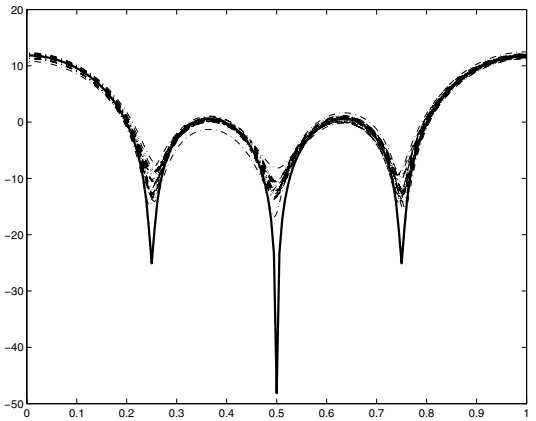


FIG. 6.4. *Comparison of MA*(6) *filters estimated using problem* $(\mathcal{M}_\lambda)$ *with enhanced covariance estimates (dashed) relative to the true system (solid),* $\lambda = 5e-4$.

TABLE 6.1
True covariance values and mean absolute values of the estimation errors and correction terms.

Covariance	$\|r_0\|$	$\|r_1\|$	$\|r_2\|$	$\|r_3\|$	$\|r_4\|$	$\|r_5\|$	$\|r_6\|$
True value	3.838	0	2.878	0	1.917	0	0.959
Est. error	0.366	0.367	0.332	0.345	0.292	0.337	0.266
Enh. est. error	0.205	0.111	0.166	0.076	0.102	0.040	0.059
Corr. terms	$\|\lambda\epsilon_0\|$	$\|\lambda\epsilon_1\|$	$\|\lambda\epsilon_2\|$	$\|\lambda\epsilon_3\|$	$\|\lambda\epsilon_4\|$	$\|\lambda\epsilon_5\|$	$\|\lambda\epsilon_6\|$
Ergodic est.	0.157	0.063	0.038	0.057	0.075	0.076	0.049
Enhanced est.	0.018	0.001	0.008	0.001	0.003	0.001	0.004

TABLE 6.2
True cepstrum values and mean absolute values of the estimation errors and correction terms.

Cepstrum	$\|c_1\|$	$\|c_2\|$	$\|c_3\|$	$\|c_4\|$	$\|c_5\|$	$\|c_6\|$
True value	0	0.998	0	0.463	0	0.331
Est. error	0.040	0.074	0.058	0.075	0.060	0.086
LPC est. error	0.047	0.088	0.036	0.053	0.039	0.089
Corr. terms	$\|\lambda\epsilon_1\|$	$\|\lambda\epsilon_2\|$	$\|\lambda\epsilon_3\|$	$\|\lambda\epsilon_4\|$	$\|\lambda\epsilon_5\|$	$\|\lambda\epsilon_6\|$
Ergodic est.	0.012	0.007	0.013	0.018	0.011	0.006
LPC cepstrum est.	0.002	0.004	0.002	0.004	0.002	0.003

can be seen that indeed the error of the estimated parameters $r_0, r_1, \ldots, r_n$ is much smaller for the enhanced estimate, and the correction terms are also smaller when using this estimate and quite negligible compared to the estimation errors.

However, the results depicted in Figure 6.3 with the cepstra obtained using the LPC cepstrum are of a similar quality to that in Figure 6.1. From this it appears that the LPC cepstrum fails to provide an enhanced estimate of the cepstrum, which is confirmed by Table 6.2, which is natural since it is not the same cepstrum. The AR approximation seems to generate a certain amount of regularization to the estimates, which can be seen from a bit less deep valleys in the spectrum of Figure 6.3 compared to Figure 6.1, and from smaller correction terms in Table 6.2.

6.2. ARMA filter design.

Experiment 2. In order to test the regularized CCM method, Experiment 1 is amended to provide an ARMA example. As before, 20 sets of data samples of length 400 are generated, this time using an ARMA filter. The zeros of the ARMA filter were located at $0.99e^{\pm i(\pi/4)}$, $0.99e^{\pm i(3\pi/4)}$, and $0.999e^{\pm i(\pi/2)}$ and the poles at $0.99e^{\pm i(\pi/6)}$, $0.99e^{\pm i(5\pi/6)}$, and $0.999e^{\pm i(2\pi/5)}$. As depicted in Figure 6.5, the peaks in the spectrum have a nice fit, but the notches are not reproduced as well. This is probably due to bad estimates of the cepstral lags. Using a higher order AR model for generating cepstrum estimates further degraded the results in this sense. If the experiment is repeated using the true cepstrum instead of the estimated cepstrum, the result in Figure 6.6 is obtained, and the very nice fit gives a strong indication that the weak point of the method is the estimation of the cepstrum. Hence the estimation of the cepstral lags needs to be improved.

Experiment 3. The scalar example in [20] is used here for comparison. Data of length 500 is generated by passing white noise through the filter $w(z) = \sigma(z)/a(z)$, where

$$\sigma(z) = z^5 - 1.051z^4 + 0.0718z^3 + 0.05164z^2 + 0.5322z - 0.5735$$

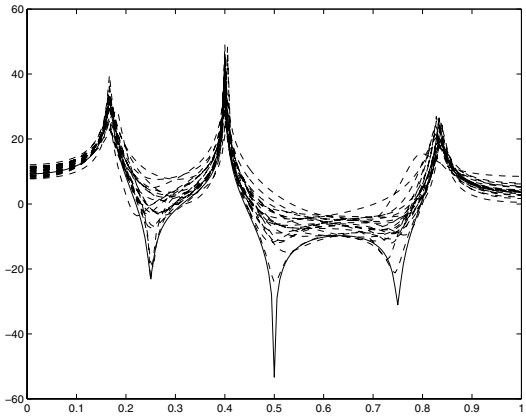


FIG. 6.5. *Comparison of ARMA(6,6) filters estimated using the regularized CCM method (dashed) relative to the true system (solid),* $\lambda = 5e-4$.

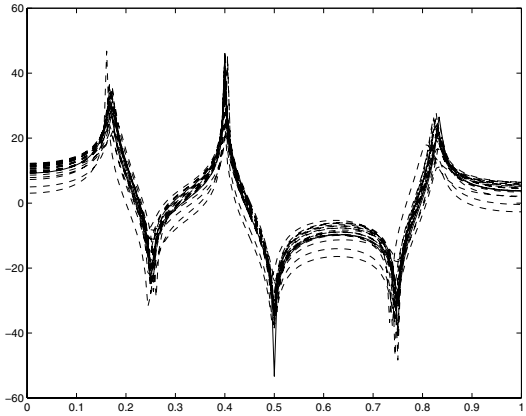


FIG. 6.6. *Comparison of ARMA(6,6) filters estimated using the regularized CCM method with true cepstral lags (dashed) relative to the true system (solid),* $\lambda = 5e{-}4$.

and

$$a(z) = z^5 - 0.8713z^4 - 1.539z^3 + 1.371z^2 + 0.6451z - 0.5827.$$

This was repeated 100 times, and the regularized CCM method was used to design an ARMA(5,5) filter for each output sequence. As in [20] the average power spectral density (in decibels) is plotted as a dashed line in Figure 6.7. The true spectral density is plotted as a solid line, and the dashed-dotted line represents the average estimated spectral density plus and minus the statistical standard deviations. The quality of the estimates is similar to the results obtained in [20].

Experiment 4: Speech. In order to show that the regularized CCM method provides good estimates for practical processes, a speech signal is considered. A sample of the phoneme "e" of length 200, corresponding to 25 milliseconds for a sampling rate of 8 kilohertz, is used for the analysis. Since there is no generic value for the orders m

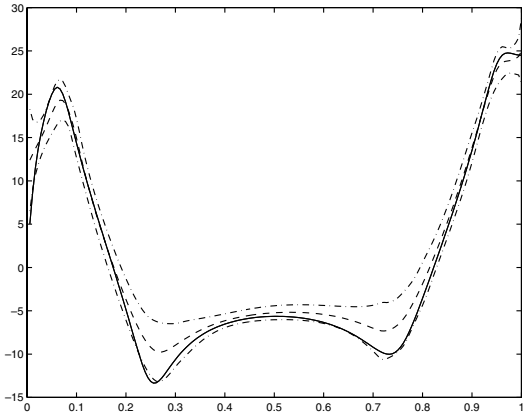


FIG. 6.7. *Comparison of ARMA*(5,5) *filters estimated using the regularized CCM method (dashed,* $\lambda = 5e{-}4$*) with standard deviations (dashed-dotted), relative to the true system (solid).*

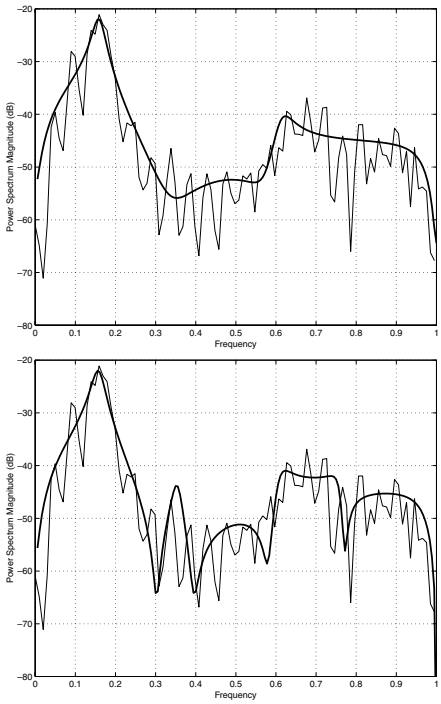


FIG. 6.8. *Comparison of ARMA*(6,6) *and ARMA*(10,10) *filters estimated using the regularized CCM method* ($\lambda = 5e-4$) *and the corresponding periodogram for the phoneme "e."*

and n, the regularized CCM method is applied for $m = n = 6$ and 10, and the results are depicted in Figure 6.8 plotted against the periodogram of the sample for reference. It is clear that the ARMA(6,6) filter describes the envelope of the periodogram well, whereas perhaps the ARMA(10,10) filter is overmodeling in that it models the fine structure of the signal.

Appendix.

LEMMA A.1. *Let* $P \in \hat{\mathcal{D}}_m^+$. *Then*

$$\left\langle \frac{1}{P} - 1, 1 \right\rangle \geq 0.$$

Proof. Consider the optimization problem

$$\min \left\{ \langle 1/P, 1\rangle \mid P \in \hat{\mathcal{D}}_m^+ \right\}.$$

Since the function $1/x$ is convex for positive x, any stationary point is minimal. Differentiating with respect to p_k, we get

$$\frac{\partial}{\partial p_k}\langle 1/P, 1\rangle = -\langle 1/P^2, z^k\rangle, \quad k = 1, \dots, m,$$

and it is clear that the gradient is zero if $P = 1$. The inequality

$$\langle 1/P, 1\rangle \geq \langle 1/1, 1\rangle$$

now follows from the minimality of $P = 1$, and this proves the lemma. □

Proof of Lemma 4.4. Consider the optimization problem

$$\max \left\{ \langle R, \log P\rangle \mid P \in \hat{\mathcal{D}}_m^+ \right\}.$$

Since the logarithm is concave for positive arguments, any stationary point is maximal. Differentiating with respect to p_k, we get

$$\frac{\partial}{\partial p_k}\langle R, \log P\rangle = \left\langle R, \frac{\cos k\theta}{P} \right\rangle, \quad k = 1, \dots, m,$$

and it is clear that the gradient is zero for $P = R$. The inequality

$$\langle R, \log R\rangle \geq \langle R, \log P\rangle$$

now follows from the optimality of $P = R$, and this proves the lemma. □

Acknowledgments. I would like to thank Dr. J. Mari for reading an early version of this paper and the referees for their reviews. Most of all I thank my advisor, Prof. A. Lindquist, for his many invaluable suggestions on how to improve this paper.

REFERENCES

[1] G. BOX AND G. JENKINS, *Time Series Analysis Forecasting and Control*, Holden–Day, San Francisco, 1970.

[2] J. BURG, *Maximum Entropy Spectral Analysis*, Ph.D. thesis, Stanford University, Stanford, CA, 1975.

[3] C. BYRNES, P. ENQVIST, AND A. LINDQUIST, *Cepstral coefficients, covariance lags and pole-zero models for finite data strings*, IEEE Trans. Signal Process., SP-50 (2001), pp. 677–693.

[4] C. I. BYRNES, P. ENQVIST, AND A. LINDQUIST, *Identifiability and well-posedness of shaping-filter parameterizations: A global analysis approach*, SIAM J. Control Optim., 41 (2002), pp. 23–59.

[5] C. Byrnes, T. Georgiou, and A. Lindquist, *A new approach to spectral estimation: A tunable high-resolution spectral estimator*, IEEE Trans. Signal Process., 48 (2000), pp. 3189–3205.
[6] C. I. Byrnes, S. V. Gusev, and A. Lindquist, *A convex optimization approach to the rational covariance extension problem*, SIAM J. Control Optim., 37 (1998), pp. 211–229.
[7] C. I. Byrnes, A. Lindquist, S. Gusev, and A. Matveev, *A complete parameterization of all positive rational extensions of a covariance sequence*, IEEE Trans. Automat. Control, 40 (1995), pp. 1841–1857.
[8] C. I. Byrnes, T. T. Georgiou, and A. Lindquist, *A generalized entropy criterion for Nevanlinna–Pick interpolation with degree constraint*, IEEE Trans. Automat. Control, AC-46 (2001), pp. 822–839.
[9] P. Caines, *Linear Stochastic Systems*, John Wiley, New York, 1987.
[10] A. Dahlén, A. Lindquist, and J. Mari, *Experimental evidence showing that stochastic subspace methods may fail*, Systems Control Lett., 34 (1998), pp. 303–312.
[11] P. Enqvist, *A homotopy approach to rational covariance extension with degree constraint*, Internat. J. Appl. Math. Comp. Sci., 11 (2001), pp. 1173–1201.
[12] P. Enqvist, *Spectral Estimation by Geometric, Topological and Optimization Methods*, Ph.D. thesis, Royal Institute of Technology, Stockholm, Sweden, 2001.
[13] Y. Ephraim and M. Rahim, *On second-order statistics and linear estimation of cepstral coefficients*, IEEE Trans. Speech Audio Process., 7 (1999), pp. 162–176.
[14] B. Friedlander and B. Porat, *A spectral matching technique for ARMA parameter estimation*, IEEE Trans. Acoust. Speech Signal Process., ASSP-32 (1984), pp. 338–343.
[15] T. Georgiou, *Realization of power spectra from partial covariance sequences*, IEEE Trans. Acoust. Speech Signal Process., ASSP-35 (1987), pp. 438–449.
[16] L. Jackson, *Digital Filters and Signal Processing: With MATLAB Exercises*, 3rd ed., Kluwer Academic Publishers, Norwell, MA, 1996.
[17] G. Jenkins and D. Watts, *Spectral Analysis and Its Applications*, Holden–Day, San Francisco, 1968.
[18] L. Ljung, *System Identification, Theory for the User*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1999.
[19] D. Luenberger, *Linear and Nonlinear Programming*, Addison–Wesley, Reading, MA, 1984.
[20] J. Mari, P. Stoica, and T. McKelvey, *Vector ARMA estimation: A reliable subspace approach*, IEEE Trans. Signal Process., 48 (2000), pp. 2092–2104.
[21] J. Moura and M. Ribeiro, *Parametric spectral estimation for ARMA processes*, in Proceedings of the Third Workshop on Spectrum Estimation and Modeling, Boston, 1986, pp. 37–40.
[22] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice–Hall, Englewood Cliffs, NJ, 1978.
[23] T. Söderström and P. Stoica, *System Identification*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[24] P. Stoica, T. McKelvey, and J. Mari, *MA estimation in polynomial time*, IEEE Trans. Signal Process., 48 (2000), pp. 1999–2012.
[25] P. van Overschee and B. de Moor, *Subspace algorithms for stochastic identification problem*, IEEE Trans. Automat. Control, 27 (1992), pp. 382–387.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 1037–1062

THE APPROXIMATE MAXIMUM PRINCIPLE IN CONSTRAINED OPTIMAL CONTROL*

BORIS S. MORDUKHOVICH† AND ILYA SHVARTSMAN‡

Abstract. This paper concerns optimal control problems for dynamical systems described by a parametric family of discrete/finite-difference approximations of continuous-time control systems. Control theory for parametric systems governed by discrete approximations plays an important role in both qualitative and numerical aspects of optimal control and occupies an intermediate position in dynamic optimization: between optimal control of discrete-time (with fixed steps) and continuous-time control systems. The central result in optimal control of discrete approximation systems is the approximate maximum principle (AMP), which gives the necessary optimality condition in a perturbed maximum principle form with no a priori convexity assumptions and thus ensures the stability of the Pontryagin maximum principle (PMP) under discrete approximation procedures. The AMP has been justified for optimal control problems of smooth dynamical systems with endpoint constraints under some properness assumption imposed on the sequence of optimal controls. In this paper we show, by a series of counterexamples, that the properness assumption is essential for the validity of the AMP, and that the AMP does not hold, in its expected (lower) subdifferential form, for nonsmooth problems. Moreover, a new upper subdifferential form of the AMP is established for ordinary and time-delay control systems. The results obtained surprisingly solve (in both negative and positive directions) a long-standing and well-recognized question about the possibility of extending the AMP to nonsmooth control problems, for which the affirmative answer has been expected in the conventional lower subdifferential form.

Key words. optimal control, discrete approximations, approximate maximum principle, stability under perturbations, nonsmooth and variational analysis, lower and upper subgradients, time delays

AMS subject classifications. 49K15, 93C55, 49M25, 49J52, 49J53

DOI. 10.1137/S0363012903433012

1. Introduction and preliminaries.

This paper is devoted to the study of optimal control problems for *discrete approximations* of continuous-time control systems that, viewed as a *parametric process* with a decreasing discretization step, occupy an *intermediate* position between control problems involving discrete-time and continuous-time systems. As the basic model for our study, we consider discrete approximations of the following Mayer-type optimal control problem governed by ordinary differential equations with endpoint constraints:

$$(P)\qquad \begin{cases} \text{minimize } J(x,u) := \varphi_0(x(t_1)) \\ \text{subject to} \\ \dot{x}(t) = f(t,x(t),u(t)) \ \text{ a.e. } \ t \in [t_0,t_1], \quad x(t_0) = x_0 \in \mathbb{R}^n, \\ u(t) \in U \ \text{ a.e. } \ t \in [t_0,t_1], \\ \varphi_i(x(t_1)) \le 0, \quad i = 1,\dots,m, \\ \varphi_i(x(t_1)) = 0, \quad i = m+1,\dots,m+r, \end{cases}$$

*Received by the editors August 8, 2003; accepted for publication (in revised form) February 2, 2004; published electronically November 9, 2004. This research was partly supported by the National Science Foundation under grants DMS-0072179 and DMS-0304989.
http://www.siam.org/journals/sicon/43-3/43301.html
†Department of Mathematics, Wayne State University, Detroit, MI 48202 (boris@math.wayne.edu).
‡Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, Exhibition Road, London SW7 2AZ, UK (ilyashv@imperial.ac.uk).

over measurable controls $u(\cdot)$ and absolutely continuous trajectories $x(\cdot)$ on the fixed time interval $T := [t_0, t_1]$. It is well known that many other control problems (of Lagrange and Bolza types, with integral constraints, on variable time intervals, etc.) reduce to the form of (P). Observe also that the results of this paper can be easily extended to control problems with a nonfixed initial vector (i.e., when φ_i in (P) depend on both endpoints $x(t_0)$ and $x(t_1)$ for all $i = 0, \dots, m + r$) as well as to problems with continuously time-dependent control sets $U = U(t)$.

Problem (P) may be treated as an infinite-dimensional optimization problem with special equality-type dynamic constraints governed by differential operators as well as with geometric constraints given by arbitrary control sets; this makes it *nonsmooth* even under all smooth functional data f and φ_i. On the other hand, it is natural to explore a different way to study continuous-time problems (P), which goes back to Leibnitz and Euler and which consists of approximating (P) by a family of discrete-time systems arising when the time-derivative $\dot{x}(t)$ is replaced with the *finite differences*

$$\dot{x}(t) \approx \frac{x(t+h) - x(t)}{h} \quad \text{as} \quad h \to 0.$$

Allowing also *perturbations* of the *endpoint constraints* (which is very essential for variational stability), problem (P) is replaced in this way by the following family of discrete-time problems (P_N) depending on the natural parameter $N = 1, 2, \dots$:

$$(P_N) \qquad \begin{cases} \text{minimize } J(x_N, u_N) := \varphi_0(x_N(t_1)) \\ \text{subject to} \\ x_N(t + h_N) = x_N(t) + h_N f(t, x_N(t), u_N(t)), \quad x_N(t_0) = x_0 \in \mathbb{R}^n, \\ u_N(t) \in U, \quad t \in T_N := \{t_0, t_0 + h_N, \dots, t_1 - h_N\}, \\ \varphi_i(x_N(t_1)) \le \gamma_{iN}, \quad i = 1, \dots, m, \\ |\varphi_i(x_N(t_1))| \le \delta_{iN}, \quad i = m+1, \dots, m+r, \\ h_N := \dfrac{t_1 - t_0}{N}, \quad N \in \mathbb{N} := \{1, 2, \dots\}, \end{cases}$$

where $\gamma_{iN} \to 0$ and $\delta_{iN} \downarrow 0$ as $N \to \infty$ for all i. For each fixed $N \in \mathbb{N}$ problem (P_N) is *finite-dimensional* and seems to be simpler than the continuous-time problem (P). Indeed, applying well-developed methods of finite-dimensional variational analysis, it is possible to derive necessary optimality conditions in problems (P_N) even with nonsmooth data and general dynamic constraints governed by discrete inclusions and then obtain the corresponding results for optimal control of differential inclusions by passing to the limit from those for discrete approximations; see [4, 6, 13] for detailed proofs and discussions. However, this approach has some limitation regarding necessary optimality conditions of the *maximum principle* type.

As is well known, the central result of the optimal control theory for continuous-time problems (P), the Pontryagin maximum principle (PMP) [11], holds with *no convexity* assumptions on the admissible velocity sets $f(t, x, U)$. This specific result, from the general viewpoint of optimization theory, is strongly due to continuous-time dynamic constraints in (P) governed by differential operators. It happens that continuous-type control systems enjoy a certain *hidden convexity*, which is deeply related to the classical Lyapunov theorem on the range convexity of nonatomic vector measures and eventually leads to the maximum principle form. It is not surprising, therefore, that an analogue of the maximum principle for optimal control problems

governed by discrete-time systems *does not generally hold* without a priori convexity assumptions. This may create troubles for applications of the PMP in numerical calculations of nonconvex continuous-time control systems, which inevitably involve finite-difference approximations via time discretization. To avoid such troubles, it is sufficient to justify not a full analogue of the PMP, with the exact maximum condition, but its *approximate* counterpart, where an error in the maximum condition depends on the discretization stepsize *tending to zero* when the latter is decreasing.

The first result of this type in the absence of convexity assumptions was given by Gabasov and Kirillova [2, 3], under the name of "quasi-maximum principle," for parametric discrete systems with smooth cost and dynamics and with no endpoint constraints. The proof of this result, purely analytic, essentially exploited the unconstrained nature of the problem.

The following *approximate maximum principle* (AMP) for the nonconvex constrained problems (P_N) was established by Mordukhovich [4, 5]. The proof in [4, 5] is geometric based on the discovered finite-difference counterpart of the hidden convexity property and the separation theorem. Denote

$$H(t,x,p,u) := \langle p, f(t,x,u)\rangle, \quad p \in \mathbb{R}^n, \tag{1.1}$$

the *Hamilton–Pontryagin function* for the dynamic constraints under consideration.

APPROXIMATE MAXIMUM PRINCIPLE. *Let the pairs $(\bar{x}_N, \bar{u}_N)$ be optimal to (P_N) for all $N \in \mathbb{N}$, where U is a compact subset of a metric space with the metric $d(\cdot,\cdot)$, where f is continuous with respect to its variables and continuously differentiable with respect to x in a tube containing the optimal trajectories $\bar{x}_N(t)$ for large N, and where each φ_i is continuously differentiable around the limiting point(s) of $\{\bar{x}_N(t_1)\}$. Impose the following assumptions.*

Assumption A. The *consistency condition* on the perturbation of the equality constraints, meaning that

$$\lim_{N\to\infty} \frac{h_N}{\delta_{iN}} = 0 \quad \text{for all} \quad i = m+1, \ldots, m+r. \tag{1.2}$$

Assumption B. The *properness* of the sequences of optimal controls $\{\bar{u}_N\}$, which means that for every increasing subsequence $\{N\}$ of natural numbers and every sequence of mesh points $\tau_{\theta(N)} \in T_N$ satisfying $\tau_{\theta(N)} = t_0 + \theta(N)h_N$, $\theta(N) = 0, 1, \ldots, N-1$, and $\tau_{\theta(N)} \to t \in [t_0, t_1]$ one has either

$$d\big(u_N(\tau_{\theta(N)}), u_N(\tau_{\theta(N)+q})\big) \to 0$$

or

$$d\big(u_N(\tau_{\theta(N)}), u_N(\tau_{\theta(N)-q})\big) \to 0$$

as $N \to \infty$ with any natural constant q.

Then there are numbers $\{\lambda_{iN} |\ i = 0, \ldots, m+r\}$ and a function $\varepsilon(t, h_N) \to 0$ as $N \to \infty$ uniformly in $t \in T_N$ such that

$$H(t, \bar{x}_N(t), p_N(t+h_N), \bar{u}_N(t)) = \max_{u\in U} H(t, \bar{x}_N(t), p_N(t+h_N), u) + \varepsilon(t, h_N) \tag{1.3}$$

for all $t \in T_N$ and

$$\lambda_{iN}(\varphi_i(\bar{x}_N(t_1)) - \gamma_{iN}) = O(h_N), \quad i = 1, \ldots, m, \tag{1.4}$$

$$\lambda_{iN} \geq 0, \quad i = 0, \ldots, m, \quad \text{and} \quad \sum_{i=0}^{m+r} \lambda_{iN}^2 = 1 \tag{1.5}$$

for all $N \in \mathbb{N}$, where $p_N(t)$, $t \in T_N \cup \{t_1\}$, is the corresponding trajectory of the adjoint system

$$p_N(t) = p_N(t + h_N) + h_N \frac{\partial H}{\partial x}(t, \bar{x}_N(t), p_N(t + h_N), \bar{u}_N(t)), \quad t \in T_N, \tag{1.6}$$

with the transversality condition

$$p_N(t_1) = -\sum_{i=0}^{m+r} \lambda_{iN} \nabla \varphi_i(\bar{x}_N(t_1)). \tag{1.7}$$

Observe that the closer h_N is to zero, the more precise the approximate maximum condition (1.3) and the approximate complementary slackness condition (1.4) are. This means that the AMP in (P_N) tends to the PMP in (P) as $N \to \infty$, which actually justifies the *stability* of the PMP with respect to discrete approximations under the assumptions made.

It has been shown in [4, 5] that the *consistency* condition in (a) is *essential* for the validity of the AMP in problems with equality constraints. The first goal of this paper is to examine the other two significant assumptions made in the above theorem: the *properness* condition in Assumption B and the *smoothness* of the initial data. We show in section 2 that *both of these assumptions are essential for the validity of the AMP*.

Note that the properness of the sequence of optimal controls in Assumption B is a *finite-difference counterpart* of the piecewise continuity (or, more generally, of *Lebesgue regular points* having full measure) for optimal controls in continuous-time systems. It turns out that the situation when sequences of optimal controls are not proper in discrete approximations is not unusual for systems with nonconvex velocities, and it leads to the violation of the AMP already in the case of smooth problems with inequality constraints.

The impact of nonsmoothness to the validity of the AMP happens to be even more striking: The AMP *does not hold* in the expected conventional subdifferential form already for minimizing *convex* cost functions in discrete approximations of linear systems with no endpoint constraints, as well as for problems with nonsmooth dynamics. It seems that the *AMP is one of very few results on necessary optimality conditions that do not have expected counterparts in nonsmooth settings*.

On the other hand, we derive the AMP in problems (P_N) with nonsmooth functions describing the objective and inequality constraints in a new *upper subdifferential* (or superdifferential) form, which is also new for necessary optimality conditions in continuous-time control systems. The main difference between the conventional subdifferential form, which does not hold for the AMP but holds for the PMP, and the new one is that the latter involves upper (not lower) subgradients of nonsmooth functions in transversality conditions. This form applies to a class of *uniformly upper subdifferentiable* functions described in this paper, which particularly contains smooth and concave continuous functions being closed with respect to taking the minimum over compact sets. The results obtained solve (in a surprising, unexpected way) a long-standing and widely discussed question about the possibility of establishing the AMP for nonsmooth control problems. We also derive the upper subdifferential form

of the AMP in discrete approximations of control systems with *time delays*, for which no results of this type have been known before. The main results of this paper have been announced in [9].

The rest of the paper is organized as follows. Section 2 contains examples on the *violation of the AMP* in smooth problems (P_N) without the properness condition as well as in problems with nonsmooth cost functions, nonsmooth dynamics, or both. In section 3 we discuss appropriate tools of nonsmooth analysis, paying attention mainly to the concepts of *upper regularity* and uniform *upper subdifferentiability*, which are new in the study of *minimization* problems. Section 4 is devoted to the derivation of the AMP for the discrete approximation problems (P_N) in the *upper subdifferential form*; it contains three slightly different modifications of this result in somewhat distinct settings. In section 5 we extend the AMP to discrete approximations of constrained *time-delay* systems, where the results obtained are new in both smooth and nonsmooth frameworks. We also present an example on the *violation* of the AMP in discrete approximations of functional-differential control systems of *neutral type*, even under smoothness assumptions in the absence of endpoint constraints.

Throughout the paper we use standard notation with some special symbols defined in the text where they are introduced.

2. Counterexamples. Let us start with an example on the violation of the AMP in discrete approximations of linear control systems with linear cost functions and linear endpoint inequality constraints but with *no properness condition*.

EXAMPLE 2.1 (AMP may not hold in smooth control problems with no properness condition). *There is a two-dimensional linear control problem with an inequality constraint such that optimal controls in the sequence of its discrete approximations are not proper and do not satisfy the AMP.*

Proof. Let us consider a linear continuous-time optimal control problem (P) with a two-dimensional state $x = (x_1, x_2) \in \mathbb{R}^2$ in the following form:

$$\begin{cases} \text{minimize } \varphi(x(1)) := -x_1(1) \\ \text{subject to} \\ \dot{x}_1 = u, \quad \dot{x}_2 = x_1 - at, \quad x_1(0) = x_2(0) = 0, \\ u(t) \in U := \{0,1\}, \quad 0 \le t \le 1, \\ x_2(1) \le -\dfrac{a-1}{2}, \end{cases} \tag{2.1}$$

where $a > 1$ is a given constant. Observe that the only "unpleasant" feature of this problem is that the control set $U = \{0,1\}$ is *nonconvex*, and hence the feasible velocity sets $f(t,x,U)$ are nonconvex as well. It is clear that $\bar{u}(t) \equiv 1$ is the unique optimal solution to problem (2.1) and that the corresponding optimal trajectory is $\bar{x}_1(t) = t$, $\bar{x}_2(t) = -\frac{a-1}{2}t^2$. Moreover, the inequality constraint is active, since $\bar{x}_2(1) = -\frac{a-1}{2}$.

Let us now discretize this problem with the stepsize $h_N := \frac{1}{2N}$, $N \in \mathbb{N}$. For notational convenience we omit the index N in what follows. Thus the discrete approximation problems (P_N) corresponding to (2.1) are written as

$$\begin{cases} \text{minimize } \varphi(x(1)) = -x_1(1) \\ \text{subject to} \\ x_1(t+h) = x_1(t) + hu(t), \quad x_1(0) = 0, \\ x_2(t+h) = x_2(t) + h\big(x_1(t) - at\big), \quad x_2(0) = 0, \\ u(t) \in \big\{0,1\big\}, \quad t \in \big\{0, h, \dots, 1-h\big\}, \\ x_2(1) \le -\dfrac{a-1}{2} + h^2; \end{cases} \tag{2.2}$$

i.e., we put $\gamma_N := h_N^2$ in the constraint perturbation for (P_N).

To proceed, we compute the trajectories of (2.2) corresponding to $u(t) \equiv 1$. It is easy to see that $x_1(t) = t$ for this u. To compute $x_2(t)$, observe that

$$\big[y(t+h) = y(t) + ht,\ y(0) = 0\big] \Longrightarrow y(t) = \frac{t^2}{2} - \frac{th}{2}.$$

Indeed, one has by the direct calculation that

$$y(t) = h\sum_{\tau=0}^{t-h} \tau = \big[\text{put } \tau = kh\big] = h^2 \sum_{k=0}^{\frac{t}{h}-1} k = h^2 \frac{\frac{t}{h}\big(\frac{t}{h}-1\big)}{2} = \frac{t^2}{2} - \frac{th}{2}.$$

Therefore, for $x_2(t)$ corresponding to $u(t) \equiv 1$ in (2.2) we have

$$x_2(t) = h\sum_{\tau=0}^{t-h} \big(\tau - a\tau\big) = -\frac{a-1}{2}t^2 + \frac{a-1}{2}ht.$$

By this calculation we see that, for h sufficiently small, $x_2(t_1)$ no longer satisfies the endpoint constraint, and thus $u(t) \equiv 1$ is not a feasible control to problem (2.2) for all h close to zero. This implies that an optimal control to (2.2) for small h, which obviously exists, must have at least one *switching point* s such that $u(s) = 0$, and hence the maximum value of the corresponding endpoint $x_1(1)$ will be less than or equal to $1 - h$. Put

$$u(t) := \begin{cases} 1, & t \neq s, \\ 0, & t = s, \end{cases}$$

and show that

$$x_2(t) = \begin{cases} -\dfrac{a-1}{2}t^2 + \dfrac{a-1}{2}ht, & t \leq s, \\ -\dfrac{a-1}{2}t^2 + \dfrac{a-1}{2}ht - h(t-s) + h^2, & t \geq s+h, \end{cases} \tag{2.3}$$

for the corresponding trajectories in (2.2) depending on h and s. We need only justify the second part of this formula. To compute $x_2(t)$ for $t \geq s+h$, substitute $x_1(t) = t-h$ into (2.2). It is easy to see that the increment $\Delta x_2(t)$ compared to the case when $u(t) \equiv 1$ is

$$h\sum_{\tau=s+h}^{t-h} (-h) = -h(t-h-s) = -h(t-s) + h^2,$$

which justifies (2.3). Now let us specify the parameters of the above control putting $a = 2$ and $s = 0.5$ for all N, i.e., considering the discrete-time function

$$\bar{u}(t) := \begin{cases} 1, & t \neq 0.5, \\ 0, & t = 0.5. \end{cases}$$

Note that the point $t = 0.5$ belongs to the grid T_N for all N due to $h_N := \frac{1}{2N}$. Observe that the sequence of these controls *does not satisfy the properness property* in Assumption B of the AMP formulated in section 1. It follows from (2.3) that the

corresponding trajectories satisfy the endpoint constraint in (2.2) for all $N \in \mathbb{N}$, since $\bar{x}_2(1) = -\frac{1}{2}t^2 + h^2$. Moreover, it is clear from the above calculations that the control $\bar{u}(t)$ is optimal to problem (P_N) in (2.2) for any N. Let us show that the sequence of optimal controls $\bar{u}(t)$ does not satisfy the approximate maximum condition (1.3) at the point of switch.

The adjoint system (1.6) for problem (2.2) with any h is

$$p(t) = p(t+h) + h\frac{\partial f}{\partial x}^*(t, \bar{x}_1, \bar{x}_2, \bar{u})p(t+h),$$

where the Jacobian matrix $\partial f/\partial x$ and its adjoint/transposed one equal

$$\frac{\partial f}{\partial x} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad \frac{\partial f}{\partial x}^* = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Thus we have the adjoint trajectories

$$p_1(t) = p_1(t+h) + hp_2(t+h) \quad \text{and} \quad p_2(t) \equiv \text{const},$$

where (p_1, p_2) satisfy the transversality condition (1.7) with the corresponding sign/nontriviality conditions (1.5) written as

$$p_1(1) = \lambda_0,\ p_2(1) = -\lambda_1; \quad \lambda_0 \geq 0,\ \lambda_1 \geq 0,\ \lambda_0^2 + \lambda_1^2 = 1.$$

This implies that $p_1(t)$ is a linear nondecreasing function. The corresponding Hamilton–Pontryagin function (1.1) equals

$$H(t, x(t), p(t+h), u(t)) = p_1(t+h)u(t) + \text{terms not depending on } u.$$

Examining the latter expression and taking into account that the optimal controls are equal to one for all t but $t = 0.5$, we conclude that the approximate maximum condition (1.3) holds only if $p_1(t)$ is either nonnegative or tends to zero everywhere except $t = 0.5$. Observe that $p_1(t) \equiv 0$ yields $\lambda_1 = \lambda_2 = 0$, which contradicts the nontriviality condition. Hence $p_1(t)$ must be positive away from $t = 0$. Therefore, a sequence of controls having a point of switch not tending to zero as $h \downarrow 0$ *cannot* satisfy the approximate maximum condition at this point. This shows that the AMP does not hold for the sequence of optimal controls to (2.2) built above. □

Many examples of this type can be constructed based on the above idea, which essentially means the following. Take a continuous-time problem with active inequality constraints and *nonconvex* admissible velocity sets $f(t, x, U)$. It often happens that after the discretization the "former" optimal control becomes not feasible in discrete approximations, and the "new" optimal control in the sequence of discrete-time problems has a singular point of switch (thus making the sequence of optimal controls not proper), where the approximate maximum condition does not hold.

The next example demonstrates that the AMP may be violated in problems of minimizing *nonsmooth cost* functions in linear systems with no endpoint constraint.

EXAMPLE 2.2 (AMP may not hold for linear systems with nonsmooth and convex cost functions). *There is a one-dimensional control problem of minimizing a nonsmooth and convex cost function over a linear system with no endpoint constraints such that a proper sequence of optimal controls to discrete approximations does not satisfy the AMP.*

Proof. Consider the following sequence of one-dimensional optimal control problem (P_N), $N \in \mathbb{N}$, for discrete-time systems:

$$\text{(2.4)} \qquad \begin{cases} \text{minimize } \ \varphi(x_N(1)) := |x_N(1) - r| \\ \text{subject to} \\ x_N(t+h_N) = x_N(t) + h_N u_N(t), \quad x_N(0) = 0, \\ u_N(t) \in U := \{0,1\}, \quad t \in T_N := \{0, h_N, \ldots, 1 - h_N\}, \end{cases}$$

where r is a positive *irrational* number less than 1 whose choice will be specified below. The dynamics in (2.4) is a discretization of the simplest ODE control system $\dot{x} = u$. Observe that, since r is irrational and h_N is rational, we have $\bar{x}_N(1) \neq r$ for the endpoint of an optimal trajectory to (2.4) as $N \in \mathbb{N}$, while obviously $\bar{x}(1) = r$ for optimal solutions to the continuous-time counterpart. It is also clear that for sufficiently small h_N an optimal control to (2.4) will be neither $u_N(t) \equiv 0$ nor $u_N(t) \equiv 1$, but it will have at least one point of switch.

Suppose that for some subsequence $N_k \to \infty$ one has $\bar{x}_{N_k}(1) > r$; put $\{N_k\} = \mathbb{N}$ without loss of generality. Let us show that in this case the approximate maximum condition does *not* hold at points $t \in T_N$ for which $\bar{u}_N(t) = 1$. Indeed, we have

$$H(\bar{x}_N(t), p_N(t+h_N), u) = p_N(t+h_N)u \quad \text{and} \quad p_N(t) \equiv -1$$

for the Hamilton–Pontryagin function and the adjoint trajectory in (1.6) and (1.7), since $\bar{x}_N(1) > r$ along the optimal solution to (2.4). Thus

$$\begin{aligned} &\max_{u \in U} H(\bar{x}_N(t), p_N(t+h_N), u) = 0 \ \text{ for all } \ t \in T_N, \\ &\text{while } H(\bar{x}_N(s), p_N(s+h_N), \bar{u}_N(s)) = -1 \end{aligned}$$

at the points $s \in T_N$ of control switch, where $\bar{u}_N(s) = 1$ regardless of h_N.

Let us specify the choice of r in (2.4) ensuring that $\bar{x}_N(1) > r$ along some subsequence of natural numbers. We claim that $\bar{x}_N(1) > r$ if $r \in (0,1)$ is an irrational number whose decimal representation contains infinitely many digits from the set $\{5,6,7,8,9\}$; e.g., $r = 0.676676667\ldots$. Indeed, put $h_N := 10^{-N}$, which is a subsequence of $h_N = N^{-1}$ as required in (2.4). It is easy to see that in this case the set of all reachable points at $t = 1$ is the set of rational numbers between 0 and 1 with exactly N digits in the fractional part of their decimal representations. In particular, for $N = 3$ this set is $\{0, 0.001, 0.002, \ldots, 0.999, 1\}$. Therefore, by the construction of r, the closest point to r from the reachable set is greater than r, and such a point must be the endpoint of the optimal trajectory $\bar{x}_N(1)$.

It remains to show that one always can choose a sequence of optimal control to (2.4) satisfying the properness condition. Taking r as above, we denote by $s(N) \in T_N$ the point of the grid closest to r. It is easy to see that the control

$$\bar{u}_N(t) := \begin{cases} 1, & t \le s(N), \\ 0, & t \ge s(N) + h_N \end{cases}$$

is optimal to (2.4) for each $N \in \mathbb{N}$, and the sequence $\{\bar{u}_N\}$ satisfies the properness condition. □

Example 2.2 contradicts the AMP with the transversality condition in the *conventional subdifferential form*, which is

$$-p_N(t_1) \in \partial\varphi(\bar{x}_N(t_1))$$

for problems with no endpoint constraints. In our example the function $\varphi(x) = |x-r|$ is *convex*, and hence the subdifferential ∂ is understood in the sense of convex analysis. Note that we actually showed that the subdifferential agrees with the gradient

$$\partial\varphi(\bar{x}_N(1)) = \big\{\nabla\varphi(\bar{x}_N(1))\big\} = \big\{1\big\} \ \text{ for all } \ N \in \mathbb{N}$$

along the optimal trajectories in (2.4) due to the choice of r. Since any reasonable (lower) subdifferential for nonsmooth functions must reduce to the convex subdifferential for convex ones, Example 2.2 proves that there is *no hope for an extension of the AMP in the conventional subdifferential form to problems with nonsmooth costs.*

The next example, complemented to Example 2.2, shows that the AMP fails even for problems with *differentiable* but *not continuously differentiable* cost functions.

EXAMPLE 2.3 (AMP may not hold for linear systems with differentiable, but not C^1, cost functions). *There is a one-dimensional control problem of minimizing a Fréchet differentiable but not continuously differentiable cost function over a linear system with no endpoint constraints such that a proper sequence of optimal controls to discrete approximations does not satisfy the AMP.*

Proof. Consider the same control system as in (2.4) and construct a minimizing function $\varphi(x)$ satisfying the above requirements. Let $\psi(x)$ be a C^1 function with the properties

$$\begin{aligned}&\psi(x) \ge 0, \quad \psi(x) = \psi(-x), \quad \psi(x) \equiv 0 \ \text{ if } \ |x| > 2,\\ &|\nabla\psi(x)| \le 1 \ \text{ for all } \ x, \quad \text{and } \ \nabla\psi(-1) = a > 0.\end{aligned}$$

Define the cost function $\varphi(x)$ by

$$\varphi(x) := \Big(x - \frac{1}{9}\Big)^2 + \sum_{n=1}^{\infty} 10^{-2n-3}\psi\left(10^{2n+3}\left(x - \sum_{k=1}^{n} 10^{-k}\right) - 1\right),$$

which is continuously differentiable around every point but $x = \frac{1}{9}$, where it is differentiable and attains its absolute minimum at $x = \frac{1}{9}$. As in Example 2.2, we put $h_N := 10^{-N}$, and then the point $x = \frac{1}{9}$ *cannot be reached* by discretization. It is not hard to check that the endpoint of the optimal trajectory $\bar{x}_N$ for each N is

$$\bar{x}_N(1) = \sum_{k=1}^{N} 10^{-k} \quad \text{with} \quad \nabla\varphi(\bar{x}_N(1)) = a + \varepsilon_N,$$

where $\varepsilon_N \downarrow 0$ as $N \to \infty$. Proceeding as in Example 2.2, with the same sequence of optimal controls, we have $H(\bar{x}_N(t), p_N(t+h_N), u) \equiv -au$, and the approximate maximum condition (1.3) does not hold at those points where $\bar{u}_N(t) = 1$. □

The last example in this section concerns systems with *nonsmooth dynamics*. We actually consider a finite-difference analogue of minimizing an integral functional over a one-dimensional control system, which is equivalent to a two-dimensional optimal control problem of the Mayer type. The discrete "integrand" in this problem is nonsmooth with respect to the state variable x; it happens to be continuously differentiable with respect to x *along* the optimal process $\{\bar{x}_N(\cdot), \bar{u}_N(\cdot)\}$ under consideration but *not uniformly* in N. Thus the example below demonstrates that the *uniform smoothness* assumption on f in a tube containing optimal trajectories is essential for the validity of the AMP formulated in section 1.

EXAMPLE 2.4 (violation of AMP for control problems with nonsmooth dynamics). *The AMP does not hold in discrete approximations of a minimization problem*

for an integral functional over a one-dimensional linear control system with no endpoint constraints such that the integrand is linear with respect to the control variable while convex and nonsmooth with respect to the state one. Moreover, the integrand in this problem happens to be C^1 with respect to the state variable along the sequence of optimal solutions to the discrete approximations (P_N) for all $N \in \mathbb{N}$ but not uniformly in N.

Proof. First we consider the following continuous-time optimal control problem:

$$\begin{cases} \text{minimize} \;\; J(x,u) := \displaystyle\int_0^{t_1} \big(u(t) + |x(t) - t^2/2|\big)dt \\ \text{subject to} \\ \dot{x} = tu, \quad x(0) = 0, \\ u(t) \in U := \big\{1, c\big\}, \quad 0 \le t \le t_1, \end{cases} \tag{2.5}$$

where the terminal time t_1 and the number $c > 1$ will be specified below. It is obvious that the optimal control to (2.5) is $\bar{u}(t) \equiv 1$ and the corresponding optimal trajectory is $\bar{x}(t) = t^2/2$.

Discretizing (2.5), we get the sequence of finite-difference control problems

$$\begin{cases} \text{minimize} \;\; J(x_N, u_N) := h_N \displaystyle\sum_{t \in T_N} \big(u_N(t) + |x_N(t) - t^2/2|\big) \\ \text{subject to} \\ x_N(t + h_N) = x_N(t) + h_N t u_N(t), \quad x_N(0) = 0, \\ u_N(t) \in U = \big\{1, c\big\}, \quad t \in T_N := \big\{k h_N\big\}_{k=0}^{N-1}. \end{cases} \tag{2.6}$$

Let us first show that $\bar{u}_N(t) \equiv 1$ remains the (unique) optimal control to (2.6) if the stepsize h_N is sufficiently small and (t_1, c) are chosen appropriately. Indeed, similarly to Example 2.1 we compute the trajectory to (2.6) corresponding to the control $u_N(t) \equiv 1$ as

$$x_N(t) = \frac{t^2}{2} - \frac{t h_N}{2} \;\; \text{for all} \;\; N \in \mathbb{N}.$$

The value $J_N(1)$ of the cost functional at $u_N(t) \equiv 1$ equals

$$J_N(1) = t_1 + h_N^2 \sum_{t \in T_N} \frac{t}{2} = t_1 + \frac{t_1^2 h_N}{4} + o(h_N). \tag{2.7}$$

If we replace $u_N(t) = 1$ by $u_N(t) = c$ at some point $t \in T_N$, then the increment of the summation $h_N \sum_{t \in T_N} u_N(t)$ equals $(c-1)h_N$. Hence

$$J(x_N, u_N) = h_N \sum_{t \in T_N} u_N(t) + h_N \sum_{t \in T_N} |x_N(t) - t^2/2| > h_N \sum_{t \in T_N} u_N(t) \ge t_1 + (c-1)h_N$$

for any feasible control $u_N(t)$ to (2.6), which is not $u_N(t) \equiv 1$. Comparing the latter with (2.7), we conclude that the control $u_N(t) \equiv 1$ is *optimal* to (2.6) if $(t_1)^2/4 < c - 1$ and N is sufficiently large.

Let us finally show that for $t_1 > 2$ and $c > t^2/4 + 1$ (e.g., for $t_1 = 3$ and $c = 4$) the sequence of optimal controls $\bar{u}_N \equiv 1$ does *not* satisfy the approximate maximum condition (1.3) at points $t \in T_N$ sufficiently close to $t = t_1/2$. Compute the

Hamilton–Pontryagin function (1.1) as a function of $t \in T_N$ and $u \in U$ at the optimal trajectory $\bar{x}_N(t)$ corresponding to the optimal control under consideration with the adjoint trajectory $p_N(t)$ to (1.6). Reducing (2.6) to the standard Mayer form and taking into account that $\bar{x}_N(t) < t^2/2$ for all $t \in T_N$ due to the above formula for the trajectory of (2.6) corresponding to $u_N(t) \equiv 1$, we get

$$\begin{aligned} H(t, \bar{x}_N(t), p_N(t+h_N), u) &= tp_N(t+h_N)u - u - |\bar{x}_N(t) - t^2/2| \\ &= (tp_N(t+h_N) - 1)u + (\bar{x}_N(t) - t^2/2), \end{aligned}$$

where $p_N(t)$ satisfies the equation

$$p_N(t) = p_N(t+h_N) + h_N, \quad p_N(t_1) = 0,$$

whose solution is $p_N(t) = t_1 - t$. Therefore,

$$\begin{aligned} H(t, \bar{x}_N(t), p_N(t+h_N), u) &= \big(t(t_1 - t + h_N) - 1\big)u + O(h_N) \\ &= (-t^2 + t_1 t - 1)u + O(h_N). \end{aligned}$$

The multiplier $-t^2 + t_1 t - 1$ is positive in the neighborhood of $t = t_1/2$ if its discriminant $t_1^2 - 4$ is positive. Thus $u = c$, but not $u = 1$, provides the maximum to the Hamilton–Pontryagin function around $t = t_1/2$ if h_N is sufficiently small. □

Observe that the constructions in Examples 2.2 and 2.4 are actually based on the same idea. The crucial point in Example 2.2 (and similarly in Example 2.3) is that, due to the *incommensurability* of the reachable set and the ideal point of minimum $x_N(1) = r$, the endpoint of the optimal trajectory $\bar{x}_N(1)$ turns out to be in the zone, where the discontinuous derivative of the cost function has the "wrong sign." A similar situation is in Example 2.4, but in this case the function $\frac{\partial H}{\partial x}$ is discontinuous with respect to x, and the optimal trajectory in the discrete problem deviates to the "wrong side" of the ideal (continuous-time) optimal trajectory.

3. Uniformly upper subdifferentiable functions. In this section we present some tools of nonsmooth analysis needed for the formulation and proofs of the main *positive* results of the paper: the AMP for ordinary and time-delay systems in the new *upper subdifferential* form. Results in this form are definitely nontraditional in optimization, since they concern *minimization* problems for which *lower* subdifferential constructions are usually employed. However, we saw in the preceding section that results of the conventional lower type simply do not hold for the AMP. In sections 4 and 5 we are going to employ *upper* subdifferential constructions for nonsmooth minimization problems of optimal control, which happen to work for a special class of *uniformly upper subdifferentiable* functions we describe and discuss in this section.

Given an extended-real-valued function $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}} := [-\infty, \infty]$ finite at $\bar{x}$, we first define its *Fréchet upper subdifferential* by

$$\widehat{\partial}^+ \varphi(\bar{x}) := \left\{ x^* \in \mathbb{R}^n \,\middle|\, \limsup_{x \to \bar{x}} \frac{\varphi(x) - \varphi(\bar{x}) - \langle x^*, x - \bar{x} \rangle}{\|x - \bar{x}\|} \le 0 \right\}. \tag{3.1}$$

This construction is known also as the "Fréchet superdifferential" or the "viscosity superdifferential"; it is extensively used in the theory of viscosity solutions. The set (3.1) is symmetric to the (lower) Fréchet subdifferential

$$\widehat{\partial}^+ \varphi(\bar{x}) = -\widehat{\partial}(-\varphi)(\bar{x}),$$

which is widely used in variational analysis under the name of "regular" or "strict" subdifferential; see, e.g., [12, 14]. The upper subdifferential (3.1) is our *primary* generalized differential construction in this paper. This set is closed and convex but may be empty for many functions useful in minimization. In fact, both $\widehat{\partial}^+\varphi(\bar{x})$ and $\widehat{\partial}\varphi(\bar{x})$ are nonempty simultaneously if and only if φ is Fréchet differentiable at $\bar{x}$ in which case

$$\widehat{\partial}^+\varphi(\bar{x}) = \widehat{\partial}\varphi(\bar{x}) = \big\{\nabla\varphi(\bar{x})\big\}.$$

Following [5], we define the *basic upper subdifferential* of φ at $\bar{x}$ by

$$\partial^+\varphi(\bar{x}) := \Big\{x^* \in \mathbb{R}^n \big| \ \exists\, x_k \to \bar{x} \ \text{ with } \ \varphi(x_k) \to \varphi(\bar{x}) \\ \text{and } \exists\, x_k^* \in \widehat{\partial}^+\varphi(x_k) \ \text{ with } \ x_k^* \to x^*\Big\}$$

and call φ *upper regular* at $\bar{x}$ if $\partial^+\varphi(\bar{x}) = \widehat{\partial}^+\varphi(\bar{x})$. This class includes, in particular, all strictly differentiable functions as well as proper concave functions. In the concave case $\widehat{\partial}^+\varphi(\bar{x})$ reduces to the upper subdifferential of convex analysis, which is nonempty whenever $\bar{x} \in \operatorname{ri}(\operatorname{dom}\varphi)$. Moreover, $\widehat{\partial}^+\varphi(\bar{x}) \neq \emptyset$ if φ is upper regular at $\bar{x}$ and Lipschitz continuous around this point. In the latter case the upper regularity of φ agrees with the subdifferential regularity of $-\varphi$ at the same point in the sense of [12]. It is interesting to observe that, for Lipschitzian upper regular functions, the Fréchet upper subdifferential (3.1) agrees with Clarke's generalized gradient $\overline{\partial}\varphi(\bar{x})$ of [1]. Indeed, one has

$$\overline{\partial}\varphi(\bar{x}) = \operatorname{co}\partial^+\varphi(\bar{x})$$

if φ is Lipschitz continuous around $\bar{x}$; see, e.g., [5, Theorem 2.1]. Since $\partial^+\varphi(\bar{x}) = \widehat{\partial}^+\varphi(\bar{x})$ for upper regular functions and since $\widehat{\partial}^+\varphi(\bar{x})$ is always convex, we arrive at $\overline{\partial}\varphi(\bar{x}) = \widehat{\partial}^+\varphi(\bar{x})$.

Let us now define a class of functions for which we obtain an extension of the AMP to nonsmooth control problems in the next section.

DEFINITION 3.1 (uniform upper subdifferentiability). *A function $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is* uniformly upper subdifferentiable *around a point $\bar{x}$, where it is finite, if there is a neighborhood V of $\bar{x}$ such that for every $x \in V$ there exists $x^* \in \mathbb{R}^n$ with the following property: Given any $\varepsilon > 0$, there is $\eta > 0$ for which*

$$\varphi(v) - \varphi(x) - \langle x^*, v - x\rangle \le \varepsilon\|v - x\| \tag{3.2}$$

whenever $v \in V$ with $\|v - x\| \le \eta$.

It is easy to check that the class of uniformly upper subdifferentiable functions includes all continuously differentiable functions and concave continuous functions, and also is closed with respect to taking the minimum over compact sets. The uniform upper subdifferentiability property of φ around $\bar{x}$ is actually a localization of the so-called weak convexity property for $-\varphi$ in the sense of [10], which has been broadly used in numerical optimization. Note that if φ is Lipschitz continuous and differentiable at some point, it may not be uniformly upper subdifferentiable around it, for example, $\varphi(x) = x^2\sin(1/x)$ for $x \neq 0$ with $\varphi(0) = 0$. The following result shows, in particular, that uniformly upper subdifferential functions enjoy upper regularity and fully describe the set of x^* satisfying (3.2).

PROPOSITION 3.2 (upper regularity of uniformly upper subdifferentiable functions). *Let φ be uniformly upper subdifferentiable around $\bar{x}$. Then it is upper regular at $\bar{x}$ and Lipschitz continuous around this point, and property* (3.2) *holds for all $x^* \in \widehat{\partial}^+\varphi(x)$ with x around $\bar{x}$.*

Proof. Denote by $G(x)$ the set of x^* for which (3.2) holds. This set is nonempty and, as directly follows from (3.2), it is closed, convex, and bounded for all $x \in V$. One can immediately observe from the comparison of (3.1) and (3.2) that $G(x) \subset \widehat{\partial}^+\varphi(x)$. Let us show that in fact $G(x) = \widehat{\partial}^+\varphi(x)$ whenever $x \in V$.

It follows from the results of [10, section 1.1] that φ is locally Lipschitzian around $\bar{x}$ and directionally differentiable on V in any direction, and its directional derivative admits the representation

$$\varphi'(x;w) = \min\big\{\langle x^*, w\rangle \big|\ x^* \in G(x)\big\} \quad \text{for all} \quad x \in V,\ w \in \mathbb{R}^n. \tag{3.3}$$

As is well known, the Fréchet upper subdifferential (3.1) of a locally Lipschitzian function φ at x is representable as

$$\widehat{\partial}^+\varphi(x) = \left\{ x^* \in \mathbb{R}^n \,\middle|\, \langle x^*, w\rangle \ge \limsup_{\tau\downarrow 0} \frac{\varphi(x+\tau w) - \varphi(x)}{\tau} \ \text{ for all } \ w \in \mathbb{R}^n \right\}.$$

Comparing the latter with (3.3), we get $G(x) = \widehat{\partial}^+\varphi(x)$ for all $x \in V$. Furthermore, it is not hard to show directly from the definition that the mapping $G\colon V \rightrightarrows \mathbb{R}^n$ is closed-graph on any compact subset of V. Finally taking into account the construction of the basic subdifferential, we conclude that $\partial^+\varphi(x) = \widehat{\partial}^+\varphi(x)$ for all $x \in V$. ∎

Note that Proposition 3.2 is in accordance with [12, Theorem 9.16], which gives a characterization of the simultaneous Lipschitz continuity and subdifferential regularity of a function on an open set via the existence of the classical directional derivative and its upper semicontinuity with respect to directions. Note also that we need an extra requirement on the *uniform* upper subdifferentiability in Definition 3.1, which essentially restricts the class of functions suitable for applications to the AMP in the upper subdifferential form, due to the *parametric* nature of finite-difference systems viewed as a process as $N \to \infty$. In particular, the Lipschitz continuity and upper regularity are *not* needed for upper subdifferential results related to the necessary optimality condition for *fixed* solutions in various problems of constrained optimization and optimal control; cf. [7, 8].

4. AMP in upper subdifferential form. This is a central section of the paper, which collects the main positive results on the fulfillment of the AMP in the upper subdifferential form for the discrete approximation problems (P_N). We derive three closely related versions of the AMP in somewhat different settings of (P_N). The first version applies to problems with *no endpoint constraints* and establishes the upper subdifferential form of the AMP with *no properness* requirement on the sequence of optimal controls and with an *error estimate* as $\varepsilon(t, h_N) = O(h_N)$ in the approximate maximum condition. The second result, with a different proof, is the major version of the AMP for the *constrained nonsmooth* problems (P_N), which extends the one formulated in section 1. The last version of the AMP concerns discrete approximation problems in the case of *incommensurability* of the time interval $t_1 - t_0$ and the discretization step h_N. This version is basic for the extension of the AMP to time-delay systems obtained in the next section.

Let us start with the upper subdifferential form of the AMP for problems with

no endpoint constraints. Throughout this section impose the following *standing assumptions* on the mapping f and the control set U:

(H1) $f = f(t,x,u)$ is continuous with respect to all its variables and continuously differentiable with respect to the state variable x in some tube containing optimal trajectories for all u from the compact set U in a metric space and for all $t \in T_N$ uniformly in $N \in \mathbb{N}$.

THEOREM 4.1 (AMP for problems with no endpoint constraints). *Let the pairs $(\bar{x}_N, \bar{u}_N)$ be optimal to problems (P_N) with $\varphi_i = 0$ for all $i = 1, \dots, m+r$. Assume in addition to* (H1) *that φ_0 is uniformly upper subdifferentiable around the limiting point(s) of the sequence $\{\bar{x}_N(t_1)\}$, $N \in \mathbb{N}$. Then for every sequence of upper subgradients $x_N^* \in \widehat{\partial}^+\varphi_0(\bar{x}_N(t_1))$ there is $\varepsilon(t,h_N) \to 0$ as $N \to \infty$ uniformly in $t \in T_N$ such that the approximate maximum condition* (1.3) *holds for all $t \in T_N$, where each $p_N(t)$ satisfies the adjoint system* (1.6) *with the transversality condition*

$$p_N(t_1) = -x_N^* \quad \text{for all} \quad N \in \mathbb{N}. \tag{4.1}$$

Moreover, $\varepsilon(t,h_N) = O(h_N)$ in (1.3) *if φ_0 is locally concave around $\bar{x}_N(t_1)$ uniformly in N while $\partial f(\cdot,u,t)/\partial x$ is locally Lipschitz around $\bar{x}_N(t)$ with a constant uniform in $u \in U$, $t \in T_N$, $N \in N$.*

Proof. Considering a sequence of optimal solutions $(\bar{x}_N, \bar{u}_N)$ to (P_N), we suppose that $\bar{x}_N(t)$ belong to the uniform neighborhoods in the assumptions made for all $N \in \mathbb{N}$. It follows from the uniform upper subdifferentiability of φ_0 by Proposition 3.2 that $\widehat{\partial}^+\varphi_0(\bar{x}_N(t_1)) \neq \emptyset$ and that inequality (3.2) holds for any $x^* \in \widehat{\partial}^+\varphi_0(\bar{x}_N(t_1))$ as $N \to \infty$. Now taking an arbitrary sequence of $x_N^* \in \widehat{\partial}^+\varphi_0(\bar{x}_N(t_1))$, we get

$$\varphi_0(x) - \varphi_0(\bar{x}_N(t_1)) \le \langle x_N^*, x - \bar{x}_N(t_1)\rangle + o(\|x - \bar{x}_N(t_1)\|), \tag{4.2}$$

where

$$\frac{o(\|x - \bar{x}_N(t_1)\|)}{\|x - \bar{x}_N(t_1)\|} \to 0$$

as $x \to x_N(t_1)$ *uniformly* in N. Moreover, one can clearly eliminate o, i.e., put $o(\|x - \bar{x}_N(t_1)\|) \equiv 0$ if φ_0 is assumed to be uniformly locally concave. Letting $p_N(t_1) := -x_N^*$ as in (4.1), we derive from (4.2) that

$$J(x_N, u_N) - J(\bar{x}_N, \bar{u}_N) \le -\big\langle p_N(t_1), \Delta x_N(t_1)\big\rangle + o(\|\Delta x_N(t_1)\|) \tag{4.3}$$

with $\Delta x_N(t) := x_N(t) - \bar{x}_N(t)$ for all feasible processes (x_N, u_N) to (P_N) whenever $x_N(t_1)$ is sufficiently close to $\bar{x}_N(t_1)$. From the identity

$$\begin{aligned}\big\langle p_N(t_1), \Delta x_N(t_1)\big\rangle &= \sum_{t \in T_N} \langle p_N(t+h_N) - p_N(t), \Delta x_N(t)\rangle \\ &\quad + \sum_{t \in T_N} \langle p_N(t+h_N), \Delta x_N(t+h_N) - \Delta x_N(t)\rangle\end{aligned}$$

and (4.3) we get

(4.4)
$$\begin{aligned}0 &\le J(x_N,u_N)-J(\bar{x}_N,\bar{u}_N)\le -\big\langle p_N(t_1),\Delta x_N(t_1)\big\rangle+o(\|\Delta x_N(t_1)\|)\\ &=-\sum_{t\in T_N}\langle p_N(t+h_N)-p_N(t),\Delta x_N(t)\rangle\\ &\quad -h_N\sum_{t\in T_N}\left\langle p_N(t+h_N),\frac{\partial f}{\partial x}(t,\bar{x}_N,\bar{u}_N)\Delta x_N(t)\right\rangle\\ &\quad -h_N\sum_{t\in T_N}\Delta_u H(t,\bar{x}_N(t),p_N(t+h_N),\bar{u}_N(t))+h_N\sum_{t\in T_N}\eta_N(t)+o(\|\Delta x_N(t_1)\|),\end{aligned}$$

where the remainder $\eta_N(t)$ is computed by

(4.5)
$$\begin{aligned}\eta_N(t)=&\left\langle\frac{\partial H}{\partial x}(t,\bar{x}_N(t),p_N(t+h_N),u_N(t))-\frac{\partial H}{\partial x}(t,\bar{x}_N(t),p_N(t+h_N),\bar{u}_N(t)),\Delta x_N(t)\right\rangle\\ &+o(\|\Delta x_N(t)\|)\end{aligned}$$

with $o(\|\Delta x_N(t)\|)$ uniform in N due to (H1), and where

$$\begin{aligned}\Delta_u H(t,\bar{x}_N(t),p_N(t+h_N),\bar{u}_N(t)) :=\ & H(t,\bar{x}_N(t),p_N(t+h_N),u_N(t))\\ & -H(t,\bar{x}_N(t),p_N(t+h_N),\bar{u}_N(t)).\end{aligned}$$

One can easily see that $o(\|\Delta x_N(t)\|)=O\big(\|\Delta x_N(t)\|^2\big)$ in (4.5), uniformly in N, under the additional Lipschitzian assumption on $\partial f(\cdot,u,t)/\partial x$ in the theorem.

Now let us consider *needle variations* of the optimal controls $\bar{u}_N$ in the form

$$u_N(t)=\begin{cases}v & \text{if } t=\tau,\\ \bar{u}_N(t) & \text{if } t\in T_N\setminus\{\tau\},\end{cases}\tag{4.6}$$

where $v\in U$ and $\tau=\tau(N)\in T_N$ as $N\in\mathbb{N}$. All the controls (4.6) are feasible to the discrete problems with no endpoint constraints. The trajectory increments corresponding to the needle variations (4.6) satisfy

$$\Delta x_N(t)=0 \ \text{ for } \ t=t_0,\dots,\tau \quad \text{ and } \quad |\Delta x_N(t)|=O(h_N) \ \text{ for } \ t=\tau+h_N,\dots,t_1.$$

Taking this into account and substituting (4.6) into (4.4), we get

$$\begin{aligned}0 &\le J(x_N,u_N)-J(\bar{x}_N,\bar{u}_N)\\ &\le -h_N\Delta_u H(\tau,\bar{x}_N(\tau),p_N(\tau+h_N),\bar{u}_N(\tau))+o(h_N),\end{aligned}\tag{4.7}$$

where one obviously has $o(h_N)=O(h_N^2)$ under the additional concavity and Lipschitzian assumptions made in the theorem. Arguing by contradiction, we derive from (4.7) the approximate maximum condition (1.3), with the specification of $\varepsilon(t,h_N)$ under the additional assumptions, and complete the proof of the theorem. □

Remark 4.1 (upper versus lower subdifferential forms of transversality conditions). The main difference between the conventional (lower) subdifferential form, which is *not* actually fulfilled in the case of the AMP, and the upper subdifferential form of Theorem 4.1 is that the transversality condition (4.1) holds for *every* upper subgradient $x_N^*\in\widehat{\partial}^+\varphi_0(\bar{x}_N(t_1))$ instead of just for *some* lower subgradient in the

conventional transversality conditions for continuous-time and discrete-time (with a fixed step) systems. In particular, for discrete-time systems with convex velocity sets both lower and upper subdifferential forms of the (exact) discrete maximum principle hold; see [8], where the upper subdifferential/superdifferential form of the discrete maximum principle has been established under milder assumptions on φ_0 in comparison with Theorem 4.1. If φ_0 is Lipschitz continuous and upper regular and hence $\widehat{\partial}^+\varphi_0(\bar{x}) = \overline{\partial}\varphi_0(\bar{x})$, which is the case for uniformly upper subdifferential functions by Proposition 3.2, there is indeed a *dramatic* difference between the upper subdifferential form of transversality conditions and a well-recognized form in terms of the Clarke subdifferential: Instead of establishing the fulfillment transversality for just some element of $\overline{\partial}\varphi_0(\bar{x}(t_1))$ we establish it for the *whole set*. A similar situation takes place for continuous-time systems, where the upper subdifferential form of transversality in the maximum principle can be proved, in problems with no endpoint constraints, in the line of arguments of Theorem 4.1. Observe, however, that there is a more subtle lower subdifferential form of transversality conditions for continuous-time and discrete-time (of a fixed step) systems that involves basic/limiting subgradients but not Clarke ones; see [5, 14]. This form is generally independent of the upper subdifferential form of transversality conditions. Note that the major drawback of the upper subdifferential form is that it applies to a restrictive class of functions. But, as seen in section 2, there is *no alternative* to this form for the AMP.

Next let us consider a sequence of the discrete approximation problems (P_N) with *endpoint constraints* of the inequality and equality types. We are going to derive an extension of the AMP formulated in section 1 to these problems involving *nonsmooth* functions that describe the cost and inequality constraints. The following upper subdifferential version of the AMP for constrained problems imposes the *uniform upper subdifferentiability* property on the cost and inequality constraint functions, the *properness* assumption on the sequence of optimal controls, and the *consistency* condition on perturbations of the equality constraints. As seen in section 2, all three requirements are essential.

The proof of the AMP for constrained problems is substantially different from the one in Theorem 4.1 and much more involved, although it employs the same approach to handle nonsmoothness. The major part of the proof goes back to the smooth setting and is based on a finite-difference counterpart of the *hidden convexity* properties for sequences of discrete approximations.

Before formulating and proving the theorem, we need an auxiliary result that actually reflects the hidden convexity property in the nonsmooth setting under consideration. Let us first recall some definitions from [4, 5]. Given a sequence of feasible solutions (x_N, u_N) to (P_N), we say that the inequality constraint

$$\varphi_i(x_N(t_1)) \le \gamma_{iN} \ \text{ with } \ i \in \{1, \dots, m\}$$

is *essential* for x_N along a subsequence $\mathcal{M}$ of natural numbers if $\varphi_i(x_N(t_1)) - \gamma_{iN} = O(h_N)$ as $h_N \to 0$; i.e., there is $K_i \ge 0$ such that

$$-K_i h_N \le \varphi_i(x_N(t_1)) - \gamma_{iN} \le 0 \ \text{ as } \ N \to \infty, \ \ N \in \mathcal{M}.$$

This constraint is *inessential* for x_N along $\mathcal{M}$ if for any $K > 0$ there is $N_0 \in \mathbb{N}$ such that

$$\varphi_i(x_N(t_1)) - \gamma_{iN} \le -K h_N \ \text{ for all } \ N \ge N_0, \ \ N \in \mathcal{M}.$$

Note that the notion of essential constraints in sequences of discrete approximations corresponds to the notion of *active* constraints in nonparametric optimization problems. Without loss of generality we suppose that for the sequence of optimal trajectories $\bar{x}_N$ to (P_N) under consideration the first $l \in \{1,\dots,m\}$ inequality constraints are essential while the other $m-l$ constraints are inessential along all natural numbers, i.e., with $\mathcal{M} = \mathcal{N}$.

Assume now that $\widehat{\partial}^+\varphi_i(\bar{x}_N(t_1)) \neq \emptyset$ for all $i = 0,\dots,l$ and $N \in \mathbb{N}$ sufficiently large and fix some sequence of upper subgradients $x^*_{iN} \in \widehat{\partial}^+\varphi_i(\bar{x}_N(t_1))$ for such i and N. Denote by $\Delta_{\tau,v}\bar{x}_N(t_1)$ the *endpoint increment* generated by the *needle variation* (4.6) of the optimal control $\bar{u}_N$ with some $\tau \in T_N$ and $v \in U$. Form the set

$$S_N := \big\{(s_0,\dots,s_l) \in \mathbb{R}^n \big|\ s_i = \big\langle x^*_{iN}, \Delta_{\tau,v}\bar{x}_N(t_1)\big\rangle,\ \tau \in T_N,\ \ v \in U\big\} \tag{4.8}$$

along the fixed sequences of the above upper subgradients x^*_{iN} and consider the *negative orthant* in R^{l+1} given by

$$\mathbb{R}^{l+1}_< := \big\{(x_0,\dots,x_l) \in \mathbb{R}^{l+1} \big|\ x_i < 0 \ \text{ for all } \ i = 0,\dots,l\big\}.$$

The following result is due to the hidden convexity property of finite-difference systems established in [4, 5] with the adjustment to nonsmoothness via uniform upper subdifferentiability.

LEMMA 4.2 (hidden convexity). *Let $(\bar{x}_N, \bar{u}_N)$ be a sequence of optimal solutions to problems (P_N) with no equality constraints and with the inequality constraints such that the first $l \in \{1,\dots,m\}$ of them are essential for the sequence of $\bar{x}_N$ while the other are inessential for this sequence. In addition to* (H1) *assume that each φ_i, $i = 0,\dots,l$, is uniformly upper subdifferentiable around the limiting point(s) of $\{\bar{x}_N(t_1)\}$, $N \in \mathbb{N}$, and that*

(H2) *the sequence of optimal controls $\{\bar{u}_N\}$ is proper.*

Then there is a sequence of $(l+1)$-dimensional quantities of order $o(h_N)$ as $h_N \downarrow 0$ such that

$$\big(\operatorname{co} S_N + o(h_N)\big) \cap \mathbb{R}^{l+1}_< = \emptyset \ \text{ for large } \ N \in \mathbb{N}, \tag{4.9}$$

where $\operatorname{co} S_N$ stands for the convex hull of the set S_N in (4.8) *built upon the given sequences of upper subgradients $x^*_{iN} \in \widehat{\partial}^+\varphi_i(\bar{x}_N(t_1))$, $i = 0,\dots,l$.*

Proof. It follows the proof of Lemma 3 in [4] based on the hidden convexity property of Theorem 1 therein (respectively, [5, Lemma 16.2 and Theorem 15.1]). The only essential difference is that the equalities

$$\begin{aligned}\varphi_i(\bar{x}_N(t_1)) \ - \ \varphi_i(\bar{x}_N(t_1)) - \big\langle \nabla\varphi_i(\bar{x}_N(t_1)), \Delta x_N(t_1)\big\rangle \\ + \ o(\|\Delta x_N(t_1)\|) = 0, \quad i = 0,\dots,l,\end{aligned}$$

in the smooth case of [4, 5] are replaced with the inequalities

$$\varphi_i(\bar{x}_N(t_1)) - \varphi_i(\bar{x}_N(t_1)) - \big\langle x^*_{iN}, \Delta x_N(t_1)\big\rangle + o(\|\Delta x_N(t_1)\|) \le 0, \quad i = 0,\dots,l,$$

due to the uniform upper subdifferentiability of φ_i. Compare this proof to the proof of Theorem 4.1. □

Based on Lemma 4.2, we get the following extension of the AMP to finite-difference problems with nonsmooth inequality and smooth equality constraints.

THEOREM 4.3 (AMP for problems with endpoint constraints). *Let the pairs $(\bar{x}_N, \bar{u}_N)$ be optimal to problems (P_N), where the first $l \in \{1,\dots,m\}$ inequality constraints are essential for $\bar{x}_N$, $N \in \mathbb{N}$. In addition to* (H1) *and* (H2) *assume that*

φ_i are uniformly upper subdifferentiable around the limiting point(s) of $\{\bar{x}_N(t_1)\}$ for $i=0,\dots,l$ and continuously differentiable around them for $i=m+1,\dots,m+r$, and that

(H3) *the consistency condition* (1.2) *holds for the perturbations δ_{iN} of the equality constraints.*

Then for any sequences of upper subgradients $x_{iN}^*\in\widehat{\partial}^+\varphi_i(\bar{x}_N(t_1))$, $i=0,\dots,l$, there are numbers $\{\lambda_{iN}|\ i=0,\dots,l,m+1,\dots,m+r\}$ and a function $\varepsilon(t,h_N)\to 0$ as $N\to\infty$ uniformly in $t\in T_N$ such that the approximate maximum condition (1.3) *is fulfilled with the adjoint trajectory $p_N(t)$ to* (1.6) *satisfying the transversality condition*

$$p_N(t_1)=-\sum_{i=0}^{l}\lambda_{iN}x_{iN}^*-\sum_{i=m+1}^{m+r}\lambda_{iN}\nabla\varphi_i(\bar{x}_N(t_1)) \tag{4.10}$$

along with

$$\varphi_i(\bar{x}_N(t_1))-\gamma_{iN}=O(h_N)\quad\text{for}\quad i=1,\dots,l, \tag{4.11}$$

$$\lambda_{iN}\ge 0\ \text{ for }\ i=0,\dots,l,\quad\text{and}\quad\sum_{i=0}^{l}\lambda_{iN}^2+\sum_{i=m+1}^{m+r}\lambda_{iN}^2=1. \tag{4.12}$$

Proof. Let us first consider the case of inequality constraints, i.e., when $\varphi_i=0$ for the indices $i=m+1,\dots,m+r$ in (P_N). Take arbitrary sequences of $x_{iN}^*\in\widehat{\partial}^+\varphi_i(\bar{x}_N(t_1))$ as $i=0,\dots,l$. By Lemma 4.2 we apply the separation theorem to the convex sets in (4.9). It follows from the structures of these sets that there are $\lambda_{iN}\ge 0$ for $i=0,\dots,l$ and all N sufficiently large satisfying $\lambda_{0N}^2+\dots+\lambda_{lN}^2=1$ and

$$\sum_{i=0}^{l}\big\langle x_{iN}^*,\Delta_{\tau,v}\bar{x}_N(t_1)\big\rangle+o(h_N)\ge 0$$

for any $\tau\in T_N$ and $v\in U$. Then considering the trajectory $p_N(t)$ of the adjoint system (1.6) with the transversality condition

$$p(t_1)=-\sum_{i=0}^{l}\lambda_{iN}x_{iN}^*$$

and arguing as in the proof of Theorem 4.1, we get the inequality

$$h_N\big[H(\tau,\bar{x}_N(\tau),p_N(\tau+h_N),v)-H(\tau,\bar{x}_N(\tau),p_N(\tau+h_N),\bar{u}_N(\tau))\big]+o(h_N)\le 0$$

held for all $\tau\in T_N$ and $v\in U$. This easily implies the approximate maximum condition (1.3). Since (4.11) just means that the inequality constraints are essential for $\bar{x}_N$ as $i=1,\dots,l$, we arrive at all the conclusions of the theorem for the case of inequality constraints. Observe that the result obtained ensures the fulfillment of the AMP with *zero multipliers* corresponding to inessential inequality constraints.

Next let us consider the general case of (P_N) when the equality constraints are present as well. Each equality constraint $\varphi_{iN}(x)\le\delta_N$ can be obviously represented as the two inequality constraints

$$\varphi_{iN}^+(x):=\varphi_i(x)-\delta_{iN}\le 0,\quad\varphi_{iN}^-(x):=-\varphi_i(x)-\delta_{iN}\le 0 \tag{4.13}$$

for $i = m+1, \ldots, m+r$. Let us show that if one of the constraints (4.13) is essential for $\bar{x}_N$ along some subsequence $\mathcal{M} \subset \mathbb{N}$, then the other is inessential along the same subsequence under the consistency condition (1.2). Indeed, suppose for definiteness that the constraint $\varphi_{iN}^+(\bar{x}_N(t_1)) \le 0$ is essential for some $i \in \{m+1, \ldots, m+r\}$ along $\mathcal{M}$. Then by (1.2) we have

$$\varphi_{iN}^-(\bar{x}_N(t_1)) = -\varphi_i(\bar{x}_N(t_1)) + \delta_{iN} - 2\delta_{iN} = -\varphi_{iN}^+(\bar{x}_N(t_1)) - 2\delta_{iN} \le Kh_N, \quad N \in \mathcal{M},$$

for any $K > 0$ as $N \to \infty$, which means that the constraint $\varphi_{iN}^-(\bar{x}_N(t_1)) \le 0$ is inessential. Since φ_i is assumed to be C^1 for $i = m+1, \ldots, m+r$, both φ_{iN}^+ and φ_{iN}^- are uniformly upper subdifferentiable around the reference points. Applying now the inequality case of the theorem that has been already proved, we find either λ_{iN}^+ or λ_{iN}^- corresponding to one of the essential constraints $\varphi_{iN}^+(\bar{x}_N(t_1)) \le 0$ and $\varphi_{iN}^-(\bar{x}_N(t_1)) \le 0$, respectively. Finally putting

$$\lambda_{iN} := \lambda_{iN}^+ \quad \text{or} \quad \lambda_{iN} := -\lambda_{iN}^-, \quad i = m+1, \ldots, m+r,$$

depending on which of these constraints is essential, we complete the proof of the theorem. □

Note that both Theorems 4.1 and 4.3 concern the discrete approximation problems (P_N) with $t_1 - t_0 = Nh_N$, i.e., when the time interval and the discretization step are commensurable. Of course, it is not always the case in applications. Moreover, to extend the AMP to time-delay systems in the next section, we reduce them to systems with no delays but with *incommensurable* $t_1 - t_0$ and h_N. To proceed in this way, one needs to use modifications of the above results in the case of incommensurability. Let us present the corresponding modification of Theorem 4.1 for problems with no endpoint constraints. For simplicity we use the notation $f(t, x_N, u_N) := f(t, x_N(t), u_N(t))$ and consider the following sequences of discrete approximations with the grid

$$T_N := \{t_0, t_0 + h_N, \ldots, t_1 - \widetilde{h}_N - h_N\}, \quad h_N := \frac{t_1 - t_0}{N},$$

$$\widetilde{h}_N := t_1 - t_0 - h_N \left[\frac{t_1 - t_0}{h_N}\right],$$

where $[a]$ stands as usual for the greatest integer less than or equal to the real number a. The modified problems are written as

$$(\widetilde{P}_N) \qquad \begin{cases} \text{minimize } J(x_N, u_N) := \varphi(x_N(t_1)) \\ \text{subject to} \\ x_N(t + h_N) = x_N(t) + h_N f(t, x_N, u_N), \quad t \in T_N, \quad x_N(t_0) = x_0 \in \mathbb{R}^n, \\ x_N(t_1) = x_N(t_1 - \widetilde{h}_N) + \widetilde{h}_N f(t_1 - \widetilde{h}_N, x_N, u_N), \\ u_N(t) \in U, \quad t \in T_N. \end{cases}$$

THEOREM 4.4 (AMP for problems with incommensurability). *Let the pairs $(\bar{x}_N, \bar{u}_N)$ be optimal to problems $(\widetilde{P}_N)$. Assume in addition to* (H1) *that φ is uniformly upper subdifferentiable around the limiting point(s) of the sequence $\{\bar{x}_N(t_1)\}$, $N \in \mathbb{N}$. Then for every sequence of upper subgradients $x_N^* \in \widehat{\partial}^+\varphi(\bar{x}_N(t_1))$ there is $\varepsilon(t, h_N) \to 0$ as $N \to \infty$ uniformly in $t \in T_N$ such that the approximate maximum condition*

$$H(t, \bar{x}_N, p_N, u_N) = \max_{u \in U} H(t, \bar{x}_N, p_N, u) + \varepsilon(t, h_N)$$

holds for all $t \in \widetilde{T}_N := T_N \cup \{t_1 - \widetilde{h}_N\}$, *where the Hamilton–Pontryagin function is defined by*

$$H(t, x_N, p_N, u) := \begin{cases} \langle p_N(t + h_N), f(t, x_N, u)\rangle & \text{if } t \in T_N, \\ \langle p_N(t), f(t - \widetilde{h}_N, x_N, u)\rangle & \text{if } t = t_1 - \widetilde{h}_N, \end{cases}$$

and where each $p_N(t)$ *satisfies the adjoint system*

$$\begin{cases} p_N(t) = p_N(t + h_N) + h_N \dfrac{\partial f}{\partial x}^*(t, \bar{x}_N, \bar{u}_N) p_N(t + h_N), \quad t \in T_N, \\ p_N(t_1 - \widetilde{h}_N) = p_N(t_1) + \widetilde{h}_N \dfrac{\partial f}{\partial x}^*(t_1 - \widetilde{h}_N, \bar{x}_N, \bar{u}_N) p_N(t_1) \end{cases}$$

with the transversality condition $p_N(t_1) = -x_N^*$.

Proof. It is similar to the proof of Theorem 4.1 with the modification of the increment formula for the minimizing functional,

$$\begin{aligned} 0 \le\ & J(x_N, u_N) - J(\bar{x}_N, \bar{u}_N) \le -\langle p_N(t_1), \Delta x_N(t_1)\rangle + o(\|\Delta x_N(t_1)\|) \\ =\ & -\sum_{t \in T_N} \langle p_N(t + h_N) - p_N(t), \Delta x_N(t)\rangle \\ & -\langle p_N(t_1) - p_N(t_1 - \widetilde{h}_N), \Delta x_N(t_1 - \widetilde{h}_N)\rangle \\ & - h_N \sum_{t \in T_N} \left\langle p_N(t + h_N), \frac{\partial f}{\partial x}(t, \bar{x}_N, \bar{u}_N) \Delta x_N(t) \right\rangle \\ & - \widetilde{h}_N \left\langle p_N(t_1), \frac{\partial f}{\partial x}(t_1 - \widetilde{h}_N, \bar{x}_N, \bar{u}_N) \Delta x_N(t_1 - \widetilde{h}_N) \right\rangle \\ & - h_N \sum_{t \in \widetilde{T}_N} \Delta_u H(t, \bar{x}_N, p_N, \bar{u}_N) + h_N \sum_{t \in \widetilde{T}_N} \eta_N(t) + o(\|\Delta x_N(t_1)\|), \end{aligned}$$

where $\Delta_u H$ and $\eta_N(t)$ are defined as above. Substituting now the adjoint trajectory into this formula and using the needle variation (4.6), we arrive at the conclusions of the theorem. □

Similar to the proof of Theorem 4.3 we can get its modification to the case of incommensurability with the transversality and related conditions (4.10)–(4.12).

5. AMP for discrete approximations of delay systems. This section is devoted to the extension of the AMP in the upper subdifferential form to finite-difference approximations of *time-delay* control systems. Actually we are not familiar with any previous results on the AMP for optimal control problems with delays, so the results obtained below seem to be new even for smooth-delay problems.

We mainly pay attention to discrete approximations of the following time-delay problem with no endpoint constraints:

$$(D) \qquad \begin{cases} \text{minimize } J(x, u) := \varphi(x(t_1)) \\ \text{subject to} \\ \dot{x}(t) = f(t, x(t), x(t - \theta), u(t)) \ \text{ a.e. } \ t \in [t_0, t_1], \\ x(t) = c(t), \quad t \in [t_0 - \theta, t_0], \\ u(t) \in U \ \text{ a.e. } \ t \in [t_0, t_1] \end{cases}$$

over measurable controls and absolute continuous trajectories, where $\theta > 0$ is the constant time delay, and where $c\colon [t_0 - \theta, t_0] \to \mathbb{R}^n$ is a given function defining the

initial "tail" condition that is necessary to start the delay system. Based on the above constructions for nondelayed systems, one can derive similar results for delay systems with endpoint constraints. We may also extend the results obtained to more complicated delay systems involving variable delays, set-valued tail conditions, etc. On the other hand, we show in the end of this section that the AMP *does not hold* for discrete approximations of functional-differential systems of *neutral type* that contain time delays not only in state variables but in velocity variables as well.

Let us build discrete approximations of the time-delay problem (D) based on the Euler finite-difference replacement of the derivative. In the case of time-delay systems we need to ensure that the point $t-\theta$ belongs to the discrete grid when t does. It can be achieved by defining the discretization step as $h_N := \frac{\theta}{N}$ in contrast to $h_N = \frac{t_1-t_0}{N}$ for the nondelayed problems (P_N). In such a scheme the length of the time interval $t_1 - t_0$ is generally *no longer commensurable* with the discretization step h_N.

To this end we consider the following sequences of discrete approximations of the delay problem (D) with the grid on the main interval $[t_0, t_1]$ given by

$$T_N := \big\{t_0, t_0 + h_N, \ldots, t_1 - \widetilde{h}_N - h_N\big\}, \quad h_N := \frac{\theta}{N}, \quad \widetilde{h}_N := t_1 - t_0 - h_N \left[\frac{t_1 - t_0}{h_N}\right]$$

but also involving the grid T_{0N} on the initial interval $[t_0 - \theta, t_0]$ as follows:

$$(D_N) \qquad \begin{cases} \text{minimize } J(x_N, u_N) := \varphi(x_N(t_1)) \\ \text{subject to} \\ x_N(t + h_N) = x_N(t) + h_N f(t, x_N(t), x_N(t - Nh_N), u_N(t)), \ t \in T_N, \\ x_N(t_1) = x_N(t_1 - \widetilde{h}_N) + \widetilde{h}_N f(t_1 - \widetilde{h}_N, x_N(t_1 - \widetilde{h}_N), u_N(t_1 - \widetilde{h}_N)), \\ x_N(t) = c(t), \quad t \in T_{0N} := \big\{t_0 - \theta, t_0 - \theta + h_N, \ldots, t_0\big\}, \\ u_N(t) \in U, \quad t \in T_N. \end{cases}$$

To derive the AMP for the sequence of problems (D_N), we reduce these problems to the ones *without delays* and employ the results of section 4. This follows the "method of steps" developed by Warga [15] in the case of delay problems for continuous-time systems. Our assumptions on the initial data of (P) are similar to those in section 4 for nondelay systems. A counterpart of (H1) is formulated as follows:

(H) $f = f(t, x, y, u)$ is continuous with respect to all its variables and continuously differentiable with respect to (x, y) in some tube containing optimal trajectories for all u from the compact set U in a metric space and for all $t \in \widetilde{T}_N := T_N \cup \{t_1 - \widetilde{h}_N\}$ uniformly in $N \in \mathbb{N}$.

For convenience we introduce the notation

$$\begin{aligned} &\xi_N(t) := (x_N(t), x_N(t-\theta)), \quad \bar{\xi}_N(t) := (\bar{x}_N(t), \bar{x}_N(t-\theta)), \\ &f(t, \xi_N, u_N) := f(t, x_N(t), x_N(t-\theta), u_N(t)), \\ &f(t, \bar{\xi}_N, u_N) := f(t, \bar{x}_N(t), \bar{x}_N(t-\theta), u_N(t)) \end{aligned}$$

in which terms the *adjoint system* to (D_N) is written as

$$\begin{aligned} p_N(t) = p_N(t + h_N) &+ h_N \frac{\partial f}{\partial x}^*(t, \bar{\xi}_N, \bar{u}_N) p_N(t + h_N) \\ &+ h_N \frac{\partial f}{\partial y}^*(t + \theta, \bar{\xi}_N, \bar{u}_N) p_N(t + \theta + h_N) \quad \text{for } \ t \in T_N, \end{aligned}$$

$$p_N(t_1 - \widetilde{h}_N) = p_N(t_1) + \widetilde{h}_N \frac{\partial f}{\partial x}^*(t_1 - \widetilde{h}_N, \bar{\xi}_N, \bar{u}_N) p_N(t_1)$$

along the optimal processes $(\bar{x}_N, \bar{u}_N)$ to the delay problems for each $N \in \mathbb{N}$. Introducing the corresponding *Hamilton–Pontryagin function*

$$H(t, x_N, y_N, p_N, u) := \begin{cases} \langle p_N(t+h_N), f(t, x_N, y_N, u)\rangle & \text{if } t \in T_N, \\ \langle p_N(t), f(t-\widetilde{h}_N, x_N, y_N, u)\rangle & \text{if } t = t_1 - \widetilde{h}_N \end{cases} \tag{5.1}$$

with $y_N(t) = x_N(t-\theta)$, we rewrite the adjoint system as

$$\begin{cases} p_N(t) = p_N(t+h_N) + h_N\left[\dfrac{\partial H}{\partial x}(t, \bar{\xi}_N, p_N, \bar{u}_N) + \dfrac{\partial H}{\partial y}(t+\theta, \bar{\xi}_N, p_N, \bar{u}_N)\right], \ t \in T_N, \\ p_N(t_1 - \widetilde{h}_N) = p_N(t_1) + \widetilde{h}_N \dfrac{\partial H}{\partial x}(t_1 - \widetilde{h}_N, \bar{\xi}_N, p_N, \bar{u}_N). \end{cases} \tag{5.2}$$

THEOREM 5.1 (AMP for delay systems). *Let the pairs $(\bar{x}_N, \bar{u}_N)$ be optimal to problems (D_N). Assume in addition to* (H) *that φ is uniformly upper subdifferentiable around the limiting point(s) of the sequence $\{\bar{x}_N(t_1)\}$, $N \in \mathbb{N}$. Then for every sequence of upper subgradients $x_N^* \in \widehat{\partial}^+\varphi(\bar{x}_N(t_1))$, the approximate maximum condition*

$$H(t, \bar{\xi}_N, p_N, \bar{u}_N) = \max_{u \in U} H(t, \bar{\xi}_N, p_N, u) + \varepsilon(t, h_N), \quad t \in \widetilde{T}_N, \tag{5.3}$$

holds with the Hamilton–Pontryagin function (5.1) *and with some $\varepsilon(t, h_N) \to 0$ as $h_N \to 0$ uniformly in $t \in \widetilde{T}_N$, where the adjoint trajectory p_N satisfies* (5.2) *and the transversality relations*

$$p_N(t_1) = -x_N^*, \qquad p_N(t) = 0 \ \ \textit{as} \ \ t > t_1. \tag{5.4}$$

Proof. Let us reduce the delay discrete approximation problems to the ones with no delay by the following multistep procedure. Denote

$$\begin{aligned} &y_{1N}(t) := x_N(t-h_N), \quad t \in \{t_0 + 2h_N, \dots, t_1 - \widetilde{h}_N\}, \\ &y_{1N}(t) := c_N(t-h_N), \quad t \in \{t_0 - \theta + h_N, \dots, t_0 + h_N\}, \\ &y_{2N}(t) := y_{1N}(t-h_N), \quad t \in \{t_0 - \theta + 2h_N, \dots, t_1 - \widetilde{h}_N\}, \\ &\quad \vdots \\ &y_{NN}(t) := y_{N-1,N}(t-h_N), \quad t \in \{t_0, \dots, t_1 - \widetilde{h}_N\}, \end{aligned}$$

and observe that the values of $y_{1N}(t_1), \dots, y_{NN}(t_1)$ can be defined arbitrarily, since they do not enter either the adjoint system or the cost function. To match the setup of Theorem 4.1, define

$$y_{1N}(t_1) := x_N(t_1 - \widetilde{h}_N), \ \ y_{2N}(t_1) := y_{1N}(t_1 - \widetilde{h}_N), \dots, y_{NN}(t_1) := y_{N-1,N}(t_1 - \widetilde{h}_N).$$

After the change of variables one has

$$y_{NN}(t) = \begin{cases} x_N(t-\theta), & t \in \{t_0 + \theta + h_N, \dots, t_1 - \widetilde{h}_N\}, \\ c(t-\theta), & t \in \{t_0, \dots, t_0 + \theta\}. \end{cases}$$

The original system in (D_N) is thereby reduced, for each $N \in \mathbb{N}$, to the following *nondelayed* system of dimension $\mathbb{R}^{(N+1)n}$:

$$\begin{cases} z_N(t+h_N) = z_N(t) + h_N g(t, z_N, u_N), \ \ t \in T_N, \\ z_N(t_1) = z_N(t_1 - \widetilde{h}_N) + \widetilde{h}_N g(t_1 - \widetilde{h}_N, z_N, u_N), \end{cases} \tag{5.5}$$

with the state vector $z_N(t) := (x_N(t), y_{1N}(t), \dots, y_{NN}(t))^*$ (the star stands for transposing) and with the mapping $g(t, z_N, u_N)$ given by

$$g(t, z_N(t), u_N(t)) := \begin{pmatrix} f(t, x_N(t), y_{NN}(t), u_N(t)) \\ \dfrac{x_N(t) - y_{1N}(t)}{h_N} \\ \cdots\cdots \\ \dfrac{y_{N-1,N}(t) - y_{NN}(t)}{h_N} \end{pmatrix},$$

where h_N should be replaced by $\widetilde{h}_N$ for $t = t_1 - \widetilde{h}_N$ in the last formula.

Let us apply Theorem 4.1 to minimizing the same functional as in (D_N) on the feasible pairs (z_N, u_N) of the nondelayed system (5.5). The adjoint system in this problem, with respect to the new adjoint variable $q \in \mathbb{R}^{(N+1)n}$, has the form

$$\text{(5.6)} \qquad \begin{cases} q_N(t) = q_N(t + h_N) + h_N \dfrac{\partial g}{\partial z}^* (t, \bar{z}_N, \bar{u}_N) q(t + h_N), \quad t \in T_N, \\ q_N(t_1 - \widetilde{h}_N) = q_N(t_1) + \widetilde{h}_N \dfrac{\partial g}{\partial z}^* (t_1 - \widetilde{h}_N, \bar{z}_N, \bar{u}_N) q_N(t_1) \end{cases}$$

with the transversality condition

$$\text{(5.7)} \qquad q_N(t_1) = -(x_N^*, 0, \dots, 0)^* \ \text{ for } \ x_N^* \in \widehat{\partial}^+ \varphi(\bar{x}_N(t_1)).$$

Computing the matrix $\frac{\partial g}{\partial z}$, we get

$$\frac{\partial g^*}{\partial z} = \frac{1}{h_N} \begin{pmatrix} h_N \dfrac{\partial f^*}{\partial x} & 1 & \dots & 0 & 0 \\ 0 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ h_N \dfrac{\partial f^*}{\partial y_{NN}} & 0 & \dots & 0 & -1 \end{pmatrix}.$$

Taking this into account and performing elementary calculations, we arrive at the adjoint system (5.2) and the transversality relations (5.4) for the first component $p_N(t)$ of the adjoint trajectory $q_N(t)$ satisfying (5.6) and (5.7). Denoting by $\widetilde{H}(t, z_N, q_N, u)$ the Hamilton–Pontryagin function (5.1) to the nondelayed system (5.5), one has

$$\begin{aligned} \widetilde{H}(t, \bar{z}_N, q_N, u) &= \big\langle q_N(t + h_N), g(t, \bar{z}_N, u) \big\rangle \\ &= \big\langle p_N(t + h_N), f(t, \bar{\xi}_N, u) \big\rangle + r(t, \bar{z}_N, q_N, h_N) \\ &= H(t, \bar{\xi}_N, p_N, u) + r(t, \bar{z}_N, q_N, h_N), \quad t \in T_N, \end{aligned}$$

and similarly for $t = t_1 - \widetilde{h}_N$, where H is given in (5.1), and where the remainder $r(t, \bar{z}_N, q_N, h_N)$ does not depend on u. Finally applying the approximate maximum condition (1.3) from Theorem 4.1 to system (5.5), we arrive at (5.3) and complete the proof of the theorem. □

Note that in the case of continuously differentiable cost functions φ around $\bar{x}_N(t_1)$ uniformly in N, the transversality relations (5.4) reduce to

$$p_N(t_1) = -\nabla \varphi(\bar{x}_N(t_1)), \qquad p_N(t) = 0 \ \text{ as } \ t > t_1.$$

Similarly to the proof of Theorem 5.1 we can deduce from Theorem 4.3 its delay counterpart for discrete approximation problems with *endpoint constraints*. In this result we add assumptions (H2) and (H3) to those in (H) and replace the transversality relations (5.4) in Theorem 5.1 with the conditions (4.10)–(4.12) with $p_N(t) = 0$ as $t > t_1$.

Finally in this section, we consider optimal control problems for finite-difference approximations of the so-called *functional-differential systems of neutral type*

$$\dot{x}(t) = f(t, x(t), x(t-\theta), \dot{x}(t-\theta), u(t)) \ \text{ a.e. } \ t \in [t_0, t_1], \tag{5.8}$$

which contain time delays not only in state but also in *velocity* variables. A finite-difference counterpart of (5.8) with the stepsize h and with the grid $T := \{t_0, t_0 + h, \dots, t_1 - h\}$ is

$$x(t+h) = x(t) + hf\left(t, x(t), x(t-\theta), \frac{x(t-\theta+h) - x(t-\theta)}{h}, u(t)\right), \quad t \in T, \tag{5.9}$$

and the adjoint system is given by

$$\begin{aligned} p(t) = p(t+h) &+ h\frac{\partial f}{\partial x}^*(t, \bar{\xi}, \bar{u})p(t+h) + h\frac{\partial f}{\partial y}^*(t+\theta, \bar{\xi}, \bar{u})p(t+\theta+h) \\ &+ h\frac{\partial f}{\partial z}^*(t+\theta-h, \bar{\xi}, \bar{u})p(t+\theta) - h\frac{\partial f}{\partial z}^*(t+\theta, \bar{\xi}, \bar{u})p(t+\theta+h), \quad t \in T, \end{aligned} \tag{5.10}$$

where $(\bar{x}, \bar{u})$ is an optimal solution to the neutral analogue of problems (D_N) and where

$$\bar{\xi}(t) := \left(\bar{x}(t), \bar{x}(t-\theta), \frac{\bar{x}(t-\theta+h) - \bar{x}(t-\theta)}{h}\right), \quad t \in T.$$

It has been proved in [8] that optimal solutions to problems like (D_N) for discrete systems of the neutral type (5.9) satisfy the *exact discrete maximum principle* with transversality conditions in the *upper subdifferential form* provided that the velocity sets $f(t, x, y, z, U)$ are *convex* around $\bar{\xi}(t)$. What about an analogue of the AMP with no convexity assumptions on the velocity sets? The following example shows that the AMP is *not fulfilled* for finite-difference neutral systems, in contrast to ordinary and delay ones, even in the case of *smooth* cost functions.

Example 5.1 (AMP does not hold for neutral systems). *There is a two-dimensional control problem of minimizing a linear function over a smooth neutral system with no endpoint constraints such that some sequence of optimal controls to discrete approximations does not satisfy the AMP regardless of the stepsize and a mesh point.*

Proof. Consider the following parametric family of discrete optimal control problems with the parameter $h > 0$:

$$\begin{cases} \text{minimize } J(x_1, x_2, u) := x_2(2) \\ \text{subject to} \\ x_1(t+h) = x_1(t) + hu(t), \ \ t \in T := \{0, h, \dots, 2-h\}, \\ x_2(t+h) = x_2(t) + h\left(\dfrac{x_1(t-1+h) - x_1(t-1)}{h}\right)^2 - hu^2(t), \ \ t \in T, \\ x_1(t) \equiv x_2(t) \equiv 0, \ \ t \in T_0 := \{-1, \dots, 0\}, \\ |u(t)| \le 1, \ \ t \in T. \end{cases} \tag{5.11}$$

It is easy to see that

$$x_2(1) = -h\sum_{t=0}^{1-h} u^2(t)$$

and

$$\begin{aligned} J(x_1,x_2,u) := x_2(2) &= x_2(1) + h\sum_{t=1}^{2-h}\left(\frac{x_1(t-1+h)-x_1(t-1)}{h}\right)^2 - h\sum_{t=1}^{2-h} u^2(t) \\ &= -h\sum_{t=0}^{1-h} u^2(t) + h\sum_{t=0}^{1-h} u^2(t) - h\sum_{t=1}^{2-h} u^2(t) = -h\sum_{t=1}^{2-h} u^2(t). \end{aligned}$$

Thus the control

$$\bar{u}(t) = \begin{cases} 0, & t \in \{0,\dots,1-h\}, \\ 1, & t \in \{1,\dots,2-h\}, \end{cases}$$

is the only optimal control to (5.11) for any h. The corresponding trajectory is

$$\bar{x}_1(t) = \begin{cases} 0, & t \in \{0,\dots,1-h\}, \\ t-1, & t \in \{1,\dots,2-h\}; \end{cases} \quad \bar{x}_2(t) = \begin{cases} 0, & t \in \{0,\dots,1-h\}, \\ -t+1, & t \in \{1,\dots,2-h\}. \end{cases}$$

Computing the partial derivatives of f in (5.11), we get

$$\frac{\partial f}{\partial x} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \frac{\partial f}{\partial y} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

and

$$\frac{\partial f}{\partial z}(t+1) = \frac{1}{h}\begin{pmatrix} 0 & 0 \\ 2(x_1(t+h)-x_1(t)) & 0 \end{pmatrix}.$$

Hence the adjoint system (5.10) reduces to

$$\begin{aligned} &p_1(t) = p_1(t+h) + 2(\bar{x}_1(t)-\bar{x}_1(t-h))p_2(t+1) - 2(\bar{x}_1(t+h)-\bar{x}_1(t))p_2(t+1+h), \\ &p_2(t) \equiv \text{const}, \quad t \in \{0,\dots,2-h\}, \end{aligned}$$

with the transversality conditions

$$p_1(2) = 0, \;\; p_2(2) = -1; \quad p_1(t) = p_2(t) = 0 \;\text{ for }\; t > 2.$$

The solution of this system is

$$p_1(t) \equiv 0, \;\; p_2(t) \equiv -1 \;\text{ for all }\; t \in \{0,\dots,2-h\}.$$

Thus the Hamilton–Pontryagin function along the optimal solution is

$$\begin{aligned} H(t,\bar{x}_1,\bar{x}_2,p_1,p_2,u) &= p_1(t+h)u + p_2(t+h)\left\{\left(\frac{x_1(t-1+h)-x_1(t-1)}{h}\right)^2 - u^2\right\} \\ &= u^2 \;\text{ for all }\; t \in \{0,\dots,1-h\}. \end{aligned}$$

This shows that the optimal control $\bar{u}(t) = 0$ does not provide the approximate maximum to the Hamilton–Pontryagin function regardless of h and the mesh point $t \in \{0,\dots,1-h\}$. Note at the same time that another sequence of optimal controls with $\bar{u}(t) = 1$ for all $t \in \{0,\dots,2-h\}$ satisfies the exact discrete maximum principle regardless of h. □

REFERENCES

[1] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[2] R. Gabasov and F. M. Kirillova, *On the extension of the maximum principle by L. S. Pontryagin to discrete systems,* Avtomat. i Telemekh., 27 (1966), pp. 46–61 (in Russian); Autom. Remote Control, 27 (1966), pp. 1878–1882 (in English).

[3] R. Gabasov and F. M. Kirillova, *Qualitative Theory of Optimal Processes,* Marcel Dekker, New York, 1976.

[4] B. S. Mordukhovich, *Approximate maximum principle for finite-difference control systems,* Comput. Math. Math. Phys., 28 (1988), pp. 106–114.

[5] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian).

[6] B. S. Mordukhovich, *Discrete approximations and refined Euler–Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.

[7] B. S. Mordukhovich, *Necessary conditions in nonsmooth minimization via lower and upper subgradients*, Set-Valued Anal., 12 (2004), pp. 163–193.

[8] B. S. Mordukhovich and I. Shvartsman, *Discrete maximum principle for nonsmooth optimal control problems with delays*, Cybernet. Systems Anal., 38 (2002), pp. 255–264.

[9] B. S. Mordukhovich and I. Shvartsman, *The approximate maximum principle for constrained control systems*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 4345–4350.

[10] E. A. Nurminskii, *Numerical Methods of Solutions to Deterministic and Stochastic Minimax Problems*, Naukova Dumka, Kiev, 1979 (in Russian).

[11] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Wiley, New York, 1962.

[12] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[13] G. V. Smirnov, *Introduction to the Theory of Differential Inclusions*, AMS, Providence, RI, 2002.

[14] R. B. Vinter, *Optimal Control*, Birkhäuser Boston, Boston, MA, 2000.

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

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 1063–1077

ELIMINATION OF STRICT CONVERGENCE IN OPTIMIZATION*

DUŠAN BEDNAŘÍK† AND KAREL PASTOR‡

Abstract. We present second-order optimality conditions (sufficient and necessary) and the characterization of convexity in terms of generalized second-order derivatives. Some conclusions obtained by R. Cominetti and R. Correa are proved under weakened assumptions.

Key words. generalized second-order directional derivative, second-order optimality conditions, convex function

AMS subject classifications. 49K35, 26B02, 26B25

DOI. 10.1137/S0363012903424174

1. Introduction and preliminaries. Second-order optimality conditions in terms of the generalized second-order derivatives are very useful in various problems connected with optimization. See, for instance, [BZ], [CC], [CHN], [GZ], [HSN], [K], [Ma], [Q], [R1], [R2], [RW], [Y1], and references therein.

Much attention has been focused on $C^{1,1}$ functions (i.e., continuously differentiable functions with a locally Lipschitz gradient) which appear in, e.g., the augmented Lagrange method, the penalty function method, and the proximal point method.

Second-order optimality conditions (both sufficient and necessary) can be classified into two groups. One type assumes the existence of some second-order derivatives (which are defined, for example, as lim); see, e.g., the classical result in smooth analysis or the generalized results given in [BZ], [JY]. The other type uses the second-order derivatives which exist always (they are defined, for example, as lim inf or lim sup) [CC], [YJ], [Pa2]. On the contrary, the conditions of the first type are often tighter than those of the second. These advantages and disadvantages are compared, for example, in [Y1, section 5].

At this point we should say that we restrict our considerations to the second type of second-order optimality conditions.

By the process of elimination of strict convergence given in the title of this paper, we mean that some results (see [CC] and notice that it has been cited 37 times so far by Web of Science) presented in terms of Cominetti–Correa generalized second-order directional derivatives can be given in tighter form in such a way that, from the formal point of view, we wipe out the strict convergence $y \to x$ in second-order conditions expressed by means of Cominetti–Correa derivatives (compare Theorems 3.1 and 3.2, Theorems 5.1 and 5.3, Theorems 6.1 and 6.4, and Theorems 7.1 and 7.5).

This process was started in [Pa2] for second-order sufficient optimality conditions. In section 3, we present a new, more precise proof of this improved result.

Connections among generalized second-order directional derivatives introduced by Cominetti–Correa and those introduced in [Pa1] are established in section 4.

*Received by the editors March 10, 2003; accepted for publication (in revised form) January 3, 2004; published electronically November 9, 2004.
http://www.siam.org/journals/sicon/43-3/42417.html
†Department of Mathematics, University of Hradec Králové, Víta Nejedlého 573, 500 03 Hradec Králové, Czech Republic (dusan.bednarik@uhk.cz).
‡Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Tř. Svobody 26, 771 46 Olomouc, Czech Republic (pastor@inf.upol.cz).

Section 5 is devoted to the characterization of convexity via generalized second-order derivatives (we note that convexity plays an important role not only in optimization theory).

A new (to the best of our knowledge) necessary second-order optimality condition introduced in section 6 is tighter than the ones in [YJ]. We finish our elimination process by studying certain minimax problems in section 7.

2. Survey of derivatives. Let X be a real Banach space with the norm $\|\cdot\|$, with X^* the topological dual space of X, and let $\langle x^*, x\rangle$ be a canonical pair, where $x^* \in X^*$, $x \in X$. S_X denotes the set $\{x \in X : \|x\| = 1\}$.

In fact, our results use only $f'^{l}_{+}(x;u,v)$ and $f'^{u}_{+}(x;u,v)$ (see below for definitions). Since we want to compare our results with those obtained before in terms of other generalized second-order directional derivatives, we give a survey of their definitions for the sake of completeness. On the other hand, there is maybe a pitfall that one might be caught in the mire of definitions. Therefore it might be helpful to skim section 2 during the first reading, then study it more carefully during the next.

Let us note, also on account of applications, that the calculus with $f'^{l}_{+}(x;u,v)$ and $f'^{u}_{+}(x;u,v)$ feels more comfortable than the calculus with some other generalized second-order directional derivatives.

For a function $f : X \to \mathbb{R}$, we denote the first-order directional derivative of f at $x \in X$ in the direction $v \in X$ by

$$f'(x;v) = \lim_{t\downarrow 0} \frac{f(x+tv) - f(x)}{t},$$

and the upper and lower Clarke directional derivatives of f at $x \in X$ in the direction $v \in X$, respectively, by

$$f^{\circ}(x;v) = \limsup_{y\to x, t\downarrow 0} \frac{f(y+tv) - f(y)}{t},$$

$$f_{\circ}(x;v) = \liminf_{y\to x, t\downarrow 0} \frac{f(y+tv) - f(y)}{t}.$$

As in [CC], we define the Cominetti–Correa upper and lower generalized second-order directional derivatives at $x \in X$ in the direction $(u,v) \in X \times X$ by, respectively,

$$f^{\infty}(x;u,v) = \limsup_{y\to x,\ s,t\downarrow 0} \frac{f(y+tu+sv) - f(y+tu) - f(y+sv) + f(y)}{st},$$

$$f_{\infty}(x;u,v) = \liminf_{y\to x,\ s,t\downarrow 0} \frac{f(y+tu+sv) - f(y+tu) - f(y+sv) + f(y)}{st}.$$

Considering a $C^{1,1}$ function $f : X \to \mathbb{R}$, the generalized upper and lower second-order directional derivatives of f at $x \in X$ in direction $(u,v) \in X \times X$ in the sense of Michel–Penot [MP] are given, respectively, as follows:

$$f^{\diamond u}(x;u,v) = \sup_{z\in X} \limsup_{t\downarrow 0} \frac{\langle \bigtriangledown f(x+tz+tu) - \bigtriangledown f(x+tz), v\rangle}{t},$$

$$f^{\diamond l}(x;u,v) = \inf_{z\in X} \liminf_{t\downarrow 0} \frac{\langle \bigtriangledown f(x+tz+tu) - \bigtriangledown f(x+tz), v\rangle}{t},$$

where the symbol $\bigtriangledown f(x)$ denotes the Gâteaux derivative of f at x.

We note that for a set-valued mapping $F : \mathbb{R} \rightsquigarrow \mathbb{R}$, it is natural to denote

$$\liminf_{t\downarrow 0} F(t) = \liminf_{t\downarrow 0} \inf\{a : a \in F(t)\}.$$

Similarly for lim sup. In [Pa1], there were introduced the following lower generalized second-order directional derivatives:

$$f'^{L}(x;u,v) = \liminf_{t\downarrow 0} \frac{\langle \partial_c f(x+tu) - \partial_c f(x), v\rangle}{t},$$

$$f^{*L}(x;u,v) = \liminf_{y\to x, t\downarrow 0} \frac{\langle \partial_c f(y+tu) - \partial_c f(y), v\rangle}{t},$$

where f is Lipschitz near $x \in X$, $(u,v) \in X \times X$, and the symbol $\partial_c f(x)$ stands for the Clarke subdifferential [Cl1] at x defined by

$$\partial_c f(x) = \{x^* \in X^* : \langle x^*, v\rangle \le f^\circ(x;v) \quad \forall v \in X\}.$$

One has that it holds

$$f^\circ(x;v) = \max\{\langle x^*, v\rangle : x^* \in \partial_c f(x)\},$$

$$f_\circ(x;v) = \min\{\langle x^*, v\rangle : x^* \in \partial_c f(x)\}.$$

Dually to $f'^{L}(x;u,v)$, we can define

$$f'^{U}(x;u,v) = \limsup_{t\downarrow 0} \frac{\langle \partial_c f(x+tu) - \partial_c f(x), v\rangle}{t}.$$

Let us recall that a locally Lipschitz function $f : X \to \mathbb{R}$ is said to be regular at $x \in X$ provided that $f'(x;v) = f^\circ(x;v)$ for every $v \in X$. It is natural to define the following upper and lower generalized derivatives at $x \in X$ in the direction $(u,v) \in X \times X$ for a function $f : X \to \mathbb{R}$ which is regularly locally Lipschitz near x by means of

$$f'^{u}(x;u,v) = \limsup_{t\downarrow 0} \frac{f'(x+tu;v) - f'(x;v)}{t}$$

and

$$f'^{l}(x;u,v) = \liminf_{t\downarrow 0} \frac{f'(x+tu;v) - f'(x;v)}{t}.$$

Finishing the survey of generalized derivatives used in the paper, we recall some of their properties and the relationships among them.

If f is regularly locally Lipschitz near $x \in X$, and $u, v \in X$, then it holds that [CC]

$$f^\infty(x;u,v) = \limsup_{y\to x, t\downarrow 0} \frac{f'(y+tu;v) - f'(y;v)}{t}, \tag{1}$$

$$f_\infty(x;u,v)=\liminf_{y\to x,t\downarrow 0}\frac{f'(y+tu;v)-f'(y;v)}{t}. \tag{2}$$

Thus, if f is continuously differentiable near x, then

$$f^\infty(x;u,v)=\limsup_{y\to x,t\downarrow 0}\frac{\langle\bigtriangledown f(y+tu)-\bigtriangledown f(y),v\rangle}{t},$$

$$f_\infty(x;u,v)=\liminf_{y\to x,t\downarrow 0}\frac{\langle\bigtriangledown f(y+tu)-\bigtriangledown f(y),v\rangle}{t}. \tag{3}$$

Further, for each continuously differentiable function f near $x\in X$, one has

$$f'^u(x;u,v)=\limsup_{t\downarrow 0}\frac{\langle\bigtriangledown f(x+tu)-\bigtriangledown f(x),v\rangle}{t}$$

and

$$f'^l(x;u,v)=\liminf_{t\downarrow 0}\frac{\langle\bigtriangledown f(x+tu)-\bigtriangledown f(x),v\rangle}{t}. \tag{4}$$

LEMMA 2.1 (see [Pa1, Proposition 2.1]). *Let $f:X\to\mathbb{R}$ be a $C^{1,1}$ function, $x\in X$, and let $(u,v)\in X\times X$. Then*

$$f'^L(x;u,v)\geq f^{\diamond l}(x;u,v)\geq f^{*L}(x;u,v).$$

LEMMA 2.2 (see [Pa1, Proposition 3.1]). *Let $f:X\to\mathbb{R}$ be a continuously differentiable function, $x,u,v\in X$. Then*

$$f_\infty(x;u,v)\leq f'^L(x;u,v).$$

3. Sufficient optimality conditions. Cominetti and Correa published the following result in 1990.

THEOREM 3.1 (see [CC, Proposition 5.2]). *Let $f:\mathbb{R}^n\to\mathbb{R}$ be a $C^{1,1}$ function near $\hat{x}\in\mathbb{R}^n$. If $\bigtriangledown f(\hat{x})=0$ and it holds that*

$$f_\infty(\hat{x};h,h)>0\quad\forall h\in S_{\mathbb{R}^n}, \tag{5}$$

then f attains a strict local minimum at $\hat{x}$.

As was shown in [Pa2, Theorem 3], the conclusion of Theorem 3.1 does not change when we wipe out the strict convergence in (5), as shown in the following. (Compare also with (3) and (4).)

THEOREM 3.2. *Let $f:\mathbb{R}^n\to\mathbb{R}$ be a $C^{1,1}$ function near $\hat{x}\in\mathbb{R}^n$. If $\bigtriangledown f(\hat{x})=0$ and it holds that*

$$f'^l(\hat{x};h,h)>0\quad\forall h\in S_{\mathbb{R}^n}, \tag{6}$$

then f attains a strict local minimum at $\hat{x}$.

The following example illustrates the advantage of Theorem 3.2.

Example 3.3. Consider a function $f:\mathbb{R}\to\mathbb{R}$ such that

$$f(x)=\begin{cases}\int_0^{\|x\|}t(\frac{5}{4}+\sin\ln t)dt & \text{if } x\neq 0,\\ 0 & \text{if } x=0.\end{cases}$$

It easy to show that f is a $C^{1,1}$ function, $\bigtriangledown f(0) = 0$, and $f'^l(0;1,1) = f'^l(0;-1,-1) = \frac{1}{4}$. On the other hand, $f_\infty(0;1,1) = \frac{5}{4} - \sqrt{2} < 0$. For the calculus, see [BP, Example]. So, Theorem 3.2 can be used to test strict local minimum at 0 in contrast to Theorem 3.1.

Now, we provide a new proof of Theorem 3.2 which appears to be more precise than those given in [Pa2]. Analogously, as in [DR, p. 229], we can prove Lemma 3.4.

LEMMA 3.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $\hat{x} \in \mathbb{R}^n$, and let*

$$\liminf_{t \downarrow 0} \frac{f(\hat{x}+th) - f(\hat{x})}{t} > 0 \quad \forall h \in S_{\mathbb{R}^n}.$$

Then f attains a strict local minimum at $\hat{x}$.

LEMMA 3.5. *Let $T : X \to X^*$ be Lipschitz near 0 and $T(0) = 0$. Then a function $g : X \to \mathbb{R}$,*

$$g(x) = \begin{cases} \left\langle Tx, \dfrac{x}{\|x\|} \right\rangle & \text{for } x \neq 0, \\ 0 & \text{for } x = 0, \end{cases}$$

is also Lipschitz near 0.

Proof. There exist $\delta > 0$ and $L > 0$ satisfying

$$\|Tx - Ty\| \leq L\|x - y\| \quad \forall x, y \in B(0,\delta),$$

where $B(x,r) = \{y \in X : \|y - x\| \leq r\}$ for an $x \in X$ and $r > 0$. We fix $x, y \in B(0,\delta)$. If $x \neq 0$ and $y \neq 0$, then let us calculate

$$\begin{aligned}
\|g(x) - g(y)\| &= \left\| \left\langle Tx, \frac{x}{\|x\|} \right\rangle - \left\langle Ty, \frac{y}{\|y\|} \right\rangle \right\| \\
&= \left\| \left\langle Tx, \frac{x}{\|x\|} \right\rangle - \left\langle Ty, \frac{x}{\|x\|} \right\rangle + \left\langle Ty, \frac{x}{\|x\|} \right\rangle - \left\langle Ty, \frac{y}{\|y\|} \right\rangle \right\| \\
&\leq \|Tx - Ty\| + \|Ty\| \left\| \frac{x}{\|x\|} - \frac{y}{\|y\|} \right\| \\
&\leq L\|x - y\| + L\|y\| \left\| \frac{x\|y\| - x\|x\| + x\|x\| - y\|x\|}{\|x\|\|y\|} \right\| \\
&= L\|x - y\| + L \left\| \frac{(\|y\| - \|x\|)x + \|x\|(x - y)}{\|x\|} \right\| \\
&\leq L\|x - y\| + L \left\| \frac{(\|y\| - \|x\|)}{\|x\|} x \right\| + L\|x - y\| \\
&\leq 3L\|x - y\|.
\end{aligned}$$

In the other cases (i.e., $x = 0, y \neq 0$ or $x = 0, y = 0$), the above formula $\|g(x) - g(y)\| \leq 3L\|x - y\|$ follows immediately. ∎

Proof of Theorem 3.2. Without any loss of generality, we can assume that $\hat{x} = 0$ and $f(0) = 0$.

Since $\bigtriangledown f(0) = 0$, one can rewrite condition (6) as

$$\liminf_{t \downarrow 0} \frac{\langle \bigtriangledown f(th), h \rangle}{t} > 0 \quad \forall h \in S_{\mathbb{R}^n}. \tag{7}$$

Let us consider a function $g : X \to \mathbb{R}$ such that

$$g(x) = \begin{cases} \left\langle \bigtriangledown f(x), \dfrac{x}{\|x\|} \right\rangle & \text{for } x \neq 0, \\ 0 & \text{for } x = 0. \end{cases}$$

The function $x \to \bigtriangledown f(x)$ is Lipschitz near 0. Subsequently, g is also Lipschitz near 0 by Lemma 3.5. Using condition (7), we have

$$\liminf_{t \downarrow 0} \frac{g(0+th) - g(0)}{t} = \liminf_{t \downarrow 0} \frac{\langle \bigtriangledown f(th), h \rangle}{t} > 0 \quad \forall h \in S_{\mathbb{R}^n}.$$

Due to Lemma 3.4, g attains a strict local minimum at 0, i.e., there exists $\delta > 0$ such that

$$\left\langle \bigtriangledown f(x), \frac{x}{\|x\|} \right\rangle > 0 \tag{8}$$

for every $x \in B(0, \delta)$. We fix $x \in B(0, \delta)$, $x \neq 0$. Applying the Lagrange mean value theorem, we get $\eta \in (0, 1)$ satisfying

$$f(x) - f(0) = \langle \bigtriangledown f(\eta x), x \rangle = \left\langle \bigtriangledown f(\eta x), \frac{\eta x}{\|\eta x\|} \right\rangle \|x\| > 0,$$

where the previous inequality follows from (8). Therefore, f attains a strict local minimum at 0. □

Unfortunately, as shown by the following counterexample, Theorem 3.2 is not true for an arbitrary continuously differentiable function.

Example 3.6. Define $f : \mathbb{R}^2 \to \mathbb{R}$, partially in polar coordinates, as shown by the following:

1. $f(x, y) = (y - x^{\frac{3}{2}})^2$ if $(x, y) \in \left\{(x, y) \in \mathbb{R}^2, x \geq 0, x^{\frac{3}{2}} \leq y \leq x^{\frac{3}{2}} + 1\right\}$;
2. $f(x, y) = 1$ if $(x, y) \in \left\{(x, y) \in \mathbb{R}^2, x \geq 0, x^{\frac{3}{2}} + 1 \leq y\right\}$;
3. $f(r \sin\varphi, r \cos\varphi) = \hat{f}(r, \varphi) = \left(r - \frac{\sin^2\varphi}{\cos^3\varphi}\right)^2$ if $\varphi \in \left(0, \frac{\pi}{2}\right)$ and $r > \frac{\sin^2\varphi}{\cos^3\varphi}$;
4. $f(r \sin\varphi, r \cos\varphi) = \hat{f}(r, \varphi) = r^2$ if $\varphi \in \left[\frac{3}{2}\pi, 2\pi\right]$;
5. $f(-x, y) = f(x, y)$.

For the main parts of the domain of f, see Figure 3.1.

First, we show that f is continuously differentiable near 0. Due to the symmetry of f, it suffices to calculate for the following subsets of the domain of f:

1. $M_1 = \left\{(x, y) \in \mathbb{R}^2, x \geq 0, x^{\frac{3}{2}} < y < x^{\frac{3}{2}} + 1\right\}$. Then

$$\bigtriangledown f(x, y) = \left[-3x^{\frac{1}{2}}\left(y - x^{\frac{3}{2}}\right), 2\left(y - x^{\frac{3}{2}}\right)\right]. \tag{9}$$

2. $M_2 = \left\{(r \cos\varphi, r \sin\varphi) \in \mathbb{R}^2, \varphi \in (0, \frac{\pi}{2}), r > \frac{\sin^2\varphi}{\cos^3\varphi}\right\}$. Then

$$\bigtriangledown \hat{f}(r, \varphi) = \left[2\left(r - \frac{\sin^2\varphi}{\cos^3\varphi}\right), -2\left(r - \frac{\sin^2\varphi}{\cos^3\varphi}\right)\frac{2\sin\varphi\cos^2\varphi + 3\sin^3\varphi}{\cos^4\varphi}\right].$$

3. $M_3 = \left\{(r \cos\varphi, r \sin\varphi) \in \mathbb{R}^2, \varphi \in \left[\frac{3}{2}\pi, 2\pi\right], r > 0\right\}$. Then

$$\bigtriangledown \hat{f}(r, \varphi) = [2r, 0].$$

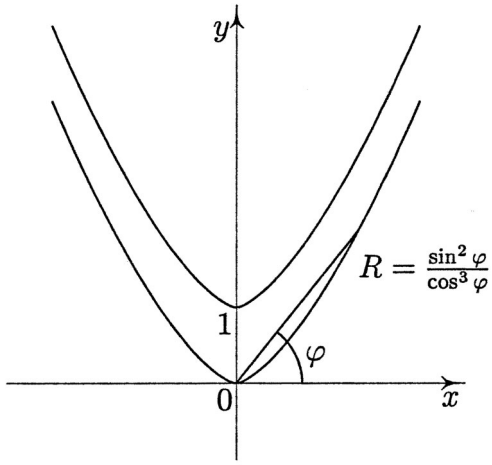


FIG. 3.1. *The division of the domain in Example* 3.6.

4. $M_4 = \{(x, y) \in \mathbb{R}^2, x \geq 0, y = x^{\frac{3}{2}}\}$. Then

$$\bigtriangledown f(x, y) = (0, 0).$$

Analyzing the previous calculation, one has that f is continuously differentiable near 0 (but is not a $C^{1,1}$ function near 0).

Now let us verify that $f'^{l}_{+}(0)(h, h) > 0$ whenever $h = (\cos\varphi, \sin\varphi)$ for $\varphi \in (0, 2\pi]$. Setting $x = r\cos\varphi, y = r\sin\varphi$ in (9) and using symmetry again, we obtain

$$f'^{l}_{+}(0)(h, h) = 2\sin^2\varphi > 0$$

whenever $h = (\cos\varphi, \sin\varphi)$ for $\varphi \in (0, \pi)$.

Finally, for $h = (\cos\varphi, \sin\varphi), \varphi \in [\pi, 2\pi]$,

$$f'^{l}_{+}(0)(h, h) = 2.$$

On the other hand, for every $(x, y) \in \left\{(x, y) \in \mathbb{R}^2, x \geq 0, y = x^{\frac{3}{2}}\right\}$, it holds that $f(x, y) = 0$. It implies that the function f does not attain a strict local minimum at 0.

4. Relations among derivatives. In order to generalize Lemma 2.2, we will employ the notion of minimal cusco mapping.

A set-valued mapping F from a topological space A into nonempty subsets of a Hausdorff locally convex space Y is called *cusco* if $F(t)$ is compact and convex for each $t \in A$ and F is upper-semicontinuous (i.e., $\{t \in A : F(t) \subset U\}$ is an open subset for each open subset U of Y).

Furthermore, F is said to be *minimal cusco* on A if its graph does not contain the graph of any other cusco on A. The Clarke subdifferential of a regularly locally Lipschitz function [Mo] or maximal monotone operator (see, e.g., [Ph]) are the well-known examples of minimal cusco mappings. We recall several basic facts concerning this notion.

As a particular case of a more general result given in [J, Corollary 4.1], the following lemma holds.

LEMMA 4.1. *Let $D \subset \mathbb{R}$ be an open set, and let $F : D \rightsquigarrow \mathbb{R}$ be minimal cusco. Then the set $\{x \in D : F$ is single-valued at $x\}$ is dense in D.*

We note that $f : A \to \mathbb{R} \cup \{+\infty\}$ (A is a topological space) is a lower-semicontinuous function provided that $\{x \in A : f(x) \leq r\}$ is closed in A for every $r \in \mathbb{R}$. By a lower hull of a densely defined function f from A into $\mathbb{R} \cup \{+\infty\}$, we mean

$$l(f) = \max\{h : A \to \mathbb{R} \cup \{+\infty\} : (h \leq \bar{f}) \text{ and } h \text{ is lower-semicontinuous}\},$$

where $\bar{f} : A \to \mathbb{R} \cup \{+\infty\}$ coincides with f on the domain of f and $\bar{f}(x) = +\infty$ otherwise.

LEMMA 4.2 (see [PB, section 2]). *Let A be a topological space and let $F : A \rightsquigarrow \mathbb{R}$ be minimal cusco. Then for arbitrary densely defined selections s and s', it holds that*

$$l(s) = l(s').$$

For more details about subdifferential mapping $x \rightsquigarrow \partial_c f(x)$ in terms of minimal cusco, see, e.g., [BM].

Now, we state the relationships among some generalized second-order directional derivatives mentioned in section 2.

By the analogous way as in [Cl2, the lemma near Theorem 2.3.7], we can obtain the following.

LEMMA 4.3. *Let $f : X \to \mathbb{R}$ be a regularly locally Lipschitz function on an open subset containing the line segment $[x, y]$. Then the function $g : [0, 1] \to \mathbb{R}$ defined by $g(t) := f(x + t(y - x))$ is regularly locally Lipschitz on $(0, 1)$, and we have*

$$\partial_c g(t) = \langle \partial_c f(x + t(y - x)), y - x \rangle.$$

PROPOSITION 4.4. *Let $f : X \to \mathbb{R}$ be a regularly locally Lipschitz function, $x, h \in X$. Then*

$$f'^L(x; h, h) = f'^l(x; h, h). \tag{10}$$

Proof. In order to prove (10), we want to show that

$$\liminf_{t \downarrow 0} \frac{f_\circ(x + th; h) - f^\circ(x; h)}{t} \geq \liminf_{t \downarrow 0} \frac{f^\circ(x + th; h) - f^\circ(x; h)}{t}. \tag{11}$$

Recall that by Lemma 4.3 the set-valued mapping $F : t \rightsquigarrow \langle \partial_c f(x + th), h \rangle$, $t \in \mathbb{R}$, is minimal cusco. Further on, we will use the notation

$$F_\circ(t) := \min\langle \partial_c f(x + th), h \rangle = f_\circ(x + th; h),$$

$$F^\circ(t) := \max\langle \partial_c f(x + th), h \rangle = f^\circ(x + th; h), \quad t \in \mathbb{R}.$$

Assuming that $f^\circ(x; h) = 0$, we can rewrite (11) as the following:

$$\liminf_{t \downarrow 0} \frac{F_\circ(t)}{t} \geq \liminf_{t \downarrow 0} \frac{F^\circ(t)}{t}. \tag{12}$$

Let us suppose for now that we have proved the following claim.

CLAIM. *If $0 < \varepsilon < t < 1$, then there exists a sequence $\{t_n\}_{n=1}^\infty$ such that $t_n \to t$ and*

$$\frac{F^\circ(t_n)}{t_n} \leq \frac{F_\circ(t)}{t} + \frac{\varepsilon}{t} \quad \text{for every } n \in \mathbb{N}.$$

Now, let $\{t_n\}_{n=1}^{\infty}$ be a sequence such that $t_n \downarrow 0$, $0 < t_n^2 < t_n < 1$ and which realizes the lim inf on the left-hand side of (12), i.e.,

$$\lim_{n\to\infty} \frac{F_\circ(t_n)}{t_n} = \liminf_{t\downarrow 0} \frac{F_\circ(t)}{t}.$$

Due to the claim above, for each $n \in \mathbb{N}$, there exists a sequence $\{t_k^n\}_{k=1}^{\infty}$ such that $\lim_{k\to\infty} t_k^n = t_n$ and such that, for each $k \in \mathbb{N}$,

$$\frac{F^\circ(t_k^n)}{t_k^n} \le \frac{F_\circ(t_n)}{t_n} + \frac{t_n^2}{t_n} = \frac{F_\circ(t_n)}{t_n} + t_n.$$

Now, for each $n \in \mathbb{N}$, we can find $k_n \in \mathbb{N}$ such that $t_{k_n}^n \downarrow 0$ as $n \to \infty$, and

$$\frac{F^\circ(t_{k_n}^n)}{t_{k_n}^n} \le \frac{F_\circ(t_n)}{t_n} + t_n \quad \forall n \in \mathbb{N}. \tag{13}$$

Letting $n \to \infty$ in (13), we obtain

$$\liminf_{t\downarrow 0} \frac{F^\circ(t)}{t} \le \liminf_{n\to\infty} \frac{F^\circ(t_{k_n}^n)}{t_{k_n}^n} \le \lim_{n\to\infty} \left(\frac{F_\circ(t_n)}{t_n} + t_n \right) = \liminf_{t\downarrow 0} \frac{F_\circ(t)}{t}.$$

Proof of the claim. Suppose, on the contrary, that there are $0 < \varepsilon < t < 1$ and $\delta > 0$ such that

$$\frac{F^\circ(t')}{t'} > \frac{F_\circ(t)}{t} + \frac{\varepsilon}{t} \quad \text{for every } t' > 0 \quad \text{such that } \|t' - t\| < \delta. \tag{14}$$

Obviously there is $\delta' > 0$ such that $0 < \delta' \le \delta$, and for any $t' > 0$, $\|t' - t\| < \delta'$ we have by (14)

$$F^\circ(t') > (F_\circ(t) + \varepsilon)\frac{t'}{t} \ge (F_\circ(t) + \varepsilon) - \frac{\varepsilon}{2} = F_\circ(t) + \frac{\varepsilon}{2}. \tag{15}$$

Let $A \subset \mathbb{R}$ denote a set of single valuedness of F which is dense according to Lemma 4.1. Now consider a densely defined selection s, s' of the mapping F, where $s(y) := F(y)$ for every $y \in A$ and

$$s'(y) = \begin{cases} s(y) & \text{if } y \in A, \\ F_\circ(t) & \text{if } y = t \end{cases}$$

(notice that by (15) $t \notin A$). Then it follows from (15) that

$$l(s')(t) = F_\circ(t) < F_\circ(t) + \frac{\varepsilon}{2} \le l(s)(t),$$

which contradicts Lemma 4.2. □

From (2) and Proposition 4.4, one can derive the extension of Lemma 2.2.

COROLLARY 4.5. *Let $f : X \to \mathbb{R}$ be a regularly locally Lipschitz function, $x, h \in X$. Then*

$$f_\infty(x; h, h) \le f'^L(x; h, h).$$

Using lim inf and lim sup calculus, Proposition 4.4 and Corollary 4.5 imply the following.

COROLLARY 4.6. *Let $f : X \to \mathbb{R}$ be a regularly locally Lipschitz function, $x, h \in X$. Then*

$$f'^U(x; h, h) = f'^u(x; h, h),$$

$$f^\infty(x; h, h) \ge f'^U(x; h, h).$$

5. Characterization of convexity. A former characterization of convexity via the Cominetti–Correa lower generalized second-order directional derivative given in [CC] for continuously differentiable and twice C-differentiable functions was proved under a more relaxed condition (when f is regularly locally Lipschitz) in [Y2, Theorem 4.1].

THEOREM 5.1. *Let $f : X \to \mathbb{R}$ be a regularly locally Lipschitz function. Then the following conditions are equivalent:*

(i) *f is convex;*
(ii) $f_\infty(x; h, h) \geq 0 \quad \forall x, h \in X$.

A characterization of convexity for arbitrary locally Lipschitz functions was obtained in [Pa1, Theorem 1.3].

THEOREM 5.2. *Let $f : X \to \mathbb{R}$ be a locally Lipschitz function. Then the following conditions are equivalent:*

(i) *f is convex;*
(ii) $f^{*L}(x; h, h) \geq 0 \quad \forall x, h \in X$;
(iii) $f'^{L}(x; h, h) \geq 0 \quad \forall x, h \in X$.

Now, Theorem 5.2 and Proposition 4.4 imply the following.

THEOREM 5.3. *Let $f : X \to \mathbb{R}$ be a regularly locally Lipschitz function. Then the following conditions are equivalent:*

(i) *f is convex;*
(ii) $f'^{l}(x; h, h) \geq 0 \quad \forall x, h \in X$.

With respect to formula (2), we eliminate the strict convergence $y \to x$ in condition (ii) of Theorem 5.1 to obtain Theorem 5.3.

We note that the calculus with $f'^{l}(x; h, h)$ is more simple than the calculus with $f_\infty(x; h, h)$.

Example 5.4. Let the convex function f be defined by

$$f(x) = \begin{cases} x^2 & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

Conditions given in Theorems 5.1 and 5.3 are satisfied. Indeed,

$$f_\infty(x; h, h) = \begin{cases} 2 & \text{if } x > 0,\ h \in \mathbb{R}, h \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f'^{l}(x; h, h) = \begin{cases} 2 & \text{if } x = 0,\ h > 0 \quad \text{or} \quad x > 0,\ h \in \mathbb{R}, h \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Another characterization of convexity for regularly locally Lipschitz functions in terms of generalized parabolic second-order directional derivatives is given in [HN, Theorem 3.3] (see also [Ch, Theorem 2.4] and [Y2, Theorem 4.1]).

6. Necessary optimality conditions. We begin this section by recalling two second-order necessary conditions.

THEOREM 6.1 (see [CC]). *Let $f : X \to \mathbb{R}$ attain a local minimum at $\hat{x} \in X$. Then*

$$f^\infty(\hat{x}; h, h) \geq 0 \quad \forall h \in X. \tag{16}$$

THEOREM 6.2 (see [YJ]). *Let $f : X \to \mathbb{R}$ be a $C^{1,1}$ function and let f attain a local minimum at $\hat{x} \in X$. Then*

$$f^{\diamond u}(\hat{x}; h, h) \geq 0 \quad \forall h \in X. \tag{17}$$

It was shown in [YJ] that condition (17) is tighter than condition (16) for $C^{1,1}$ functions.

For the proof of the following new second-order necessary condition, we use the Lebourg mean value theorem.

THEOREM 6.3 (see [L]). *Let $f : X \to \mathbb{R}$ be Lipschitz on an open convex set $U \subset X$, $x, y \in U$. Then there exists a point $u \in (x, y)$ with the property*

$$f(y) - f(x) \in \langle \partial_c f(u), y - x \rangle.$$

THEOREM 6.4. *Let $f : X \to \mathbb{R}$ be a locally Lipschitz function and let f attain a local minimum at $\hat{x}$. Then*

$$f'^{U}(\hat{x}; h, h) \geq 0 \quad \forall h \in X. \tag{18}$$

Proof. Inequality (18) is true for $h = 0$ because $f'^{U}(\hat{x}, 0, 0) = 0$. Now, we fix arbitrary $h \in X$, $h \neq 0$. By Theorem 6.3, for every $t > 0$, there exist $0 < \lambda(t) < t$ and $p \in \partial_c f(\hat{x} + \lambda(t)h)$ satisfying

$$f(\hat{x} + th) - f(\hat{x}) = \langle p, th \rangle.$$

Since $\hat{x}$ is a local minimum of f, one has $0 \in \partial_c f(\hat{x})$ and, moreover, it holds that $f(\hat{x} + th) - f(\hat{x}) \geq 0$ for every sufficiently small $t > 0$. Thus

$$f'^{U}(\hat{x}; h, h) \geq 0. \qquad \square$$

Lemma 2.1 and the calculus for lim sup and lim inf yield that (18) is tighter than (17) for $C^{1,1}$ functions. Moreover, using Corollary 4.6, we have that (18) is also tighter than (16) for regularly locally Lipschitz functions. (Compare formula (1) and Corollary 4.6 to see our process of elimination of strict convergence.)

We illustrate the results by giving examples.

Example 6.5. Consider a function $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = [\max\{x, 0\}]^2 \quad \forall x \in \mathbb{R}.$$

We can verify (see also [Y1, Example 5.2]) that f is $C^{1,1}$, 0 is a (local) minimum of f, $\bigtriangledown f(0) = 0$, and that for every $h \in \mathbb{R}$,

$$f^{\infty}(0; h, h) = f^{\diamond u}(0; h, h) = 2h^2 \geq 0.$$

Further, one has that it holds that

$$f'^{U}(0; h, h) = f'^{u}(0; h, h) = \begin{cases} 2h^2 & \text{if } h \geq 0, \\ 0 & \text{if } h < 0. \end{cases}$$

Thus, (16), (17), and (18) hold, but (18) is tighter than (16) and (17) because

$$f'^{U}(0; h, h) < f^{\infty}(0; h, h) = f^{\diamond u}(0; h, h)$$

for each $h < 0$.

Example 6.6. Let $f : \mathbb{R} \to \mathbb{R}$ be defined as

$$f(x) = \begin{cases} \displaystyle\int_0^{\|x\|} t\left(-\frac{4}{3} + \sin(\ln t)\right) dt & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

It can be shown (see also Example 3.3) that f is $C^{1,1}$, $\bigtriangledown f(0) = 0$,

$$f^\infty(0; h, h) = -\frac{4}{3}h^2 + \sqrt{2}h^2 \geq 0 \quad \forall h \in \mathbb{R},$$

and

$$f'^U(0; h, h) = f'^u(0; h, h) = -\frac{1}{3}h^2 < 0 \quad \text{for } h \neq 0.$$

Condition (18) excludes the possibility so that 0 may be a local minimum of f in contrast to condition (16).

It remains an open problem whether Corollary 4.6 is true also for arbitrary locally Lipschitz functions (a positive answer would mean that (18) is tighter than (16) not only for regularly locally Lipschitz functions).

7. Max-functions. In [CC, pp. 800–802], there was considered a finite family of C^2 functions $f_i : X \to \mathbb{R}$ for $i \in I = \{1, \ldots, n\}$ and the mapping $f : X \to \mathbb{R}$ given by

$$f(x) = \max\{f_i(x) : i \in I\}. \tag{19}$$

Before we recall the result stated in [CC], we denote $I(x) = \{i \in I : f_i(x) = f(x)\}$, $D^2g(x; u, v)$ as the usual second-order directional derivative at x with respect to the directions u, v. Finally, say that index $i \in I$ is essential at x if there exists a net $x_\alpha \to x$ with $I(x_\alpha) = \{i\}$, and denote by $I^*(x)$ the set of essential indexes at x.

THEOREM 7.1 (see [CC, Proposition 3.8]). *With the above notation we have*

$$f^\infty(x; u, v) \geq \max_{i \in I^*(x)} D^2 f_i(x; u, v) \quad \forall u, v \in X.$$

Now, we weaken assumptions setting on the regularity of functions f_i. Let $f_i : X \to \mathbb{R}$, $i \in I = \{1, \ldots, n\}$ be a collection of regular, locally Lipschitz functions. Let us consider a max-function $f : X \to \mathbb{R}$ defined as in (19). We recall some known results.

LEMMA 7.2. *The max-function f is locally Lipschitz provided that each of the functions f_i, $i \in I$, is locally Lipschitz.*

Proof. We can express the max-function f as the composition $f(x) = g(f_1(x), \ldots, f_n(x))$, where $g : \mathbb{R}^n \to \mathbb{R}$ and $g(\lambda_1, \ldots, \lambda_n) = \max\{\lambda_i : i \in I\}$. Clearly g is locally Lipschitz. Thus the composition of g and the maps $x \to (f_1(x), \ldots, f_n(x))$, which are locally Lipschitz from X into $\mathbb{R}^n$, must in turn also be locally Lipschitz. □

LEMMA 7.3. *The max-function f is regular provided that each of the functions f_i, $i \in I$, is regular.*

Proof. We can use an idea from the previous proof and [CLSW, Theorem 4.5]. □

Now, again put $I(x) := \{i \in \mathbb{N} : f_i(x) = f(x)\}$ for $x \in X$. We say that a locally Lipschitz function $g : X \to \mathbb{R}$ has a *directional second-order derivative at $x \in X$ along a vector $h \in S_X$* if there exists a finite limit

$$\lim_{t \downarrow 0} \frac{g^\circ(x + th; h) - g_\circ(x; h)}{t}.$$

Remark 7.4. Note that if $g : X \to \mathbb{R}$ is a C^2 function at $x \in X$, then it has a directional second-order derivative along $h \in S_X$ for every $h \in S_X$.

THEOREM 7.5. *Suppose that functions $f_i : X \to \mathbb{R}$, $i \in I$, are regularly locally Lipschitz and that each has a directional second-order derivative at $x \in X$ along $h \in S_X$. If f is defined as in* (19)*, then the following inequality holds:*

$$f'^u(x; h, h) \geq \max_{i \in I(x)} f_i'^u(x; h, h).$$

For the proof of this main result of section 7, we use Lemma 7.6.

LEMMA 7.6. *In the situation of Theorem* 7.5*, for every $i \in I(x)$, there exists a sequence $\{t_n\}_{n=1}^{+\infty}$ such that $t_n \downarrow 0$ and for any $n \in \mathbb{N}$ we have*

$$f'(x + t_n h; h) \geq f_i'(x + t_n h; h). \tag{20}$$

Proof. Due to Lemma 7.3 the max-function has a directional derivative because of regularity. Also notice that $f_i(x) = f(x)$ since $i \in I(x)$. Now suppose on the contrary that there is a $\delta > 0$ such that

$$f'(x + th; h) < f_i'(x + th; h) \tag{21}$$

for every $t \in (0, \delta)$.

The condition (21) implies that for every $t \in (0, \delta)$, we have

$$(f_i - f)'(x + th; h) > 0.$$

Hence, a function $t \to (f_i - f)(x + th)$ is increasing on $(0, \delta)$, which is contrary to the definition of f. □

Proof of Theorem 7.5. Fix an arbitrary $i \in I(x)$. By Lemma 7.6, there is a sequence $\{t_n\}_{n=1}^{+\infty}$ such that $t_n \downarrow 0$ and which holds condition (20). Notice also that

$$f^\circ(x; -h) = f'(x; -h) \geq f_i'(x; -h) = f_i^\circ(x; -h).$$

As a consequence we obtain

$$(f_i)_\circ(x; h) \geq f_\circ(x; h). \tag{22}$$

Now, with respect to (20) and (22), we have for any $n \in \mathbb{N}$

$$\frac{f^\circ(x + t_n h; h) - f_\circ(x; h)}{t_n} \geq \frac{f_i^\circ(x + t_n h; h) - (f_i)_\circ(x; h)}{t_n}. \tag{23}$$

By our assumption, the limit on the right side exists for $n \to +\infty$. Thus, by passing to the limit, we come to an inequality,

$$f'^U(x; h, h) \geq f_i'^u(x; h, h).$$

Using Lemma 7.3 and Corollary 4.6 on the left side of (23) and the fact that $i \in I(x)$ was arbitrary on the right side, one has

$$f'^u(x; h, h) \geq \max_{i \in I(x)} f_i'^u(x; h, h). \qquad \square$$

With respect to Remark 7.4, Theorem 7.5 generalizes Theorem 7.1 (according to formula (1), also in an elimination sense).

Acknowledgments. The authors are grateful to the referees for their detailed comments and valuable suggestions, which have contributed to the final preparation of the paper.

REFERENCES

[BM] J. M. Borwein and W. B. Moors, *Essentially strictly differentiable Lipschitz functions*, J. Funct. Anal., 149 (1997), pp. 305–351.

[BP] D. Bednařík and K. Pastor, *A technical note concerning one example*, Nonlinear Anal., 55 (2003), pp. 187–189.

[BZ] A. Ben-Tal and J. Zowe, *Necessary and sufficient optimality conditions for a class of nonsmooth minimization problems*, Math. Program., 24 (1982), pp. 70–91.

[CC] R. Cominetti and R. Correa, *A generalized second-order derivative in nonsmooth optimization*, SIAM J. Control Optim., 28 (1990), pp. 789–809.

[Ch] R. W. Chaney, *Second-order sufficiency conditions for nondifferentiable programming problems*, SIAM J. Control Optim., 20 (1982), pp. 20–33.

[CHN] W. L. Chan, L. R. Huang, and K. F. Ng, *On generalized second-order derivatives and Taylor expansions in nonsmooth optimization*, SIAM J. Control Optim., 32 (1994), pp. 591–611.

[Cl1] F. H. Clarke, *Necessary conditions for nonsmooth problems in optimal control and the calculus of variations*, Ph.D thesis, University of Washington, Seattle, WA, 1973.

[Cl2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[CLSW] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, Berlin, 1998.

[DR] V. F. Demyanov and A. M. Rubinov, *Constructive Nonsmooth Analysis*, in Approx. Optim. 7, Lang, Frankfurt am Main, Germany, 1995.

[GZ] P. G. Georgiev and N. P. Zlateva, *Second-order subdifferentials of $C^{1,1}$ functions and optimality conditions*, Set-Valued Anal., 4 (1996), pp. 101–117.

[HN] L. R. Huang and K. F. Ng, *On lower bounds of the second-order directional derivatives of Ben-Tal-Zowe and Chaney*, Math. Oper. Res., 22 (1997), pp. 747–753.

[HSN] J. B. Hiriart–Urruty, J. J. Strodiot, and V. H. Nguyen, *Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data*, Appl. Math. Optim., 11 (1984), pp. 43–56.

[J] L. Jokl, *Minimal convex-valued weak* usco correspondences and the Radon-Nikodym property*, Comment. Math. Univ. Carolina., 28 (1987), pp. 353–376.

[JY] V. Jeyakumar and X. Q. Yang, *Appropriate generalized Hessians and Taylor's expansions for continuously Gâteaux differentiable functions*, Nonlinear Anal., 36 (1999), pp. 353–368.

[K] H. Kawasaki, *An envelope-like effect of infinitely many inequality constraints on second-order necessary conditions for minimization problems*, Math. Program., 41 (1988), pp. 73–96.

[L] G. Lebourg, *Generic differentiability of Lipschitz functions*, Trans. Amer. Math. Soc., 256 (1979), pp. 125–144.

[Ma] Y. Maruyama, *Second-order necessary conditions for nonlinear optimization problems in Banach spaces and their application to an optimal control problem*, Math. Oper. Res., 15 (1990), pp. 467–482.

[Mo] W. B. Moors, *A characterization of minimal subdifferential mappings of locally Lipschitz functions*, Set-Valued Anal., 3 (1995), pp. 129–141.

[MP] P. Michel and J. P. Penot, *Calcul sous-différential pour des fonctions Lipschitzieness et non-Lipschitzieness*, C. R. Acad. Sci. Paris Sér. I Math., 298 (1985), pp. 269–272.

[Pa1] K. Pastor, *On relations among the generalized second-order directional derivatives*, Discuss. Math. Differ. Incl. Control Optim., 21 (2001), pp. 235–247.

[Pa2] K. Pastor, *Generalized second-order sufficient optimality conditions*, J. Electr. Eng., 53 (12/s) (2002), pp. 76–78.

[PB] K. Pastor and D. Bednařík, *On monotone minimal cuscos*, Acta Univ. Palacki. Olomuc. Fac. Rerum. Nat., Math., 40 (2001), pp. 185–194.

[Ph] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Math., Springer-Verlag, Berlin, 1993.

[Q] L. Qi, *Superlinearly convergent approximate Newton methods for LC^1 optimization problems*, Math. Program., 64 (1994), pp. 277–294.

[R1] R. T. Rockafellar, *First- and second-order epi-differentiability in nonlinear program-*

ming, Trans. Amer. Math. Soc., 307 (1988), pp. 75–108.

[R2] R. T. Rockafellar, *Second-order optimality conditions in nonlinear programming obtained by way of epi-derivatives*, Math. Oper. Res., 14 (1989), pp. 462–484.

[RW] R. T. Rockafellar and J. B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1998.

[Y1] X. Q. Yang, *On second-order directional derivatives*, Nonlinear Anal., 26 (1996), pp. 55–66.

[Y2] X. Q. Yang, *On relations and applications of generalized second-order directional derivatives*, Nonlinear Anal., 36 (1999), pp. 595–614.

[YJ] X. Q. Yang and V. Jeyakumar, *Generalized second-order directional derivatives and application with $C^{1,1}$ functions*, Optimization, 26 (1992), pp. 165–185.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 1078–1093

ROBUST ROOT-CLUSTERING OF A MATRIX IN INTERSECTIONS OR UNIONS OF REGIONS*

OLIVIER BACHELIER†, DIDIER HENRION‡, BERNARD PRADIN§, AND DRISS MEHDI†

Abstract. This paper considers robust stability analysis for a matrix affected by unstructured complex uncertainty. A method is proposed to compute a bound on the amount of uncertainty ensuring robust root-clustering in a combination (intersection and/or union) of several possibly nonsymmetric half planes, discs, and outsides of discs. In some cases to be detailed, this bound is not conservative. The conditions are expressed in terms of linear matrix inequalities (LMIs) and derived through Lyapunov's second method. As a distinctive feature of the approach, the Lyapunov matrices proving robust root-clustering (one per subregion) are not necessarily positive definite but have prescribed inertias depending on the number of roots in the corresponding subregions. As a special case, when root-clustering in a single half plane, disc, or outside of a disc is concerned, the whole clustering region reduces to only one convex subregion, and the corresponding unique Lyapunov matrix has to be positive definite as usual.

Key words. robust stability analysis, unstructured uncertainty, matrix root-clustering, $\mathcal{D}$-stability radius, inertia of a matrix, LMIs

AMS subject classifications. 93D09, 34D20, 15A39

DOI. 10.1137/S0363012903432365

1. Introduction.

Robust stability has been raising much interest in the last three decades. Indeed, in a linear state-space context, it matters to attest whether an uncertain state matrix has its eigenvalues in the open left half plane (OLHP) for continuous-time analysis or in the open unit disc (OUD) for discrete-time analysis. More precisely, assuming nominal stability, it can be useful to estimate the maximal size of the uncertainty domain for which stability is preserved.

The way to estimate this size obviously depends on the structure of the uncertainty. The structured (parametric) case should be distinguished from the unstructured (nonparametric) one as pointed out in one of the first contributions due to Patel and Toda [20]. The present contribution is restricted to an unstructured uncertainty, namely, the so-called norm-bounded uncertainty [22]. In this context, the maximal acceptable size of uncertainty was clearly defined, in continuous time, as the stability radius [13]. Such a stability radius was shown to equal the reciprocal of the $\mathcal{H}_\infty$-norm of a strictly proper transfer in [16] and thus also appears to be the reciprocal of the maximal structured singular value μ [9] along frequency. The discrete-time counterpart is described in [19]. In these references, the computation of the stability radius could be carried out with iterative resolutions of Riccati equations [6, 7]. Another technique consists in computing μ while sweeping frequencies. With the emergence of convex optimization over linear matrix inequalities (LMIs), the stability radius can be computed owing to the LMI version of the bounded real lemma [1].

*Received by the editors July 28, 2003; accepted for publication (in revised form) February 24, 2004; published electronically, November 9, 2004.

http://www.siam.org/journals/sicon/43-3/43236.html

†LAII-ESIP, Bâtiment de Mécanique, 40 Avenue du Recteur Pineau, 86022 Poitiers CEDEX, France (Olivier.Bachelier@esip.univ-poitiers.fr, Driss.Mehdi@esip.univ-poitiers.fr).

‡LAAS-CNRS, 7, Avenue du Colonel Roche, 31077 Toulouse CEDEX, France. Department of Control Engineering, Czech Technical University in Prague, Technická, 16627, Praha 6, Czech Republic (henrion@laas.fr). The work of this author has been supported by the Ministry of Education of the Czech Republic under contract ME 698.

§LAAS-CNRS, 7, Avenue du Colonel Roche, 31077 Toulouse CEDEX, France (pradin@laas.fr).

It is also important to distinguish between the complex stability radius and the real one. The former concerns a complex uncertainty and can be computed with LMI software as just mentioned. The latter takes the realness of the uncertainty into account (what is more discerning to analyze practical plants in automatic control) and is a bit more difficult to obtain, involving a one-dimensional sweeping over the second largest singular value of a parametrized matrix [23, 12]. The present contribution is restricted to the complex case.

When further performances are required, such as transient ones, it might be shrewd to consider a more sophisticated region for the state matrix root-clustering, different from the OLHP or the OUD. Based on the notions of Ω-regions and generalized Lyapunov equations (GLEs) due to Gutman and Jury [10], Yedavalli has proposed significant robustness bounds [31] (later improved by other authors), but the results are still conservative. The reader is also invited to see [29]. Moreover, the considered regions are usually connected, which might not be suitable for plants with separate dynamics or with specified robust damping ratios. One of the first attempts to consider unions of regions is provided in [2]. The concept of $\mathcal{D}_U$-stability (root-clustering in some region $\mathcal{D}_U$ whose form encompasses many unions of possible disjoint and nonsymmetric subregions) enables more general results [3]. However, these results remain quite conservative.

This paper is an attempt to consider sophisticated clustering regions by extending the notion of a complex stability radius to some combinations (unions and/or intersections) of half planes, discs, and exteriors of discs. Besides, the conservatism of the previously proposed methods is reduced. In some typical cases to be further detailed, the exact value of the complex radius is reached.

The paper is organized as follows: section 2 states the problem, presenting the clustering regions and extending the concept of a complex $\mathcal{D}$-stability radius to the case where $\mathcal{D}$ is some combination of regions. Section 3 introduces the notion of $\partial\mathcal{D}$-regularity of a nominal matrix, which is the nonmembership of the matrix eigenvalues to a geometric curve $\partial\mathcal{D}$. Such a property can often be checked through the derivation of a Lyapunov matrix which is not necessarily positive definite but is just nonsingular with constant inertia. When $\mathcal{D}$-stability is concerned, the Lyapunov matrix is required to be strictly positive or negative definite. In section 4, the considered matrix is affected by an additive norm-bounded uncertainty. Based upon the notion of $\partial\mathcal{D}$-regularity, a method to reach the complex $\mathcal{D}$-stability radius is proposed. In some cases to be detailed, the exact value is obtained. Numerical examples are provided in section 5 before the conclusion.

Notation. M' denotes the transpose conjugate of matrix M. Hence, s' is the conjugate of complex number s. M^H is the Hermitian matrix $M + M'$. The 2-norm of M induced by the Euclidean vector norm (maximal singular value) is denoted by $||M||_2$. $\mathbb{I}_n$ is the identity matrix of order n, and $\mathbb{O}$ is a null matrix of appropriate dimensions. Matrix inequalities are considered in the sense of Löwner, i.e., > 0 (resp., < 0) means positive (resp., negative) definite. The symbol $\mathbf{i}$ denotes the imaginary unit and $\lambda(A)$ denotes the spectrum of square matrix A. $\mathfrak{I}$ and $\mathfrak{C}$ are the imaginary axis and the unit circle, respectively. At last, the vector $\mathrm{In}(M) = [n_+ \; n_- \; n_0]$ is the inertia of a square matrix M if n_+, n_-, n_0 are the numbers of eigenvalues of M with positive, negative, and zero real part, respectively.

2. Problem statement. First, the form of the uncertain matrix to be analyzed is given. Then, the clustering region is introduced. At last, the problem to be solved is stated.

2.1. The uncertain matrix. The considered matrix reads

$$A_c = A + B\Delta C \in \mathbb{C}^{n\times n}. \tag{1}$$

In the above expression, Δ is constant, unknown, and assumed to belong to $\mathcal{B}(\rho)$, the ball of all matrices $\Delta \in \mathbb{C}^{q\times r}$ satisfying $||\Delta||_2 \leq \rho$. Matrices $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times q}$, and $C \in \mathbb{C}^{r\times n}$ are perfectly known. Such a description is referred to as norm-bounded uncertainty [22].

Remark 1. In this work, all the matrices are assumed to be complex. However, in practice, when transient performances of a linear state-space model are to be analyzed, state matrix A_c is real. In that case, Δ is then real and $\mathcal{B}(\rho)$ must be restricted to real matrices. Actually, this restriction is more difficult to take into account and this case is not investigated in the paper although some interesting results exist [23, 12].

2.2. Clustering region $\mathcal{D}$. Consider the following geometric curves:

$$\begin{cases} \partial\mathcal{D}_k = \left\{ s \in \mathbb{C} | f_k(s) = r_{k_{00}} + (r_{k_{10}} s)^H + r_{k_{11}} s' s = 0 \right\} \\ \\ \{ r_{k_{00}}, r_{k_{10}}, r_{k_{11}} \} \in \mathbb{R} \times \mathbb{C} \times \mathbb{R} \end{cases} \quad \forall k \in \{1, \ldots, m\}. \tag{2}$$

Relevant curves are lines or circles. Each curve $\partial\mathcal{D}_k$ enables us to define an associated region:

$$\mathcal{D}_k = \{ s \in \mathbb{C} \,|\, f_k(s) < 0 \} \quad \forall k \in \{1, \ldots, m\}. \tag{3}$$

Clearly, $\mathcal{D}_k$ denotes either one side or the other side of the boundary $\partial\mathcal{D}_k$. It can then be a half plane, a disc, or the outside of a disc. It is an open region (i.e., not including $\partial\mathcal{D}_k$) in order to encompass the concept of asymptotic stability for linear time invariant (LTI) systems. $\mathcal{D}_k$ can actually correspond to the scalar case of regions defined in [21] or to a special case of second order Ω-regions [10].

Also define the region $\mathcal{D}$ as a combination, i.e., any union and/or intersection of the various subregions $\mathcal{D}_k$. Such a formulation of $\mathcal{D}$ clearly enables a very large choice of clustering regions.

2.3. Problem statement. This contribution aims at computing the complex $\mathcal{D}$-stability radius.

More precisely, assume that A is $\mathcal{D}$-stable, i.e., that the whole of its spectrum lies in $\mathcal{D}$. Define $r_{\mathcal{D}}$ as the largest value of ρ, the radius of $\mathcal{B}(\rho)$, such that A_c defined in (1) remains $\mathcal{D}$-stable for any $\Delta \in \mathcal{B}(\rho)$. Such a value is the so-called complex $\mathcal{D}$-stability radius. A lower bound $\rho^\star$ of $r_{\mathcal{D}}$, as tight as possible, is to be computed. For this purpose, the concept of $\partial\mathcal{D}$-regularity is introduced in the next section.

3. $\partial\mathcal{D}$-regularity. In this section, only nominal matrices are considered. The concepts of matrix $\partial\mathcal{D}$-regularity and matrix $\partial\mathcal{D}$-singularity are introduced. A necessary and sufficient condition for matrix $\partial\mathcal{D}$-regularity to be satisfied when $\partial\mathcal{D}$ is defined as in (2) is expressed in terms of an LMI. After preliminary notions and assumptions in subsection 3.1, subsection 3.2 presents this condition through a first theorem. In subsection 3.3, the distribution of the matrix eigenvalues with respect to the boundary $\partial\mathcal{D}$ is connected to the inertia of the solution to the LMI, owing to a second theorem. Subsection 3.4 is devoted to a discussion of these theorems.

3.1. Preliminaries.

DEFINITION 1. *Let $\partial\mathcal{D}$ be any curve in the complex plane; then matrix $A \in \mathbb{C}^{n\times n}$ is called*

- *$\partial\mathcal{D}$-singular when* $\lambda(A) \cap \partial\mathcal{D} \neq \emptyset$;
- *$\partial\mathcal{D}$-regular when* $\lambda(A) \cap \partial\mathcal{D} = \emptyset$.

Remark 2. Assume that $\partial\mathcal{D}$ is a boundary separating two open regions $\mathcal{D}$ and $\bar{\mathcal{D}}^C$ (then $\mathbb{C} = \mathcal{D} \cup \partial\mathcal{D} \cup \bar{\mathcal{D}}^C$). Matrix A is $\mathcal{D}$-stable if and only if it is $\partial\mathcal{D}$-regular and the whole of its spectrum lies in $\mathcal{D}$. Otherwise, it is $\mathcal{D}$-unstable.

Now assume that $\partial\mathcal{D}$ here reduces to only one curve as defined in (2), i.e.,

$$\partial\mathcal{D} = \{s \in \mathbb{C} \,|\, f(s) = r_{00} + (r_{10}s)^H + r_{11}s's = 0\}. \tag{4}$$

In parallel with the work of Hill [11], we state two theorems in the next two parts. The result provided in [11] is based on Ostrowski and Schneider's theorem [18] and on Frobenius's theorem. Our contribution is more part of the framework relevant to Lyapunov's second method [17] and its extensions to root-clustering [10].

3.2. LMI condition for $\partial\mathcal{D}$-regularity.

THEOREM 1. *Let A and $\partial\mathcal{D}$ be, respectively, a complex square matrix of dimension n and a curve of the complex plane as defined in* (4). *Matrix A is $\partial\mathcal{D}$-regular if and only if there exists a Hermitian matrix $P \in \mathbb{C}^{n\times n}$ such that*

$$F(A,P) = r_{00}P + (r_{10}PA)^H + r_{11}A'PA < 0. \tag{5}$$

Proof. Some arguments are inspired from [14, 8].

Sufficiency. First assume that there exists a suitable P such that (5) holds. Also assume that A has $\bar{n} \leq n$ distinct eigenvalues. It suffices to prove the nonmembership of these $\bar{n}$ eigenvalues to $\partial\mathcal{D}$. Let λ_j denote the jth eigenvalue of A. There exists a nonzero vector v_j such that $Av_j = \lambda_j v_j$, so it follows that

$$v_j'F(A,P)v_j \;<\; 0 \quad \forall j \in \{1,\ldots,\bar{n}\}$$

$$\Leftrightarrow\; r_{00}v_j'Pv_j + (r_{10}v_j'PAv_j)^H + r_{11}v_j'A'PAv_j < 0 \quad \forall j \in \{1,\ldots,\bar{n}\}$$

$$\Leftrightarrow\; (r_{00} + (r_{10}\lambda_j)^H + r_{11}\lambda_j'\lambda_j)(v_j'Pv_j) < 0 \quad \forall j \in \{1,\ldots,\bar{n}\},$$

which implies

$$f(\lambda_j) = r_{00} + (r_{10}\lambda_j)^H + r_{11}\lambda_j'\lambda_j \neq 0 \quad \forall j \in \{1,\ldots,\bar{n}\},$$

and hence

$$\lambda_j \notin \partial\mathcal{D} \quad \forall j \in \{1,\ldots,\bar{n}\} \Leftrightarrow \lambda_i \notin \partial\mathcal{D} \quad \forall i \in \{1,\ldots,n\}.$$

Matrix A is then $\partial\mathcal{D}$-regular.

Necessity. Now assume that A is $\partial\mathcal{D}$-regular. Then

$$f(\lambda_i) \neq 0 \quad \forall i \in \{1,\ldots,n\}. \tag{6}$$

It means that $f(\lambda_i)$ is either strictly positive or strictly negative. Define matrix Λ as

$$\Lambda \;=\; \underset{i=1,\ldots,n}{\operatorname{diag}}\{\lambda_i\} \;=\; \begin{bmatrix} \Lambda_1 & \mathbb{O} \\ \mathbb{O} & \Lambda_2 \end{bmatrix},$$

where $\Lambda_1 \in \mathbb{C}^{n_-}$ contains the various λ_i for which $f(\lambda_i) > 0$ and $\Lambda_2 \in \mathbb{C}^{n_+}$ contains those for which $f(\lambda_i) < 0$. With

$$\mathcal{I} = \begin{bmatrix} -\mathbb{I}_{n_-} & \mathbb{O} \\ \mathbb{O} & \mathbb{I}_{n_+} \end{bmatrix}, \tag{7}$$

it holds that

$$F(\Lambda, \mathcal{I}) = r_{00} \begin{bmatrix} -\mathbb{I}_{n_-} & \mathbb{O} \\ \mathbb{O} & \mathbb{I}_{n_+} \end{bmatrix} + \left(r_{10} \begin{bmatrix} -\Lambda_1 & \mathbb{O} \\ \mathbb{O} & \Lambda_2 \end{bmatrix} \right)^H + r_{11} \begin{bmatrix} -\Lambda_1'\Lambda_1 & \mathbb{O} \\ \mathbb{O} & \Lambda_2'\Lambda_2 \end{bmatrix} < 0. \tag{8}$$

Let $\mathcal{J}$ be the Jordan canonical form of A. There exists a sequence of full rank matrices T_l such that [8]

$$\lim_{l \to \infty} (T_l \mathcal{J} T_l^{-1}) = \Lambda \tag{9}$$

(for example, if $\mathcal{J}$ is a single Jordan block $\left[\begin{smallmatrix} \lambda & 1 \\ 0 & \lambda \end{smallmatrix}\right]$, T_l equals $\left[\begin{smallmatrix} \frac{1}{l} & 0 \\ 0 & 1 \end{smallmatrix}\right]$). Since F is continuous with respect to its first argument, from (8) it follows that

$$\lim_{l \to \infty} F(T_l \mathcal{J} T_l^{-1}, \mathcal{I}) < 0. \tag{10}$$

Let $\mathcal{T}$ denote T_l for a sufficiently large $l < \infty$ (i.e., such that $\mathcal{T}^{-1}$ exists) so that

$$F(\mathcal{T}\mathcal{J}\mathcal{T}^{-1}, \mathcal{I}) < 0.$$

Applying congruence on the previous inequality leads to

$$\mathcal{T}' F(\mathcal{T}\mathcal{J}\mathcal{T}^{-1}, \mathcal{I})\mathcal{T} = F(\mathcal{J}, \mathcal{T}'\mathcal{I}\mathcal{T}) < 0.$$

Now assume that V denotes a matrix such that $\mathcal{J} = VAV^{-1}$. Left and right multiplying the previous inequality, respectively, by V' and V leads to (5) with

$$P = V'\mathcal{T}'\mathcal{I}\mathcal{T}V. \tag{11}$$

Note that if there exists a matrix V such that $\Lambda = VAV^{-1}$ (or, equivalently, $\Lambda = \mathcal{J}$), then the above reasoning stands with $T_l = \mathbb{I}_n \, \forall l$. This completes the proof. ∎

3.3. Root-distribution and the inertia of P.

THEOREM 2. *Let A and $\partial\mathcal{D}$ be, respectively, a complex square matrix of dimension n and a curve of the complex plane as defined in* (4). *Matrix A is $\partial\mathcal{D}$-regular with n_+ eigenvalues in $\mathcal{D}$ and n_- eigenvalues outside $\mathcal{D}$ if and only if any Hermitian solution P to LMI* (5) *has inertia* $\mathrm{In}(P) = [n_+,\ n_-,\ 0]$.

Proof. Some possible expression of P is given by (11). Since inertia is invariant under congruence, $\mathrm{In}(P) = \mathrm{In}(\mathcal{I}) = [n_+,\ n_-,\ 0]$, where $\mathcal{I}$ is as defined in (7). The question is to know if it is possible that there exists some other solution to LMI (5) with a different inertia. If it is possible, by continuity of the convex set of solutions, there exists a singular X such that

$$F(A, X) < 0.$$

Let $\mathcal{J}$ be a Jordan canonical form similar to A and V be the associated similarity matrix. By congruence, it follows that

$$F(\mathcal{J}, Z) < 0,$$

where $Z = (V')^{-1}XV^{-1}$ is also a singular matrix. Taking the sequence of full rank matrices T_l introduced in (9) into account enables us to deduce

$$F\left(\lim_{l\to\infty}(T_l^{-1}\Lambda T_l), Z\right) < 0,$$

which by continuity of F with respect to its first argument leads to

$$\lim_{l\to\infty} F(T_l^{-1}\Lambda T_l, Z) < 0.$$

Once again, let $\mathcal{T}$ denote T_l for a sufficiently large value of $l < \infty$ so that

$$F(\mathcal{T}^{-1}\Lambda\mathcal{T}, Z) < 0.$$

Applying congruence on the previous inequality leads to

$$F(\Lambda, Y) < 0, \tag{12}$$

where

$$Y = \begin{bmatrix} Y_1 & Y_3 \\ Y_3' & Y_2 \end{bmatrix} = (\mathcal{T}')^{-1}Z\mathcal{T}^{-1}$$

is singular. At this stage a distinction is made between two cases:

• $r_{11} = 0$. In this case, $\mathcal{D}$ and $\bar{\mathcal{D}}^C$ are half planes. By Schur factorization, there exists a unitary matrix U such that $Y = U'SU$, where S is a diagonal matrix whose diagonal entries are the eigenvalues of Y. With no loss of generality, Y being singular, the first diagonal entry is assumed to be zero. By congruence, one gets

$$F(U\Lambda U', S) < 0.$$

The previous inequality never holds since the first diagonal entry of the left member is zero. Hence, in this case, Y or X cannot be singular, and the inertia of the solution is invariant.

• $r_{11} \neq 0$. In this case, first note that $\partial\mathcal{D}$ can be defined another way:

$$\partial\mathcal{D} = \left\{ s \in \mathbb{C} \,|\, f(s) = \tilde{r}_{00} + r_{11}\left(s' + \frac{r_{10}}{r_{11}}\right)\left(s + \frac{r_{10}'}{r_{11}}\right) = 0 \right\} \tag{13}$$

with

$$\tilde{r}_{00} = r_{00} - \frac{r_{10}r_{10}'}{r_{11}}.$$

Note that $\frac{\tilde{r}_{00}}{r_{11}}$ is strictly negative; otherwise, $\partial\mathcal{D}$ reduces to $\emptyset$. In terms of matrix inequality, it leads to rewriting (12) as follows:

$$\tilde{F}(\tilde{\Lambda}, Y) = \tilde{r}_{00}Y + r_{11}\tilde{\Lambda}'Y\tilde{\Lambda} < 0,$$

where $\tilde{\Lambda} = \Lambda + \frac{r_{10}'}{r_{11}}\mathbb{I}_n$ is $\partial\tilde{\mathcal{D}}$-regular, $\partial\tilde{\mathcal{D}}$ being the circle centered around the new origin and of radius $\sqrt{-\tilde{r}_{00}/r_{11}}$. Two diagonal blocks can be extracted from the previous inequality:

$$\begin{cases} \tilde{F}(\tilde{\Lambda}_1, Y_1) < 0, \\ \tilde{F}(\tilde{\Lambda}_2, Y_2) < 0. \end{cases}$$

Clearly, $\tilde{\Lambda}_1 = \Lambda_1 + \frac{r'_{10}}{r_{11}}\mathbb{I}_{n_-}$ and $\tilde{\Lambda}_2 = \Lambda_2 + \frac{r'_{10}}{r_{11}}\mathbb{I}_{n_+}$ are such that $\tilde{\Lambda}_2$ is $\tilde{\mathcal{D}}$-stable and $\tilde{\Lambda}_1$ has its whole spectrum outside $\tilde{\mathcal{D}}$. It is clear, from the beginning of the proof, that there exists a Hermitian matrix $\tilde{Y}_1 < 0$ such that $\tilde{F}(\tilde{\Lambda}_1, \tilde{Y}_1) < 0$. It will now be shown that any other solution Y_1 remains negative definite by continuity, i.e., never becomes singular. First note that since $\tilde{\Lambda}_1$ has its spectrum outside of the disc $\tilde{\mathcal{D}}$, it can be inverted and it follows that

$$r_{11}Y_1 + \tilde{r}_{00}(\tilde{\Lambda}_1^{-1})'Y_1\tilde{\Lambda}_1^{-1} < 0.$$

Assume that Y_1 is negative semidefinite. Achieve Schur factorization of Y_1 such that $Y_1 = U_1'S_1U_1$. The diagonal entries of S_1 are negative, excepting the first one which is zero. By congruence, it follows that

$$S_1 + \frac{\tilde{r}_{00}}{r_{11}}U_1(\tilde{\Lambda}_1^{-1})'U_1'S_1U_1\tilde{\Lambda}_1^{-1}U_1' < 0.$$

Since $\frac{\tilde{r}_{00}}{r_{11}} < 0$, the first diagonal entry of the left-hand side member is positive or zero, and thus inequality never holds. Y_1 is then necessarily strictly negative definite.

In the same way, there also exists some Hermitian matrix $\tilde{Y}_2 > 0$ such that $\tilde{F}(\tilde{\Lambda}_2, \tilde{Y}_2) < 0$. Is it possible that, by continuity, some other solution Y_2 becomes singular, i.e., positive semidefinite? Assume that such a Y_2 exists and by Schur factorization $Y_2 = U_2'S_2U_2$ (S_2 is diagonal with the first diagonal entry zero and the other ones positive) and by congruence we have

$$\tilde{F}(U_2\tilde{\Lambda}_2U_2', S_2) < 0.$$

Since $\tilde{\Lambda}_2$ and S_2 are diagonal matrices and since the first diagonal entry of S_2 is zero, then the first diagonal entry of the left-hand side member is positive or zero, which leads to a contradiction. Thus, Y_2 is necessarily strictly positive definite.
Now, coming back to Y, using Schur complement, one gets

$$\mathrm{In}(Y) = \mathrm{In}\left(\left[\begin{array}{c|c} Y_1 & \mathbb{O} \\ \hline \mathbb{O} & Y_2 - Y_3'Y_1^{-1}Y_3 \end{array}\right]\right).$$

Clearly, since $Y_2 > 0$ and $Y_1 < 0$, none of the diagonal blocks of the above matrix can be singular. Hence, $\mathrm{In}(Y) = \mathrm{In}(X) = \mathrm{In}(P)$ and the proof is complete. ∎

As mentioned in subsection 3.1, our result is an alternative to results by Hill [11]. However, our result is summarized under the form of one single theorem valid both for any line and any circle. It does not require any preliminary mathematical lemma except Sylvester's well-known theorem ($\mathrm{In}(H) = \mathrm{In}(MHM')$ for any nonsingular matrix M). Although it is formulated as in Lyapunov's framework, the proofs basically require simple algebraic manipulations which, the authors hope, will help extension to other curves. In that sense, it is closer to the result of [14]. Nevertheless, some differences are pointed out in the discussion provided in the next subsection.

COROLLARY 1. *Let A and $\partial\mathcal{D}$ be, respectively, a complex square matrix of dimension n and a curve of the complex plane as defined in* (4). *Also let $\mathcal{D}$ and $\bar{\mathcal{D}}^C$ be the regions defined by $f(s) < 0$ and $f(s) > 0$, respectively. Matrix A is $\mathcal{D}$-stable (resp., $\bar{\mathcal{D}}^C$-stable) if and only if there exists a Hermitian positive (resp., negative) definite matrix $P \in \mathbb{C}^{n\times n}$ such that* (5) *holds.*

Proof. The proof follows directly from Theorem 2 considering inertia $\mathrm{In}(P) = [n\ 0\ 0]$ (resp., $\mathrm{In}(P) = [0\ n\ 0]$). ∎

As special cases, $\mathcal{D}$ can be the OLHP or the OUD. In those cases, Lyapunov's theorem [17] and Stein's theorem [26] are, respectively, recovered.

Remark 3. Since any solution to (5) is necessarily nonsingular, there is no need to specify it as a constraint, and (5) is a simple LMI in P easy to test with various existing LMI software. Note that the nonstrict LMI cannot be considered here because it would allow $\partial\mathcal{D}$-singularity.

3.4. Discussion. With appropriate changes, the above proof could be adapted to any second order Ω-transformable region. In that sense, this could be seen as a special case of [14]. However, although [14, Theorem 1] seems suitable to prove the first statement in Theorem 1, we do not agree with the proof of [14, Theorem 2] related to the inertia of P. Indeed after having proven that some solution P to an LMI exists and has expression [14, equation (6)], it is claimed that for any choice of the negative definite right member of the associated equality, the solution to this equality keeps the same expression. We do not agree with that point. Perhaps the same doubt led Jury to achieve a special proof for the nonsingularity of the solution of a GLE [15, Theorem 3.16]. The notion of Ω-transformability was required to prove this nonsingularity (see also [10, Theorem 12]). It could at first sight directly be derived from [14, Theorem 2], but we also think that this special proof was necessary.

Going on with Ω-transformability, it is interesting to see that although transformability seemed to be required to prove the *nonsingularity* of the solution to a GLE [10], we show here that, owing to the notion of $\partial\mathcal{D}$-regularity (rather than just $\mathcal{D}$-stability), $\mathcal{D}$ and $\bar{\mathcal{D}}^C$ are considered altogether (the reader is here reminded of the fact that $\mathcal{D}$-stability and $\bar{\mathcal{D}}^C$-stability are only special cases of $\partial\mathcal{D}$-regularity as stated in Corollary 1). As a consequence, the outside of a disc is a non-Ω-transformable region for which it is impossible to find a singular solution to a corresponding GLE. Otherwise, it would be in contradiction with Remark 3. In other words, any solution to a GLE attesting matrix root-clustering is necessarily nonsingular (existence and uniqueness of the solution to a GLE is another problem; see [10]). The outside of the disc is then a region for which Ω-transformability is not required to guarantee the nonsingularity of the solution to a corresponding GLE. It is what was illustrated by an example proposed in [28].

Apart from our doubt of the proof of [14, Theorem 2], we would like to add that this contribution seems to have been overlooked. Actually, [14, Theorem 1] is nothing but an LMI test for matrix root-clustering in an Ω-region. In 1971, such a test was not tractable from a computational point of view (at about the same time, Willems was just beginning to warn the control community about the great interest in handling LMIs [30]). For this reason, it mattered to "convert" this LMI test into a GLE [10]. Now that LMIs have become classical tools, although some significant contributions enabled us to test matrix root-clustering via LMI conditions [8], many authors should remember the pioneer work [14].

Furthermore, if solving an LMI is not an obstacle, results in [14] can be used to consider problem 85 formulated by Wang in the electronic book proposed by Blondel and Megretski [5]. In problem 85, the analysis of root-clustering in Ω-transformable regions of order greater than 2 is concerned and expressed in terms of GLEs. Of course, it is an interesting challenge from a mathematical point of view. At first sight, as far as root-clustering analysis is concerned, the LMI approach, as used in this paper or in [14], could be exploited to solve the problem and the GLE approach could seem useless. However, it is not really true because the problem is not purely theoretical. When problems involving very high dimensions are to be solved, the LMI

approach generates heavy computations and can lead to numerical failure while some P actually exists. This is why the GLE approach preserves all its interest. Solving a GLE requires choosing a right-hand side member in order to get a simple linear system. Thus it is important to be sure that this member can be arbitrarily chosen and will always lead to a unique nonsingular solution P. Such a possible arbitrary choice is probably strongly related to Ω-transformability and that is exactly the point presented in problem 85. To sum up this part of the discussion, for low-dimensional problems, the LMI approach might enable us to be free of the transformability assumption and of GLEs, but for high-dimensional problems, those notions are still fundamental.

Chapter 17 of [4] describes the guardian map as an alternative tool for checking $\mathcal{D}$-stability or $\partial\mathcal{D}$-regularity. The guardian map is a mapping ν such that $\nu(A) \neq 0$ if $A \in \mathcal{A}$ and $\nu(A) = 0$ if $A \notin \mathcal{A}$. For example, the mapping $\nu(A) = \det A$ guards the sets of nonsingular matrices, and $\nu(A) = \det H(p)$ guards the sets of Hurwitz stable matrices, where $H(p)$ is the Hurwitz matrix of characteristic polynomial $p(s) = \det(sI - A)$. As shown in [4], checking properties like robust regularity or stability then amounts to computing eigenvalues of matrices related to guardian maps. Our approach can be viewed as an alternative to guardian map techniques, in the sense that the robustness certificate for regularity or stability is the existence of a Lyapunov matrix with given inertia solving a given LMI.

4. Complex $\mathcal{D}$-stability radius. In this section, the uncertain case is studied. First, a necessary and sufficient condition for the uncertain A_c defined in (1) to be $\partial\mathcal{D}$-regular when $\partial\mathcal{D}$ complies with (4) is given. This condition is then used to compute the $\mathcal{D}$-stability radius when $\mathcal{D}$ is some combination of regions as defined in subsection 2.2.

THEOREM 3. *Let A_c and $\partial\mathcal{D}$ be, respectively, an uncertain matrix as defined in* (1) *and a geometric curve as defined in* (4). *Assume that nominal matrix A is $\partial\mathcal{D}$-regular. Matrix A_c is robustly $\partial\mathcal{D}$-regular against the ball $\mathcal{B}(\rho)$ if and only if there exists some Hermitian matrix P with inertia* $\mathrm{In}(P) = [n_+,\ n_-,\ 0]$ *such that*

$$\mathcal{Q}(P,\gamma) = \begin{bmatrix} r_{00P} + (r_{10}PA)^H + r_{11}A'PA + C'C & r_{10}PB + r_{11}A'PB \\ r'_{10}B'P + r_{11}B'PA & r_{11}B'PB - \gamma\mathbb{I}_q \end{bmatrix} < 0 \tag{14}$$

with $\gamma = \rho^{-2}$. In this event, A_c keeps n_+ eigenvalues inside $\mathcal{D}$ and n_- outside $\mathcal{D}$.

Proof. First assume that $r_{11} = 0$. Then $\partial\mathcal{D}$ is a line. By appropriate rotation and shifting, it can be transformed into the imaginary axis $\partial\hat{\mathcal{D}} = \mathfrak{I}$. Apply the same mapping on matrices A, and then the necessary and sufficient condition for the new matrix $\hat{A}_c = \hat{A} + B\Delta C$ to be $\partial\hat{\mathcal{D}}$-regular can be deduced from Lemma 1 recalled in the appendix. Indeed, consider Lemma 1. Condition (14) corresponds to (28) with A substituted by $\hat{A}$ and

$$M = \begin{bmatrix} C'C & \mathbb{O} \\ \mathbb{O} & -\gamma\mathbb{I}_q \end{bmatrix}, \tag{15}$$

which is equivalent to (27). It follows that

$$(C(\mathbf{i}\omega\mathbb{I} - \hat{A})^{-1}B)'(C(\mathbf{i}\omega\mathbb{I} - \hat{A})^{-1}B) < \gamma\mathbb{I}_q \quad \forall\omega \in \mathbb{R}. \tag{16}$$

In the same way, if $r_{11} \neq 0$, then $\partial\mathcal{D}$ is a circle. With appropriate scaling and shifting, $\partial\mathcal{D}$ can be transformed into the unit circle $\partial\hat{\mathcal{D}} = \mathfrak{C}$. Apply the same mapping on matrix A in order to get the new matrix $\hat{A}$. Uncertain matrix $\hat{A}_c = \hat{A} + B\Delta C$ has to be $\mathfrak{C}$-regular, which can be tested owing to Lemma 2 given in the appendix. Condition

(14) is equivalent to (30) with $A = \hat{A}$ and M given by (15), which is equivalent to (29). It follows that

$$(C(e^{\mathbf{i}\omega}\mathbb{I} - \hat{A})^{-1}B)'(C(e^{\mathbf{i}\omega}\mathbb{I} - \hat{A})^{-1}B) < \gamma\mathbb{I}_q \quad \forall\omega \in \mathbb{R}. \tag{17}$$

Both cases then lead to the equivalent condition

$$\Leftrightarrow ||C(s\mathbb{I}_n - \hat{A})^{-1}B||_2 < \sqrt{\gamma} = \rho^{-1} \quad \forall s \in \partial\hat{\mathcal{D}}.$$

Using classical arguments on singular values yields

$$[\inf\{||\Delta||_2 | \Delta \in \mathbb{C}^{q\times r} \text{ and } \det(\mathbb{I}_q - \Delta C(s\mathbb{I}_n - \hat{A})^{-1}B) = 0\}]^{-1} < \rho^{-1} \quad \forall s \in \partial\hat{\mathcal{D}},$$

which is the definition of the structured singular value. Using properties of det(.), the rewriting of the above expression leads to

$$[\inf\{||\Delta||_2 | \Delta \in \mathbb{C}^{q\times r} \text{ and } \det(s\mathbb{I}_n - (\hat{A} + B\Delta C)) = 0\}]^{-1} < \rho^{-1} \quad \forall s \in \partial\hat{\mathcal{D}} \Leftrightarrow$$

$$\eta^{-1}(s) = [\inf\{||\Delta||_2 | \Delta \in \mathbb{C}^{q\times r} \text{ and } s \in \lambda(\hat{A} + B\Delta C)\}]^{-1} < \rho^{-1} \ \forall s \in \partial\hat{\mathcal{D}} \tag{18}$$

$$\Leftrightarrow \eta(s) > \rho \quad \forall s \in \partial\hat{\mathcal{D}}.$$

The above inequality shows that any uncertain matrix Δ inducing the $\partial\hat{\mathcal{D}}$-singularity of matrix $\hat{A}_c$ is such that $||\Delta||_2 > \rho$. As a consequence, $\hat{A}_c$ is $\partial\hat{\mathcal{D}}$-regular over $\mathcal{B}(\rho)$. Because each eigenvalue of A is implicitly subject to the same mapping as A, the root-distribution of A with respect to $\partial\mathcal{D}$ is the same as the root-distribution of $\hat{A}$ with respect to $\partial\hat{\mathcal{D}}$. From the block $(1,1)$ in (14), it can be seen that P guarantees the $\partial\mathcal{D}$-regularity of A. By virtue of Theorem 2, the root-distribution of A with respect to $\partial\mathcal{D}$ is given by the inertia of P. Since A_c never becomes $\partial\mathcal{D}$-singular over $\mathcal{B}(\rho)$, the inertia of P gives the root-distribution of A_c over $\mathcal{B}(\rho)$. □

The value of ρ obtained when minimizing γ under LMI constraints (14) is the largest acceptable value of ρ. Thus, it is what can be called the complex $\partial\mathcal{D}$-regularity radius. It will be denoted by $\varrho_{\partial\mathcal{D}}$ in what follows.

Now come back to region $\mathcal{D}$ defined as a combination of several regions $\mathcal{D}_k$ (see subsection 2.2). Referring to previous works on stability radii [13, 23], the complex $\mathcal{D}$-stability radius can also be defined as follows:

$$r_{\mathcal{D}} = \inf\{||\Delta||_2 \,|\, \Delta \in \mathbb{C}^{q\times r} : \ A + B\Delta C \text{ is } \mathcal{D}\text{-unstable}\}. \tag{19}$$

The complex $\partial\mathcal{D}$-regularity of a complex matrix A (not necessarily $\mathcal{D}$-stable) is here defined by

$$\varrho_{\partial\mathcal{D}} = \inf\{||\Delta||_2 \,|\, \Delta \in \mathbb{C}^{q\times r} : \ \lambda(A + B\Delta C) \cap \partial\mathcal{D} \neq \emptyset\}, \tag{20}$$

where $\partial\mathcal{D}$ is the boundary of $\mathcal{D}$. The formulation of $r_{\mathcal{D}}$ implicitly assumes that A is $\mathcal{D}$-stable. That is the basic difference from the formulation of $\varrho_{\partial\mathcal{D}}$ for which no assumption of nominal $\mathcal{D}$-stability is made. For this reason, it follows that

$$A \text{ is } \mathcal{D}\text{-stable } \Rightarrow \ \varrho_{\partial\mathcal{D}} = r_{\mathcal{D}}. \tag{21}$$

Also define the set $\partial\mathcal{U}$ by

$$\partial\mathcal{U} = \bigcup_{k=1}^{m} \partial\mathcal{D}_k. \tag{22}$$

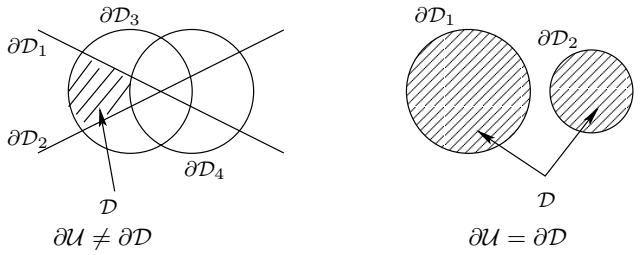


FIG. 1. $\varrho_{\partial\mathcal{D}} = \varrho_{\partial\mathcal{U}}$?

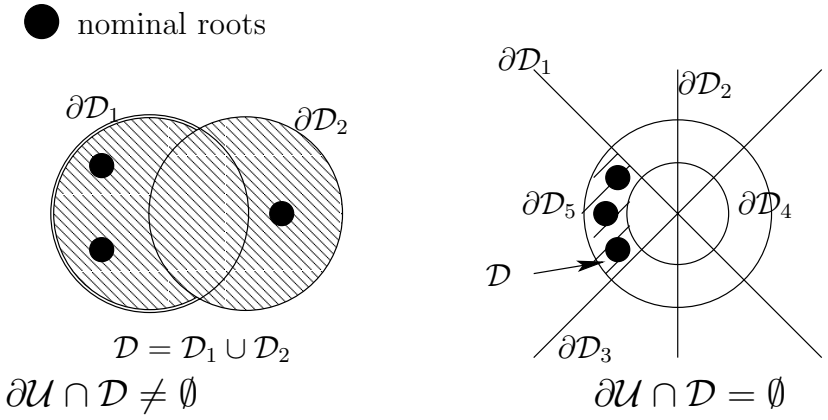


FIG. 2. $\varrho_{\partial\mathcal{U}} = r_{\mathcal{D}}$?

It is clear that $\partial\mathcal{D} \subset \partial\mathcal{U}$. Hence, A_c has no eigenvalue on $\partial\mathcal{D}$ if it has no eigenvalue on $\partial\mathcal{U}$, so we have

$$\varrho_{\partial\mathcal{U}} = \min_{k\in\{1,\ldots,m\}} \varrho_{\partial\mathcal{D}_k} \leq \varrho_{\partial\mathcal{D}}. \tag{23}$$

Moreover, if $\partial\mathcal{U} = \partial\mathcal{D}$, the above inequality becomes an equality. Besides, if A is assumed to be $\mathcal{D}$-stable and if $\partial\mathcal{U} \cap \mathcal{D} = \emptyset$, it can be assessed that

$$\varrho_{\partial\mathcal{U}} = r_{\mathcal{D}}. \tag{24}$$

To make the previous reasoning clear, it is illustrated by Figures 1 and 2 where region $\mathcal{D}$ is hatched.

Figure 1 shows the two cases where $\partial\mathcal{D}$ is either different from $\partial\mathcal{U}$ (then $\varrho_{\partial\mathcal{D}} \neq \varrho_{\partial\mathcal{U}}$) or equal to $\partial\mathcal{U}$ (then $\varrho_{\partial\mathcal{D}} = \varrho_{\partial\mathcal{U}}$). Figure 2 highlights the fact that if A is $\mathcal{D}$-stable, and if $\partial\mathcal{U} \cap \mathcal{D}$ is empty, then $\varrho_{\partial\mathcal{U}}$ equals $r_{\mathcal{D}}$.

From the previous reasoning, the next theorem is deduced.

THEOREM 4. *Let A_c, $\mathcal{D}$, $\partial\mathcal{D}$ be, respectively, an uncertain matrix as defined in* (1)*, a clustering region as defined in subsection* 2.2*, and its boundary. Then A_c is robustly $\partial\mathcal{D}$-regular against $\mathcal{B}(\rho)$ if there exist m Hermitian matrices P_k, $k = 1, \ldots, m$, such that*

$$\mathcal{Q}_k(P_k, \gamma) < 0 \quad \forall k \in \{1, \ldots, m\}, \tag{25}$$

where

$$\mathcal{Q}_k(P_k,\gamma)=\begin{bmatrix} r_{k_{00}}P_k+(r_{k_{10}}P_kA)^H+r_{k_{11}}A'PA+C'C & r_{k_{10}}PB+r_{k_{11}}A'PB \\ r'_{k_{10}}B'P+r_{k_{11}}B'PA & r_{k_{11}}B'PB-\gamma\mathbb{I}_q \end{bmatrix}$$

and $\gamma=\rho^{-2}$.

In this event, A_c keeps the same number of eigenvalues inside $\mathcal{D}$ for any $\Delta\in\mathcal{B}(\rho)$.

Moreover, if the boundary of $\mathcal{D}$ can be identified with the union of the boundaries of all subregions (i.e., $\partial\mathcal{D}=\partial\mathcal{U}$), then condition (25) *is also necessary.*

Proof. LMI system (25) is found feasible if and only if A_c is $\partial\mathcal{D}_k$-regular for any k in $\{1,\dots,m\}$, by virtue of Theorem 3. Hence, it is equivalent to $\partial\mathcal{U}$-regularity. Another consequence of Theorem 3 is that A_c has the same root-distribution as A with respect to $\partial\mathcal{D}_k$ (and a fortiori with respect to $\partial\mathcal{D}$). This proves the first and second statements in Theorem 4. Besides, if $\partial\mathcal{U}=\partial\mathcal{D}$ as in the right part of Figure 1, then $\partial\mathcal{U}$-regularity is equivalent to $\partial\mathcal{D}$-regularity, which proves the third statement. ∎

In light of this theorem, the following statements, which can be seen as corollaries, can be formulated:

- γ can be minimized under LMI constraints (25) down to $\gamma^\star$, and then $\rho^\star=(\gamma^\star)^{-1/2}$ actually equals $\varrho_{\partial\mathcal{U}}$. It is the largest robust $\partial\mathcal{D}$-regularity bound provided by this approach. If A is $\mathcal{D}$-stable, then $\rho^\star$ is a robust $\mathcal{D}$-stability bound.
- If $\partial\mathcal{U}=\partial\mathcal{D}$, then $\rho^\star$ equals $\varrho_{\partial\mathcal{D}}$, the complex $\partial\mathcal{D}$-regularity radius.
- If A is $\mathcal{D}$-stable and if $\partial\mathcal{U}\cap\mathcal{D}=\emptyset$, then $\rho^\star$ equals both $\varrho_{\partial\mathcal{D}}$ and $r_{\mathcal{D}}$, the complex $\mathcal{D}$-stability radius.

5. Numerical illustration. In this section, we propose a simple illustration of the presented technique in order to highlight its relevance. Computations are performed on a PC Pentium 1.7 MHz with MATLAB LMI CONTROL TOOLBOX.

5.1. First example. This model is borrowed from [25]. The lateral dynamic of an aircraft is modeled by state and input matrices:

$$A_0=\begin{bmatrix} -0.3400 & 0.0517 & 0.0010 & -0.9970 & 0.0000 \\ 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ -2.6900 & 0.0000 & -1.1500 & 0.7380 & 0.0000 \\ 5.9100 & 0.0000 & 0.1380 & -0.5060 & 0.0000 \\ -0.3400 & 0.0517 & 0.0010 & 0.0031 & 0.0000 \end{bmatrix};$$

$$B_0=\begin{bmatrix} 0.0755 & 0.0000 & 0.0246 \\ 0.0000 & 0.0000 & 0.0000 \\ 4.4800 & 5.2200 & -0.7420 \\ -5.0300 & 0.0998 & 0.9848 \\ 0.0755 & 0.0000 & 0.0246 \end{bmatrix}.$$

A static state feedback control law associated with matrix

$$K=\begin{bmatrix} -3.9063 & -0.2869 & 0.0006 & -1.5109 & -1.8135 \\ 0.8077 & -2.4178 & -0.9356 & -0.2877 & -0.0296 \\ -21.5282 & -1.2212 & -0.0424 & -10.1518 & -11.1581 \end{bmatrix}$$

is applied to the previous model so that

$$\lambda(A=A_0+B_0K)=\{-0.5;-2\pm 2\mathbf{i};-3\pm 2\mathbf{i}\}. \tag{26}$$

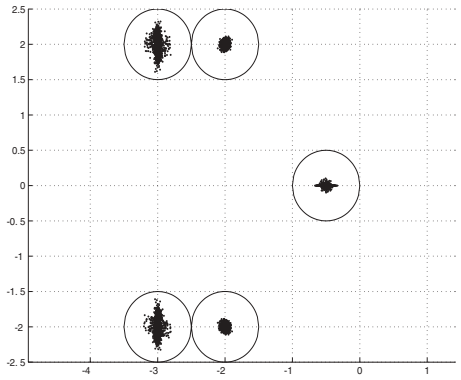


FIG. 3. *Pole migration of closed-loop aircraft model with norm-bounded complex uncertainty.*

Matrix A is assumed to be subject to an additive uncertainty, but no particular structure of this uncertainty is considered, so B and C are both assumed to equal $\mathbb{I}_5$ in (1) or, equivalently, $A_c = A + \Delta$. To analyze the robustness of the pole location in the presence of the uncertainty, $\mathcal{D}$ is chosen as the union of 5 discs $\mathcal{D}_k$, $k = 1, \ldots, 5$, centered around the nominal eigenvalues given by (26) and all of radius 0.5. In [2], a robust $\mathcal{D}$-stability bound is computed: $\rho^\star = 0.0729$. This value is improved owing to our approach. Indeed, since $\mathcal{D}_2$ and $\mathcal{D}_3$ (resp., $\mathcal{D}_4$ and $\mathcal{D}_5$) are symmetric to each other with respect to the real axis, it is only required to compute three values, namely, $\varrho_{\partial\mathcal{D}_1}$, $\varrho_{\partial\mathcal{D}_2} = \varrho_{\partial\mathcal{D}_3}$, $\varrho_{\partial\mathcal{D}_4} = \varrho_{\partial\mathcal{D}_5}$. After `2.58s` of computation time, we have

$$\rho^\star = \min_{k\in\{1,2,4\}} \varrho_{\partial\mathcal{D}_k} = \min\{0.3012; 0.1433; 0.1262\} = 0.1262 = r_{\partial\mathcal{D}}.$$

The bound of [2] is then increased by 73%. If many spectra are plotted for various values of complex Δ as in the first example, Figure 3 is obtained (to proceed, it suffices to generate many random complex matrices using standard MATLAB functions such as `rand` and to scale its entries so that its 2-norm is a random value in the range $[0; \rho^\star]$). It might seem that the bound is a bit conservative, but this example is not trivial, and since Δ is of dimension (5×5), it is hard to generate a random instance of Δ that corresponds to the worst case. However, the pole migration is not very far from $\partial\mathcal{D}$, especially from $\partial\mathcal{D}_4$ and $\partial\mathcal{D}_5$, and since A is $\mathcal{D}$-stable and since the five open subregions are disjoint (which implies that $\partial\mathcal{U} \cap \mathcal{D} = \emptyset$), $\rho^\star$ actually equals $r_{\mathcal{D}}$. Hence, in this case, the exact value of the complex $\mathcal{D}$-stability radius is obtained, without conservatism.

Computing $\rho^\star$ leads to deriving five Lyapunov matrices P_k, $i = 1, \ldots, 5$. Each matrix P_k is such that $\mathrm{In}(P_k) = [1,\ 4,\ 0]$ since, for any $\Delta \in \mathcal{B}(\rho^\star)$, A_c has one root inside $\mathcal{D}_k$ and four roots outside $\mathcal{D}_k$.

5.2. Second example. This example is simpler but involves a clustering region that is of interest for the analysis of some plant models. It is inspired by [27] and is relevant to the short-period approximation of the F-16 dynamics (see [27] for further details). The open-loop model is

$$A_0 = \begin{bmatrix} -1.0188 & 0.0905 \\ 4.0639 & -0.7701 \end{bmatrix}; \quad B_0 = \begin{bmatrix} -0.0021 \\ -0.1692 \end{bmatrix}.$$

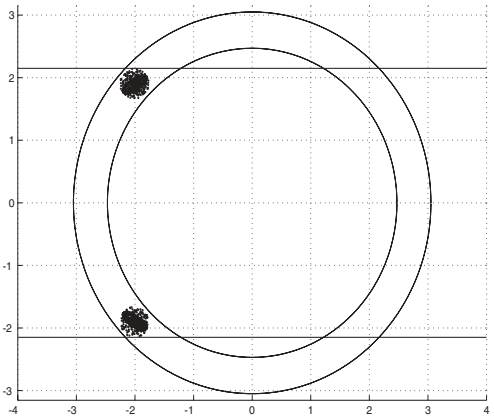


FIG. 4. *Pole migration of closed-loop model with norm-bounded complex uncertainty.*

A static state feedback control law associated with a matrix K is applied such that the spectrum of the closed-loop state matrix

$$A = A_0 + B_0 K = \begin{bmatrix} -1.1319 & 0.8790 \\ -4.9643 & -2.8683 \end{bmatrix}$$

is $\sigma(A) = \{-2 \pm 1.9\mathbf{i}\}$. We aim at analyzing the nonfragility induced by K. More precisely, assuming that K can be affected by an additive uncertainty Δ, the uncertain closed-loop state matrix A_c is given by expression (1) with $B = B_0$ and $C = \mathbb{I}_2$. The transient performances are supposed to be preserved when the spectrum of A_c remains in the clustering region $\mathcal{D}$ defined by the intersection of the OLHP, the ring centered around the origin and of interior and exterior radii 2.47 and 3.05, and the symmetric horizontal strip with a half width equaling 2.15. Such a region is the so-called good ride quality region of an aircraft [27]. The present approach can be used to compute the $\mathcal{D}$-stability radius in `1.52s`:

$$\rho^\star = \varrho_{\partial\mathcal{D}} = r_{\mathcal{D}} = 2.39.$$

Plotting the spectrum of A_c for many random instances of Δ highlights the relevance of our approach to computing this nonfragility criterion (see Figure 4).

6. Conclusion. In this paper, the concept of matrix $\partial\mathcal{D}$-regularity, i.e., absence of eigenvalues along a curve $\partial\mathcal{D}$ of the complex plane, was used to compute, through a Lyapunov approach, a robust $\mathcal{D}$-stability bound, i.e., a bound on an uncertainty affecting a matrix ensuring that the eigenvalues of the uncertain matrix remain inside a subregion $\mathcal{D}$ of the complex plane.

An original point in the paper is the wide class of allowed clustering regions since $\mathcal{D}$ can be a combination (i.e., the union and/or intersection) of several possibly nonsymmetric half planes, discs, and outsides of discs. Such an originality in the choice of the region is made possible by using Lyapunov matrices which are not necessarily positive definite but preserve their inertia over the uncertainty domain. When root-clustering in a single disc, a half plane, or the outside of a disc is to be analyzed, the Lyapunov matrix has to be strictly positive or negative definite. The computed bound proves to be not very conservative. When the boundaries of various subregions do not intersect $\mathcal{D}$, this bound turns out to be the exact complex $\mathcal{D}$-stability radius, that

is, the largest bound on a complex uncertainty preserving $\mathcal{D}$-stability, thus enabling efficient robust matrix root-location analysis. The bound is deduced from the solution of a very simple LMI problem.

As an extension of our work, a parametric structured uncertainty could therefore be very easily taken into account, especially through a polytopic description. Another challenge is the use of such an approach in a design context. More precisely, the question is to know if the inertia of various Lyapunov matrices P_k can be discerningly exploited to distribute the poles in various subregions while determining a control law.

Appendix. Two lemmas are here provided to the reader. They are exploited in the proof of Theorem 3.

LEMMA 1 (see [24]). *Given $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times m}$, and $M = M' \in \mathbb{C}^{(n+m)\times(n+m)}$, with $\det(\mathbf{i}\omega\mathbb{I} - A) \neq 0\ \forall\omega \in \mathbb{R}$ (i.e., A is $\mathfrak{I}$-regular), the following two statements are equivalent:*

(i)

$$\begin{bmatrix} (\mathbf{i}\omega\mathbb{I} - A)^{-1}B \\ \mathbb{I} \end{bmatrix}' M \begin{bmatrix} (\mathbf{i}\omega\mathbb{I} - A)^{-1}B \\ \mathbb{I} \end{bmatrix} < 0 \quad \omega \in \mathbb{R}. \tag{27}$$

(ii) *There exists a matrix $P = P' \in \mathbb{C}^{n\times n}$ such that*

$$M + \begin{bmatrix} A'P + PA & PB \\ B'P & \mathbb{O} \end{bmatrix} < 0. \tag{28}$$

The above lemma is known as the Kalman–Yakubovich–Popov lemma [30]. It is expressed here with complex matrices.

LEMMA 2 (see [24]). *Given $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times m}$, and $M = M' \in \mathbb{C}^{(n+m)\times(n+m)}$, with $\det(e^{\mathbf{i}\omega}\mathbb{I} - A) \neq 0\ \forall\omega \in \mathbb{R}$ (i.e., A is $\mathfrak{C}$-regular), the following two statements are equivalent:*

(i)

$$\begin{bmatrix} (e^{\mathbf{i}\omega}\mathbb{I} - A)^{-1}B \\ \mathbb{I} \end{bmatrix}' M \begin{bmatrix} (e^{\mathbf{i}\omega}\mathbb{I} - A)^{-1}B \\ \mathbb{I} \end{bmatrix} < 0 \quad \omega \in \mathbb{R} \cup \infty. \tag{29}$$

(ii) *There exists a matrix $P = P' \in \mathbb{C}^{n\times n}$ such that*

$$M + \begin{bmatrix} A'PA - P & PB \\ B'P & B'PB \end{bmatrix} < 0. \tag{30}$$

The above lemma is the discrete-time counterpart of the Kalman–Yakubovich–Popov lemma.

Acknowledgment. The authors thank an anonymous reviewer for the quality of his comments.

REFERENCES

[1] P. APKARIAN AND P. GAHINET, *A linear matrix inequality approach to H_∞ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.

[2] C. R. ASHOKKUMAR AND R. K. YEDAVALLI, *Eigenstructure perturbation analysis in disjointed domains for linear uncertain systems*, Internat. J. Control, 67 (1997), pp. 887–899.

[3] O. BACHELIER AND D. MEHDI, *Robust D_U-stability analysis*, Internat. J. Robust Nonlinear Control, 13 (2003), pp. 533–558.

[4] B. R. Barmish, *New Tools for Robustness of Linear Systems*, Macmillan, New York, 1994.
[5] V. D. Blondel and A. Megretski, eds., *Unsolved Problems in Mathematical Systems and Control Theory*, Princeton University Press, Princeton, NJ, 2004. Also available online from http://www.inma.ucl.ac.be/~blondel.
[6] S. Boyd and V. Balakrishnan, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its L_∞-norm*, Systems Control Lett., 15 (1990), pp. 1–17.
[7] N. A. Bruinsma and M. Steinbuch, *A fast algorithm to compute the H_∞-norm of a transfer function matrix*, Systems Control Lett., 14 (1990), pp. 287–293.
[8] M. Chilali and P. Gahinet, *H_∞ design with pole placement constraints: An LMI approach*, IEEE Trans. Automat. Control, 41 (1996), pp. 358–367.
[9] J. C. Doyle, *Analysis of feedback systems with structured uncertainties*, Proc. IEE-D, 129 (1982), pp. 242–250.
[10] S. Gutman and E. I. Jury, *A general theory for matrix root-clustering in subregions of the complex plane*, IEEE Trans. Automat. Control, 26 (1981), pp. 853–863.
[11] R. D. Hill, *Eigenvalue location using certain matrix functions and geometric curves*, Linear Algebra Appl., 16 (1977), pp. 83–91.
[12] D. Hinrichsen and B. Kelb, *Stability radii and spectral value sets for real matrix perturbations*, in Systems and Networks: Mathematical Theory and Applications, Vol. II, Invited and Contributed Papers, Akademie-Verlag, Berlin, 1994, pp. 217–220.
[13] D. Hinrichsen and A. J. Pritchard, *Stability radii of linear systems*, Systems Control Lett., 7 (1986), pp. 1–10.
[14] J. L. Howland, *Matrix equations and the separation of matrix eigenvalues*, J. Math. Anal. Appl., 33 (1971), pp. 683–691.
[15] E. I. Jury, *Inners and Stability of Dynamic Systems*, Wiley, New York, 1974.
[16] P. Khargonekar, I. R. Petersen, and K. Zhou, *Robust stabilization of uncertain linear systems: Quadratic stabilizability and H_∞ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356–361.
[17] A. Lyapunov, *Problème général de la stabilité du mouvement*, Ann. Fac. Sci. Toulouse, 1907. Translated into French from the original Russian text, Kharkov, 1892.
[18] A. Ostrowski and H. Schneider, *Some theorems on the inertia of general matrices*, J. Math. Anal. Appl., 4 (1962), pp. 72–84.
[19] A. Packard and J. Doyle, *Quadratic stability with real and complex perturbations*, IEEE Trans. Automat. Control, 35 (1990), pp. 198–201.
[20] R. V. Patel and M. Toda, *Quantitative measures of robustness for multivariable systems*, in Proceedings of the Joint Automatic Control Conference, TP-8A, American Institute of Chemical Engineers, San Francisco, CA, 1980.
[21] D. Peaucelle, D. Arzelier, O. Bachelier, and J. Bernussou, *A new robust D-stability condition for real convex polytopic uncertainty*, Systems Control Lett., 40 (2000), pp. 21–30.
[22] I. R. Petersen, *A stabilization algorithm for a class of uncertain linear systems*, Systems Control Lett., 8 (1987), pp. 351–357.
[23] L. Qiu, B. Bernhardsson, A. Rantzer, E. J. Davison, P. M. Young, and J. C. Doyle, *A formula for computation of the real stability radius*, Automatica J. IFAC, 31 (1995), pp. 879–890.
[24] A. Rantzer, *On the Kalman-Yakubovich-Popov lemma*, Systems Control Lett., 28 (1996), pp. 7–10.
[25] K. M. Sobel and F. J. Lallman, *Eigenstructure assignment for the control of highly augmented aircraft*, J. Guidance Control Dynamics, 12 (1989), pp. 318–324.
[26] P. Stein, *Some general theorems on iterants*, J. Research of the National Bureau of Standards, 48 (1952), pp. 82–83.
[27] S. Wang, *Robust pole clustering in a good ride quality region of aircraft for matrices with structured unertainties*, Automatica J. IFAC, 39 (2003), pp. 525–532.
[28] S.-G. Wang, *Comments on "Perturbation bounds for root-clustering of linear systems in a specified second order subregion,"* IEEE Trans. Automat. Control, 41 (1996), pp. 766–767.
[29] S.-G. Wang and L. S. Shieh, *A general theory for analysis and design of robust pole clustering in subregions of the complex plane*, in Proceedings of the American Control Conference, Baltimore, MD, 1994, pp. 627–631.
[30] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.
[31] Y. K. Yedavalli, *Robust root clustering for linear uncertain systems using generalized Lyapunov theory*, Automatica J. IFAC, 29 (1993), pp. 237–240.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 1094–1119

THE PONTRYAGIN MAXIMUM PRINCIPLE AND TRANSVERSALITY CONDITIONS FOR A CLASS OF OPTIMAL CONTROL PROBLEMS WITH INFINITE TIME HORIZONS*

SERGEI M. ASEEV† AND ARKADY V. KRYAZHIMSKIY†

Abstract. This paper suggests some further developments in the theory of first-order necessary optimality conditions for problems of optimal control with infinite time horizons. We describe an approximation technique involving auxiliary finite-horizon optimal control problems and use it to prove new versions of the Pontryagin maximum principle. Special attention is paid to the behavior of the adjoint variables and the Hamiltonian. Typical cases, in which standard transversality conditions hold at infinity, are described. Several significant earlier results are generalized.

Key words. optimal control, infinite horizon, Pontryagin maximum principle, transversality conditions, optimal economic growth

AMS subject classifications. 49K15, 91B62

DOI. 10.1137/S0363012903427518

1. Introduction. We deal with the following infinite-horizon optimal control problem (P):

$$\dot{x}(t) = f(x(t), u(t)), \qquad u(t) \in U; \tag{1.1}$$

$$x(0) = x_0; \tag{1.2}$$

$$\text{maximize } J(x,u) = \int_0^\infty e^{-\rho t} g(x(t), u(t))dt. \tag{1.3}$$

Here $x(t) = (x^1(t), \dots, x^n(t)) \in \mathbb{R}^n$ and $u(t) = (u^1(t), \dots, u^m(t)) \in \mathbb{R}^m$ are the current values of the system's states and controls; U is a nonempty convex compactum in $\mathbb{R}^m$; x_0 is a given initial state; and $\rho \geq 0$ is a discount parameter. The functions $f : G \times U \mapsto \mathbb{R}^n$, $g : G \times U \mapsto \mathbb{R}^1$, the matrix $\partial f/\partial x = (\partial f^i/\partial x^j)_{i,j=1,\dots,n}$, and the gradient $\partial g/\partial x = (\partial g/\partial x^1, \dots, \partial g/\partial x^n)$ are assumed to be continuous on $G \times U$. Here G is an open set in $\mathbb{R}^n$ such that $x_0 \in G$. As usual an admissible control in system (1.1) is identified with an arbitrary measurable function $u : [0, \infty) \mapsto U$. A trajectory corresponding to a control u is a Carathéodory solution x to (1.1), which satisfies the initial condition (1.2). We assume that, for any control u, a trajectory x corresponding to u exists on $[0, \infty)$ and takes values in G (due to the continuous differentiability of f, the trajectory x is unique). Any pair (u, x), where u is a control and x the trajectory corresponding to u, will be called an admissible pair.

Problems of this type naturally arise in the studies on optimization of economic growth (see [1], [2], [14], [23], [27], [33], [39]). Progress in this field of economics was initiated by Ramsey in the 1920s [35].

*Received by the editors May 12, 2003; accepted for publication (in revised form) February 20, 2004; published electronically November 9, 2004. This work was supported by the Fujitsu Research Institute (IIASA-FRI contract 01-109).

http://www.siam.org/journals/sicon/43-3/42751.html

†International Institute for Applied Systems Analysis, Schlossplatz 1, Laxenburg, A-2361, Austria and Steklov Institute of Mathematics, Gubkina str. 8, Moscow, 119991, Russia (aseev@iiasa.ac.at, aseev@mi.ras.ru; kryazhim@mtu-net.ru). The first author was partially supported by the Russian Foundation for Basic Research (project 99-01-01051). The second author was partially supported by the Russian Foundation for Basic Research (project 03-01-00737).

Our basic assumptions are the following.

(A1) There exists a $C \geq 0$ such that

$$\langle x, f(x,u)\rangle \leq C(1+\|x\|^2) \quad \text{for all} \quad x \in G \quad \text{and all} \quad u \in U.$$

(A2) For each $x \in G$, the function $u \mapsto f(x,u)$ is affine, i.e.,

$$f(x,u) = f_0(x) + \sum_{i=1}^{m} f_i(x)u^i \quad \text{for all} \quad x \in G \quad \text{and all} \quad u \in U,$$

where $f_i : G \mapsto \mathbb{R}^n$, $i = 0, 1, \ldots, m$, are continuously differentiable.

(A3) For each $x \in G$, the function $u \mapsto g(x,u)$ is concave.

(A4) There exist positive-valued functions μ and ω on $[0,\infty)$ such that $\mu(t) \to 0$, $\omega(t) \to 0$ as $t \to \infty$, and for any admissible pair (u,x),

$$e^{-\rho t} \max_{u \in U} |g(x(t),u)| \leq \mu(t) \quad \text{for all} \quad t > 0;$$

$$\int_T^\infty e^{-\rho t} |g(x(t),u(t))| dt \leq \omega(T) \quad \text{for all} \quad T > 0.$$

Assumption (A1) is conventionally used in existence theorems in the theory of optimal control (see [19], [22]). Assumptions (A2) and (A3) imply that problem (P) is "linear-convex" in control; the "linear-convex" structure is important for the implementation of approximation techniques. The second condition in (A4) implies that the integral (1.3) converges absolutely for any admissible pair (u,x), which excludes any ambiguity in interpreting problem (P). As shown in [13, Theorem 3.6], assumptions (A1)–(A4) guarantee the existence of an admissible optimal pair in problem (P).

In this paper, we develop first-order necessary optimality conditions for problem (P). Note that, for infinite-horizon optimal control problems without a discounting factor ($\rho = 0$), the Pontryagin maximum principle was stated in [34]. For problems involving a positive discounting factor ($\rho > 0$), a general statement on the Pontryagin maximum principle was given in [24]. However, both statements establish the "core" relations of the Pontryagin maximum principle only and do not suggest any analogue of the transversality conditions, which constitute an immanent component of the Pontryagin maximum principle for classical finite-horizon optimal control problems with nonconstrained terminal states. The issue of transversality conditions for problem (P) is the focus of our study.

Introduce the Hamilton–Pontryagin function $\mathcal{H} : G \times [0,\infty) \times U \times \mathbb{R}^n \times \mathbb{R}^1 \mapsto \mathbb{R}^1$ and the Hamiltonian $H : G \times [0,\infty) \times \mathbb{R}^n \times \mathbb{R}^1 \mapsto \mathbb{R}^1$ for problem (P):

$$\mathcal{H}(x,t,u,\psi,\psi^0) = \langle f(x,u), \psi\rangle + \psi^0 e^{-\rho t} g(x,u);$$

$$H(x,t,\psi,\psi^0) = \sup_{u \in U} \mathcal{H}(x,t,u,\psi,\psi^0).$$

The Pontryagin maximum principle involves an admissible pair (u_*, x_*) and a pair (ψ, ψ^0) of adjoint variables associated with (u_*, x_*); here ψ is a solution to the adjoint equation

$$\dot{\psi}(t) = -\left[\frac{\partial f(x_*(t), u_*(t))}{\partial x}\right]^* \psi(t) - \psi^0 e^{-\rho t} \frac{\partial g(x_*(t), u_*(t))}{\partial x} \tag{1.4}$$

on $[0,\infty)$, and ψ^0 is a nonnegative real; (ψ,ψ^0) is said to be nontrivial if

$$\|\psi(0)\|+\psi^0>0. \tag{1.5}$$

We shall use the following definition. We shall say that an admissible pair (u_*,x_*) satisfies the core Pontryagin maximum principle (in problem (P)), together with a pair (ψ,ψ^0) of adjoint variables associated with (u_*,x_*), if (ψ,ψ^0) is nontrivial and the following maximum condition holds:

$$\mathcal{H}(x_*(t),t,u_*(t),\psi(t),\psi^0)=H(x_*(t),t,\psi(t),\psi^0)\quad\text{for a.a.}\quad t\geq 0. \tag{1.6}$$

Of special interest is the case where problem (P) is not abnormal, i.e., when the Lagrange multiplier ψ^0 in the core Pontryagin maximum principle does not vanish. In this case we do not lose generality if we set $\psi^0=1$. Accordingly, we define the normal-form Hamilton–Pontryagin function $\tilde{\mathcal{H}}:G\times[0,\infty)\times U\times\mathbb{R}^n\mapsto\mathbb{R}^1$ and the normal-form Hamiltonian $\tilde{H}:G\times[0,\infty)\times\mathbb{R}^n\mapsto\mathbb{R}^1$ as follows:

$$\tilde{\mathcal{H}}(x,t,u,\psi)=\mathcal{H}(x,t,u,\psi,1)=\langle f(x,u),\psi\rangle+e^{-\rho t}g(x,u);$$

$$\tilde{H}(x,t,\psi)=H(x,t,\psi,1)=\sup_{u\in U}\tilde{\mathcal{H}}(x,t,u,\psi).$$

Given an admissible pair (u_*,x_*), introduce the normal-form adjoint equation

$$\dot{\psi}(t)=-\left[\frac{\partial f(x_*(t),u_*(t))}{\partial x}\right]^*\psi(t)-e^{-\rho t}\frac{\partial g(x_*(t),u_*(t))}{\partial x}. \tag{1.7}$$

Any solution ψ to (1.7) on $[0,\infty)$ will be called an adjoint variable associated with (u_*,x_*). We shall say that an admissible pair (u_*,x_*) satisfies the normal-form core Pontryagin maximum principle together with an adjoint variable ψ associated with (u_*,x_*) if the following normal-form maximum condition holds:

$$\tilde{\mathcal{H}}(x_*(t),t,u_*(t),\psi(t))=\tilde{H}(x_*(t),t,\psi(t))\quad\text{for a.a.}\quad t\geq 0. \tag{1.8}$$

In the context of problem (P), [24] states the following (see also [17]).

THEOREM 1. *If an admissible pair (u_*,x_*) is optimal in problem* (P)*, then (u_*,x_*) satisfies relations* (1.4)–(1.6) *of the core Pontryagin maximum principle together with some pair (ψ,ψ^0) of adjoint variables associated with (u_*,x_*).*

Qualitatively, this formulation is weaker than the corresponding statement known for finite-horizon optimal control problems with nonconstrained terminal states. Indeed, consider the following finite-horizon counterpart of problem (P).

Problem (P_T):

$$\dot{x}(t)=f(x(t),u(t)),\qquad u(t)\in U;$$

$$x(0)=x_0;$$

$$\text{maximize } J_T(x,u)=\int_0^T e^{-\rho t}g(x(t),u(t))dt;$$

here $T>0$ is a fixed positive real. The classical theory [34] says that if an admissible pair (u_*,x_*) is optimal in problem (P_T), then (u_*,x_*) satisfies the core Pontryagin

maximum principle together with some pair (ψ, ψ^0) of adjoint variables associated with (u_*, x_*), and, moreover, (ψ, ψ^0) satisfies the transversality conditions

$$\psi^0 = 1, \qquad \psi(T) = 0. \tag{1.9}$$

In Theorem 1 any analogue of the transversality conditions (1.9) is missing.

There were numerous attempts to find specific situations in which the infinite-horizon Pontryagin maximum principle holds together with additional boundary conditions at infinity (see [12], [15], [16], [21], [26], [31], [36], [38]). However, the major results were established under rather severe assumptions of linearity or full convexity, which made it difficult to apply them to particular meaningful problems (see, e.g., [28] discussing the application of the Pontryagin maximum principle to a particular infinite-horizon optimal control problem).

In this paper we follow the approximation approach suggested in [9], [10], and [11]. We approximate problem (P) by a sequence of finite-horizon optimal control problems $\{(\mathrm{P}_k)\}$ $(k = 1, 2, \ldots)$ whose horizons go to infinity. Problems (P_k) $(k = 1, 2, \ldots)$ impose no constraints on the terminal states; in this sense, they inherit the structure of problem (P); on the other hand, problems (P_k) are not plain "restrictions" of problem (P) to finite intervals like problem (P_T): the goal functionals in problems (P_k) include special penalty terms associated with a certain control optimal in problem (P). This approach allows us to find limit forms of the classical transversality conditions for problems (P_k) as $k \to \infty$ and formulate conditions that complement the core Pontryagin maximum principle and hold with a necessity for every admissible pair optimal in problem (P). The results presented here generalize [9], [10], [11], and [12].

Earlier, a similar approximation approach was used to derive necessary optimality conditions for various nonclassical optimal control problems (see, e.g., [3], [4], [5], [7], [32], and also survey [6]). Based on relevant approximation techniques and the methodology presented here, one can extend the results of this paper to more complex infinite-horizon problems of optimal control (e.g., problems with nonsmooth data). In this paper, our primary goal is to show how the approximation approach allows us to resolve the major singularity emerging due to the unboundedness of the time horizon. Therefore, we restrict our consideration to the relatively simple nonlinear infinite-horizon problem (P), which is smooth, "linear-convex" in control, and free from any constraints on the system's states.

Finally, we note that the suggested approximation methodology, appropriately modified, can be used directly in analysis of particular nonstandard optimal control problems with infinite time horizons (see, e.g., [8]).

2. Transversality conditions: Counterexamples. Considering problem (P) as the "limit" of finite-horizon problems (P_T) whose horizons T tend to infinity, one can expect the following "natural" transversality conditions for problem (P):

$$\psi^0 = 1, \qquad \lim_{t\to\infty} \psi(t) = 0; \tag{2.1}$$

here (ψ, ψ^0) is a pair of adjoint variables satisfying the core Pontryagin maximum principle together with an admissible pair (u_*, x_*) optimal in problem (P). The relations

$$\psi^0 = 1, \qquad \lim_{t\to\infty} \langle \psi(t), x_*(t) \rangle = 0 \tag{2.2}$$

represent alternative transversality conditions for problem (P), which are frequently used in economic applications (see, e.g., [14]).

The interpretation of (2.2) as transversality conditions for problem (P) is also motivated by Arrow's statement on sufficient conditions of optimality (see [1], [2], and [36]), which (under some additional assumptions) asserts that if (2.2) holds for an admissible pair (u_*, x_*) and a pair (ψ, ψ^0) of adjoint variables, jointly satisfying the core Pontryagin maximum principle, then (u_*, x_*) is optimal in problem (P), provided the superposition $H(x, t, \psi(t), \psi^0)$ is concave in x. Another type of transversality condition formulated in terms of stability theory was proposed in [38]. In [12], global behavior of the adjoint variable associated with an optimal admissible pair was characterized in terms of appropriate integral functionals. In this paper, we concentrate on the derivation of pointwise transversality conditions of types (2.1) and (2.2).

Note that, generally, for infinite-horizon optimal control problems neither transversality condition (2.1) nor (2.2) is valid. For the case of no discounting ($\rho = 0$), illustrating counterexamples were given in [24] and [37], and for problems with discounting ($\rho > 0$), some examples were given in [12] and [31]. In particular, [31] presents an example showing that an infinite-horizon optimal control problem with a positive discount can be abnormal; i.e., in the core Pontryagin maximum principle, the Lagrange multiplier ψ^0 may necessarily vanish (which contradicts both (2.1) and (2.2)).

Here, we provide further counterexamples for problem (P) in the case where discount parameter ρ is positive.

Example 1 shows that for problem (P), the limit relation in (2.1) may be violated, whereas the alternative transversality conditions (2.2) may hold.

Example 1. Consider the optimal control problem

$$\dot{x}(t) = u(t) - x(t), \qquad u(t) \in U = [0, 1];$$

$$x(0) = \frac{1}{2};$$

$$\text{maximize } J(x, u) = \int_0^\infty e^{-t} \ln \frac{1}{x(t)} dt.$$

We set $G = (0, \infty)$ and treat the above problem as problem (P). Assumptions (A1)–(A4) are, obviously, satisfied. For an arbitrary trajectory x, we have $e^{-t}/2 \le x(t) < 1$ for all $t \ge 0$. Hence, (u_*, x_*), where $u_*(t) \overset{a.e.}{=} 0$ and $x_*(t) = e^{-t}/2$ for all $t \ge 0$, is the unique optimal admissible pair. The Hamilton–Pontryagin function is given by

$$\mathcal{H}(x, t, u, \psi, \psi^0) = (u - x)\psi - \psi^0 e^{-t} \ln x.$$

Let (ψ, ψ^0) be an arbitrary pair of adjoint variables such that (u_*, x_*) satisfies the core Pontryagin maximum principle together with (ψ, ψ^0). The adjoint equation (1.4) has the form

$$\dot{\psi}(t) = \psi(t) + \psi^0 e^{-t} \frac{1}{x_*(t)} = \psi + 2\psi^0,$$

and the maximum condition (1.6) implies

$$\psi(t) \le 0 \quad \text{for all} \quad t \ge 0. \tag{2.3}$$

Assume $\psi^0 = 0$. Then $\psi(0) < 0$ and $\psi(t) = e^t \psi(0) \to -\infty$ as $t \to \infty$; i.e., the limit relation in (2.1) does not hold. Let $\psi^0 > 0$. Without loss of generality (or multiplying

both ψ and ψ^0 by $1/\psi^0$), we assume $\psi^0 = 1$. Then $\psi(t) = (\psi(0) + 2)e^t - 2$. By (2.3), only two cases are admissible: (a) $\psi(0) = -2$ and (b) $\psi(0) < -2$. In case (a) $\psi(t) \equiv -2$, and in case (b) $\psi(t) \to -\infty$ as $t \to \infty$. In both situations the limit relation in (2.1) is violated. Note that $\psi(t) \equiv -2$ ($t \geq 0$) and $\psi^0 = 1$ satisfy (2.2).

The next example is complementary to Example 1; it shows that for problem (P), the limit relation in (2.2) may be violated, whereas (2.1) may hold.

Example 2. Consider the following optimal control problem:

$$\dot{x}(t) = u(t), \qquad u(t) \in U = \left[\frac{1}{2}, 1\right]; \tag{2.4}$$

$$x(0) = 0;$$

$$\text{maximize } J(x,u) = \int_0^\infty e^{-t}(1 + \gamma(x(t)))u(t)dt. \tag{2.5}$$

Here γ is a nonnegative continuously differentiable real function such that

$$I = \int_0^\infty e^{-t}\gamma(t)dt < \infty. \tag{2.6}$$

We set $G = \mathbb{R}^1$. Clearly, assumptions (A1)–(A3) are satisfied. Below, we specify the form of γ and show that assumption (A4) is satisfied too.

The admissible pair (u_*, x_*), where $u_*(t) \overset{a.e.}{=} 1$ and $x_*(t) = t$ for all $t \geq 0$, is optimal. Indeed, let (u, x) be an arbitrary admissible pair. Observing (2.4), we find that $\dot{x}(t) > 0$ for a.a. $t \geq 0$. Taking $\tau(t) = x(t)$ for a new integration variable in (2.5), we get $d\tau = u(t)dt$ and

$$t(\tau) = \int_0^\tau \frac{1}{u(t(s))}ds \quad \text{for all} \quad \tau \geq 0.$$

As far as

$$\int_0^\tau \frac{1}{u(t(s))}ds \geq \tau,$$

we get

$$J(x,u) = \int_0^\infty e^{-t}(1 + \gamma(x(t)))u(t)dt = \int_0^\infty e^{-\int_0^\tau \frac{1}{u(t(s))}ds}(1 + \gamma(\tau))d\tau$$

$$\leq \int_0^\infty e^{-\tau}(1 + \gamma(\tau))d\tau = J(u_*, x_*).$$

Hence, (u_*, x_*) is an optimal admissible pair. It is easy to see that there are no other optimal admissible pairs. The Hamilton–Pontryagin function has the form

$$\mathcal{H}(x,t,u,\psi,\psi^0) = u\psi + \psi^0 e^{-t}(1 + \gamma(x))u.$$

Let (ψ, ψ^0) be an arbitrary pair of adjoint variables such that (u_*, x_*) satisfies the core Pontryagin maximum principle together with (ψ, ψ^0). The adjoint equation (1.4) has the form

$$\dot{\psi}(t) = -\psi^0 \dot{\gamma}(t)e^{-t}.$$

If $\psi^0 = 0$, then the maximum condition (1.6) implies $\psi(t) \equiv \psi(0) > 0$; hence, $\psi(t)x_*(t) = \psi(0)t \to \infty$ as $t \to \infty$, and the limit relation in (2.2) is violated.

Suppose $\psi^0 > 0$, or, equivalently, $\psi^0 = 1$. Then, due to (1.4), we have

$$\psi(t) = \psi(0) - \int_0^t \dot{\gamma}(s)e^{-s}ds.$$

The limit relation in (2.2) has the form $\lim_{t\to\infty} t\psi(t) = 0$. Let us show that one can define γ so that the latter relation is violated; i.e., for any $\psi(0) \in \mathbb{R}^1$,

$$p(t) \not\to 0 \quad \text{as} \quad t \to \infty, \tag{2.7}$$

where $p(t) = t\psi(t)$. We represent $p(t)$ as follows:

$$\begin{aligned} p(t) &= t\psi(0) - t\int_0^t \dot{\gamma}(s)e^{-s}ds = t\psi(0) - t\left[\gamma(s)e^{-s}|_0^t + \int_0^t \gamma(s)e^{-s}ds\right] \\ &= t\psi(0) - t\gamma(t)e^{-t} + t\gamma(0) - tI(t), \end{aligned}$$

where

$$I(t) = \int_0^t \gamma(s)e^{-s}ds.$$

Introducing $\nu(t) = \gamma(t)e^{-t}$, rewrite

$$I(t) = \int_0^t \nu(s)ds; \tag{2.8}$$

$$p(t) = t\psi(0) - t\nu(t) + t\nu(0) - tI(t). \tag{2.9}$$

Due to (2.6),

$$\lim_{t\to\infty} I(t) = I. \tag{2.10}$$

Now let us specify the form of ν. For each natural k, we fix a positive $\varepsilon_k < 1/2$ and denote by Δ_k the ε_k-neighborhood of k. Clearly, $\Delta_k \cup \Delta_j = \emptyset$ for $k \neq j$. We set

$$\nu(k) = \frac{1}{k} \quad \text{for} \quad k = 1, 2, \ldots;$$

$$\nu(t) = 0 \quad \text{for} \quad t \notin \cup_{k=1}^{\infty} \Delta_k;$$

$$\nu(t) \in \left[0, \frac{1}{k}\right] \quad \text{for} \quad t \in \Delta_k \quad (k = 1, 2, \ldots).$$

Moreover, we require that

$$\sum_{k=j}^{\infty} \int_{\Delta_k} \nu(t)dt \le \frac{1}{j^2}. \tag{2.11}$$

This can be achieved, for example, by letting $\frac{2\varepsilon_k}{k} \le \frac{a_k}{k^2}$, where $\sum_{k=1}^{\infty} a_k = 1$, $a_k > 0$. Indeed, in this case

$$\sum_{k=j}^{\infty} \int_{\Delta_k} \nu(t)dt \le \sum_{k=j}^{\infty} \frac{2\varepsilon_k}{k} \le \sum_{k=j}^{\infty} \frac{a_k}{k^2} \le \frac{1}{j^2} \sum_{k=j}^{\infty} a_k \le \frac{1}{j^2};$$

i.e., (2.11) holds. Note that, for $j = 1$, the left-hand side in (2.11) equals I (see (2.6)); thus, (2.11) implies that assumption (2.6) holds.

Another fact following from (2.11) is that

$$\lim_{t\to\infty} t(I - I(t)) = 0. \tag{2.12}$$

Indeed, by (2.8), $I(j+\varepsilon_j) = \sum_{k=1}^{j} \int_{\Delta_k} \nu(t)dt$; hence, due to (2.11),

$$I - I(j+\varepsilon_j) = \sum_{k=j+1}^{\infty} \int_{\Delta_k} \nu(t)dt \le \frac{1}{(j+1)^2}.$$

For $t \in [j+\varepsilon_j,\, j+1+\varepsilon_{j+1}]$, we have $I(j+\varepsilon_j) \le I(t) \le I$; therefore, for $t \ge 1$,

$$0 \le I - I(t) \le \frac{1}{(j+1)^2} \le \frac{1}{(t-\varepsilon_{j+1})^2} \le \frac{1}{(t-1/2)^2},$$

which yields (2.12). The given definition of ν is equivalent to defining γ by

$$\begin{gathered} \gamma(k) = \frac{e^k}{k} \quad \text{for} \quad k = 1, 2, \ldots; \\ \gamma(t) = 0 \quad \text{for} \quad t \notin \cup_{k=1}^{\infty} \Delta_k; \\ \gamma(t) \in \left[0, \frac{e^k}{k}\right] \quad \text{for} \quad t \in \Delta_k \quad (k = 1, 2, \ldots) \end{gathered} \tag{2.13}$$

and requiring (2.11). Let us show that assumption (A4) is satisfied. Let (u, x) be an arbitrary admissible pair. By (2.4), $t/2 \le x(t) \le t$ for all $t \ge 0$. Hence, by the definition of ν, we have $\nu(x(t)) \le \left(\frac{t}{2} - 1\right)^{-1} = \frac{2}{(t-2)}$ for all $t > 2$. Hence,

$$0 \le e^{-\rho t} \max_{u\in U}[(1 + \gamma(x(t))u] \le \mu(t) = e^{-\rho t} + \frac{2}{(t-2)} \to 0 \quad \text{as} \quad t \to \infty.$$

Thus, the first condition in (A4) holds. Furthermore, introducing the integration variable $\tau(t) = x(t)$, we get

$$\begin{aligned} \int_T^{\infty} e^{-t}(1 + \gamma(x(t)))u(t)dt &= \int_{x(T)}^{\infty} e^{-\int_0^{\tau} \frac{1}{u(t(s))}ds}(1 + \gamma(\tau))d\tau \\ &\le \int_{x(T)}^{\infty} e^{-\tau}(1 + \gamma(\tau))d\tau \le \omega(T) \\ &= \int_{\frac{T}{2}}^{\infty} e^{-t}(1 + \gamma(t))dt \to 0 \quad \text{as} \quad T \to \infty. \end{aligned}$$

Hence, the second condition in (A4) holds. We stated the validity of assumption (A4).

By the definition of γ, for $t \in \Delta_k$, $k = 1, 2, \ldots$, we have

$$0 \le t\nu(t) \le \frac{k + \varepsilon_k}{k} \le 1 + \frac{1}{k}.$$

Hence,

$$0 \le t\nu(t) \le 2 \quad \text{for all} \quad t \ge 0; \tag{2.14}$$

i.e., the function $t\nu(t)$ is bounded. Furthermore, $k\nu(k) = 1$, and due to (2.13) for any sequence $t_k \to \infty$ such that $t_k \in [k, k+1] \setminus (\Delta_k \cup \Delta_{k+1})$, we have $t_k\nu(t_k) = 0$. Therefore, $\lim_{t\to\infty} t\nu(t)$ does not exist.

Using $\nu(0) = 0$, we specify (2.9) as

$$p(t) = t\psi(0) - t\nu(t) - tI(t). \tag{2.15}$$

If $\psi(0) > I$, then, in view of (2.10), $\lim_{t\to\infty} t(\psi(0) + I(t)) = \infty$, which implies $\lim_{t\to\infty} p(t) = \infty$, since $t\nu(t)$ is bounded. Similarly, we find that if $\psi(0) < I$, then $\lim_{t\to\infty} p(t) = -\infty$. Let, finally, $\psi(0) = I$. Then,

$$\lim_{t\to\infty} t(\psi(0) - I(t)) = \lim_{t\to\infty} t(I - I(t)) = 0,$$

as follows from (2.12). Thus, in the right-hand side of (2.15) the sum of the first and third terms has the zero limit at infinity, whereas the second term, $t\nu(t)$, has no limit at infinity, as we noted earlier. Consequently, $p(t)$, the left-hand side in (2.15), has no limit at infinity. We showed that (2.7) holds for every $\psi(0) \in \mathbb{R}^1$.

Thus, the limit relation in the transversality conditions (2.2) is violated. Note that setting $\psi^0 = 1$ and $\psi(0) = I$, we make the adjoint variable ψ satisfy the transversality conditions (2.1). Indeed, in this case $\psi(t) = p(t)/t = \psi(0) - I - \nu(t)$ for all $t > 0$, and the conditions $\psi(0) = I$ and (2.14) imply that $\psi(t) \to 0$ as $t \to \infty$.

Examples 1 and 2 show that assumptions (A1)–(A4) are insufficient for the validity of the core Pontryagin maximum principle together with the transversality conditions (2.1) or (2.2) as necessary conditions of optimality in problem (P). Below, we find mild additional assumptions that guarantee that necessary conditions of optimality in problem (P) include the core Pontryagin maximum principle and transversality conditions of type (2.1) or of type (2.2).

3. Basic constructions. In this section, we define a sequence of finite-horizon optimal control problems $\{(\mathrm{P}_k)\}$ $(k = 1, 2, \ldots)$ with horizons $T_k \to \infty$; we treat problems (P_k) as approximations to the infinite-horizon problem (P).

Let us describe the data defining problems (P_k) $(k = 1, 2, \ldots)$. Given a control u_* optimal in problem (P), we fix a sequence of continuously differentiable functions $z_k : [0, \infty) \to \mathbb{R}^m$ $(k = 1, 2, \ldots)$ and a sequence of positive σ_k $(k = 1, 2, \ldots)$ such that

$$\sup_{t\in[0,\infty)} \|z_k(t)\| \le \max_{u\in U} \|u\| + 1; \tag{3.1}$$

$$\int_0^\infty e^{-(\rho+1)t} \|z_k(t) - u_*(t)\|^2 dt \le \frac{1}{k}; \tag{3.2}$$

$$\sup_{t\in[0,\infty)} \|\dot z_k(t)\| \le \sigma_k < \infty; \tag{3.3}$$

$$\sigma_k \to \infty \quad \text{as} \quad k \to \infty$$

(obviously, such sequences exist). Next, we take a monotonically increasing sequence of positive T_k such that $T_k \to \infty$ as $k \to \infty$ and

$$\omega(T_k) \le \frac{1}{k(1 + \sigma_k)} \quad \text{for all} \quad k = 1, 2, \ldots; \tag{3.4}$$

recall that ω is defined in (A4). For every $k = 1, 2, \ldots$, we define problem (P_k) as follows.

Problem (P_k):

$$\dot{x}(t) = f(x(t), u(t)), \qquad u(t) \in U;$$

$$x(0) = x_0;$$

$$\text{maximize } J_k(x,u) = \int_0^{T_k} e^{-\rho t} g(x(t),u(t))dt - \frac{1}{1+\sigma_k}\int_0^{T_k} e^{-(\rho+1)t}\|u(t)-z_k(t)\|^2 dt.$$

By Theorem 9.3.i of [19], for every $k = 1, 2, \ldots$ there exists an admissible pair (u_k, x_k) optimal in problem (P_k).

The above-defined sequence of problems, $\{(\mathrm{P}_k)\}$ $(k = 1, 2, \ldots)$, will be said to be associated with the control u_*.

We are ready to formulate our basic approximation lemma.

LEMMA 1. *Let assumptions* (A1)–(A4) *be satisfied; let u_* be a control optimal in problem* (P)*; let $\{(\mathrm{P}_k)\}$ $(k = 1, 2, \ldots)$ be the sequence of problems associated with u_*; and for every $k = 1, 2, \ldots$, let u_k be a control optimal in problem (P_k). Then, for every $T > 0$, it holds that $u_k \to u_*$ in $L^2([0,T], \mathbb{R}^m)$ as $k \to \infty$.*

Proof. Take a $T > 0$. Let k_1 be such that $T_{k_1} \geq T$. For every $k \geq k_1$, we have

$$J_k(x_k,u_k) = \int_0^{T_k} e^{-\rho t}\left[g(x_k(t),u_k(t)) - e^{-t}\frac{\|u_k(t)-z_k(t)\|^2}{1+\sigma_k}\right]dt$$

$$\leq \int_0^{T_k} e^{-\rho t} g(x_k(t),u_k(t))dt - \frac{e^{-(\rho+1)T}}{1+\sigma_k}\int_0^{T}\|u_k(t)-z_k(t)\|^2 dt,$$

where x_k is the trajectory corresponding to u_k. Hence, introducing the trajectory x_* corresponding to u_* and taking into account the optimality of u_k in problem (P_k), optimality of u_* in problem (P), assumption (A4), and conditions (3.2) and (3.4), we find that, for all sufficiently large k,

$$\frac{e^{-(\rho+1)T}}{1+\sigma_k}\int_0^{T}\|u_k(t)-z_k(t)\|^2 dt \leq \int_0^{T_k} e^{-\rho t} g(x_k(t),u_k(t))dt - J_k(x_*,u_*)$$

$$\leq \int_0^{T_k} e^{-\rho t} g(x_k(t),u_k(t))dt - J(x_*,u_*)$$

$$+ \omega(T_k) + \int_0^{\infty}\frac{e^{-(\rho+1)t}}{1+\sigma_k}\|u_*(t)-z_k(t)\|^2 dt$$

$$\leq \int_0^{T_k} e^{-\rho t} g(x_k(t),u_k(t))dt - J(x_*,u_*) + \frac{2}{k(1+\sigma_k)}$$

$$\leq J(x_k,u_k) - J(x_*,u_*) + \frac{3}{k(1+\sigma_k)} \leq \frac{3}{k(1+\sigma_k)}.$$

Hence,

$$\|u_k - z_k\|_{L^2}^2 \leq \frac{3e^{(\rho+1)T}}{k}.$$

Then, in view of (3.2),

$$\|u_k - u_*\|_{L^2} \le \left(\int_0^T \|u_*(t) - z_k(t)\|^2 dt\right)^{1/2} + \left(\int_0^T \|u_k(t) - z_k(t)\|^2 dt\right)^{1/2}$$

$$\le \left(\frac{e^{(\rho+1)T}}{k}\right)^{1/2} + \left(\frac{3e^{(\rho+1)T}}{k}\right)^{1/2} = (1+\sqrt{3})\left(\frac{e^{(\rho+1)T}}{k}\right)^{1/2}.$$

Therefore, for any $\epsilon > 0$, there exists a $k_2 \ge k_1$ such that $\|u_k - u_*\|_{L^2} \le \epsilon$ for all $k \ge k_2$. □

Now, based on Lemma 1, we derive a limit form of the classical Pontryagin maximum principle for problems $(\mathrm{P}_k)\,(k = 1, 2, \ldots)$, which leads us to the core Pontryagin maximum principle for problem (P).

We use the following formulation of the Pontryagin maximum principle [34] for problems $(\mathrm{P}_k)\,(k = 1, 2, \ldots)$. Let an admissible pair (u_k, x_k) be optimal in problem (P_k) for some k. Then there exists a pair (ψ_k, ψ_k^0) of adjoint variables associated with (u_k, x_k) such that (u_k, x_k) satisfies relations (1.4)–(1.6) of the core Pontryagin maximum principle (in problem (P_k)) together with (ψ_k, ψ_k^0) and, moreover, $\psi_k^0 > 0$ and the transversality condition

$$\psi_k(T_k) = 0 \tag{3.5}$$

holds; recall that ψ_k is a solution on $[0, T_k]$ to the adjoint equation associated with (u_k, x_k) in problem (P_k), i.e.,

$$\dot{\psi}_k(t) \overset{a.e.}{=} -\left[\frac{\partial f(x_k(t), u_k(t))}{\partial x}\right]^* \psi_k(t) - \psi^0 e^{-\rho t} \frac{\partial g(x_k(t), u_k(t))}{\partial x}, \tag{3.6}$$

and the core Pontryagin maximum principle satisfied by (u_k, x_k), together with (ψ_k, ψ_k^0), implies that the following maximum condition holds:

$$\mathcal{H}_k(x_k(t), t, u_k(t), \psi_k(t), \psi_k^0) \overset{a.e.}{=} H_k(x_k(t), t, \psi_k(t), \psi_k^0); \tag{3.7}$$

here $\mathcal{H}_k$ and H_k, given by

$$\mathcal{H}_k(x, t, u, \psi, \psi^0) = \langle f(x, u), \psi\rangle + \psi^0 e^{-\rho t} g(x, u) - \psi^0 e^{-(\rho+1)t} \frac{\|u - z_k(t)\|^2}{1 + \sigma_k}; \tag{3.8}$$

$$H_k(x, t, \psi, \psi^0) = \sup_{u \in U} \mathcal{H}_k(x, t, u, \psi, \psi^0),$$

are, respectively, the Hamilton–Pontryagin function and the Hamiltonian in problem (P_k); note that in [34] it is shown that (3.6) and (3.7) imply

$$\frac{d}{dt} H_k(x_k(t), t, \psi_k(t), \psi_k^0) \overset{a.e.}{=} \frac{\partial \mathcal{H}_k}{\partial t}(x_k(t), t, u_k(t), \psi_k(t), \psi_k^0). \tag{3.9}$$

Lemma 2. *Let assumptions* (A1)–(A4) *be satisfied; let* (u_*, x_*) *be an admissible pair optimal in problem* (P)*; let* $\{(\mathrm{P}_k)\}\,(k = 1, 2, \ldots)$ *be the sequence of problems associated with* u_**; for every* $k = 1, 2, \ldots$, *let* (u_k, x_k) *be an admissible pair optimal in problem* (P_k)*; for every* $k = 1, 2, \ldots$, *let* (ψ_k, ψ_k^0) *be a pair of adjoint variables associated with* (u_k, x_k) *in problem* (P_k) *such that* (u_k, x_k) *satisfies relations* (3.6)

and (3.7) *of the core Pontryagin maximum principle in problem* (P_k) *together with* (ψ_k, ψ_k^0)*; and for every* $k = 1, 2, \ldots$*, one has* $\psi_k^0 > 0$*, and the transversality condition* (3.5) *holds. Finally, let the sequences* $\{\psi_k(0)\}$ *and* $\{\psi_k^0\}$ *be bounded and*

$$\|\psi_k(0)\| + \psi_k^0 \geq a \quad (k = 1, 2, \ldots) \tag{3.10}$$

for some $a > 0$*. Then there exists a subsequence of* $\{(u_k, x_k, \psi_k, \psi_k^0)\}$*, denoted again as* $\{(u_k, x_k, \psi_k, \psi_k^0)\}$*, such that*

(i) *for every* $T > 0$*,*

$$u_k(t) \to u_*(t) \quad \textit{for a.a.} \quad t \in [0, T] \quad \textit{as} \quad k \to \infty; \tag{3.11}$$

$$x_k \to x_* \quad \textit{uniformly on} \quad [0, T] \quad \textit{as} \quad k \to \infty; \tag{3.12}$$

(ii)

$$\psi_k^0 \to \psi^0 \quad \textit{as} \quad k \to \infty \tag{3.13}$$

and for every $T > 0$*,*

$$\psi_k \to \psi \quad \textit{uniformly on} \quad [0, T] \quad \textit{as} \quad k \to \infty, \tag{3.14}$$

where (ψ, ψ^0) *is a nontrivial pair of adjoint variables associated with* (u_*, x_*)*;*

(iii) (u_*, x_*) *satisfies relations* (1.4)–(1.6) *of the core Pontryagin maximum principle in problem* (P) *together with* (ψ, ψ^0)*;*

(iv) *the stationarity condition holds:*

$$H(x_*(t), t, \psi(t), \psi^0) = \psi^0 \rho \int_t^\infty e^{-\rho s} g(x_*(s), u_*(s)) ds \quad \textit{for all} \quad t \geq 0. \tag{3.15}$$

Proof. Lemma 1 and the Ascoli theorem (see, e.g., [19]) imply that, selecting a subsequence if needed, we get (3.11) and (3.12) for every $T > 0$. By assumption, the sequence $\{\psi_k^0\}$ is bounded; therefore, selecting a subsequence if needed, we obtain (3.13) for some $\psi^0 \geq 0$.

Now, our goal is to select a subsequence of $\{(u_k, x_k, \psi_k)\}$ such that for every $T > 0$, (3.14) holds and (ψ, ψ^0) is a nontrivial pair of adjoint variables associated with (u_*, x_*) (we do not change notation after the selection of a subsequence).

Consider the sequence $\{\psi_k\}$ restricted to $[0, T_1]$. Observing (3.6), taking into account the boundedness of the sequence $\{\psi_k(0)\}$ (see the assumptions of this lemma), using the Gronwall lemma (see, e.g., [25]), and selecting if needed a subsequence denoted further as $\{\psi_k^1\}$, we get that $\psi_k^1 \to \psi^1$ uniformly on $[0, T_1]$ and $\dot{\psi}_k^1 \to \dot{\psi}^1$ weakly in $L^1[0, T_1]$ as $k \to \infty$ for some absolutely continuous $\psi^1 : [0, T_1] \to \mathbb{R}^n$; here and in what follows $L^1[0, T] = L^1([0, T], \mathbb{R}^n)$ $(T > 0)$.

Now consider the sequence $\{\psi_k^1\}$ restricted to $[0, T_2]$. Taking if necessary a subsequence $\{\psi_k^2\}$ of $\{\psi_k^1\}$, we get that $\psi_k^2 \to \psi^2$ uniformly on $[0, T_2]$ and $\dot{\psi}_k^2 \to \dot{\psi}^2$ weakly in $L^1[0, T_2]$ as $k \to \infty$ for some absolutely continuous $\psi^2 : [0, T_2] \to \mathbb{R}^n$ whose restriction to $[0, T_1]$ coincides with ψ^1.

Repeating this procedure sequentially for $[0, T_i]$ with $i = 3, 4, \ldots$, we find that there exist absolutely continuous $\psi^i : [0, T_i] \to \mathbb{R}^n$ $(i = 1, 2, \ldots)$ and $\psi_k^i : [0, T_i] \to \mathbb{R}^n$ $(i, k = 1, 2, \ldots)$ such that for every $i = 1, 2, \ldots$, the restriction of ψ^{i+1} to $[0, T_i]$ is ψ^i, the restriction of the sequence $\{\psi_k^{i+1}\}$ to $[0, T_i]$ is a subsequence of $\{\psi_k^i\}$, and, moreover, $\psi_k^i \to \psi$ uniformly on $[0, T_i]$ and $\dot{\psi}_k^i \to \dot{\psi}^i$ weakly in $L^1[0, T_i]$ as $k \to \infty$.

Define $\psi : [0,\infty) \mapsto \mathbb{R}^n$ so that the restriction of ψ to $[0,T_i]$ is ψ^i for every $i = 1,2,\ldots$. Clearly, ψ is absolutely continuous. Furthermore, without changing notation, for every $i = 1,2,\ldots$ and every $k = 1,2,\ldots$, we extend ψ_k^i to $[0,\infty)$ so that the extended function is absolutely continuous and, moreover, the family $\dot{\psi}_k^i$ $(i,k = 1,2,\ldots)$ is bounded in $L^1[0,T]$ for every $T > 0$. Since $T_i \to \infty$ as $i \to \infty$, for every $T > 0$, we get that ψ_k^k converges to ψ uniformly on $[0,T]$ and $\dot{\psi}_k^k \to \dot{\psi}$ weakly in $L^1[0,T]$ as $k \to \infty$. Simplifying notation, we again write ψ_k instead of ψ_k^k and note that for ψ_k, (3.6) holds $(k = 1,2,\ldots)$. Thus, for every $T > 0$, we have (3.14) and also get that $\dot{\psi}_k \to \dot{\psi}$ weakly in $L^1[0,T]$ as $k \to \infty$. These convergences together with equalities (3.6) and convergences (3.11) and (3.12) (holding for every $T > 0$) yield that ψ solves the adjoint equation (1.4). Thus, (ψ,ψ^0) is a pair of adjoint variables associated with (u_*,x_*) in problem (P). The nontriviality of (ψ,ψ^0) (see (1.5)) is ensured by (3.10).

For every $k = 1,2,\ldots$, consider the maximum condition (3.7) and specify it as

$$\langle f(x_k(t),u_k(t)),\psi_k(t)\rangle + \psi_k^0 e^{-\rho t} g(x_k(t),u_k(t)) - \psi_k^0 e^{-(\rho+1)t}\frac{\|u_k(t)-z_k(t)\|^2}{1+\sigma_k}$$

$$\overset{a.e.}{=} \max_{u\in U}\left[\langle f(x_k(t),u),\psi_k(t)\rangle + \psi_k^0 e^{-\rho t} g(x_k(t),u) - \psi_k^0 e^{-(\rho+1)t}\frac{\|u-z_k(t)\|^2}{1+\sigma_k}\right].$$

Taking into account that $T_k \to \infty$ and $\sigma_k \to \infty$ as $k \to \infty$ and using convergences (3.13), (3.14), (3.11), and (3.12) (holding for every $T > 0$), we obtain the maximum condition (1.6) as the limit of (3.7). Thus, (u_*,x_*) satisfies the core Pontryagin maximum principle together with the pair (ψ,ψ^0) of adjoint variables associated with (u_*,x_*).

Now we specify (3.9) using the form of $\mathcal{H}_k$ (see (3.9)). We get

$$\frac{d}{dt}H_k(x_k(t),t,\psi_k(t),\psi_k^0) \overset{a.e.}{=} \frac{\partial \mathcal{H}_k}{\partial t}(x_k(t),t,u_k(t),\psi_k(t),\psi_k^0)$$

$$\overset{a.e.}{=} -\psi_k^0\rho e^{-\rho t}\left[g(x_k(t),u_k(t)) + (\rho+1)e^{-(\rho+1)t}\frac{\|u_k(t)-z_k(t)\|^2}{1+\sigma_k}\right]$$

$$+ 2\psi_k^0 e^{-(\rho+1)t}\frac{\langle u_k(t)-z_k(t),\dot{z}_k(t)\rangle}{1+\sigma_k}.$$

Take an arbitrary $t > 0$ and an arbitrary k such that $T_k > t$ and integrate the last equality over $[t,T_k]$ taking into account the boundary condition (3.5). We arrive at

$$H_k(x_k(t),t,\psi_k(t),\psi_k^0) = \psi_k^0 e^{-\rho T_k}\max_{u\in U}\left[g(x_k(T_k),u) - e^{-\rho T_k}\frac{\|u-z_k(T_k)\|^2}{1+\sigma_k}\right]$$

$$- \psi_k^0\rho\int_t^{T_k} e^{-\rho s} g(x_k(s),u_k(s))ds$$

$$+ \psi_k^0(\rho+1)\int_t^{T_k} e^{-(\rho+1)s}\frac{\|u_k(s)-z_k(s)\|^2}{1+\sigma_k}ds$$

$$+ 2\psi_k^0\int_t^{T_k} e^{-(\rho+1)s}\frac{\langle u_k(s)-z_k(s),\dot{z}_k(s)\rangle}{1+\sigma_k}ds.$$

Now, we take the limit using convergences (3.13), (3.14), (3.11), and (3.12) (holding for every $T > 0$) and also estimates (3.1)–(3.3). We end up with (3.15). □

Corollary 1 below specifies Lemma 2 for the case where the Pontryagin maximum principle for problems $(\mathrm{P}_k)\,(k = 1, 2, \ldots)$ is taken in the normal form. We use the following formulation of the normal-form Pontryagin maximum principle for problems $(\mathrm{P}_k)\,(k = 1, 2, \ldots)$. Let an admissible pair (u_k, x_k) be optimal in problem (P_k) for some k. Then there exists an adjoint variable ψ_k associated with (u_k, x_k) such that (u_k, x_k) satisfies the normal-form core Pontryagin maximum principle (in problem (P_k)) together with ψ_k, and the transversality condition (3.5) holds; here ψ_k is a solution on $[0, T_k]$ of the normal-form adjoint equation associated with (u_k, x_k) in problem (P_k), i.e.,

$$\dot{\psi}_k(t) \overset{a.e.}{=} -\left[\frac{\partial f(x_k(t), u_k(t))}{\partial x}\right]^* \psi_k(t) - e^{-\rho t}\frac{\partial g(x_k(t), u_k(t))}{\partial x}, \tag{3.16}$$

and the fact that (u_k, x_k) satisfies the normal-form core Pontryagin maximum principle, together with ψ_k, implies that the following maximum condition holds:

$$\tilde{\mathcal{H}}_k(x_k(t), t, u_k(t), \psi(t)) = \tilde{H}_k(x_k(t), t, \psi_k(t)) \quad \text{for a.a.} \quad t \in [0, T_k]; \tag{3.17}$$

here $\tilde{\mathcal{H}}_k$ and $\tilde{H}_k$, given by

$$\tilde{\mathcal{H}}_k(x, t, u, \psi) = \langle f(x, u), \psi\rangle + e^{-\rho t} g(x, u) - e^{-(\rho+1)t}\frac{\|u - z_k(t)\|^2}{1 + \sigma_k};$$

$$\tilde{H}_k(x, t, \psi) = \sup_{u \in \tilde{U}} \tilde{\mathcal{H}}_k(x, t, \tilde{u}, \psi),$$

are, respectively, the normal-form Hamilton–Pontryagin function and normal-form Hamiltonian in problem (P_k).

Corollary 1. *Let assumptions* (A1)–(A4) *be satisfied; let* (u_*, x_*) *be an admissible pair optimal in problem* (P)*; let* $\{(P_k)\}\,(k = 1, 2, \ldots)$ *be the sequence of problems associated with* u_**; for every* $k = 1, 2, \ldots$*, let* (u_k, x_k) *be an admissible pair optimal in problem* (P_k)*; and for every* $k = 1, 2, \ldots$*, let* ψ_k *be an adjoint variable associated with* (u_k, x_k) *in problem* (P_k) *such that* (u_k, x_k) *satisfies relations* (3.16) *and* (3.17) *of the normal-form core Pontryagin maximum principle in problem* (P_k) *together with* ψ_k*, and the transversality condition* (3.5) *holds. Finally, let the sequence* $\{\psi_k(0)\}$ *be bounded. Then there exists a subsequence of* $\{(u_k, x_k, \psi_k)\}$*, denoted again as* $\{(u_k, x_k, \psi_k)\}$*, such that*

(i) *for every* $T > 0$, (3.11) *and* (3.12) *hold;*

(ii) *for every* $T > 0$, (3.14) *holds where* ψ *is an adjoint variable associated with* (u_*, x_*) *in problem* (P)*;*

(iii) (u_*, x_*) *satisfies relations* (1.7) *and* (1.8) *of the normal-form core Pontryagin maximum principle in problem* (P) *together with* ψ*;*

(iv) *the normal-form stationarity condition holds:*

$$\tilde{H}(x_*(t), t, \psi(t)) = \rho\int_t^\infty e^{-\rho s} g(x_*(s), u_*(s))ds \quad \text{for all} \quad t \geq 0. \tag{3.18}$$

Corollary 2. *Let assumptions* (A1)–(A4) *be satisfied and let* (u_*, x_*) *be an admissible pair optimal in problem* (P)*. Then there exists a pair* (ψ, ψ^0) *of adjoint variables associated with* (u_*, x_*) *such that*

(i) (u_*, x_*) *satisfies relations* (1.4)–(1.6) *of the core Pontryagin maximum principle together with* (ψ, ψ^0)*, and*

(ii) (u_*, x_*) *and* (ψ, ψ^0) *satisfy the stationarity condition* (3.15)*.*

Proof. Let $\{(\mathrm{P}_k)\}$ $(k = 1, 2, \ldots)$ be the sequence of problems associated with u_*, and for every $k = 1, 2, \ldots$, let (u_k, x_k) be an admissible pair optimal in problem (P_k). In accordance with the classical formulation of the Pontryagin maximum principle, for every $k = 1, 2, \ldots$, there exists a pair (ψ_k, ψ_k^0) of adjoint variables associated with (u_k, x_k) in problem (P_k) such that (u_k, x_k) satisfies the core Pontryagin maximum principle together with (ψ_k, ψ_k^0) and for every $k = 1, 2, \ldots$, $\psi_k^0 > 0$, and the transversality condition (3.5) holds.

Since $\psi_k^0 > 0$, the value $c_k = \|\psi_k(0)\| + \psi_k^0$ is positive. We keep the notation ψ_k and ψ_k^0 for the normalized elements ψ_k/c_k and ψ_k^0/c_k, thus achieving $\|\psi_k(0)\| + \psi_k^0 = 1$ and, clearly, preserving the transversality condition (3.5) and the fact that (u_k, x_k) satisfies the core Pontryagin maximum principle (in problem (P_k)) together with (ψ_k, ψ_k^0) $(k = 1, 2, \ldots)$. Now, the sequences $\{\psi_k(0)\}$ and $\{\psi_k^0\}$ are bounded and (3.10) holds with $a = 1$. Thus, the sequence $\{(u_k, x_k, \psi_k, \psi_k^0)\}$ satisfies all the assumptions of Lemma 2. By Lemma 2 there exists a subsequence of $\{(u_k, x_k, \psi_k, \psi_k^0)\}$, denoted again as $\{(u_k, x_k, \psi_k, \psi_k^0)\}$, such that for the pairs (ψ_k, ψ_k^0) of adjoint variables, convergences (3.13) and (3.14) hold with an arbitrary $T > 0$; the limit element (ψ, ψ^0) is a nontrivial pair of adjoint variables associated with (u_*, x_*) in problem (P); (u_*, x_*) satisfies the core Pontryagin maximum principle in problem (P) together with (ψ, ψ^0); and, finally, (u_*, x_*) and (ψ, ψ^0) satisfy the stationarity condition (3.15). □

It is easy to see that in the framework of problem (P) the necessary optimality conditions given by Corollary 2 are equivalent to those stated in [31]. Indeed, relation (3.15) implies the asymptotic stationarity condition introduced in [31]:

$$\lim_{t \to \infty} H(x_*(t), t, \psi(t), \psi^0) = 0; \tag{3.19}$$

on the other hand, if in problem (P), (u_*, x_*) satisfies the core Pontryagin maximum principle together with (ψ, ψ^0), then (3.19) implies (3.15). One can, however, anticipate that beyond setting (P) (for example, for problems with nonsmooth data), condition (3.15) complementing the core Pontryagin maximum principle can be substantially stronger than (3.19).

Note that in the problem considered in Example 1 the usage of the core Pontryagin maximum principle (Theorem 1) does not lead to the specification of an optimal control, whereas it can be shown that the latter control is determined uniquely if one applies the core Pontryagin maximum principle together with (3.15) (Corollary 2).

As an example given in [31] shows, under the assumptions of Corollary 2, the nontriviality condition (1.5) can hold with $\psi^0 = 0$; i.e., problem (P) can be abnormal. Below we find additional assumptions excluding abnormality of problem (P).

4. Normal-form maximum principle with positive adjoint variables. In this section, we suggest an assumption that excludes abnormality of problem (P), i.e., ensures that for problem (P), the normal-form Pontryagin maximum principle (see section 1) provides a necessary condition of optimality. Moreover, the basic result of this section formulated in Theorem 2 states that all the coordinates of the adjoint variable ψ in the Pontryagin maximum principle are necessarily positive-valued. Based on Theorem 2, we formulate conditions ensuring that the core Pontryagin maximum principle is complemented by the transversality conditions discussed in section 2. The proof of Theorem 2 is based on Corollary 1.

In what follows, the notation $a > 0$ (respectively, $a \geq 0$) for a vector $a \in \mathbb{R}^n$ designates that all coordinates of a are positive (respectively, nonnegative). Similarly, the notation $A > 0$ (respectively, $A \geq 0$) for a matrix A designates that all elements of A are positive (respectively, nonnegative).

The assumption complementing (A1)–(A4) is the following.

(A5) For every admissible pair (u, x) and for a.a. $t \geq 0$, one has

$$\frac{\partial g(x(t), u(t))}{\partial x} > 0 \quad \text{and} \quad \frac{\partial f^i(x(t), u(t))}{\partial x^j} \geq 0 \quad \text{for all} \quad i, j : i \neq j.$$

In typical models of regulated economic growth the coordinates of the state vector x represent positive-valued production factors. Normally it is assumed that the utility flow and the rate of growth of every production factor increase as all the production factors grow. In terms of problem (P), this implies that the integrand $g(x, u)$ in the goal functional (1.3), together with every coordinate of the right-hand side $f(x, u)$ of the system equation (1.1), is monotonically increasing in every coordinate of x. These monotonicity properties (specified so that $g(x, u)$ is strictly increasing in every coordinate of x) imply that assumption (A5) is satisfied. Note that the utility flow and the rates of growth of the production factors are normally positive, implying $g(x, u) > 0$ and $f(x, u) > 0$. The latter assumptions as well as the assumption $x > 0$ mentioned earlier appear in different combinations in the formulations of the results of this section.

The next theorem strengthens Theorem 1 under assumption (A5) and some positivity assumptions for f.

THEOREM 2. *Let assumptions* (A1)–(A5) *be satisfied. There exists a* $u_0 \in U$ *such that* $f(x_0, u_0) > 0$, *and for every admissible pair* (u, x), *it holds that* $f(x(t), u(t)) \geq 0$ *for a.a.* $t \geq 0$. *Let* (u_*, x_*) *be an admissible pair optimal in problem* (P). *Then there exists an adjoint variable* ψ *associated with* (u_*, x_*) *such that*

(i) (u_*, x_*) *satisfies relations* (1.7) *and* (1.8) *of the normal-form core Pontryagin maximum principle together with* ψ;

(ii) (u_*, x_*) *and* ψ *satisfy the normal-form stationarity condition* (3.18);

(iii)

$$\psi(t) > 0 \quad \textit{for all} \quad t \geq 0. \tag{4.1}$$

Proof. Let $\{(\mathrm{P}_k)\}\,(k = 1, 2, \ldots)$ be the sequence of problems associated with u_* and for every $k = 1, 2, \ldots$, let (u_k, x_k) be an admissible pair optimal in problem (P_k). In accordance with the classical formulation of the normal-form Pontryagin maximum principle, for every $k = 1, 2, \ldots$, there exists an adjoint variable ψ_k associated with (u_k, x_k) in problem (P_k) such that (u_k, x_k) satisfies the normal-form core Pontryagin maximum principle (in problem (P_k)) together with ψ_k and for every $k = 1, 2, \ldots$, the transversality condition (3.5) holds.

Observing assumption (A5), the adjoint equation resolved by ψ_k (see (3.16)), and transversality condition (3.5) for ψ_k, we easily find that $\psi_k(t) > 0$ for all t sufficiently close to T_k. Let us show that

$$\psi_k(t) > 0 \quad \text{for all} \quad t \in [0, T_k). \tag{4.2}$$

Suppose the contrary. Then, for some k, there exists a $\tau \in [0, T_k)$ such that at least one coordinate of the vector $\psi_k(\tau)$ vanishes. Let ξ be the maximum of all such $\tau \in [0, T_k)$, and let $i \in \{1, 2, \ldots, n\}$ be such that $\psi_k^i(\xi) = 0$. Then,

$$\psi_k(t) > 0 \quad \text{for all} \quad t \in (\xi, T_k) \tag{4.3}$$

and for all $t \in [\xi, T_k]$, we have

$$\psi_k^i(t) = -\int_\xi^t \left\langle \frac{\partial f^i(x_k(s), u_k(s))}{\partial x}, \psi_k(s) \right\rangle ds - \int_\xi^t e^{-\rho s} \frac{\partial g^i(x_k(s), u_k(s))}{\partial x} ds. \tag{4.4}$$

The latter equation and assumption (A5) imply that there is an $\epsilon > 0$ such that $\psi_k^i(t) < 0$ for all $t \in (\xi, \xi+\epsilon)$, which contradicts (4.3). The contradiction proves (4.2).

Let us show that the sequence $\{\psi_k(0)\}$ is bounded. The equation for ψ_k (see (3.16)) and maximum condition (3.17) yield

$$\begin{aligned}\frac{d}{dt}\tilde{H}_k(x_k(t),t,\psi_k(t)) &\overset{a.e.}{=} \frac{\partial\tilde{\mathcal{H}}_k}{\partial t}(x_k(t),t,\tilde{u}_k(t),\psi_k(t))\\ &\overset{a.e.}{=} -\rho e^{-\rho t}g(x_k(t),u_k(t)) + (\rho+1)e^{-(\rho+1)t}\frac{\|u_k(t)-z_k(t)\|^2}{1+\sigma_k}\\ &\quad + 2e^{-(\rho+1)t}\frac{\langle u_k(t)-z_k(t),\dot{z}_k(t)\rangle}{1+\sigma_k}.\end{aligned}$$

Integrating over $[0,T_k]$ and using the transversality condition (3.5), we arrive at

$$\begin{aligned}\tilde{H}_k(x_0,0,\psi_k(0)) &= e^{-\rho T_k}\max_{u\in U}\left[g(x_k(T_k),u) - e^{-T_k}\frac{\|u-z_k(T_k)\|^2}{1+\sigma_k}\right]\\ &\quad + \rho\int_0^{T_k} e^{-\rho t}g(x_k(t),u_k(t))dt\\ &\quad - (\rho+1)\int_0^{T_k} e^{-(\rho+1)t}\frac{\|u_k(t)-z_k(t)\|^2}{1+\sigma_k}dt\\ &\quad - 2\int_0^{T_k} e^{-(\rho+1)t}\frac{\langle u_k(t)-z_k(t),\dot{z}_k(t)\rangle}{1+\sigma_k}dt.\end{aligned}$$

This, together with (3.1)–(3.3), implies that $\tilde{H}_k(x_0,0,\psi_k(0)) \le M$ for some $M > 0$ and all $k = 1,2,\dots$. Hence, by virtue of

$$\langle f(x_0,u_0),\psi_k(0)\rangle + g(x_0,u_0) - \frac{\|u_0-z_k(0)\|^2}{1+\sigma_k} \le \tilde{H}_k(x_0,0,\psi_k(0)),$$

we have

$$\langle f(x_0,u_0),\psi_k(0)\rangle \le M + |g(x_0,u_0)| + (2|U|+1)^2,$$

where $|U| = \max_{u\in U}\|u\|$. The latter estimate, assumption $f(x_0,u_0) > 0$, and (4.2) yield that the sequence $\{\psi_k(0)\}$ is bounded.

Therefore, the sequence $\{(u_k,x_k,\psi_k)\}$ satisfies all the assumptions of Corollary 1. By Corollary 1, there exists a subsequence of $\{(u_k,x_k,\psi_k)\}$, denoted again as $\{(u_k,x_k,\psi_k)\}$, such that for every $T > 0$, one has convergence (3.14) for the adjoint variables ψ_k, where the limit element ψ is an adjoint variable associated with (u_*,x_*) in problem (P); (u_*,x_*) satisfies the normal-form core Pontryagin maximum principle in problem (P) together with ψ; and, finally, (u_*,x_*) and ψ satisfy the normal-form asymptotic stationarity condition (3.18). Thus, for (u_*,x_*) and ψ, statements (i) and (ii) are proved.

From (3.14) and (4.2) it follows that $\psi(t) \ge 0$ for all $t \ge 0$. Assumption (A5) and the fact that ψ solves the adjoint equation (1.7) imply (4.1), thus proving (iii). ∎

Now, we formulate conditions coupling the normal-form core Pontryagin maximum principle and the transversality conditions discussed in section 2.

COROLLARY 3. *Let the assumptions of Theorem 2 be satisfied and*

$$f(x_*(t), u_*(t)) \geq a_1 \quad \text{for a.a.} \quad t \geq 0, \tag{4.5}$$

where $a_1 > 0$. Then there exists an adjoint variable ψ associated with (u_*, x_*) such that statements (i), (ii), *and* (iii) *of Theorem 2 hold true and, moreover, ψ satisfies the transversality condition*

$$\lim_{t\to\infty} \psi(t) = 0. \tag{4.6}$$

Proof. By Theorem 2, there exists an adjoint variable ψ associated with (u_*, x_*) such that statements (i), (ii), and (iii) of Theorem 2 hold true. Let us prove (4.6). From (3.15) and (4.5) we get

$$\lim_{t\to\infty} \langle a_1, \psi(t)\rangle \leq \lim_{t\to\infty} \max_{u\in U} \langle f(x_*(t), u), \psi(t)\rangle = 0;$$

the latter, together with (4.1), implies (4.6). □

COROLLARY 4. *Let the assumptions of Theorem 2 be satisfied, and let*

$$x_0 \geq 0; \tag{4.7}$$

$$g(x_*(t), u_*(t)) \geq 0 \quad \text{for a.a.} \quad t \geq 0; \tag{4.8}$$

and

$$\frac{\partial f(x_*(t), u_*(t))}{\partial x} \geq A \quad \text{for a.a.} \quad t \geq 0, \tag{4.9}$$

where A is a matrix of order n such that $A > 0$. Then there exists an adjoint variable ψ associated with (u_*, x_*) such that statements (i), (ii), *and* (iii) *of Theorem 2 hold true and, moreover, ψ satisfies the transversality condition*

$$\lim_{t\to\infty} \langle x_*(t), \psi(t)\rangle = 0. \tag{4.10}$$

Proof. By Theorem 2, there exists an adjoint variable ψ associated with (u_*, x_*) such that statements (i), (ii), and (iii) of Theorem 2 hold true. Let us prove (4.10). The system equation (1.1) and normal-form adjoint equation (1.7) yield

$$\begin{aligned} \frac{d}{dt}\langle x_*(t), \psi(t)\rangle &= \langle f(x_*(t), u_*(t)), \psi(t)\rangle \\ &\quad - \left\langle x_*(t), \left[\frac{\partial f(x_*(t), u_*(t))}{\partial x}\right]^* \psi(t)\right\rangle \\ &\quad - e^{-\rho t}\left\langle x_*(t), \frac{\partial g(x_*(t), u_*(t))}{\partial x}\right\rangle \quad \text{for a.a.} \quad t \geq 0. \end{aligned} \tag{4.11}$$

From (4.7), assumption (A5), and (4.8), it follows that

$$-e^{-\rho t}\left\langle x_*(t), \frac{\partial g(x_*(t), u_*(t))}{\partial x}\right\rangle \leq 0 \leq e^{-\rho t} g(x_*(t), u_*(t)).$$

Taking this into account and using assumption (A5), the normal-form maximum condition (1.8), and assumption (4.9), we continue (4.11) as follows:

$$\begin{aligned}\frac{d}{dt}\langle x_*(t),\psi(t)\rangle &\le \langle f(x_*(t),u_*(t)),\psi(t)\rangle\\ &\quad - \left\langle x_*(t), \left[\frac{\partial f(x_*(t),u_*(t))}{\partial x}\right]^*\psi(t)\right\rangle + e^{-\rho t}g(x_*(t),u_*(t))\\ &\le -\langle Ax_*(t),\psi(t)\rangle + \tilde{H}(x_*(t),t,\psi(t)) \quad \text{for a.a.} \quad t\ge 0.\end{aligned}$$

Therefore, by (4.9), for some $\theta > 0$, we have

$$\frac{d}{dt}\langle x_*(t),\psi(t)\rangle \le -\theta\langle x_*(t),\psi(t)\rangle + \alpha(t),$$

where

$$\alpha(t) = \tilde{H}(x_*(t),t,\psi(t)) \to 0 \quad \text{as} \quad t\to\infty$$

(see (3.18)). Then, taking into account (4.7) and (4.1), we get

$$0 \le \langle x_*(t),\psi(t)\rangle \le e^{-\theta t}\langle x_0,\psi(0)\rangle + e^{-\theta t}\int_0^t e^{\theta s}\alpha(s)ds. \tag{4.12}$$

Furthermore,

$$\begin{aligned}\dot{\alpha}(t) &= \frac{d}{dt}\tilde{H}(x_*(t),t,\psi(t)) = \frac{\partial}{\partial t}\tilde{\mathcal{H}}(x_*(t),t,u_*(t),\psi(t))\\ &= -\rho e^{-\rho t}g(x_*(t),u_*(t)) \le 0 \quad \text{for a.a.} \quad t\ge 0\end{aligned}$$

(here we used (4.8)). Therefore,

$$\int_0^t e^{\theta s}\alpha(s)ds = \frac{1}{\theta}[e^{\theta t}\alpha(t) - \alpha(0)] + \frac{1}{\theta}\int_0^t e^{\theta s}\dot{\alpha}(s)ds \le \frac{1}{\theta}(e^{\theta t}\alpha(t) - \alpha(0)).$$

Substituting this estimate into (4.12), we get

$$0 \le \langle x_*(t),\psi(t)\rangle \le e^{-\theta t}\langle x_0,\psi(0)\rangle + e^{-\theta t}\frac{1}{\theta}[e^{\theta t}\alpha(t) - \alpha(0)] \to 0 \quad \text{as} \quad t\to\infty. \qquad \square$$

The next theorem is, to a certain extent, an inversion of Theorem 2. It adjoins works treating the Pontryagin maximum principle as a key component in sufficient conditions of optimality. Within the finite-horizon setting, this line of analysis was initiated in [30]. In [1] the approach was extended to infinite-horizon optimal control problems.

THEOREM 3. *Let assumptions* (A1)–(A5) *be satisfied,* $x_0 \ge 0$*, and for every admissible pair* (u,x)*, it holds that* $f(x(t),u(t)) \ge 0$ *and* $g(x(t),u(t)) \ge 0$ *for a.a.* $t\ge 0$*. Let* (u_*,x_*) *be an admissible pair satisfying* (4.9) *with some* $A > 0$*, and there exists an adjoint variable* ψ *associated with* (u_*,x_*) *such that statements* (i), (ii), *and* (iii) *of Theorem* 2 *hold true. Finally, let the set* G *be convex and the function* $x \mapsto \tilde{H}(x,t,\psi(t)) : G \mapsto \mathbb{R}^1$ *be concave for every* $t \ge 0$*. Then, the admissible pair* (u_*,x_*) *is optimal in problem* (P).

We omit the proof, which is similar to the proofs given in [2] and [36].

Combining Corollary 4 and Theorem 3, we arrive at the following optimality criterion for problem (P).

COROLLARY 5. *Let assumptions* (A1)–(A5) *be satisfied,* $x_0 \geq 0$*, and there exists a* $u_0 \in U$ *such that* $f(x_0, u_0) > 0$*. For every admissible pair* (u, x) *it holds that* $f(x(t), u(t)) \geq 0$*,* $g(x(t), u(t)) \geq 0$*, and* $\partial f(x(t), u(t))/\partial x \geq A$ *for a.a.* $t \geq 0$ *with some* $A > 0$*. Let, finally, the set* G *be convex and the function* $x \mapsto \tilde{H}(x, t, \psi) : G \mapsto \mathbb{R}^1$ *be concave for every* $t \geq 0$ *and for every* $\psi > 0$*. Then, an admissible pair* (u_*, x_*) *is optimal in problem* (P) *if and only if there exists an adjoint variable* ψ *associated with* (u_*, x_*) *such that statements* (i)*,* (ii)*, and* (iii) *of Theorem* 2 *hold true and the transversality condition* (4.10) *is satisfied.*

5. Case of dominating discount. In [12], infinite-horizon necessary optimality conditions involving the normal-form core Pontryagin maximum principle and an integral characterization of global behavior of the adjoint variable were stated for problems with sufficiently large (dominating) discount factors; in this work the control system was assumed to be linear. In this section, we consider problem (P) for the nonlinear control system (1.1) and apply the approximation scheme developed in section 3 in the case of the dominating discount. If system (1.1) is linear, the basic statement of this section (Theorem 4) leads to a formulation of the Pontryagin maximum principle (Corollary 7), which is stronger than that given in [12].

Following [12], we posit the next growth constraint on g.

(A6) There exist a $\kappa \geq 0$ and an $r \geq 0$ such that

$$\left\| \frac{\partial g(x, u)}{\partial x} \right\| \leq \kappa(1 + \|x\|^r) \quad \text{for all} \quad x \in G \quad \text{and for all} \quad u \in U.$$

Given an admissible pair (u, x), we denote by $Y_{(u,x)}$ the normalized fundamental matrix for the linear differential equation

$$\dot{y}(t) = \frac{\partial f(x(t), u(t))}{\partial x} y(t); \tag{5.1}$$

more specifically, $Y_{(x,u)}$ is the $n \times n$ matrix-valued function on $[0, \infty)$ whose columns y_i $(i = 1, \ldots, n)$ are the solutions to (5.1) such that $y_i^j(0) = \delta_{i,j}$ $(i, j = 1, \ldots, n)$, where $\delta_{i,i} = 1$ and $\delta_{i,j} = 0$ for $i \neq j$; for every $t \geq 0$, $\|Y_{(u,x)}(t)\|$ stands for the standard norm of $Y_{(u,x)}(t)$ as a linear operator in $\mathbb{R}^n$. Similarly, given an admissible pair (u, x), we denote by $Z_{(u,x)}$ the normalized fundamental matrix for the linear differential equation

$$\dot{z}(t) = -\left[\frac{\partial f(x(t), u(t))}{\partial x} \right]^* z(t).$$

Note that

$$[Z_{(u,x)}(t)]^{-1} = [Y_{(u,x)}(t)]^*. \tag{5.2}$$

We introduce the following growth assumption.

(A7) There exist a $\lambda \in \mathbb{R}^1$, a $C_1 \geq 0$, a $C_2 \geq 0$, and a $C_3 \geq 0$ such that for every admissible pair (u, x), one has

$$\|x(t)\| \leq C_1 + C_2 e^{\lambda t} \quad \text{for all} \quad t \geq 0$$

and

$$\|Y_{(u,x)}(t)\| \leq C_3 e^{\lambda t} \quad \text{for all} \quad t \geq 0.$$

It is easily seen that assumption (A6) implies that there exist a $C_4 \geq 0$ and a $C_5 \geq 0$ such that for every admissible pair (u, x), one has

$$|g(x(t), u(t))| \leq C_4 + C_5 \|x(t)\|^{r+1} \quad \text{for all} \quad t \geq 0. \tag{5.3}$$

Furthermore, (A7) and (5.3) imply that

$$e^{-\rho t}|g(x(t), u(t))| \leq C_6 e^{-\rho t} + C_7 e^{-(\rho-(r+1)\lambda)t}$$

holds for every admissible pair (u, x) with $C_6 \geq 0$ and $C_7 \geq 0$ not depending on (u, x). Therefore, if $\rho > 0$, then assumptions (A6) and (A7) imply (A4), provided $\rho > (r+1)\lambda$. The latter inequality implies that the discount parameter ρ in the goal functional (1.3) dominates the growth parameters r and λ (see (A6) and (A7)), which is a counterpart of a condition assumed in [12].

The proof of the next result is based on Corollary 1.

THEOREM 4. *Let assumptions* (A1)–(A4), (A6), *and* (A7) *be satisfied and let* $\rho > (r+1)\lambda$. *Let* (u_*, x_*) *be an admissible pair optimal in problem* (P). *Then there exists an adjoint variable* ψ *associated with* (u_*, x_*) *such that*

(i) (u_*, x_*) *satisfies relations* (1.7) *and* (1.8) *of the normal-form core Pontryagin maximum principle together with* ψ;

(ii) (u_*, x_*) *and* ψ *satisfy the normal-form stationarity condition* (3.18);

(iii) *for every* $t \geq 0$, *the integral*

$$I_*(t) = \int_t^\infty e^{-\rho s}[Z_*(s)]^{-1} \frac{\partial g(x_*(s), u_*(s))}{\partial x} ds, \tag{5.4}$$

where $Z_* = Z_{(u_*, x_*)}$, *converges absolutely and*

$$\psi(t) = Z_*(t) I_*(t). \tag{5.5}$$

Proof. Let $\{(\mathrm{P}_k)\}$ $(k = 1, 2, \ldots)$ be the sequence of problems associated with u_* and for every $k = 1, 2, \ldots$, let (u_k, x_k) be an admissible pair optimal in problem (P_k). In accordance with the classical formulation of the normal-form Pontryagin maximum principle, for every $k = 1, 2, \ldots$, there exists an adjoint variable ψ_k associated with (u_k, x_k) in problem (P_k) such that (u_k, x_k) satisfies the normal-form core Pontryagin maximum principle (in problem (P_k)) together with ψ_k and for every $k = 1, 2, \ldots$, the transversality condition (3.5) holds.

Let us show that the sequence $\{\psi_k(0)\}$ is bounded. Using the standard representation of the solution ψ_k to the linear normal-form adjoint equation (3.16) with the zero-boundary condition (3.5) through the fundamental matrix $Z_k = Z_{(u_k, x_k)}$ of the corresponding linear homogeneous equation (see, e.g., [25]), we get

$$\psi_k(0) = \int_0^{T_k} e^{-\rho s}[Z_k(s)]^{-1} \frac{\partial g(x_k(s), u_k(s))}{\partial x} ds.$$

Therefore, due to (5.2),

$$\|\psi_k(0)\| \leq \int_0^{T_k} e^{-\rho s} \|Y_{(x_k, u_k)}(s)\| \left\| \frac{\partial g(x_k(s), u_k(s))}{\partial x} \right\| ds,$$

and due to assumptions (A6) and (A7),

$$\|\psi_k(0)\| \leq \int_0^{T_k} (C_8 e^{-(\rho-\lambda)s} + C_9 e^{-(\rho-(r+1)\lambda)s}) ds,$$

where $C_8 \geq 0$ and $C_9 \geq 0$ do not depend on k. Now the assumption $\rho > (r+1)\lambda$ implies that the sequence $\{\psi_k(0)\}$ is bounded.

Therefore, the sequence $\{(u_k, x_k, \psi_k)\}$ satisfies all the assumptions of Corollary 1. By Corollary 1, there exists a subsequence of $\{(u_k, x_k, \psi_k)\}$, denoted again as $\{(u_k, x_k, \psi_k)\}$, such that for every $T > 0$, one has convergences (3.11) and (3.12) for the admissible pairs (u_k, x_k) and convergence (3.14) for the adjoint variables ψ_k, where the limit element ψ is an adjoint variable associated with (u_*, x_*) in problem (P); (u_*, x_*) satisfies the normal-form core Pontryagin maximum principle in problem (P) together with ψ; and, finally, (u_*, x_*) and ψ satisfy the normal-form stationarity condition (3.18). Thus, for (u_*, x_*) and ψ, statements (i) and (ii) are proved.

Consider the integral $I_*(t)$ in (5.4) for an arbitrary $t \geq 0$. Convergences (3.11) and (3.12) imply

$$Z_k(s) \to Z_*(s) \quad \text{for all} \quad s \geq 0. \tag{5.6}$$

Hence,

$$I_*(t) = \lim_{T\to\infty} \int_t^T e^{-\rho s}[Z_*(s)]^{-1} \frac{\partial g(x_*(s), u_*(s))}{\partial x} ds$$

$$= \lim_{T\to\infty} \lim_{k\to\infty} \int_t^T e^{-\rho s}[Z_k(s)]^{-1} \frac{\partial g(x_k(s), u_k(s))}{\partial x} ds.$$

Furthermore, from (5.2) and (A7) it follows that for all $s \geq 0$,

$$e^{-\rho t}\|[Z_k(s)]^{-1}\| \left\| \frac{\partial g(x_k(s), u_k(s))}{\partial x} \right\| \leq C_{10} e^{-(\rho-\lambda)s} + C_{11} e^{-(\rho-(r+1)\lambda)s}$$

with some positive C_{10} and C_{11}. Therefore, $I_*(t)$ converges absolutely. Let us prove (5.5). Integrate the adjoint equation for ψ_k (see (3.16)) over $[t, T_k]$ assuming that k is large enough (i.e., $T_k \geq t$) and taking into account the transversality condition (3.5). We get

$$\psi_k(t) = Z_k(t) \int_t^{T_k} e^{-\rho s} Z_k^{-1}(s) \frac{\partial g(x_k(s), u_k(s))}{\partial x} ds. \tag{5.7}$$

Convergences (3.11) and (3.12) (holding for every $T > 0$) imply that $x_k(s) \to x_*(s)$ for all $s \geq 0$ and $u_k(s) \to u_*(s)$ for a.a. $s \geq 0$. The latter convergences, convergences (5.6) and (3.14), and the absolute convergence of the integral $I_*(t)$ yield that the desired equality (5.5) is the limit of (5.7) with $k \to \infty$. Statement (iii) is proved. □

Now we recall some facts from the stability theory (see [18], [20]).

Consider a linear differential equation

$$\dot{x}(t) = A(t)x(t). \tag{5.8}$$

Here $t \in [0, \infty)$, $x \in \mathbb{R}^n$, and the components of the real $n \times n$ matrix function A are measurable and bounded.

Let x be a nonzero solution to (5.8). Then, a number

$$\lambda = \limsup_{t\to\infty} \frac{1}{t} \ln \|x(t)\|$$

is said to be the Lyapunov characteristic number of x. Note that the characteristic number λ was defined by A. M. Lyapunov with the opposite sign [29].

The Lyapunov characteristic number of any nonzero solution to (5.8) is finite. The set of the characteristic numbers for all nonzero solutions to (5.8) is called the Lyapunov spectrum of (5.8). The Lyapunov spectrum of (5.8) has no more than n elements. A fundamental system $x_1, \dots, x_n$ of solutions to (5.8) is said to be normal if the sum of their characteristic numbers is minimal in the set of all fundamental systems of solutions to (5.8). A normal fundamental system always exists. If $x_1, \dots, x_n$ is a normal fundamental system of solutions to (5.8), then the characteristic numbers for $x_1, \dots, x_n$ cover the Lyapunov spectrum of (5.8) (for different x_j and x_k the Lyapunov characteristic numbers may coincide). Any normal fundamental system contains the same number n_s of solutions to (5.8) with characteristic number λ_s, $1 \le s \le l$, $l \le n$, from the Lyapunov spectrum of (5.8).

Let $\sigma = \sum_{s=1}^{l} n_s \lambda_s$ be the sum of all numbers $\lambda_1, \dots, \lambda_l$ from the Lyapunov spectrum of (5.8) taken according to their multiplicity. Equation (5.8) is called regular if

$$\sigma = \liminf_{t \to \infty} \frac{1}{t} \int_0^t \mathrm{tr} A(s) ds,$$

where $\mathrm{tr}A(s)$ is the trace of matrix $A(s)$. If (5.8) is regular, then for every $\epsilon > 0$, the Cauchy matrix $(s,t) \mapsto K(s,t)$ for (5.8) satisfies

$$\|K(s,t)\| \le \kappa_1 e^{\bar{\lambda}(s-t)+\epsilon s} \quad \text{for all} \quad t \ge 0 \quad \text{and all} \quad s \ge t, \tag{5.9}$$

where $\bar{\lambda}$ is the maximum element in the Lyapunov spectrum of (5.8) and $\kappa_1 \ge 0$ is a constant depending only on ϵ (see [20]).

Corollary 6. *Let the assumptions of Theorem 4 be satisfied, let the linear differential equation*

$$\dot{y}(t) = \frac{\partial f(x_*(t), u_*(t))}{\partial x} y(t) \tag{5.10}$$

be regular, and let $\lambda \ge \bar{\lambda}$, where $\bar{\lambda}$ is the maximum element in the Lyapunov spectrum of (5.10). *Then for every $\epsilon > 0$, it holds that*

$$\|\psi(t)\| \le \kappa_2 (e^{-\rho t + \epsilon t} + e^{-(\rho - r\lambda)t + \epsilon t}) \quad \textit{for all} \quad t \ge 0, \tag{5.11}$$

where $\kappa_2 \ge 0$ is a constant depending only on ϵ.

Proof. By (5.5) and (5.2),

$$\begin{aligned}
\psi(t) &= \int_t^\infty e^{-\rho s} [[Y_*(t)]^*]^{-1} [Y_*(s)]^* \frac{\partial g(x_*(s), u_*(s))}{\partial x} ds \\
&= \int_t^\infty e^{-\rho s} [Y_*(s)[Y_*(t)]^{-1}]^* \frac{\partial g(x_*(s), u_*(s))}{\partial x} ds \\
&= \int_t^\infty e^{-\rho s} [K(s,t)]^* \frac{\partial g(x_*(s), u_*(s))}{\partial x} ds,
\end{aligned}$$

where $Y_* = Y_{(u_*, x_*)}$ is a normalized fundamental solution matrix of (5.10) and $K(s,t) = Y_*(s)[Y_*(t)]^{-1}$ is the Cauchy matrix of (5.10). Hence, by (A6), (A7), and

(5.9), for any $0 < \epsilon < \min\{\rho - \lambda,\ \rho - (r+1)\lambda\}$, we have

$$\begin{aligned}\|\psi(t)\| &\le \int_t^\infty e^{-\rho s}\|[K(s,t)]^*\| \left\| \frac{\partial g(x_*(s), u_*(s))}{\partial x} \right\| ds \\ &\le C_{12} \int_t^\infty e^{-\rho s} e^{\bar\lambda(s-t)} e^{\epsilon s}(1 + e^{r\lambda s}) ds \\ &\le \kappa_2 (e^{-\rho t + \epsilon t} + e^{-(\rho - r\lambda)t + \epsilon t}),\end{aligned}$$

where $C_{12} \ge 0$ and $\kappa_2 \ge 0$ depend only on ϵ. Hence, estimate (5.11) holds for any $\epsilon > 0$. □

Note that if $\rho > 0$, then Corollary 6 implies the validity of both (2.1) and (2.2).

Now, let us consider the situation where the control system (1.1) is linear and stationary. Problem (P) is specified as follows.

Problem (P1):

$$\dot x(t) = Fx(t) + u(t), \qquad u(t) \in U; \tag{5.12}$$

$$x(0) = x_0;$$

$$\text{maximize } J(x,u) = \int_0^\infty e^{-\rho t} g(x(t), u(t)) dt;$$

here F is a real $n \times n$ matrix.

Let λ_F be the maximum of the real parts of the eigenvalues of F. Then λ_F is the maximal number from the Lyapunov spectrum of the linear homogenous differential equation corresponding to (5.12), and for any $\epsilon > 0$, we have

$$\|e^{Ft}\| \le C_{13} e^{(\lambda_F + \epsilon)t} \quad \text{for all} \quad t \ge 0.$$

Here, e^{Ft} is the exponential of matrix F and $C_{13} \ge 0$ is a constant depending only on ϵ (see [20]). A standard representation of a solution to (5.12) through the matrix exponential e^{Ft} (see [25]) implies that for any $\epsilon > 0$ and for any admissible trajectory x of system (5.12), it holds that

$$\|x(t)\| \le C_{14} + C_{15} e^{(\lambda_F + \epsilon)t} \quad \text{for all} \quad t \ge 0,$$

where $C_{14} \ge 0$ and $C_{15} \ge 0$ depend only on ϵ. Thus, for an arbitrary $\lambda > \lambda_F$, assumption (A7) is satisfied. The latter observation and the fact that every linear stationary equation is regular [20] imply the following specification of Corollary 6 for problem (P1).

COROLLARY 7. *Let assumptions* (A3), (A4), *and* (A6) *be satisfied and let* $\rho > (r+1)\lambda_F$. *Let* (u_*, x_*) *be an admissible pair optimal in problem* (P1). *Then there exists an adjoint variable* ψ *associated with* (u_*, x_*) *such that*

(i) (u_*, x_*) *satisfies relations* (1.7) *and* (1.8) *of the normal-form core Pontryagin maximum principle together with* ψ;

(ii) (u_*, x_*) *and* ψ *satisfy the normal-form stationarity condition* (3.18);

(iii) *for every* $\epsilon > 0$, *it holds that*

$$\|\psi(t)\| \le \kappa_3 (e^{-\rho t + \epsilon t} + e^{-(\rho - r\lambda_F)t + \epsilon t}) \quad \textit{for all} \quad t \ge 0,$$

where $\kappa_3 \ge 0$ *is a constant depending only on* ϵ.

Note that if $\rho > 0$, then Corollary 7 implies the validity of both (2.1) and (2.2).

REFERENCES

[1] K. Arrow, *Application of control theory to economic growth*, in Mathematics of the Decision Sciences, Part 2, AMS, Providence, RI, 1968, pp. 85–119.

[2] K. Arrow and M. Kurz, *Public Investment, the Rate of Return, and Optimal Fiscal Policy*, The Johns Hopkins University Press, Baltimore, MD, 1970.

[3] A. V. Arutyunov, *Perturbations of extremal problems with constraints and necessary optimality conditions*, in Mathematical Analysis, Vol. 27, Itogi Nauki i Tekhniki, VINITI, Moscow, 1989, pp. 147–235 (in Russian); J. Soviet Math., 54 (1991), pp. 1342–1400 (in English).

[4] A. V. Arutyunov and S. M. Aseev, *Investigation of the degeneracy phenomenon of the maximum principle for optimal control problems with state constraints*, SIAM J. Control Optim., 35 (1997), pp. 930–952.

[5] S. M. Aseev, *The method of smooth approximations in the theory of necessary optimality conditions for differential inclusions*, Izv. Math., 61 (1997), pp. 235–258.

[6] S. M. Aseev, *Methods of regularization in nonsmooth problems of dynamic optimization*, J. Math. Sci. (New York), 94 (1999), pp. 1366–1393.

[7] S. M. Aseev, *Extremal problems for differential inclusions with state constraints*, Proc. Steklov Inst. Math., 233 (2001), pp. 1–63.

[8] S. Aseev, G. Hutschenreiter, and A. Kryazhimskii, *A Dynamic Model of Optimal Allocation of Resources to R&D*, IIASA Interim Report IR-02-16, Laxenburg, Austria, 2002.

[9] S. M. Aseev, A. V. Kryazhimskii, and A. M. Tarasyev, *First Order Necessary Optimality Conditions for a Class of Infinite-Horizon Optimal Control Problems*, IIASA Interim Report IR-01-007, Laxenburg, Austria, 2001.

[10] S. M. Aseev, A. V. Kryazhimskii, and A. M. Tarasyev, *The Pontryagin maximum principle and transversality conditions for an optimal control problem with infinite time interval*, Proc. Steklov Inst. Math., 233 (2001), pp. 64–80.

[11] S. M. Aseev, A. V. Kryazhimskii, and A. M. Tarasyev, *The maximum principle and transversality conditions for a class of optimal economic growth problems*, in Proceedings of the 5th IFAC Symposium on Nonlinear Control Systems (St.-Petersburg, Russia), A. L. Fradkov and A. B. Kurzhanski, eds., Elsevier, New York, 2002, pp. 64–68.

[12] J. P. Aubin and F. H. Clarke, *Shadow prices and duality for a class of optimal control problems*, SIAM J. Control Optim., 17 (1979), pp. 567–586.

[13] E. J. Balder, *An existence result for optimal economic growth problems*, J. Math. Anal. Appl., 95 (1983), pp. 195–213.

[14] R. Barro and X. Sala-i-Martin, *Economic Growth*, McGraw–Hill, New York, 1995.

[15] L. M. Benveniste and J. A. Scheinkman, *Duality theory for dynamic optimization models of economics: The continuous time case*, J. Econom. Theory, 27 (1983), pp. 1–19.

[16] J. Blot and P. Michel, *First-order necessary conditions for infinite-horizon variational problems*, J. Optim. Theory Appl., 88 (1996), pp. 339–364.

[17] D. A. Carlson, A. B. Haurie, and A. Leizarowitz, *Infinite Horizon Optimal Control. Deterministic and Stochastic Systems*, Springer-Verlag, Berlin, 1991.

[18] L. Cesari, *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, Springer-Verlag, Berlin, 1959.

[19] L. Cesari, *Optimization—Theory and Applications. Problems with Ordinary Differential Equations*, Springer-Verlag, New York, 1983.

[20] B. P. Demidovich, *Lectures on Mathematical Stability Theory*, The Moscow State University Press, Moscow, 1998 (in Russian).

[21] I. Ekeland, *Some variational problems arising from mathematical economics*, in Mathematical Economics, Lecture Notes in Math. 1330, Springer-Verlag, Berlin, 1988, pp. 1–18.

[22] A. F. Filippov, *On certain questions in the theory of optimal control*, Vestn. Mosk. Univ. Ser. Mat. Mekh. Astron. Fiz. Khom., 2 (1959), pp. 25–32 (in Russian); J. Soc. Indust. Appl. Math. Ser. A: Control, 1 (1962), pp. 76–84 (in English).

[23] G. M. Grossman and E. Helpman, *Innovation and Growth in the Global Economy*, MIT Press, Cambridge, MA, 1991.

[24] H. Halkin, *Necessary conditions for optimal control problems with infinite horizons*, Econometrica, 42 (1974), pp. 267–272.

[25] P. Hartman, *Ordinary Differential Equations*, John Wiley and Sons, New York, 1964.

[26] T. Kamihigashi, *Necessity of transversality conditions for infinite horizon problems*, Econometrica, 69 (2001), pp. 995–1012.

[27] T. Koopmans, *Objectives, constraints, and outcomes in optimal growth models*, Econometrica, 35 (1967), pp. 1–15.

[28] S. F. Leung, *Transversality condition and optimality in a class of infinite horizon continuous*

time economic models, J. Econom. Theory, 54 (1991), pp. 224–233.
[29] A. M. Lyapunov, *Probléme général de la stabilité du mouvement*, Ann. of Math. Stud. 17, Princeton University Press, Princeton, NJ, 1947.
[30] O. L. Mangasarian, *Sufficient conditions for the optimal control of nonlinear systems*, SIAM J. Control, 4 (1966), pp. 139–152.
[31] P. Michel, *On the transversality conditions in infinite horizon optimal problems*, Econometrica, 50 (1982), pp. 975–985.
[32] B. Sh. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian).
[33] W. D. Nordhaus, *Managing the Global Commons. The Economics of Climate Change*, MIT Press, Cambridge, MA, 1994.
[34] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience Publishers, John Wiley and Sons, New York, 1962.
[35] F. P. Ramsey, *A mathematical theory of saving*, Econom. J., 38 (1928), pp. 543–559.
[36] A. Seierstad and K. Sydsæter, *Optimal Control Theory with Economic Applications*, North–Holland, Amsterdam, 1987.
[37] K. Shell, *Applications of Pontryagin's maximum principle to economics*, in Mathematical Systems Theory and Economics I, H. W. Kuhn and G. P. Szegö, eds., Springer-Verlag, Berlin, 1969, pp. 241–292.
[38] G. V. Smirnov, *Transversality conditions for infinite horizon problems*, J. Optim. Theory Appl., 88 (1996), pp. 671–688.
[39] R. M. Solow, *Growth Theory: An Exposition*, Oxford University Press, New York, 1970.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 3, pp. 1120–1150

STOCHASTIC LINEAR-QUADRATIC CONTROL WITH CONIC CONTROL CONSTRAINTS ON AN INFINITE TIME HORIZON*

XI CHEN† AND XUN YU ZHOU‡

Abstract. This paper is concerned with a stochastic linear-quadratic (LQ) control problem in the infinite time horizon where the control is constrained in a given, arbitrary closed cone, the cost weighting matrices are allowed to be indefinite, and the state is scalar-valued. First, the (mean-square, conic) stabilizability of the system is defined, which is then characterized by a set of simple conditions involving linear matrix inequalities (LMIs). Next, the issue of well-posedness of the underlying optimal LQ control, which is necessitated by the indefiniteness of the problem, is addressed in great detail, and necessary and sufficient conditions of the well-posedness are presented. On the other hand, to address the LQ optimality two new algebraic equations à la Riccati, called extended algebraic Riccati equations (EAREs), along with the notion of their stabilizing solutions, are introduced for the first time. Optimal feedback control as well as the optimal value are explicitly derived in terms of the stabilizing solutions to the EAREs. Moreover, several cases when the stabilizing solutions do exist are discussed and algorithms of computing the solutions are presented. Finally, numerical examples are provided to illustrate the theoretical results established.

Key words. stochastic linear-quadratic control, infinite time horizon, constrained control, cone, (conic) stabilizability, well-posedness, extended algebraic Riccati equations

AMS subject classifications. 93E20, 93E15, 49K40

DOI. 10.1137/S0363012903429529

1. Introduction.

Linear-quadratic (LQ) control, pioneered by Kalman [17] for deterministic systems and extended to stochastic systems by Wonham [31, 32] and Bismut [5], constitutes, in both theory and applications, an extremely important class of control problems. In recent years, there has been considerable renewed interest in stochastic LQ control. In particular, the notion of mean-square stabilizability and detectability was introduced in [12]. On the other hand, initiated by Chen, Li, and Zhou [10], extensive research has been carried out in the so-called *indefinite* stochastic LQ control, where, quite contrary to the conventional belief, the cost weighting matrices are allowed to be indefinite; see [11, 2, 1, 33]. Moreover, this new theory has found applications in financial portfolio selection; see [35, 19, 18]. For systematic accounts of the deterministic and stochastic LQ theory, refer to [4] and [34], respectively.

A key assumption in the LQ theory at large, deterministic and stochastic alike, is that the control variable is unconstrained. This assumption renders the feedback control constructed via the Riccati equation *automatically* admissible, and in turn (along with the underlying LQ structure) makes possible the elegant explicit solution to the optimal LQ control problem. Because of this, the whole conventional LQ approach would collapse in the presence of any control constraint.

*Received by the editors June 9, 2003; accepted for publication (in revised form) March 8, 2004; published electronically November 17, 2004.

http://www.siam.org/journals/sicon/43-3/42952.html

†Center for Intelligent and Networked Systems, Tsinghua University, Beijing, China (bjchenxi@tsinghua.edu.cn). The work of this author was done while she was a Postdoctoral Fellow at the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong.

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong (xyzhou@se.cuhk.edu.hk). This author's work was supported by RGC Earmarked grants CUHK4175/00E and CUHK 4234/01E.

On the other hand, from a practical point of view, LQ control with control constraints is a well-defined and sensible problem. For example, in many real applications the control variable is required to take only nonnegative values. The mean-variance portfolio selection problem with short-selling prohibition exemplifies such problems. Other applications include models in medicine, chemistry, and economics where system inputs are inherently constrained.

There have been some attempts in dealing with *deterministic* LQ problems with positive controls or, more general, with controls contained in a given cone. For example, controllability for linear system $\dot{x} = Ax + Bu$ with positive/conic controls was studied in [26, 7, 25, 14, 8, 22]. These papers investigated the necessary and sufficient conditions of different types of controllability (null-controllability, global controllability, differential controllability, etc.). Later, conic stabilization was addressed in [23]. In a recent work on positive feedback stabilization [30], a stabilizing positive feedback controller was derived based on the pole placement technique.

Deterministic continuous-time LQ optimal control problems with positive controllers were studied in [21, 9, 13]. Discrete-time versions can be found in [27, 28]. In these works, however, only some necessary and sufficient conditions for optimality were derived based on Pontryagin's maximum principle and/or Bellman's dynamic programming, and some numerical schemes were suggested. The special LQ structure was not fully taken advantage of, and no explicit result comparable to those of unconstrained control was obtained.

As for the constrained *stochastic* LQ control, to the best of our knowledge it has never been studied by anyone else in the literature. A related, albeit specific, problem is the mean-portfolio selection model with no-shorting constraint solved in Li, Zhou, and Lim [18], which was formulated as a stochastic LQ control problem with positive controls in a finite time horizon. The approach developed in [18] is nevertheless rather ad hoc (via the Hamilton–Jacobi–Bellman equation and viscosity solution theory) and by no means suggests a remedy for a more general problem. More recently, a stochastic LQ control problem with conic control constraint, random coefficients, as well as possibly singular cost weighting matrices in a finite time horizon was solved by Hu and Zhou [15], with explicit solutions based on Tanaka's formula and the backward stochastic differential equation theory.

In this paper, we study stochastic LQ control in the infinite time horizon, where the control variable is constrained in a cone (which includes the problem with positive controls as a special case). Moreover, the problem is allowed to be indefinite in the sense that the cost weighting matrices are possibly indefinite. A main assumption of the paper is that the state variable is scalar-valued. Note that this assumption is valid in many meaningful practical applications, in particular in the area of finance where the one-dimensional wealth process is typically taken as the state. The investigation in this paper centers around several key issues associated with the problem, namely, conic stabilizability, well-posedness, and optimality. Conic stabilizability refers to the question of whether the system can be stabilized by a control satisfying the given conic constraint. It arises from the infinity of the time horizon under consideration, and is quite different from the normal stabilizability for unconstrained control. In this paper we will derive simple necessary and sufficient conditions for the conic stabilizability. The second issue, well-posedness of the LQ problem, becomes an *issue* because the problem is indefinite. To ensure well-posedness the problem data must coordinate well, which will be characterized in this paper by the nonemptiness of certain sets in the real space. Finally, for optimality, we aim to obtain *explicit* solutions comparable

to those classical unconstrained-control counterparts. To this end, we will introduce, for the first time in this paper, two algebraic equations termed the extended algebraic Riccati equations (EAREs) along with the notion of their stabilizing solutions. Then it will be shown that the existence of the stabilizing solutions is *sufficient* for the existence of optimal feedback control of the constrained LQ problem, and explicit forms of the optimal feedback control as well as the optimal cost value will be derived in terms of the stabilizing solutions. Furthermore, several important cases, including that of the definite LQ control, will be discussed where stabilizing solutions to the EAREs do exist, and algorithms for computing these solutions will be presented. To demonstrate the theoretical results obtained, numerical examples will be given.

The rest of the paper is organized as follows. In section 2, the constrained stochastic LQ control problem is formulated and conic stabilizability of the system defined. As a prelude to the main analysis, two technical lemmas are presented in section 3. The subsequent three sections, sections 4, 5, and 6, are devoted to the three major issues, namely, stabilizability, well-posedness, and optimality, respectively. Numerical examples are reported in section 7. Finally, section 8 concludes the paper.

2. Problem formulation.

Notation. We make use of the following basic notation in this paper:

$\mathbf{R}^n$	:	the set of n-dimensional column vectors.
$\mathbf{R}$	:	$= \mathbf{R}^1$.
$\mathbf{R}^n_+$	:	the subset of $\mathbf{R}^n$ consisting of elements with nonnegative components.
$\mathbf{R}^{m\times n}$	:	the set of all $m \times n$ matrices.
$\mathbf{S}^{n\times n}$	:	the set of all $n \times n$ symmetric matrices.
$\lvert v\rvert$	:	$= \sqrt{v'v}$, $v \in \mathbf{R}^n$.
x^+	:	$= \max\{x, 0\}$, $x \in \mathbf{R}$.
x^-	:	$= \max\{-x, 0\}$, $x \in \mathbf{R}$.
M'	:	the transpose of a matrix M.
$M > 0$	:	the square matrix M is positive definite.
$M \geq 0$	:	the square matrix M is positive semidefinite.
$\mathbf{E}[x]$	:	the expectation of a random variable x.

Let $(\Omega, \mathcal{F}, \mathbf{P}; \mathcal{F}_t)$ be a given filtered probability space with a standard $\mathcal{F}_t$-adapted, k-dimensional Brownian motion $w(t) \equiv (w_1(t), w_2(t), \ldots, w_k(t))'$ on $[0, +\infty)$. Let $\Gamma \subseteq \mathbf{R}^m$ be a given closed cone; i.e., Γ is closed, and if $u \in \Gamma$, then $\alpha u \in \Gamma$ $\forall \alpha \geq 0$. Typical examples of such a cone are $\Gamma = \mathbf{R}^m_+$, $\Gamma = \{u \in \mathbf{R}^m | Mu \leq 0\}$, and $\Gamma = \{u \in \mathbf{R}^m | Mu = 0\}$, where $M \in \mathbf{R}^{n\times m}$, or the so-called second-order cone (cf., e.g., [20, p. 221]) $\Gamma = \{(t, u) \in \mathbf{R} \times \mathbf{R}^{m-1} | t \geq |u|\}$. Next, for $M \in \mathbf{S}^{m\times m}$ if there is $\delta > 0$ so that $K'MK \geq \delta |K|^2$ $\forall 0 \neq K \in \Gamma$, then we denote $M|_\Gamma > 0$. Similarly we write $M|_\Gamma \geq 0$ if $K'MK \geq 0$ $\forall K \in \Gamma$. Clearly $M > 0$ (respectively, $M \geq 0$) implies $M|_\Gamma > 0$ (respectively, $M|_\Gamma \geq 0$). Finally, define the following Hilbert space:

$$L^2_{\mathcal{F}}(\Gamma) = \Big\{ \phi(\cdot,\cdot) : [0,+\infty) \times \Omega \to \Gamma \Big| \phi(\cdot,\cdot) \text{ is } \mathcal{F}_t\text{-adapted, measurable, and } \mathbf{E}\int_0^{+\infty} |\phi(t,\omega)|^2 dt < +\infty \Big\}$$

with the norm $\|\phi(\cdot,\cdot)\| := (\mathbf{E}\int_0^{+\infty} |\phi(t,\omega)|^2 dt)^{\frac{1}{2}}$.

Consider the Itô stochastic differential equation (SDE)

$$\begin{cases} dx(t) = [Ax(t) + Bu(t)]dt + \sum_{j=1}^{k} [C_j x(t) + D_j u(t)]dw_j(t), \quad t \in [0, \infty), \\ x(0) = x_0 \in \mathbf{R}, \end{cases} \tag{1}$$

where $A, C_j \in \mathbf{R}$ and $B, D_j \in \mathbf{R}^{1\times m}$. A process $u(\cdot)$ is called a control (with conic constraint) if $u(\cdot) \in L^2_{\mathcal{F}}(\Gamma)$.

DEFINITION 2.1. *A control $u(\cdot) \in L^2_{\mathcal{F}}(\Gamma)$ is called (mean-square, conic) stabilizing with respect to x_0 if the corresponding state $x(\cdot)$ of* (1) *with the initial state x_0 satisfies* $\lim_{t\to+\infty} \mathbf{E}|x(t)|^2 = 0$.

DEFINITION 2.2. *System* (1) *is said to be (mean-square, conic) stabilizable if there is a feedback control of the form $u(t) = K_+x^+(t) + K_-x^-(t)$, where K_+ and K_- are constant vectors with $K_+ \in \Gamma$ and $K_- \in \Gamma$, which is (mean-square, conic) stabilizing with respect to* every *initial state x_0.*

Now, for any $x_0 \in \mathbf{R}$, we define the set of admissible controls

$$\mathcal{U}_{x_0} := \{u(\cdot) \in L^2_{\mathcal{F}}(\Gamma) | u(\cdot) \text{ is stabilizing with respect to } x_0\}. \tag{2}$$

If $u(\cdot) \in \mathcal{U}_{x_0}$ and $x(\cdot)$ is the corresponding solution of (1), then $(x(\cdot), u(\cdot))$ is called an admissible pair (with respect to x_0). For each $(x_0, u(\cdot)) \in \mathbf{R} \times \mathcal{U}_{x_0}$ the associated cost to system (1) is

$$J(x_0; u(\cdot)) := \mathbf{E}\int_0^{+\infty} [Qx(t)^2 + u(t)'Ru(t)]dt, \tag{3}$$

where $Q \in \mathbf{R}$ and $R \in \mathbf{S}^{m\times m}$. Note that here Q and R are *not* assumed to be nonnegative/positive semidefinite. As a result, $J(x_0; u(\cdot))$ is not necessarily bounded below.

The indefinite LQ control problem with conic constraint entails minimizing the cost functional (3), for a given x_0, subject to (1) and $u(\cdot) \in \mathcal{U}_{x_0}$. Such a problem is denoted as problem (LQ). An admissible control $u(\cdot) \in \mathcal{U}_{x_0}$ is called optimal (with respect to x_0) if $u(\cdot)$ achieves the infimum of (3), and in this case problem (LQ) is also referred to as attainable (with respect to x_0).

The value function V is defined as

$$V(x_0) := \inf_{u(\cdot)\in\mathcal{U}_{x_0}} J(x_0; u(\cdot)), \quad x_0 \in \mathbf{R}, \tag{4}$$

where $V(x_0)$ is set to be $+\infty$ in the case when $\mathcal{U}_{x_0}$ is empty.

DEFINITION 2.3. *Problem* (LQ) *is called well-posed if*

$$V(x_0) > -\infty \quad \forall x_0 \in \mathbf{R}. \tag{5}$$

It is well known that V is a continuous, though not necessarily differentiable, function when problem (LQ) is well-posed. Also note that a well-posed problem is not necessarily attainable with respect to any x_0 (see Example 6.2).

3. Two lemmas. In this section we present two lemmas that are useful in what follows.

LEMMA 3.1 (Tanaka's formula). *Let $X(t)$ be a continuous semimartingale. Then*

$$\begin{aligned} dX^+(t) &= 1_{(X(t)>0)}dX(t) + \frac{1}{2}dL(t), \\ dX^-(t) &= -1_{(X(t)\le 0)}dX(t) + \frac{1}{2}dL(t), \end{aligned} \tag{6}$$

where $L(\cdot)$ is an increasing continuous process, called the local time of $X(\cdot)$ at 0, satisfying

$$\int_0^t |X(s)|dL(s) = 0, \quad \mathbf{P}\text{-a.s.} \tag{7}$$

In particular, $X^+(t)$ and $X^-(t)$ are semimartingales.

Proof. See, for example, [24, Chapter VI, Theorem 1.2 and Proposition 1.3]. □

LEMMA 3.2. *Let constants $N_+, N_- \in \mathbf{R}$ be given. Then for any admissible pair $(x(\cdot), u(\cdot))$ with respect to x_0, we have, for every $t \ge 0$,*

$$\begin{aligned} &\mathbf{E}\int_0^t [Qx(s)^2 + u(s)'Ru(s)]ds \\ &= N_+(x_0^+)^2 + N_-(x_0^-)^2 - \mathbf{E}[N_+x^+(t)^2] - \mathbf{E}[N_-x^-(t)^2] \\ &\quad + \mathbf{E}\int_0^t \left\{ Qx(s)^2 + \left(2A + \sum_{j=1}^k C_j^2\right) N_+x^+(s)^2 + \left(2A + \sum_{j=1}^k C_j^2\right) N_-x^-(s)^2 \right. \\ &\quad + u(s)' \left[R + 1_{(x(s)>0)}N_+ \sum_{j=1}^k D_j'D_j + 1_{(x(s)\le 0)}N_- \sum_{j=1}^k D_j'D_j \right] u(s) \\ &\quad \left. + 2\left(B + \sum_{j=1}^k C_jD_j\right) u(s)N_+x^+(s) - 2\left(B + \sum_{j=1}^k C_jD_j\right) u(s)N_-x^-(s) \right\} ds. \end{aligned} \tag{8}$$

Proof. Let $x(\cdot)$ be the solution of (1) under an arbitrary $u(\cdot) \in \mathcal{U}_{x_0}$. By Lemma 3.1, we have

$$\begin{aligned} dx^+(t) &= 1_{(x(t)>0)}[Ax(t) + Bu(t)]dt + 1_{(x(t)>0)} \sum_{j=1}^k [C_jx(t) + D_ju(t)]dw_j(t) \\ &\quad + \frac{1}{2}dL(t), \\ dx^-(t) &= -1_{(x(t)\le 0)}[Ax(t) + Bu(t)]dt - 1_{(x(t)\le 0)} \sum_{j=1}^k [C_jx(t) + D_ju(t)]dw_j(t) \\ &\quad + \frac{1}{2}dL(t). \end{aligned}$$

In the above equations, $L(\cdot)$ is the local time as specified in Lemma 3.1. Applying Itô's formula, we get

$$\begin{aligned}
&d[N_+x^+(t)^2]\\
&= 2N_+x^+(t)\left\{1_{(x(t)>0)}[Ax(t)+Bu(t)]dt+1_{(x(t)>0)}\sum_{j=1}^{k}[C_jx(t)+D_ju(t)]dw_j(t)+\frac{1}{2}dL(t)\right\}\\
&\quad+1_{(x(t)>0)}N_+\sum_{j=1}^{k}[C_jx(t)+u(t)'D_j'][C_jx(t)+D_ju(t)]dt\\
&=\left\{N_+[2Ax^+(t)^2+2Bu(t)x^+(t)+1_{(x(t)>0)}\sum_{j=1}^{k}(C_jx(t)+u(t)'D_j')(C_jx(t)+D_ju(t))]\right\}dt\\
&\quad+N_+\sum_{j=1}^{k}[2C_jx^+(t)^2+2D_ju(t)x^+(t)]dw_j(t)\\
&=\left[\left(2A+\sum_{j=1}^{k}C_j^2\right)N_+x^+(t)^2+2Bu(t)N_+x^+(t)+2\sum_{j=1}^{k}C_jD_ju(t)N_+x^+(t)\right.\\
&\quad\left.+1_{(x(t)>0)}N_+u(t)'\sum_{j=1}^{k}D_j'D_ju(t)\right]dt+\left[N_+(t)\sum_{j=1}^{k}(2C_jx^+(t)^2+2D_ju(t)x^+(t))\right]dw_j(t),
\end{aligned}\tag{9}$$

where we have used the fact that $x^+(t)dL(t)=0$ by virtue of (7). Similarly, for the constant $N_-\in\mathbf{R}$,

$$\begin{aligned}
&d[N_-x^-(t)^2]\\
&=\left\{N_-\left[2Ax^-(t)^2-2Bu(t)x^-(t)+1_{(x(t)\le0)}\sum_{j=1}^{k}(C_jx(t)+u(t)'D_j')(C_jx(t)+D_ju(t))\right]\right\}dt\\
&\quad+N_-\sum_{j=1}^{k}[2C_jx^-(t)^2-2D_ju(t)x^-(t)]dw_j(t)\\
&=\left[\left(2A+\sum_{j=1}^{k}C_j^2\right)N_-x^-(t)^2-2Bu(t)P_-x^-(t)-2\sum_{j=1}^{k}C_jD_ju(t)N_-x^-(t)\right.\\
&\quad\left.+1_{(x(t)\le0)}N_-u(t)'\sum_{j=1}^{k}D_j'D_ju(t)\right]dt+\left[N_-(t)\sum_{j=1}^{k}(2C_jx^-(t)^2-2D_ju(t)x^-(t))\right]dw_j(t).
\end{aligned}\tag{10}$$

Fix $t\ge0$ and define a sequence of stopping times

$$\begin{aligned}
\tau_n:=\inf\left\{r\in[0,t]:\int_0^r\left|N_+(s)\sum_{j=1}^{k}(2C_jx^+(s)^2+2D_ju(s)x^+(s))\right|^2ds\right.\\
\left.+\int_0^r\left|N_-(s)\sum_{j=1}^{k}(2C_jx^-(s)^2-2D_ju(s)x^-(s))\right|^2ds\ge n\right\},\quad n=1,2,\ldots,
\end{aligned}$$

where $\inf \emptyset := t$. It is clear that $\tau_n \uparrow t$ as $n \to +\infty$, due to $E\int_0^t (|x(s)|^2 + |u(s)|^2)ds < +\infty$. Now, summing up (9) and (10), taking integration from 0 to τ_n and then, taking expectation, we obtain (8) with t replaced by τ_n. Thus, (8) follows by sending $n \to +\infty$ together with Fatou's lemma. □

4. Conic stabilizability. In this section we address the issue of the conic stabilizability of system (1). Notice that conic stabilizability is different from the usual stabilizability with unconstrained controls, for clearly the former requires more stringent conditions. Here we will give a *complete* characterization of conic stabilizability in terms of simple conditions involving linear matrix inequalities (LMIs).

Introduce a pair of functions F_+, F_- from Γ to $\mathbf{R}$:

$$F_+(K) := 2A + \sum_{j=1}^k C_j^2 + 2\left(B + \sum_{j=1}^k C_j D_j\right) K + K' \sum_{j=1}^k D_j' D_j K, \tag{11}$$

and

$$F_-(K) := 2A + \sum_{j=1}^k C_j^2 - 2\left(B + \sum_{j=1}^k C_j D_j\right) K + K' \sum_{j=1}^k D_j' D_j K. \tag{12}$$

THEOREM 4.1. *The following assertions are equivalent.*

(i) *System* (1) *is mean-square conic stabilizable.*

(ii) *There exist* $K_+ \in \Gamma$ *and* $K_- \in \Gamma$ *such that* $F_+(K_+) < 0$ *and* $F_-(K_-) < 0$. *In this case the feedback control* $u(t) = K_+ x^+(t) + K_- x^-(t)$ *is stabilizing.*

(iii) *There exist* $K_+ \in \Gamma$ *and* $K_- \in \Gamma$ *such that*

$$\begin{pmatrix} 2A + \sum_{j=1}^k C_j^2 + 2\left(B + \sum_{j=1}^k C_j D_j\right) K_+ & K_+' D' \\ DK_+ & -I \end{pmatrix} < 0, \tag{13}$$

$$\begin{pmatrix} 2A + \sum_{j=1}^k C_j^2 - 2\left(B + \sum_{j=1}^k C_j D_j\right) K_- & K_-' D' \\ DK_- & -I \end{pmatrix} < 0, \tag{14}$$

where $D \in \mathbf{R}^{m\times m}$ *satisfies* $D'D = \sum_{j=1}^k D_j' D_j$. *In this case the feedback control* $u(t) = K_+ x^+(t) + K_- x^-(t)$ *is stabilizing.*

Proof. Take a feedback control $u(t) = K_+ x^+(t) + K_- x^-(t)$ and consider the corresponding state $x(\cdot)$ with an initial state x_0. Note that by standard SDE theory (cf., e.g., [16]) such $x(\cdot)$ uniquely exists. Moreover,

$$\mathbf{E} \sup_{0\le t\le T} |x(t)|^p dt < +\infty \ \ \forall T > 0 \ \ \forall p \ge 0. \tag{15}$$

Making use of (9) with $N_+ = 1$ and $u(t) = K_+x^+(t) + K_-x^-(t)$, we obtain

$$dx^+(t)^2 = \left[2A + \sum_{j=1}^k C_j^2 + 2\left(B + \sum_{j=1}^k C_jD_j\right)K_+ + K_+' \sum_{j=1}^k D_j'D_jK_+\right] x^+(t)^2dt$$
$$+ \sum_{j=1}^k (2C_j + 2D_jK_+)x^+(t)^2dw_j(t)$$
$$= F_+(K_+)x^+(t)^2dt + \sum_{j=1}^k (2C_j + 2D_jK_+)x^+(t)^2dw_j(t).$$

Taking integration and then expectation yields (after a localization argument as in the proof of Lemma 3.2)

$$\frac{d\mathbf{E}[x^+(t)^2]}{dt} = F_+(K_+)\mathbf{E}[x^+(t)^2], \tag{16}$$

where the expectation of the Itô integral vanishes due to (15). Similarly, we have

$$\frac{d\mathbf{E}[x^-(t)^2]}{dt} = F_-(K_-)\mathbf{E}[x^-(t)^2]. \tag{17}$$

Hence, the equivalence between the assertions (i) and (ii) is evident noting that $\lim_{t\to+\infty} \mathbf{E}|x(t)|^2 = 0$ if and only if $\lim_{t\to+\infty} \mathbf{E}|x^+(t)|^2 = 0$ and $\lim_{t\to+\infty} \mathbf{E}|x^-(t)|^2 = 0$.

Finally, the equivalence between the assertions (ii) and (iii) follows from Schur's lemma [6]. □

Conditions (13) and (14) are in terms of LMIs. On the other hand, some of the conic constraint can also be expressed as LMIs, e.g., when $\Gamma = \mathbf{R}_+^m$ or when Γ is a second-order cone. Hence, Theorem 4.1(iii) provides an easy way of numerically checking the stabilizability of system (1) due to the availability of many LMI solvers. Note that in general there may exist many feasible solutions to LMIs.

Denote the two sets of feedback gains as

$$\mathcal{K}_+ := \{K \in \Gamma | F_+(K) < 0\}, \quad \mathcal{K}_- := \{K \in \Gamma | F_-(K) < 0\}.$$

Then Theorem 4.1 implies the following

THEOREM 4.2. *System* (1) *is conic stabilizable if and only if* $\mathcal{K}_+ \neq \emptyset$ *and* $\mathcal{K}_- \neq \emptyset$.

COROLLARY 4.1. *If* $2A + \sum_{j=1}^k C_j^2 < 0$, *then system* (1) *is conic stabilizable.*

Proof. In this case $0 \in \mathcal{K}_+ \cap \mathcal{K}_-$. □

PROPOSITION 4.1. *We have the following results.*

(i) $\mathcal{K}_+$ *and* $\mathcal{K}_-$ *are convex sets if* Γ *is a convex set.*

(ii) $\mathcal{K}_+$ *and* $\mathcal{K}_-$ *are bounded if* $\sum_{j=1}^k D_j'D_j|_\Gamma > 0$.

Proof. (i) Define the following operator M_+ from Γ to $\mathbf{S}^{(m+1)\times(m+1)}$:

$$M_+(K) = \begin{pmatrix} 2A + \sum_{j=1}^k C_j^2 + 2\left(B + \sum_{j=1}^k C_jD_j\right)K & K'D' \\ DK & -I \end{pmatrix},$$

where $D'D = \sum_{j=1}^k D_j'D_j$. By Theorem 4.1, $\mathcal{K}_+$ can be equivalently represented as $\mathcal{K}_+ = \{K \in \Gamma | M_+(K) < 0\}$. Thus the convexity of $\mathcal{K}_+$ follows from that of Γ together

with the fact that the operator M_+ is affine. Similarly we can show the convexity of $\mathcal{K}_-$.

(ii) If $\mathcal{K}_+$ is unbounded, then there is a sequence $\{K_n\} \subset \mathcal{K}_+$ so that $|K_n| \to +\infty$. Since $\sum_{j=1}^k D_j' D_j|_\Gamma > 0$, we have $K_n' \sum_{j=1}^k D_j' D_j K_n \geq \delta |K_n|^2 \to +\infty$ as $n \to +\infty$, where $\delta > 0$ is some constant. Therefore $F_+(K_n) \to +\infty$ as $n \to +\infty$. This contradicts $F_+(K_n) < 0 \ \forall n$. Hence $\mathcal{K}_+$ is bounded. Similarly, $\mathcal{K}_-$ is bounded. □

5. Well-posedness. Since the cost weighting matrices Q and R are allowed to be indefinite, the well-posedness of the problem is no longer automatic or trivial (as opposed to the classical definite case when $Q \geq 0$ and $R > 0$). In fact, the well-posedness for an indefinite LQ control problem is a prerequisite for the optimality and is an interesting problem in its own right. In this section we will carry out an extensive investigation on the well-posedness and some related issues, including necessary and sufficient conditions for the well-posedness in terms of the nonemptiness of certain sets.

5.1. Representation of value function. In this subsection we present the following representation result, which is a key to many results of this paper.

PROPOSITION 5.1. *Problem* (LQ) *is well-posed if and only if the value function can be represented as*

$$V(x_0) = P_+(x_0^+)^2 + P_-(x_0^-)^2 \quad \forall x_0 \in \mathbf{R} \tag{18}$$

for some $P_+, P_- \in \mathbf{R}$.

Proof. We prove only the "only if" part, as the "if" part is evident.

Assume that problem (LQ) is well-posed. Fix any $x > 0, y > 0$. Since $V(y) > -\infty$, for any $\varepsilon > 0$ there is $u^\varepsilon(\cdot) \in \mathcal{U}_y$ along with the corresponding state $x^\varepsilon(\cdot)$ (with the initial state y) satisfying

$$V(y) \geq J(y; u^\varepsilon(\cdot)) - \varepsilon = \mathbf{E} \int_0^{+\infty} [Q x^\varepsilon(t)^2 + u^\varepsilon(t)' R u^\varepsilon(t)] dt - \varepsilon. \tag{19}$$

Now, as $x > 0$, the linearity of the dynamics (1) and the conic control constraint ensure that $x u^\varepsilon(\cdot) \in \mathcal{U}_{xy}$ with the corresponding state $x x^\varepsilon(\cdot)$. Hence it follows from (19) that

$$V(y) \geq \frac{1}{x^2} \mathbf{E} \int_0^{+\infty} [Q|x x^\varepsilon(t)|^2 + (x u^\varepsilon(t))' R(x u^\varepsilon(t))] dt - \varepsilon \geq \frac{1}{x^2} V(xy) - \varepsilon. \tag{20}$$

Sending $\varepsilon \to 0$ we obtain

$$V(xy) \leq V(y) x^2 \quad \forall x > 0, y > 0. \tag{21}$$

Similarly, one can show that

$$V(xy) \leq V(-y) x^2 \quad \forall x < 0, y > 0. \tag{22}$$

Now for any $x > 0$, by (21) we have $V(x) \leq V(1)x^2$. On the other hand, (21) also implies $V(1) = V(x\frac{1}{x}) \leq \frac{1}{x^2} V(x)$. So we have shown that

$$V(x) = V(1) x^2 \quad \forall x > 0. \tag{23}$$

Similarly, in view of (22) we can prove that

$$V(x) = V(-1) x^2 \quad \forall x < 0. \tag{24}$$

Finally, the continuity of the value function along with (23) yields

$$V(0) = 0. \tag{25}$$

The desired result (18) thus follows from (23)–(25) with $P_+ := V(1)$ and $P_- := V(-1)$. □

Remark 5.1. The preceding proposition, which suggests the form of the value function when the underlying LQ problem is well-posed, is crucial for all the main results in this paper. In fact, the proofs for the stabilizability in the previous section, the characterization of the well-posedness in this section, and the optimality in the next section are *all* inspired by this result. This also explains why one needs to apply Tanaka's formula to evaluate $dx^+(t)$ and $dx^-(t)$, as we have seen in the previous section and will continue to see in the subsequent sections.

Remark 5.2. We saw that the value function for the constrained LQ problem is *not* smooth.

5.2. Characterization of well-posedness. Define the following functions from $\mathbf{R}$ to $\mathbf{R} \cup \{-\infty\}$:

$$\Phi_+(P) := \inf_{K\in\Gamma} \left[K' \left(R + P\sum_{j=1}^k D_j'D_j \right) K + 2 \left(B + \sum_{j=1}^k C_jD_j \right) PK \right],$$

$$\Phi_-(P) := \inf_{K\in\Gamma} \left[K' \left(R + P\sum_{j=1}^k D_j'D_j \right) K - 2 \left(B + \sum_{j=1}^k C_jD_j \right) PK \right].$$

Remark 5.3. Since $0 \in \Gamma$, we must have

$$\Phi_+(P) \le 0 \quad \Phi_-(P) \le 0 \ \ \forall P \in \mathbf{R}. \tag{26}$$

On the other hand, $\Phi_+(P)$ and $\Phi_-(P)$ have finite values if $(R+P\sum_{j=1}^k D_j'D_j)|_\Gamma > 0$. Indeed, in this case there exist constants $\alpha_1 = \alpha_1(P) > 0$ and $\alpha_2 = \alpha_2(P) > 0$ such that

$$\begin{aligned} K' \left(R + P\sum_{j=1}^k D_j'D_j \right) K + 2 \left(B + \sum_{j=1}^k C_jD_j \right) PK &\ge \alpha_1|K|^2 - \alpha_2|K| \\ &= \alpha_1|K| \left(|K| - \frac{\alpha_2}{\alpha_1} \right) \qquad \forall K \in \Gamma. \end{aligned}$$

If $|K| > \frac{\alpha_2}{\alpha_1}$, then the above expression is positive. Taking (26) into consideration we conclude

$$\Phi_+(P) = \inf_{K\in\Gamma, |K| \le \frac{\alpha_2}{\alpha_1}} \left[K' \left(R + P\sum_{j=1}^k D_j'D_j \right) K + 2 \left(B + \sum_{j=1}^k C_jD_j \right) PK \right] > -\infty.$$

Hence, $\Phi_+(P)$ is finite. The same is true for $\Phi_-(P)$.

Next, we define the following two sets:

$$\mathcal{P}_+ := \left\{ P \in \mathbf{R} \,\middle|\, \left(2A + \sum_{j=1}^k C_j^2 \right) P + Q + \Phi_+(P) \ge 0, \ \left(R + P\sum_{j=1}^k D_j'D_j \right)\Big|_\Gamma \ge 0 \right\},$$

$$\mathcal{P}_- := \left\{ P \in \mathbf{R} \,\middle|\, \left(2A + \sum_{j=1}^k C_j^2 \right) P + Q + \Phi_-(P) \ge 0, \ \left(R + P\sum_{j=1}^k D_j'D_j \right)\Big|_\Gamma \ge 0 \right\}.$$

The following is the main result of the section, which characterizes the well-posedness of problem (LQ) by the nonemptiness of the sets $\mathcal{P}_+$ and $\mathcal{P}_-$.

THEOREM 5.1. *Assume that system* (1) *is conic stabilizable. Then problem* (LQ) *is well-posed if and only if* $\mathcal{P}_+ \neq \emptyset$ *and* $\mathcal{P}_- \neq \emptyset$. *Moreover, in this case*

$$V(x_0) \geq P_+(x_0^+)^2 + P_-(x_0^-)^2 \quad \forall x_0 \in \mathbf{R}, \quad \forall P_+ \in \mathcal{P}_+, \quad P_- \in \mathcal{P}_-. \tag{27}$$

Proof. First we prove the "if" part. For any $x_0 \in \mathbf{R}$ let $x(\cdot)$ be the solution of (1) under an arbitrary $u(\cdot) \in \mathcal{U}_{x_0}$. Pick any $P_+ \in \mathcal{P}_+$ and $P_- \in \mathcal{P}_-$. By Lemma 3.2, we have

$$\begin{aligned}
&\mathbf{E}\int_0^t [Qx(s)^2 + u(s)'Ru(s)]ds \\
&= P_+(x_0^+)^2 + P_-(x_0^-)^2 - \mathbf{E}[P_+x^+(t)^2] - \mathbf{E}[P_-x^-(t)^2] \\
&\quad + \mathbf{E}\int_0^t \left\{ Qx(s)^2 + \left(2A + \sum_{j=1}^k C_j^2\right) P_+x^+(s)^2 + \left(2A + \sum_{j=1}^k C_j^2\right) P_-x^-(s)^2 \right. \\
&\quad + u(s)' \left[R + 1_{(x(s)>0)} P_+ \sum_{j=1}^k D_j'D_j + 1_{(x(s)\leq 0)} P_- \sum_{j=1}^k D_j'D_j \right] u(s) \\
&\quad \left. + 2\left(B + \sum_{j=1}^k C_jD_j\right) u(s)P_+x^+(s) - 2\left(B + \sum_{j=1}^k C_jD_j\right) u(s)P_-x^-(s) \right\} ds.
\end{aligned} \tag{28}$$

Denote by $\psi(x(s), u(s))$ the integrand on the right-hand side of (28) and fix $s \in [0,t]$. If $x(s) > 0$, then write $u(s) = Kx(s)$ (note that K may depend on s). Since $u(s) \in \Gamma$ and Γ is a cone, we have $K \in \Gamma$. Hence at s, bearing in mind that $1_{(x(s)\leq 0)} = 0$, we have

$$\begin{aligned}
&\psi(x(s), u(s)) \\
&= Qx(s)^2 + \left(2A + \sum_{j=1}^k C_j^2\right) P_+x(s)^2 + K' \left[R + P_+ \sum_{j=1}^k D_j'D_j \right] Kx(s)^2 \\
&\quad + 2\left(B + \sum_{j=1}^k C_jD_j\right) KP_+x(s)^2 \\
&= \left[\left(2A + \sum_{j=1}^k C_j^2\right) P_+ + Q + K'\left(R + P_+ \sum_{j=1}^k D_j'D_j\right) K + 2\left(B + \sum_{j=1}^k C_jD_j\right) KP_+ \right] x(s)^2 \\
&\geq \left[\left(2A + \sum_{j=1}^k C_j^2\right) P_+ + Q + \Phi_+(P_+) \right] x(s)^2 \\
&\geq 0.
\end{aligned} \tag{29}$$

If $x(s) < 0$, then write $u(s) = -Kx(s)$. Again $K \in \Gamma$. An argument similar to that above yields

$$\psi(x(s), u(s)) \geq \left[\left(2A + \sum_{j=1}^k C_j^2\right) P_- + Q + \Phi_-(P_-) \right] x(s)^2 \geq 0$$

at s. Finally, if $x(s) = 0$ at s, then

$$\psi(x(s), u(s)) = u(s)' \left[R + P_- \sum_{j=1}^{k} D_j' D_j \right] u(s) \geq 0.$$

The preceding analysis shows that it always holds that $\psi(x(s), u(s)) \geq 0 \ \forall s \in [0, t]$. Consequently, it follows from (28) that

$$\begin{aligned} \mathbf{E} \int_0^t [Qx(s)^2 + u(s)' R u(s)] ds \geq P_+ (x_0^+)^2 + P_- (x_0^-)^2 - \mathbf{E}[P_+ x^+(t)^2] \\ - \mathbf{E}[P_- x^-(t)^2]. \end{aligned}$$

Letting $t \to +\infty$ and noting that $u(\cdot)$ is conic stabilizing, we obtain

$$\begin{aligned} J(x_0; u(\cdot)) &= \lim_{t \to +\infty} \mathbf{E} \int_0^t [Qx(s)^2 + u(s)' R u(s)] ds \\ &\geq P_+ (x_0^+)^2 + P_- (x_0^-)^2. \end{aligned}$$

Since $u(\cdot) \in \mathcal{U}_{x_0}$ is arbitrary, we conclude $V(x_0) \geq P_+ (x_0^+)^2 + P_- (x_0^-)^2 > -\infty$. Hence problem (LQ) is well-posed.

To prove the "only if" part, suppose that the LQ problem is well-posed. Then by Proposition 5.1 the value function has the following representation:

$$V(x) = P_+ (x^+)^2 + P_- (x^-)^2 \ \ \forall x \in \mathbf{R}.$$

We want to show that $P_+ \in \mathcal{P}_+, \ \ P_- \in \mathcal{P}_-$. To this end, applying the optimality principle of dynamic programming and noting the time-invariance of the underlying system, we obtain

$$\begin{aligned} & P_+ (x_0^+)^2 + P_- (x_0^-)^2 \\ & \leq \mathbf{E} \left\{ \int_0^h [Qx(t)^2 + u(t)' R u(t)] dt + P_+ [x^+(h)]^2 + P_- [x^-(h)]^2 \right\} \\ & \forall h > 0, \ \ \forall u(\cdot) \in \mathcal{U}_{x_0}, \ \ \forall x_0 \in \mathbf{R}. \end{aligned} \tag{30}$$

Using the above and applying Lemma 3.2, we obtain

$$\mathbf{E} \int_0^h \psi(x(s), u(s)) ds \geq 0 \quad \forall h > 0, \ \ \forall u(\cdot) \in \mathcal{U}_{x_0}, \ \ \forall x_0 \in \mathbf{R}, \tag{31}$$

where, as before, the mapping ψ is defined via the integrand on the right-hand side of (28). Set $x_0 > 0$ and take the following control

$$u_1(t) := \begin{cases} K, & 0 \leq t < 1, \\ K_+ x^+(t) + K_- x^-(t), & t \geq 1, \end{cases}$$

where $K \in \Gamma$ is arbitrarily fixed and $K_+ x^+(t) + K_- x^-(t)$ is a conic stabilizing feedback control which exists due to the stabilizability assumption. Clearly $u_1(\cdot) \in \mathcal{U}_{x_0}$, and

let $x_1(\cdot)$ be the corresponding state. Since $x_1(s) \to x_0$ and $u_1(s) \to K$ as $s \to 0$, **P**-a.s., we have

$$\begin{aligned}&\psi(x_1(s), u_1(s))\\ &\to \left[\left(2A + \sum_{j=1}^{k} C_j^2\right) P_+ + Q\right] x_0^2 + 2\left(B + \textstyle\sum_{j=1}^{k} C_j D_j\right) K P_+ x_0\\ &\quad + K'\left(R + P_+ \sum_{j=1}^{k} D_j' D_j\right) K \quad \text{as } s \to 0, \quad \mathbf{P}\text{-a.s.}\end{aligned}$$

Thus appealing to (31), the dominated convergence theorem yields

$$\begin{aligned}0 &\le \frac{1}{h}\mathbf{E}\int_0^h \psi(x_1(s), u_1(s))ds\\ &\to \left[\left(2A + \sum_{j=1}^{k} C_j^2\right) P_+ + Q\right] x_0^2 + 2\left(B + \textstyle\sum_{j=1}^{k} C_j D_j\right) K P_+ x_0\\ &\quad + K'\left(R + P_+ \sum_{j=1}^{k} D_j' D_j\right) K \quad \text{as } h \to 0.\end{aligned}$$

Letting $x_0 \to 0$ we obtain $K'(R + P_+ \sum_{j=1}^{k} D_j' D_j)K \ge 0$. The arbitrariness of $K \in \Gamma$ then implies $(R + P_+ \sum_{j=1}^{k} D_j' D_j)|_\Gamma \ge 0$. On the other hand, take $x_0 > 0$ and consider the following feedback control

$$u_2(t) := \begin{cases} Kx^+(t) + \tilde{K}x^-(t), & 0 \le t < 1,\\ K_+x^+(t) + K_-x^-(t), & t \ge 1,\end{cases}$$

where $K \in \Gamma$ and $\tilde{K} \in \Gamma$ are arbitrarily fixed. Let $x_2(\cdot)$ be the state under $u_2(\cdot)$. Noting $x_2(s) \to x_0$ and $u_2(s) = Kx_2^+(s) + \tilde{K}x_2^-(s) \to Kx_0$ as $s \to 0$, **P**-a.s., we obtain

$$\begin{aligned}&\psi(x_2(s), u_2(s))\\ &\to \left[\left(2A + \sum_{j=1}^{k} C_j^2\right) P_+ + Q + K'\left(R + P_+ \sum_{j=1}^{k} D_j' D_j\right) K + 2\left(B + \sum_{j=1}^{k} C_j D_j\right) K P_+\right] x_0^2\\ &\quad \text{as } s \to 0, \quad \mathbf{P}\text{-a.s.}\end{aligned}$$

An analysis similar to the preceding one leads to

$$\left(2A + \sum_{j=1}^{k} C_j^2\right) P_+ + Q + K'\left(R + P_+ \sum_{j=1}^{k} D_j' D_j\right) K + 2\left(B + \sum_{j=1}^{k} C_j D_j\right) K P_+ \ge 0.$$

Since $K \in \Gamma$ is arbitrary, we arrive at $(2A + \sum_{j=1}^{k} C_j^2)P_+ + Q + \Phi_+(P_+) \ge 0$. So far we have shown $P_+ \in \mathcal{P}_+$. Similarly, we can prove $P_- \in \mathcal{P}_-$.

Finally, the inequality (27) has been proved in the proof of the "if" part. □

Remark 5.4. From the above proof we see that the stabilizability assumption is in fact *not* necessary for the "if" part of the theorem.

Remark 5.5. The above theorem tells that the positive definiteness or positive semidefiniteness of Q and R are *not* necessary for problem (LQ) to be well-posed.

COROLLARY 5.1. *If* $R|_\Gamma \geq 0$ *and* $Q \geq 0$, *then problem* (LQ) *is well-posed.*

Proof. In this case $0 \in \mathcal{P}_+ \cap \mathcal{P}_-$. □

PROPOSITION 5.2. $\mathcal{P}_+$ *and* $\mathcal{P}_-$ *are both convex sets (hence they are both intervals). Moreover, if system* (1) *is stabilizable and problem* (LQ) *is well-posed, then* $\mathcal{P}_+$ *and* $\mathcal{P}_-$ *each has a finite maximum element.*

Proof. The convexity of $\mathcal{P}_+$ and $\mathcal{P}_-$ are clear noticing that the functions $\Phi_+(P)$ and $\Phi_-(P)$ are concave in P. To prove the existence of the finite maximum elements, we note that Proposition 5.1 provides

$$V(x_0) = P_+^*(x_0^+)^2 + P_-^*(x_0^-)^2 \quad \forall x_0 \in \mathbf{R} \tag{32}$$

for some $P_+^*, P_-^* \in \mathbf{R}$. Moreover, the proof of Theorem 5.1 implies $P_+^* \in \mathcal{P}_+$ and $P_-^* \in \mathcal{P}_-$. Hence it follows from (27) that P_+^* and P_-^* are the maximum elements of $\mathcal{P}_+$ and $\mathcal{P}_-$, respectively. □

Remark 5.6. Proposition 5.2 indicates that if system (1) is stabilizable and problem (LQ) is well-posed, then the infimum value of problem (LQ) or, equivalently, the P_+ and P_- in the representation of the value function as stipulated in Proposition 5.1 can be obtained by solving the following two mathematical programming problems, respectively:

$$\begin{array}{ll} \text{maximize} & P \\ \text{subject to} & \left\{ \begin{array}{l} \left(2A + \displaystyle\sum_{j=1}^{k} C_j^2 \right) P + Q + \Phi_+(P) \geq 0, \\ \left(R + P \displaystyle\sum_{j=1}^{k} D_j' D_j \right) \Big|_\Gamma \geq 0, \end{array} \right. \end{array} \tag{33}$$

and

$$\begin{array}{ll} \text{maximize} & P \\ \text{subject to} & \left\{ \begin{array}{l} \left(2A + \displaystyle\sum_{j=1}^{k} C_j^2 \right) P + Q + \Phi_-(P) \geq 0, \\ \left(R + P \displaystyle\sum_{j=1}^{k} D_j' D_j \right) \Big|_\Gamma \geq 0. \end{array} \right. \end{array} \tag{34}$$

5.3. An algorithm. Theorem 5.1 stipulates that it suffices to check the nonemptiness of $\mathcal{P}_+$ and $\mathcal{P}_-$ or the feasibility of the problems (33) and (34) in order to verify the well-posedness of a given LQ problem. However, it is sometimes hard to check numerically the aforementioned feasibility because, on one hand, the functions $\Phi_+(\cdot)$ and $\Phi_-(\cdot)$ in general do not have analytical forms, and on the other hand, the constraint $(R + P\sum_{j=1}^{k} D_j' D_j)|_\Gamma \geq 0$ is usually very hard to verify for a general cone Γ (except second-order cones for which the constraint can be reformulated as an LMI; see [29, Theorem 1]).

In this subsection we give an algorithm that can check the well-posedness more directly. First we need a lemma.

LEMMA 5.1. *Assume that problem* (LQ) *is well-posed. Given* $K_+ \in \mathcal{K}_+$ *and* $K_- \in \mathcal{K}_-$, *set* $\tilde{P}_+ := -\frac{Q+K_+'RK_+}{F_+(K_+)}$ *and* $\tilde{P}_- := -\frac{Q+K_-'RK_-}{F_-(K_-)}$. *Then*

$$P_+ \le \tilde{P}_+ \ \forall P_+ \in \mathcal{P}_+ \ \text{ and } \ P_- \le \tilde{P}_- \ \forall P_- \in \mathcal{P}_-. \tag{35}$$

Proof. By their definitions $\tilde{P}_+$ and $\tilde{P}_-$ satisfy, respectively,

$$\left(2A+\sum_{j=1}^k C_j^2\right)\tilde{P}_+ + Q + K_+'\left(R+\tilde{P}_+\sum_{j=1}^k D_j'D_j\right)K_+ + 2\left(B+\sum_{j=1}^k C_jD_j\right)K_+\tilde{P}_+ = 0 \tag{36}$$

and

$$\left(2A+\sum_{j=1}^k C_j^2\right)\tilde{P}_- + Q + K_-'\left(R+\tilde{P}_-\sum_{j=1}^k D_j'D_j\right)K_- - 2\left(B+\sum_{j=1}^k C_jD_j\right)K_-\tilde{P}_- = 0. \tag{37}$$

Take a feedback control $u(t) = K_+x^+(t) + K_-x^-(t)$, which is stabilizing by Theorem 4.1 (bearing in mind the definitions of $\mathcal{K}_+$ and $\mathcal{K}_-$), and let $x(\cdot)$ be the corresponding state with $x(0) = x_0$. Then a similar calculation as in (28) yields

$$\begin{aligned}&\mathbf{E}\int_0^t [Qx(s)^2 + u(s)'Ru(s)]ds \\ &= \tilde{P}_+(x_0^+)^2 + \tilde{P}_-(x_0^-)^2 - \mathbf{E}[\tilde{P}_+x^+(t)^2] - \mathbf{E}[\tilde{P}_-x^-(t)^2] + \mathbf{E}\int_0^t \psi(x(s),u(s))ds,\end{aligned} \tag{38}$$

where $\psi(x(s),u(s))$ is as the integrand on the right-hand side of (28), with P_+ and P_- replaced by $\tilde{P}_+$ and $\tilde{P}_-$, respectively. However, $u(s) = K_+x(s)$ whenever $x(s) > 0$; hence

$$\begin{aligned}&\psi(x(s),u(s)) \\ &= \left[\left(2A+\sum_{j=1}^k C_j^2\right)\tilde{P}_+ + Q + K_+'\left(R+\tilde{P}_+\sum_{j=1}^k D_j'D_j\right)K_+ + 2\left(B+\sum_{j=1}^k C_jD_j\right)K\tilde{P}_+\right]x(s)^2 \\ &= 0\end{aligned}$$

in view of (36). Similarly, based on (37) one can show that $\psi(x(s),u(s)) = 0$ whenever $x(s) \le 0$. It then follows from (38) that

$$\mathbf{E}\int_0^t [Qx(s)^2 + u(s)'Ru(s)]ds = \tilde{P}_+(x_0^+)^2 + \tilde{P}_-(x_0^-)^2 - \mathbf{E}[\tilde{P}_+x^+(t)^2] - \mathbf{E}[\tilde{P}_-x^-(t)^2].$$

Since $u(\cdot)$ is stabilizing, we have

$$J(x_0;u(\cdot)) = \lim_{t\to+\infty}\mathbf{E}\int_0^t [Qx(s)^2 + u(s)'Ru(s)]ds = \tilde{P}_+(x_0^+)^2 + \tilde{P}_-(x_0^-)^2.$$

By virtue of Theorem 5.1, we conclude

$$\begin{aligned}\tilde{P}_+(x_0^+)^2 + \tilde{P}_-(x_0^-)^2 &= J(x_0; u(\cdot)) \\ &\geq V(x_0) \geq P_+(x_0^+)^2 + P_-(x_0^-)^2 \quad \forall P_+ \in \mathcal{P}_+,\ \forall P_- \in \mathcal{P}_-.\end{aligned}$$

This proves (35). □

We now assume that (1) is conic stabilizable. According to Theorem 4.1 there exist $K_+ \in \Gamma$ and $K_- \in \Gamma$ such that $F_+(K_+) < 0$ and $F_-(K_-) < 0$. Take $\delta > 0$ sufficiently small with $\delta < \min\{-F_+(K_+), -F_-(K_-)\}$. Calculate

$$P_+^{(1)} := \min_{F_+(K) \leq -\delta} \Big\{ -\frac{Q + K'RK}{F_+(K)} \Big\},\ \ P_-^{(1)} := \min_{F_-(K) \leq -\delta} \Big\{ -\frac{Q + K'RK}{F_-(K)} \Big\}. \tag{39}$$

In view of Lemma 5.1, $P_+^{(1)}$ (respectively, $P_-^{(1)}$) is a (very tight) upper bound of $\mathcal{P}_+$ (respectively, $\mathcal{P}_-$) under the well-posedness assumption. As a consequence, if problem (LQ) is well-posed, then it is necessary that $(R + P_+^{(1)} \sum_{j=1}^k D_j' D_j)|_\Gamma \geq 0$ and $(R + P_-^{(1)} \sum_{j=1}^k D_j' D_j)|_\Gamma \geq 0$. Now, calculate

$$P^{(0)} := \min_{(R + P \sum_{j=1}^k D_j' D_j)|_\Gamma \geq 0} P. \tag{40}$$

Then we know that points of $\mathcal{P}_+$ (respectively, $\mathcal{P}_-$), if any, must lie between $P^{(0)}$ and $P_+^{(1)}$ (respectively, $P_-^{(1)}$). Inspired by the above discussion, we have the following algorithm.

Step 1. Apply Theorem 4.1(iii) to obtain $K_+,\ K_- \in \Gamma$ with $F_+(K_+) < 0$ and $F_-(K_-) < 0$. Set $\delta := \min\{\varepsilon, -F_+(K_+), -F_-(K_-)\}$, where ε is a very small number allowed by the computer.

Step 2. Calculate $P_+^{(1)}$ and $P_-^{(1)}$ via (39). If either $(R + P_+^{(1)} \sum_{j=1}^k D_j' D_j)|_\Gamma < 0$ or $(R + P_-^{(1)} \sum_{j=1}^k D_j' D_j)|_\Gamma < 0$ holds, stop, and problem (LQ) is not well-posed.

Step 3. Calculate $P^{(0)}$ via (40).

Step 4. If there exists a $P \in [P^{(0)}, P_+^{(1)})$ satisfying $L_+(P) \geq 0$, then go to Step 5; otherwise, stop, and problem (LQ) is not well-posed.

Step 5. If there exists a $P \in [P^{(0)}, P_-^{(1)})$ satisfying $L_-(P) \geq 0$, then stop, and problem (LQ) is well-posed; otherwise, stop, and problem (LQ) is not well-posed.

5.4. Well-posedness margin. In view of Remark 5.5, problem (LQ) may still be well-posed when R is indefinite or even negative definite. That said, it is clear that R cannot be *too* negative for the well-posedness. Therefore, it is interesting to study the *range* of R over which problem (LQ) is well-posed, given that all the other data is fixed. Specifically, define

$$r^* := \inf\{r \in \mathbf{R} \,\big|\, \text{problem (LQ) is well-posed for any } R \in \mathbf{S}^{m\times m} \text{ with } R > rI\}, \tag{41}$$

where $\inf \emptyset := +\infty$. The value r^* is called the well-posedness margin. By its very definition, r^* has the following interpretation: Problem (LQ) is well-posed if the smallest eigenvalue of R, $\lambda_{min}(R)$, is such that $\lambda_{min}(R) > r^*$, and is not well-posed if the largest eigenvalue of R, $\lambda_{max}(R)$, is such that $\lambda_{max}(R) < r^*$.

It follows from Theorem 5.1 that, provided that system (1) is stabilizable, the well-posedness margin r^* can be obtained by solving the following nonlinear program

(with P and r being the decision variables):

$$\begin{array}{ll} \text{minimize} & r \\ \text{subject to} & \begin{cases} \left(2A+\sum_{j=1}^{k} C_j^2\right) P + Q + \Phi_+(P,r) \geq 0, \\ \left(2A+\sum_{j=1}^{k} C_j^2\right) P + Q + \Phi_-(P,r) \geq 0, \\ \left(rI + P\sum_{j=1}^{k} D_j' D_j\right)\Big|_{\Gamma} \geq 0, \end{cases} \end{array} \tag{42}$$

where

$$\Phi_+(P,r) := \inf_{K\in\Gamma}\left[K'\left(rI+P\sum_{j=1}^{k} D_j'D_j\right)K + 2\left(B+\sum_{j=1}^{k} C_jD_j\right)PK\right],$$

$$\Phi_-(P,r) := \inf_{K\in\Gamma}\left[K'\left(rI+P\sum_{j=1}^{k} D_j'D_j\right)K - 2\left(B+\sum_{j=1}^{k} C_jD_j\right)PK\right].$$

Notice, again, that it is hard to solve the preceding mathematical program as, in addition to the difficulty associated with the last constraint, $\Phi_+(P,r)$ and $\Phi_+(P,r)$ in general do not have analytical forms. In the following, we provide an explicit lower bound of r^*.

THEOREM 5.2. *Assume that system* (1) *is stabilizable. Then* $\bar{r} := \max\{\bar{r}_+, \bar{r}_-\}$ *is a lower bound of the well-posedness margin, where*

$$\bar{r}_+ := \begin{cases} \dfrac{\lambda Q}{\inf_{K\in\mathcal{K}_+}\{F_+(K)-\lambda K'K\}} & \text{if } Q \geq 0, \\ \dfrac{\lambda Q}{\sup_{K\in\mathcal{K}_+}\{F_+(K)-\lambda K'K\}} & \text{if } Q < 0 \end{cases} \tag{43}$$

and

$$\bar{r}_- := \begin{cases} \dfrac{\lambda Q}{\inf_{K\in\mathcal{K}_-}\{F_-(K)-\lambda K'K\}} & \text{if } Q \geq 0, \\ \dfrac{\lambda Q}{\sup_{K\in\mathcal{K}_-}\{F_-(K)-\lambda K'K\}} & \text{if } Q < 0, \end{cases} \tag{44}$$

with $\lambda := \inf_{K\in\Gamma,|K|=1} K'\sum_{j=1}^{k} D_j'D_jK \geq 0$.

Proof. Suppose problem (LQ) is well-posed with $R = -rI$, $r \in \mathbf{R}$. Then $\mathcal{P}_+ \neq \emptyset$. Since system (1) is stabilizable, we can take a $K \in \mathcal{K}_+$. It follows from Lemma 5.1 that $P := -\frac{Q+rK'K}{F_+(K)}$ is an upper bound of the nonempty set $\mathcal{P}_+$. Because $(rI + P_+\sum_{j=1}^{k} D_j'D_j)|_\Gamma \geq 0$ $\forall P_+ \in \mathcal{P}_+$, we conclude $(rI + P\sum_{j=1}^{k} D_j'D_j)|_\Gamma \geq 0$, which is equivalent to $r + PK'\sum_{j=1}^{k} D_j'D_jK \geq 0$ for any $K \in \Gamma, |K| = 1$. Hence $r + P\lambda \geq 0$. Substituting $P = -\frac{Q+rK'K}{F_+(K)}$ we obtain $r \geq \frac{\lambda Q}{F_+(K)-\lambda K'K}$. Since $K \in \mathcal{K}_+$ is arbitrary, we easily obtain that $r \geq \bar{r}_+$. Similarly, we have $r \geq \bar{r}_-$. Our analysis implies that problem (LQ) is *not* well-posed whenever the largest eigenvalue of R, $\lambda_{max}(R)$, is such that $\lambda_{max}(R) < \bar{r}$. Hence $\bar{r}$ is a lower bound of the well-posedness margin r^*. ∎

6. Optimality. This section is devoted to solving the optimal LQ control problem under consideration. We will first introduce two algebraic equations, in the spirit of the classical Riccati equation (for the unconstrained LQ problem), along with the notion of the so-called stabilizing solution. Then the optimality of problem (LQ) is addressed via the stabilizing solutions of the two algebraic equations.

We impose the following assumptions on the rest of the paper.

ASSUMPTION 6.1. *System* (1) *is conic stabilizable.*

ASSUMPTION 6.2. *Problem* (LQ) *is well-posed.*

6.1. Extended algebraic Riccati equations. In this subsection we define the two algebraic equations that play a key role in solving problem (LQ). Denote $\mathcal{R} := \{P \in \mathbf{R} | (R + P\sum_{j=1}^{k} D_j' D_j)|_\Gamma > 0\}$ and consider the following two functions from $\mathcal{R}$ to $\mathbf{R}$:

$$\xi_+(P) := \arg\min_{K\in\Gamma} \left[K' \left(R + P\sum_{j=1}^{k} D_j' D_j \right) K + 2 \left(B + \sum_{j=1}^{k} C_j D_j \right) PK \right],$$

$$\xi_-(P) := \arg\min_{K\in\Gamma} \left[K' \left(R + P\sum_{j=1}^{k} D_j' D_j \right) K - 2 \left(B + \sum_{j=1}^{k} C_j D_j \right) PK \right].$$

Note that the minimizers above are *uniquely* achievable due to a similar argument in Remark 5.3 and the fact that Γ is closed. Moreover, it is evident that both $\xi_+(\cdot)$ and $\xi_-(\cdot)$ are continuous on $\mathcal{R}$.

Define a pair of functions L_+ and L_- from $\mathcal{R}$ to $\mathbf{R}$:

$$L_+(P) := \left(2A + \sum_{j=1}^{k} C_j^2 \right) P + Q + \Phi_+(P),$$

$$L_-(P) := \left(2A + \sum_{j=1}^{k} C_j^2 \right) P + Q + \Phi_-(P).$$

The two equations

$$L_+(P) = 0, \quad \left(R + P\sum_{j=1}^{k} D_j' D_j \right)\Big|_\Gamma > 0, \tag{45}$$

$$L_-(P) = 0, \quad \left(R + P\sum_{j=1}^{k} D_j' D_j \right)\Big|_\Gamma > 0 \tag{46}$$

are called extended algebraic Riccati equations (EAREs). Note that the constraint $(R + P\sum_{j=1}^{k} D_j' D_j)|_\Gamma > 0$ is *part* of each of the two equations; so the EAREs are not exactly equations in a strict sense. Also, being an algebraic equation, each of them may admit more than one solution, or may admit no solution at all. Note that the EAREs introduced here both reduce to the *same* stochastic algebraic Riccati equation extensively studied in [2].

DEFINITION 6.1. *A solution P of the EARE* (45) *(respectively,* (46)*) is called a stabilizing solution if $\xi_+(P) \in \mathcal{K}_+$ (respectively, $\xi_-(P) \in \mathcal{K}_-$).*

It should be noted that the EAREs may not admit any stabilizing solution (see Proposition 6.1).

Before we conclude this subsection, we will present several lemmas.

LEMMA 6.1. *We have the inequalities*

$$(P_2 - P_1)[F_+(\xi_+(P_2)) - F_+(\xi_+(P_1))] \le 0, \tag{47}$$

$$(P_2 - P_1)[F_-(\xi_-(P_2)) - F_-(\xi_-(P_1))] \le 0, \tag{48}$$

$$L_+(P_2) - L_+(P_1) \le (P_2 - P_1)F_+(\xi_+(P_1)), \tag{49}$$

$$L_-(P_2) - L_-(P_1) \le (P_2 - P_1)F_-(\xi_-(P_1)) \tag{50}$$

for any $P_1, P_2 \in \mathcal{R}$.

Proof. Denoting $v_1 := \xi_+(P_1)$ and $v_2 := \xi_+(P_2)$, we have

$$\begin{aligned} v_2' \left(R + P_1 \sum_{j=1}^k D_j' D_j \right) v_2 + 2 \left(B + \sum_{j=1}^k C_j D_j \right) P_1 v_2 - \Phi_+(P_1) \ge 0, \\ v_1' \left(R + P_2 \sum_{j=1}^k D_j' D_j \right) v_1 + 2 \left(B + \sum_{j=1}^k C_j D_j \right) P_2 v_1 - \Phi_+(P_2) \ge 0. \end{aligned}$$

Then, adding the two inequalities, we get

$$\begin{aligned} &\left[v_2' \left(R + P_1 \sum_{j=1}^k D_j' D_j \right) v_2 + 2 \left(B + \sum_{j=1}^k C_j D_j \right) P_1 v_2 \right] - \Phi_+(P_2) \\ &\ge \Phi_+(P_1) - \left[v_1' \left(R + P_2 \sum_{j=1}^k D_j' D_j \right) v_1 + 2 \left(B + \sum_{j=1}^k C_j D_j \right) P_2 v_1 \right]. \end{aligned}$$

Recall that the infimum in $\Phi_+(P_i)$ is achieved by v_i, $i = 1, 2$; hence the above yields

$$\begin{aligned} &(P_2 - P_1) \left[v_2' \sum_{j=1}^k D_j' D_j v_2 + 2 \left(B + \sum_{j=1}^k C_j D_j \right) v_2 \right] \\ &\le (P_2 - P_1) \left[v_1' \sum_{j=1}^k D_j' D_j v_1 + 2 \left(B + \sum_{j=1}^k C_j D_j \right) v_1 \right]. \end{aligned}$$

This is equivalent to (47). Similarly we can prove (48).

Next, we calculate

$$\begin{aligned}
&L_+(P_2) - L_+(P_1)\\
&= \left(2A + \sum_{j=1}^{k} C_j^2\right)(P_2 - P_1) + \Phi_+(P_2) - \Phi_+(P_1)\\
&\le \left(2A + \sum_{j=1}^{k} C_j^2\right)(P_2 - P_1) + v_1' \left(R + P_2 \sum_{j=1}^{k} D_j' D_j\right) v_1\\
&\quad + 2\left(B + \sum_{j=1}^{k} C_j D_j\right) P_2 v_1 - \Phi_+(P_1)\\
&= (P_2 - P_1)\left[\left(2A + \sum_{j=1}^{k} C_j^2\right) + v_1' \sum_{j=1}^{k} D_j' D_j v_1 + 2\left(B + \sum_{j=1}^{k} C_j D_j\right) v_1\right]\\
&= (P_2 - P_1) F_+(\xi_+(P_1)).
\end{aligned}$$

This proves (49). Similarly we can show (50). □

LEMMA 6.2. *Assume that $P_1 \in \mathcal{R}$ and $P_2 \ge P_1$. If $\xi_+(P_1) \in \mathcal{K}_+$ (respectively, $\xi_-(P_1) \in \mathcal{K}_-$), then $\xi_+(P_2) \in \mathcal{K}_+$ (respectively, $\xi_-(P_2) \in \mathcal{K}_-$).*

Proof. Since $P_2 \ge P_1$ it follows from (47) of Lemma 6.1 that

$$F_+(\xi_+(P_2)) - F_+(\xi_+(P_1)) \le 0.$$

As $F_+(\xi_+(P_1)) < 0$ we have $F_+(\xi_+(P_2)) < 0$, implying $\xi_+(P_2) \in \mathcal{K}_+$. Similarly we can prove the assertion for $\mathcal{K}_-$. □

LEMMA 6.3. *If there exists $P_+^{(0)} \in \mathcal{R}$ (respectively, $P_-^0 \in \mathcal{R}$) with $\xi_+(P_+^{(0)}) \in \mathcal{K}_+$ (respectively, $\xi_-(P_-^0) \in \mathcal{K}_-$), then $L_+(\cdot)$ (respectively, $L_-(\cdot)$) is strictly decreasing on $[P_+^{(0)}, +\infty)$ (respectively, $[P_-^0, +\infty)$).*

Proof. Take $P_2 > P_1 \ge P_+^{(0)}$. It follows from Lemma 6.2 that $F_+(\xi_+(P_1)) < 0$. On the other hand, it is clear that $P_1, P_2 \in \mathcal{R}$. Hence Lemma 6.1 yields

$$L_+(P_2) - L_+(P_1) \le (P_2 - P_1) F_+(\xi_+(P_1)) < 0.$$

This proves that $L_+(P_2) < L_+(P_1)$. We can prove the assertion for $L_-(P)$ in a similar manner. □

6.2. Optimality of the LQ problem via EAREs. In this subsection we prove that stabilizing solutions of (45) and (46), if any, lead to a complete and explicit solution to problem (LQ).

THEOREM 6.2. *If the EAREs* (45) *and* (46) *admit stabilizing solutions P_+^* and P_-^* respectively, then the feedback control*

$$u^*(t) = \xi_+(P_+^*) x^+(t) + \xi_-(P_-^*) x^-(t) \tag{51}$$

is optimal for problem (LQ) *with respect to any initial state x_0. Moreover, the value function is*

$$V(x_0) = P_+^*(x_0^+)^2 + P_-^*(x_0^-)^2 \quad \forall x_0 \in \mathbf{R}. \tag{52}$$

Proof. Since P_+^* solves (45), we have

$$P_+^* = -\frac{Q + \xi_+(P_+^*)' R \xi_+(P_+^*)}{F_+(\xi_+(P_+^*))}.$$

Similarly,

$$P_-^* = -\frac{Q + \xi_-(P_-^*)' R \xi_-(P_-^*)}{F_-(\xi_-(P_-^*))}.$$

Moreover, $\xi_+(P_+^*) \in \mathcal{K}_+, \xi_-(P_-^*) \in \mathcal{K}_-$ as both P_+^* and P_-^* are stabilizing solutions. Thus the proof of Lemma 5.1 yields $V(x_0) \le J(x_0; u^*(\cdot)) = P_+^*(x_0^+)^2 + P_-^*(x_0^-)^2$. On the other hand, $P_+^* \in \mathcal{P}_+$ and $P_-^* \in \mathcal{P}_-$. Hence it follows from Theorem 5.1 that $V(x_0) \ge P_+^*(x_0^+)^2 + P_-^*(x_0^-)^2$. Therefore, $V(x_0) = J(x_0; u^*(\cdot)) = P_+^*(x_0^+)^2 + P_-^*(x_0^-)^2$. □

The above proof has also shown the following result.

COROLLARY 6.1. *If the EAREs* (45) *and* (46) *admit stabilizing solutions P_+^* and P_-^*, respectively, then $P_+^* = \max\{P | P \in \mathcal{P}_+\}$ and $P_-^* = \max\{P | P \in \mathcal{P}_-\}$. As a result,* (45) *and* (46) *each has at most one stabilizing solution.*

Corollary 6.1 guarantees that any stabilizing solution is the maximal solution of the respective EAREs. This result is in parallel with the unconstrained case (see, e.g., [3, Theorem 2.3]).

Note that the converse of Theorem 6.2 is not necessarily true. The following example shows that the existence of a solution to the EAREs is not *necessary* for the LQ problem to be attainable with respect to any initial state.

Example 6.1. Consider the LQ problem

$$(53)\qquad \begin{aligned} &\text{minimize} && J(x_0; u(\cdot)) = \mathbf{E}\int_0^{+\infty} [|x(t)|^2 - |u(t)|^2]dt \\ &\text{subject to} && \begin{cases} dx(t) = [-x(t) + u(t)]dt + [-x(t) + u(t)]dw(t), \\ x(0) = x_0, \end{cases} \end{aligned}$$

where all the variables are scalar-valued and $\Gamma = \mathbf{R}$. This example was originally discussed in [33, Example 6.1, p. 817]. It was verified in [33] that the system is stabilizable, and the LQ problem is attainable with respect to any x_0 (in fact there are infinitely many optimal feedback controls). But both EAREs (45) and (46) in this case reduce to $-p + 1 = 0,\ -1 + p > 0$, which clearly admits no solution at all.

In spite of the preceding remarks and example, the following result shows that under an additional assumption, the EAREs indeed admit stabilizing solutions if problem (LQ) is attainable.

THEOREM 6.3. *Assume that there exist $P_+ \in \mathcal{P}_+$ and $P_- \in \mathcal{P}_-$ such that $(R + P\sum_{j=1}^k D_j' D_j)|_\Gamma > 0$ for $P = P_+, P_-$. If problem* (LQ) *is attainable with respect to any $x_0 \in \mathbf{R}$, then the EAREs* (45) *and* (46) *admit stabilizing solutions P_+^* and P_-^*, respectively. Moreover, any optimal control with respect to a given x_0 must be unique and represented by the feedback control* (51).

Proof. The proof of Proposition 5.2 yields that, under Assumptions 6.1 and 6.2, the value function can be represented as (32), where P_+^* and P_-^* are the maximum elements of $\mathcal{P}_+$ and $\mathcal{P}_-$, respectively. Moreover, by the assumption we have

$$(54)\qquad \left(R + P_+^* \sum_{j=1}^k D_j' D_j\right)\Bigg|_\Gamma > 0, \quad \left(R + P_-^* \sum_{j=1}^k D_j' D_j\right)\Bigg|_\Gamma > 0.$$

Now, for any $x_0 \in \mathbf{R}$ let $x^*(\cdot)$ be the solution of (1) under an optimal control $u^*(\cdot) \in \mathcal{U}_{x_0}$ which exists by the attainability assumption. Then a similar calculation to that of (28) leads to

$$\begin{aligned}&\mathbf{E}\int_0^t [Qx^*(s)^2 + u^*(s)'Ru^*(s)]ds \\ &= P_+^*(x_0^+)^2 + P_-^*(x_0^-)^2 - \mathbf{E}[P_+^* x^{*+}(t)^2] - \mathbf{E}[P_-^* x^{*-}(t)^2] + \mathbf{E}\int_0^t \psi(x^*(s), u^*(s))ds,\end{aligned} \tag{55}$$

where $\psi(x^*(s), u^*(s))$ is the same as the integrand on the right-hand side of (28), with P_+ and P_- replaced by P_+^* and P_-^*, respectively. Letting $t \to +\infty$ and noting that $u^*(\cdot)$ is stabilizing, we obtain

$$\begin{aligned}V(x_0) \equiv J(x_0; u^*(\cdot)) &= \lim_{t\to+\infty} \mathbf{E}\int_0^t [Qx^*(s)^2 + u^*(s)'Ru^*(s)]ds \\ &= P_+^*(x_0^+)^2 + P_-^*(x_0^-)^2 + \mathbf{E}\int_0^{+\infty} \psi(x^*(s), u^*(s))ds.\end{aligned}$$

Recalling that $V(x_0) = P_+^*(x_0^+)^2 + P_-^*(x_0^-)^2$ and that $\psi(x^*(s), u^*(s)) \geq 0$ (via the same proof of Theorem 5.1), we conclude

$$\psi(x^*(s), u^*(s)) = 0, \ \text{ a.e. } s \in [0, +\infty), \ \mathbf{P}\text{-a.s.}$$

Fix $s \in [0, +\infty)$, satisfying the above equality. If $x^*(s) > 0$, then we can write $u^*(s) = K(s)x^*(s)$, where $K(s) \in \Gamma$. Going through the same analysis as in (29), we obtain

$$\begin{aligned}0 &= \psi(x^*(s), u^*(s)) \\ &= \left[\left(2A + \sum_{j=1}^k C_j^2\right) P_+^* + Q + K(s)'\left(R + P_+^* \sum_{j=1}^k D_j' D_j\right) K(s)\right. \\ &\quad \left. + 2\left(B + \sum_{j=1}^k C_j D_j\right) K(s) P_+^*\right] x^*(s)^2 \\ &\geq \left[\left(2A + \sum_{j=1}^k C_j^2\right) P_+^* + Q + \Phi_+(P_+^*)\right] x^*(s)^2 \\ &\geq 0.\end{aligned}$$

Thus, all the inequalities above become equalities and, noting that $x^*(s) \neq 0$, one has $K(s) = \xi_+(P_+^*)$ and $L_+(P_+^*) = 0$. As a result, $u^*(s) \equiv K(s)x^*(s) = \xi_+(P_+^*)x^*(s)$ at s when $x^*(s) > 0$. Similarly, we can prove that $u^*(s) = -\xi_-(P_-^*)x^*(s)$ at s when $x^*(s) \leq 0$, and $L_-(P_-^*) = 0$. To summarize, we have shown that any optimal control $u^*(\cdot)$ can be represented by (51), and hence the uniqueness of optimal control follows. On the other hand, we have also proved that P_+^* and P_-^* are solutions to the EAREs (45) and (46), respectively. Moreover, they must be stabilizing solutions because $u^*(\cdot)$, which is now represented by (51), is a stabilizing control. □

Remark 6.1. Theorem 6.3 shows that the existence of stabilizing solutions to the EAREs (45) and (46) is *almost* necessary for the attainability of problem (LQ). The only exception, as also demonstrated by Example 6.1, is the "singular" case when $(R + P\sum_{j=1}^k D_j' D_j)|_\Gamma = 0$ for *all* elements P in at least one of the sets $\mathcal{P}_+$ and $\mathcal{P}_-$.

6.3. Existence of stabilizing solutions to EAREs. Theorem 6.2 asserts that if one can find stabilizing solutions to the EAREs, then the original optimal LQ control can be solved completely and explicitly in terms of obtaining the optimal feedback control as well as the value function. The next natural questions are, then, *when* do the EAREs admit stabilizing solutions, and *how* do we find them? These are the issues that we are going to address in this subsection. Indeed, we will identify and discuss three cases when the EAREs do have the stabilizing solutions.

THEOREM 6.4. *If there exist* $P_+^{(0)}$, $P_-^{(0)}$ *satisfying*

$$L_+(P_+^{(0)}) \geq 0, \quad \left(R + P_+^{(0)} \sum_{j=1}^{k} D_j' D_j\right)\Big|_\Gamma > 0, \tag{56}$$

$$L_-(P_-^{(0)}) \geq 0, \quad \left(R + P_-^{(0)} \sum_{j=1}^{k} D_j' D_j\right)\Big|_\Gamma > 0, \tag{57}$$

and

$$F_+(\xi_+(P_+^{(0)})) < 0, \quad F_-(\xi_-(P_-^{(0)})) < 0, \tag{58}$$

then the EAREs (45) *and* (46) *admit unique stabilizing solutions* P_+^* *and* P_-^**, respectively.*

Proof. If $L_+(P_+^{(0)}) = 0$, then $P_+^{(0)}$ is the stabilizing solution to (45) and we are done. So let us assume that $L_+(P_+^{(0)}) \equiv (2A + \sum_{j=1}^k C_j^2)P_+^{(0)} + Q + \Phi_+(P_+^{(0)}) > 0$, namely,

$$F_+(\xi_+(P_+^{(0)}))P_+^{(0)} + Q + \xi_+(P_+^{(0)})' R \xi_+(P_+^{(0)}) > 0.$$

Since $F_+(\xi_+(P_+^{(0)})) < 0$, we have

$$P_+^{(0)} < -\frac{Q + \xi_+(P_+^{(0)})' R \xi_+(P_+^{(0)})}{F_+(\xi_+(P_+^{(0)}))} := P_+^{(1)}. \tag{59}$$

By Lemma 6.2, we have $\xi_+(P_+^{(1)}) \in \mathcal{K}_+$ because $P_+^{(1)} > P_+^{(0)}$. Moreover,

$$\begin{aligned} L_+(P_+^{(1)}) &\equiv \left(2A + \sum_{j=1}^{k} C_j^2\right) P_+^{(1)} + Q + \Phi_+(P_+^{(1)}) \\ &\leq \left(2A + \sum_{j=1}^{k} C_j^2\right) P_+^{(1)} + Q + \xi_+(P_+^{(0)})' \left(R + P_+^{(1)} \sum_{j=1}^{k} D_j' D_j\right) \xi_+(P_+^{(0)}) \\ &\quad + 2\left(B + \sum_{j=1}^{k} C_j D_j\right) P_+^{(1)} \xi_+(P_+^{(0)}) \\ &= 0, \end{aligned} \tag{60}$$

where the last inequality was due to the very definition of $P_+^{(1)}$. Now, if $L_+(P_+^{(1)}) = 0$, then $P_+^{(1)}$ is the stabilizing solution of (45). If $L_+(P_+^{(1)}) < 0$, then, noting that $L_+(\cdot)$ is a strictly decreasing (by Lemma 6.3) continuous function on the interval $[P_+^{(0)}, P_+^{(1)}]$ along with $L_+(P_+^{(0)}) > 0$, we conclude that there exists a unique $P_+^* \in (P_+^{(0)}, P_+^{(1)})$ such that $L_+(P_+^*) = 0$. Clearly $(R + P_+^* \sum_{j=1}^k D_j' D_j)|_\Gamma > 0$, and $\xi_+(P_+^*) \in \mathcal{K}_+$ thanks to Lemma 6.2. Hence P_+^* is the stabilizing solution of (45).

The proof for the existence of the stabilizing solution to the EARE (46) is completely analogous. Finally, the uniqueness of the stabilizing solutions follows from Corollary 6.1. □

THEOREM 6.5. *If there exist* $P_+^{(0)}$, $P_-^{(0)}$ *satisfying*

$$L_+(P_+^{(0)}) > 0, \quad \left(R + P_+^{(0)} \sum_{j=1}^k D_j' D_j \right) \Big|_\Gamma > 0, \tag{61}$$

and

$$L_-(P_-^{(0)}) > 0, \quad \left(R + P_-^{(0)} \sum_{j=1}^k D_j' D_j \right) \Big|_\Gamma > 0, \tag{62}$$

then the EAREs (45) *and* (46) *admit unique stabilizing solutions* P_+^* *and* P_-^**, respectively.*

Proof. Take $K_+ \in \mathcal{K}_+$, which exists by the stabilizability assumption. Set

$$P_+^{(1)} := -\frac{Q + K_+' R K_+}{F_+(K_+)}. \tag{63}$$

It follows from Lemma 5.1 that $P_+^{(1)} \geq P_+^{(0)}$ since $P_+^{(0)} \in \mathcal{P}_+$. Hence

$$\left(R + P_+^{(1)} \sum_{j=1}^k D_j' D_j \right) \Big|_\Gamma > 0.$$

Moreover,

$$\begin{aligned} L_+(P_+^{(1)}) &\equiv \left(2A + \sum_{j=1}^k C_j^2 \right) P_+^{(1)} + Q + \Phi_+(P_+^{(1)}) \\ &\leq \left(2A + \sum_{j=1}^k C_j^2 \right) P_+^{(1)} + Q + K_+' \left(R + P_+^{(1)} \sum_{j=1}^k D_j' D_j \right) K_+ \\ &\quad + 2 \left(B + \sum_{j=1}^k C_j D_j \right) P_+^{(1)} K_+ \\ &= 0. \end{aligned} \tag{64}$$

Applying Lemma 6.1 and noting that $L_+(P_+^{(0)}) > 0$ we get

$$0 < L_+(P_+^{(0)}) - L_+(P_+^{(1)}) \leq (P_+^{(0)} - P_+^{(1)}) F_+(\xi_+(P_+^{(1)})). \tag{65}$$

Hence $F_+(\xi_+(P_+^{(1)})) < 0$ or $\xi_+(P_+^{(1)}) \in \mathcal{K}_+$. Now set

$$P_+^{(2)} := -\frac{Q + \xi_+(P_+^{(1)})' R \xi_+(P_+^{(1)})}{F_+(\xi_+(P_+^{(1)}))}. \tag{66}$$

Again by Lemma 5.1 we obtain $P_+^{(2)} \geq P_+^{(0)}$ and, therefore, $(R+P_+^{(2)} \sum_{j=1}^k D_j' D_j)|_\Gamma > 0$. On the other hand, the fact that $L_+(P_+^{(1)}) \leq 0$ can be rewritten as $P_+^{(2)} \leq P_+^{(1)}$ in view of the relation (66). Moreover, analysis similar to that for $P_+^{(1)}$ above leads to $L_+(P_+^{(2)}) \leq 0$, and $\xi_+(P_+^{(2)}) \in \mathcal{K}_+$ or $F_+(\xi_+(P_+^{(2)})) < 0$.

In general, we construct iteratively the following sequence:

$$P_+^{(i+1)} := -\frac{Q + \xi_+(P_+^{(i)})' R \xi_+(P_+^{(i)})}{F_+(\xi_+(P_+^{(i)}))}, \quad i = 1, 2, \ldots. \tag{67}$$

An induction argument shows that $P_+^{(0)} \leq \cdots \leq P_+^{(i+1)} \leq P_+^{(i)} \cdots \leq P_+^{(1)}$, $(R + P_+^{(i)} \sum_{j=1}^k D_j' D_j)|_\Gamma > 0$, $L_+(P_+^{(i)}) \leq 0$, and $\xi_+(P_+^{(i)}) \in \mathcal{K}_+$, $i = 1, 2, \ldots$. Since the sequence $\{P_+^{(i)}\}$ is decreasing with a lower bound $P_+^{(0)}$, there exists $P_+^* \in \mathbf{R}$ so that $P_+^* = \lim_{i\to\infty} P_+^{(i)}$. Moreover it is clear that $(R + P_+^* \sum_{j=1}^k D_j' D_j)|_\Gamma \geq (R + P_+^{(0)} \sum_{j=1}^k D_j' D_j)|_\Gamma > 0$. On the other hand, $L_+(P_+^*) \leq 0$ since each $L_+(P_+^{(i)}) \leq 0$. Thus an argument similar to (65) yields $\xi_+(P_+^*) \in \mathcal{K}_+$ or $F_+(\xi_+(P_+^*)) < 0$. As a result, we can pass the limit in (67) to obtain $P_+^* = -\frac{Q+\xi_+(P_+^*)' R \xi_+(P_+^*)}{F_+(\xi_+(P_+^*))}$, which is equivalent to $L_+(P_+^*) = 0$. This shows that P_+^* is the stabilizing solution to (45). Similarly we can prove that (46) admits the stabilizing solution. □

Remark 6.2. Recall that Theorem 5.1 characterizes the well-posedness of problem (LQ) by the nonemptiness of the sets $\mathcal{P}_+$ and $\mathcal{P}_-$. Theorems 6.4 and 6.5 spell out two important cases when the EAREs (45) and (46) have stabilizing solutions and, therefore, problem (LQ) can be completely solved with explicit solutions. These two cases are specified in terms of the existence of certain "special elements" of the sets $\mathcal{P}_+$ and $\mathcal{P}_-$. Specifically, the case with Theorem 6.4 is one when each of $\mathcal{P}_+$ and $\mathcal{P}_-$ has a "stabilizing element" in the sense that (58) holds. On the other hand, Theorem 6.5 asserts that the nonemptiness of the *interiors* of $\mathcal{P}_+$ and $\mathcal{P}_-$ is sufficient for the existence of stabilizing solutions to the EAREs. In view of the fact that the nonemptiness of $\mathcal{P}_+$ and $\mathcal{P}_-$ is the minimum requirement for the underlying LQ problem to be meaningful, the sufficient conditions respectively given in Theorems 6.4 and 6.5 are very mild indeed.

Remark 6.3. The proof of Theorem 6.5 constitutes an algorithm for finding the stabilizing solutions to the EAREs. In fact it is given by the iterative scheme (67) with an initial point (63). On the other hand, although the proof of Theorem 6.4 has not given an explicit algorithm for computing the stabilizing solutions, one can use a middle-point algorithm to find them based on the proof. Alternatively, one may use the same iterative scheme (67) with the initial point, $P_+^{(1)}$, given by (59). It can be proved, using almost the same analysis as that in the proof of Theorem 6.5, that the constructed sequence converges to the desired point, P_+^*. The only argument that needs to be modified is that for proving $\xi_+(P_+^*) \in \mathcal{K}_+$. In this case, $\xi_+(P_+^*) \in \mathcal{K}_+$ is seen from the fact that $P_+^* \geq P_+^{(0)}$ and $\xi_+(P_+^{(0)}) \in \mathcal{K}_+$ as well as from Lemma 6.2.

Finally we present the results on the definite case $Q \geq 0$ and $R|_\Gamma \geq 0$ (including the so-called singular case when R is allowed to be *singular*).

THEOREM 6.6. *Assume $Q \geq 0$ and $R|_\Gamma \geq 0$. Then the EAREs* (45) *and* (46) *admit unique stabilizing solutions P_+^* and P_-^*, respectively, under one of the following additional conditions:*

(i) *$Q > 0$ and $R|_\Gamma > 0$.*

(ii) *$Q = 0$, $R|_\Gamma > 0$, and $2A + \sum_{j=1}^k C_j^2 \neq 0$.*

(iii) *$Q > 0$, $R|_\Gamma \geq 0$, and $\sum_{j=1}^k D_j' D_j|_\Gamma > 0$.*

Proof. (i) In this case $P_+^{(0)} = P_-^{(0)} = 0$ satisfy the assumption of Theorem 6.5.

(ii) If $2A + \sum_{j=1}^k C_j^2 < 0$, then take $P_+^{(0)} = P_-^{(0)} = 0$. We see that $L_+(P_+^{(0)}) = L_-(P_-^{(0)}) = Q = 0$ and $F_+(\xi_+(P_+^{(0)})) = F_-(\xi_-(P_-^{(0)})) = 2A + \sum_{j=1}^k C_j^2 < 0$. Thus the assumption of Theorem 6.4 is satisfied.

If $2A + \sum_{j=1}^k C_j^2 > 0$, due to the stabilizability assumption there is $0 \neq K_+ \in \mathcal{K}_+$. Set $P_+^{(1)} := -\frac{K_+' R K_+}{F_+(K_+)}$. As $F_+(K_+) < 0$, $R|_\Gamma > 0$, and $K_+ \neq 0$, we have $P_+^{(1)} > 0$. Define

$$P_+^{(i+1)} := -\frac{\xi_+(P_+^{(i)})' R \xi_+(P_+^{(i)})}{F_+(\xi_+(P_+^{(i)}))}, \quad i = 1, 2, \ldots. \tag{68}$$

Then an analysis similar to that in the proof of Theorem 6.5 leads to $0 \leq \cdots \leq P_+^{(i+1)} \leq P_+^{(i)} \cdots \leq P_+^{(1)}$, $(R + P_+^{(i)} \sum_{j=1}^k D_j' D_j)|_\Gamma > 0$, $L_+(P_+^{(i)}) \leq 0$, and $\xi_+(P_+^{(i)}) \in \mathcal{K}_+$, $i = 1, 2, \ldots$. Hence there exists $P_+^* \geq 0$ so that $P_+^* = \lim_{i\to\infty} P_+^{(i)}$. Moreover $(R + P_+^* \sum_{j=1}^k D_j' D_j)|_\Gamma \geq R|_\Gamma > 0$. Note that at this point we can no longer apply the same argument as that used in the proof of Theorem 6.5 to conclude $F_+(\xi_+(P_+^*)) < 0$, because the element 0, which substitutes the point $P_+^{(0)}$ of Theorem 6.5, does not satisfy $L_+(0) > 0$. To get around this, let us suppose $F_+(\xi_+(P_+^*)) = 0$ (recall that it always holds that $F_+(\xi_+(P_+^*)) \leq 0$ since $F_+(\xi_+(P_+^{(i)})) < 0$). Multiplying (68) by $F_+(\xi_+(P_+^{(i)}))$ and then passing to the limit, we obtain $\xi_+(P_+^*)' R \xi_+(P_+^*) = 0$, resulting in $\xi_+(P_+^*) = 0$. Thus $F_+(\xi_+(P_+^*)) = 2A + \sum_{j=1}^k C_j^2 > 0$, which is a contradiction. This proves that $F_+(\xi_+(P_+^*)) < 0$. The rest of the proof is the same as that of Theorem 6.5.

(iii) First note that for any $P > 0$,

$$\begin{aligned} 0 &\geq \Phi_+(P) \\ &\geq \inf_{K\in\Gamma} K' R K + P \inf_{K\in\Gamma} \left[K' \sum_{j=1}^k D_j' D_j K + 2 \left(B + \sum_{j=1}^k C_j D_j \right) K \right] \\ &= P \inf_{K\in\Gamma} \left[K' \sum_{j=1}^k D_j' D_j K + 2 \left(B + \sum_{j=1}^k C_j D_j \right) K \right]. \end{aligned}$$

Hence $\lim_{P\to 0+} \Phi_+(P) = 0$. Since $Q > 0$, we have, by the definition of $L_+(\cdot)$, that there exists $P_+^{(0)} > 0$ so that $L_+(P_+^{(0)}) > 0$. On the other hand, it is clear that $(R + P_+^{(0)} \sum_{j=1}^k D_j' D_j)|_\Gamma \geq P_+^{(0)} \sum_{j=1}^k D_j' D_j \Big|_\Gamma > 0$. Consequently Theorem 6.5 applies. □

One may be curious about what happens if $Q = 0$, $R|_\Gamma > 0$, and $2A + \sum_{j=1}^k C_j^2 = 0$ (refer to Theorem 6.6(ii)). It turns out that in this case the EAREs *never* admit stabilizing solutions.

PROPOSITION 6.1. *Neither* (45) *nor* (46) *admits any stabilizing solution when* $Q = 0$, $R|_\Gamma > 0$, *and* $2A + \sum_{j=1}^k C_j^2 = 0$.

Proof. By Assumption 6.1, there exists $0 \neq K_+ \in \mathcal{K}_+$ and $0 \neq K_- \in \mathcal{K}_-$ since $F_+(0) = F_-(0) = 0$. Denote $K_+^\varepsilon := \varepsilon K_+$ and $K_-^\varepsilon := \varepsilon K_-$ for $\varepsilon \in (0, 1]$. Then $K_+^\varepsilon \in \Gamma$ and $K_-^\varepsilon \in \Gamma$.

For any fixed $P_+ > 0$, set $\varepsilon := \min\left\{\frac{-P_+F_+(K_+)}{2K_+'RK_+}, 1\right\} \in (0, 1]$. Then

$$\begin{aligned}
L_+(P_+) &= \Phi_+(P_+) \\
&\leq (K_+^\varepsilon)' \left(R + P_+ \sum_{j=1}^k D_j' D_j\right) K_+^\varepsilon + 2P_+ \left(B + \sum_{j=1}^k C_j D_j\right) K_+^\varepsilon \\
&\leq \varepsilon \left\{\varepsilon K_+' R K_+ + P_+ \left[K_+' \sum_{j=1}^k D_j' D_j K_+ + 2\left(B + \sum_{j=1}^k C_j D_j\right) K_+\right]\right\} \\
&\leq \varepsilon \left[\frac{-P_+F_+(K_+)}{2K_+'RK_+}(K_+'RK_+) + P_+F_+(K_+)\right] \\
&= \frac{\varepsilon}{2} P_+ F_+(K_+) \\
&< 0.
\end{aligned}$$

This implies that there exists no *positive* solution to the EARE (45).

Next, for any fixed $P_+ < 0$ with $(R + P_+ \sum_{j=1}^k D_j' D_j)|_\Gamma > 0$, set

$$\varepsilon := \min\left\{\frac{-|P_+|F_-(K_-)}{2K_-'RK_-}, 1\right\} \in (0, 1].$$

Then

$$\begin{aligned}
L_+(P_+) &= \Phi_+(P_+) \\
&\leq (K_-^\varepsilon)' \left(R + P_+ \sum_{j=1}^k D_j' D_j\right) K_-^\varepsilon + 2P_+ \left(B + \sum_{j=1}^k C_j D_j\right) K_-^\varepsilon \\
&\leq \varepsilon \left\{\varepsilon K_-' R K_- + |P_+| \left[K_-' \sum_{j=1}^k D_j' D_j K_- - 2\left(B + \sum_{j=1}^k C_j D_j\right) K_-\right]\right\} \\
&\leq \varepsilon \left[\frac{-|P_+|F_-(K_-)}{2K_-'RK_-}(K_-'RK_-) + |P_+|F_-(K_-)\right] \\
&= \frac{\varepsilon}{2}|P_+|F_-(K_-) \\
&< 0.
\end{aligned}$$

Hence there is no *negative* solution to the EARE (45).

Finally, when $P_+ = 0$, we do have $L_+(P_+) = 0$ but $F_+(\xi_+(P_+)) = F_+(0) = 0$. So $P_+ = 0$ is not a stabilizing solution either.

Similarly, we can prove the nonexistence of a stabilizing solution to (46). □

Although the conclusion of Proposition 6.1 does not necessarily lead to the nonexistence of optimal feedback control for the corresponding LQ problem (refer to section 6.2), the following example shows that the latter could indeed occur.

Example 6.2. Consider the LQ problem

$$\begin{aligned} &\text{minimize} && J(x_0;u(\cdot)) = \mathbf{E}\int_0^{+\infty} r|u(t)|^2 dt \\ &\text{subject to} && \begin{cases} dx(t) = [ax(t)+bu(t)]dt + cx(t)dw(t), \\ x(0) = x_0, \end{cases} \end{aligned} \tag{69}$$

where all the variables are scalar-valued, $\Gamma = \mathbf{R}$, $2a + c^2 = 0$, $b > 0$, and $r > 0$. It is easy to verify that the problem is stabilizable and well-posed. Take a feedback control $u^\varepsilon(t) = -\varepsilon x^\varepsilon(t)$ for $\varepsilon > 0$. Under this control the state satisfies

$$\mathbf{E}|x^\varepsilon(t)|^2 = e^{(2a+c^2-2b\varepsilon)t}x_0^2 = e^{-2b\varepsilon t}x_0^2.$$

Hence u^ε is stabilizing. Moreover, the cost under this control is

$$J(x_0;u^\varepsilon(\cdot)) = \varepsilon^2 r^2 \mathbf{E}\int_0^{+\infty} |x^\varepsilon(t)|^2 dt = \frac{r^2 x_0^2}{2b}\varepsilon.$$

Letting $\varepsilon \to 0$ we see that $V(x_0) = 0$ $\forall x_0 \in \mathbf{R}$. Note that this value cannot be attained if $x_0 \neq 0$, for whenever $\mathbf{E}\int_0^{+\infty} r|u^*(t)|^2 dt = 0$ it is necessary that $u^*(t) = 0$, a.e. $t \geq 0$. However, this control, $u^*(t)$, is *not* stabilizable when $x_0 \neq 0$. In other words, the LQ problem is not attainable with respect to $x_0 \neq 0$.

Remark 6.4. Theorem 6.6(i),(ii) and Proposition 6.1 together with Example 6.2 give a complete answer to the question of optimality for problem (LQ) in the classical definite case $Q \geq 0$ and $R|_\Gamma > 0$. Moreover, Theorem 6.6(iii) addresses the case when R is possibly singular. Note that this case occurs often in financial applications (where typically $R = 0$).

Remark 6.5. In view of Theorem 6.2, under the respective assumptions of Theorems 6.4, 6.5, and 6.6, problem (LQ) has the optimal feedback control (51) and the value function (52). Moreover, as per Remark 5.6, in these cases the stabilizing solutions P_+^* and P_-^* can also be obtained, in addition to the preceding algorithms, by solving the mathematical programs (33) and (34) if the corresponding constraints are tractable.

7. Numerical examples. To numerically calculate the optimal solution to problem (LQ) one needs to carry out two steps: the first is to check the conic stabilizability and the well-posedness, and the second is to find the stabilizing solutions to the EAREs. The procedures for the first step are depicted in sections 4 and 5.3, whereas that for the second part is described in section 6. Here we give an example to illustrate the whole process (where we used the computing tool *Scilab* to carry out all the calculations).

Example 7.1. Consider problem (LQ) with $m = k = 3$, $\Gamma = \mathbf{R}^3_+$, and the dynamics coefficients as follows:

$$\begin{aligned} &A = 2.00, && B = (-50 \quad -100 \quad 200), \\ &C_1 = -0.84, && D_1 = (6.85 \quad -8.78 \quad 0.68), \\ &C_2 = -3.78, && D_2 = (11.22 \quad 13.24 \quad 14.53), \\ &C_3 = 0.849, && D_3 = (-1.98 \quad -5.44 \quad -2.32). \end{aligned}$$

The eigenvalues of $\sum_{j=1}^3 D_j' D_j$ are 1.5880509, 126.23912, and 547.84943. So $\sum_{j=1}^3 D_j' D_j$ is positive definite in this case. The cost parameters are

$$Q = 10, \quad R = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -5.0 & 0 \\ 0 & 0 & 4.0 \end{pmatrix}.$$

Hence this is an indefinite LQ problem.

To solve this problem, we first apply Theorem 4.1(iii) (note that in this example the constraint $K \in \mathbf{R}^3_+$ is also an LMI) to obtain stabilizing feedback gains

$$K_+ = \begin{pmatrix} 0.5435353 \\ 0.5307289 \\ 0.0701460 \end{pmatrix}, \quad K_- = \begin{pmatrix} 0.0816967 \\ 0.0562570 \\ 0.7185273 \end{pmatrix}.$$

Next we use the algorithm in section 5.3 to find out that problem (LQ) is well-posed. Furthermore, when $P_+^{(0)} = P_-^{(0)} = 0.1$, the eigenvalues of $R + 0.1 * \sum_{j=1}^{3} D_j' D_j$ are 1.641309, 10.293806, and 54.632545; hence $R + 0.1 * \sum_{j=1}^{3} D_j' D_j > 0$. On the other hand, $L_+(0.1) = 1.6073676 > 0$ and $L_-(0.1) = 4.0651986 > 0$. According to Theorem 6.5, problem (LQ) has an optimal feedback control. Now we use the algorithm given in the proof of Theorem 6.5 to obtain the optimal control and optimal value. First, set the initial values $P_+^{(1)}$ and $P_-^{(1)}$ by using K_+, K_- and formulas similar to (63), respectively. They are

$$P_+^{(1)} = 1.1814412, \quad P_-^{(1)} = 13.167238.$$

By the iterative formula (67), we obtain

$$P_+^* = 0.1225762, \quad \xi_+(P_+^*) = \begin{pmatrix} 0.2889240 \\ 0.4918755 \\ 0 \end{pmatrix},$$

$$P_-^* = 0.1561608, \quad \xi_-(P_-^*) = \begin{pmatrix} 0 \\ 0 \\ 0.5875815 \end{pmatrix}.$$

Therefore, the optimal feedback control is

$$u^*(t) = \begin{pmatrix} 0.2889240 \\ 0.4918755 \\ 0 \end{pmatrix} x^+(t) + \begin{pmatrix} 0 \\ 0 \\ 0.5875815 \end{pmatrix} x^-(t),$$

with the optimal cost

$$J^*(x_0) = 0.1225762 * (x_0^+)^2 + 0.1561608 * (x_0^-)^2.$$

In the next example we demonstrate the calculation of a lower bound of the well-posedness margin (refer to section 5.4).

Example 7.2. Using the same values of the coefficients A, B, C_j, D_j, $j = 1, 2, 3$, and Q as in Example 7.1, we want to compute a lower bound of the well-posedness margin r^*.

According to Theorem 5.2, we first calculate $\lambda = 176.7313$. Next we have $\bar{r}_+ = -9.434553$ and $\bar{r}_- = -6.4967141$. Hence, the lower bound of the well-posedness margin is $\bar{r} = -6.4967141$. Note that, as seen in Example 7.1, the problem is still well-posed when one of the eigenvalues of R is -5.

8. Conclusion. In this paper, we studied an indefinite stochastic LQ control problem in the infinite time horizon with conic control constraint. Several key issues, including conic stabilizability, well-posedness, and optimality were addressed

with complete solutions. In particular, two algebraic equations, the EAREs, were newly introduced, in lieu of the classical algebraic Riccati equation, whose stabilizing solutions give rise to the explicit forms of the optimal feedback control and the value function. It was also seen that the representation of the value function given by Proposition 5.1 served as the technical key to all the main results of this paper, which motivated the utilization of the cerebrated Tanaka formula.

It should be stressed again that the approach of this paper crucially depended on the special structure of the problem. One main assumption is that the state of the system is one-dimensional. While the conclusion of Proposition 5.1 appears to hold, mutatis mutandis, for the problem with multidimensional state variable, it seems that an analogy of Lemma 3.2, if any, would be far more complicated. This makes the multidimensional problem very challenging. Another structural property of the model is that the dynamics of the system is homogeneous (in state and control) and the cost contains no first-order term of the state variable as well as no control-state cross term. As a result, our approach will fail, say, for the case when there is no state and control dependent noise, and with fixed variance. Solving these kinds of problems calls for different techniques. Finally, an even more difficult problem is the stochastic LQ control with state constraint.

Acknowledgment. We thank the three anonymous reviewers for their careful reading of an earlier version of the paper and for their constructive comments that led to an improved version.

REFERENCES

[1] M. Ait Rami, X. Chen, J. B. Moore, and X. Y. Zhou, *Solvability and asymptotic behavior of generalized Riccati equations arising in indefinite stochastic LQ controls*, IEEE Trans. Automat. Control, AC-46 (2001), pp. 428–440.

[2] M. Ait Rami and X. Y. Zhou, *Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic control*, IEEE Trans. Automat. Control, AC-45 (2000), pp. 1131–1143.

[3] M. Ait Rami, X. Y. Zhou, and J. B. Moore, *Well-posedness and attainability of indefinite stochastic linear quadratic control in infinite time horizon*, Systems Control Lett., 41 (2000), 123–133.

[4] B. D. O. Anderson and J. B. Moore, *Optimal Control—Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[5] J.-M. Bismut, *Linear quadratic optimal stochastic control with random coefficients*, SIAM J. Control Optim., 14 (1976), pp. 419–444.

[6] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.

[7] R. F. Brammer, *Controllability in linear autonomous systems with positive controllers*, SIAM J. Control, 10 (1972), pp. 339–353.

[8] R. F. Brammer, *Differential controllability and the solution of linear inequalities-Part* I, IEEE Trans. Automat. Control, AC-20 (1975), pp. 128–131.

[9] S. L. Campbell, *On positive controllers and linear quadratic optimal control problems*, Internat. J. Control, 36 (1982), pp. 885–888.

[10] S. Chen, X. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[11] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems*, Appl. Math. Optim., 43 (2001), pp. 21–45.

[12] M. D. Fragoso, O. L. V. Costa, and C. E. de Souza, *A new approach to linearly perturbed Riccati equations arising in stochastic control*, Appl. Math. Optim., 37 (1998), pp. 99–126.

[13] W. P. M. H. Heemels, S. J. L. van Eijndhoven, and A. A. Stoorvogel, *Linear quadratic regulator problem with positive controls*, Internat. J. Control, 70 (1998), pp. 551–578.

[14] M. Heymann and R. J. Stern, *Controllability of linear systems with positive controls: Geometric considerations*, J. Math. Anal. Appl., 52 (1975), pp. 36–41.

[15] Y. Hu and X. Y. Zhou, *Constrained Stochastic LQ Control with Random Coefficients, and Application to Mean-Variance Portfolio Selection*, Preprint, 2003.
[16] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North-Holland/Kodansha, Tokyo, 1981.
[17] R. E. Kalman, *Contribution to the theory of optimal control*, Bol. Soc. Mat. Mexicana (2), 5 (1960), pp. 102–119.
[18] X. Li, X. Y. Zhou, and A. E. B. Lim, *Dynamic mean-variance portfolio selection with no-shorting constraints*, SIAM J. Control Optim., 40 (2002), pp. 1540–1555.
[19] A. E. B. Lim and X. Y. Zhou, *Mean-variance portfolio selection with random parameters in a complete market*, Math. Oper. Res., 27 (2002), pp. 101–120.
[20] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.
[21] M. Pachter, *The linear-quadratic optimal control problem with positive controllers*, Internat. J. Control, 32 (1980), pp. 589–608.
[22] M. Pachter and D. H. Jacobson, *Control with conic control constraint sets*, J. Optim. Theory Appl., 25 (1978), pp. 117–123.
[23] M. Pachter and D. H. Jacobson, *Stabilization conic control constraint sets*, Internat. J. Control, 29 (1979), pp. 125–132.
[24] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, 3rd ed., Springer-Verlag, Berlin, 1999.
[25] S. H. Saperstone, *Global controllability of linear systems with positive controls*, SIAM J. Control, 11 (1973), pp. 417–423.
[26] S. H. Saperstone and J. A. Yorke, *Controllability of linear oscillatory systems using positive controls*, SIAM J. Control, 9 (1971), pp. 253–262.
[27] H. Sissaoui and W. D. Collins, *The numerical solution of the linear regulator problem with positive control*, IMA J. Math. Control Inform., 5 (1988), pp. 191–201.
[28] P. O. M. Sockaert and J. B. Rawlings, *Constrained linear quadratic regulation*, IEEE Trans. Automat. Control, AC-43 (1998), pp. 1163–1169.
[29] J. F. Sturm and S. Zhang, *On cones of nonnegative quadratic function*, Math. Oper. Res., 28 (2003), pp. 246–267.
[30] F. Willems, W. P. M. H. Heemels, B. de Jager, and A. A. Stoorvogel, *Positive feedback stabilization of centrifugal compressor surge*, Automatica, 38 (2002), pp. 311–318.
[31] W. M. Wonham, *Optimal stationary control of a linear system with state-dependent noises*, SIAM J. Control, 5 (1967), pp. 486–500.
[32] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697; erratum, SIAM J. Control, 7 (1969), p. 365.
[33] D. D. Yao, S. Zhang, and X. Y. Zhou, *Stochastic linear-quadratic control via semidefinite programming*, SIAM J. Control Optim., 40 (2001), pp. 801–823.
[34] J. Yong and X. Y. Zhou, *Stochastic Controls. Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.
[35] X. Y. Zhou and D. Li, *Continuous-time mean-variance portfolio selection: A stochastic LQ framework*, Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1151–1173

OPTIMAL MOMENTUM HEDGING VIA HYPOELLIPTIC REDUCED MONGE–AMPÈRE PDEs*

SRDJAN STOJANOVIC†

Abstract. The celebrated optimal portfolio theory of Merton was successfully extended by the author to assets that do not obey Log-Normal price dynamics in [S. Stojanovic, *Computational Financial Mathematics Using Mathematica®: Optimal Trading in Stocks and Options*, Birkhäuser Boston, Boston, 2003]. Namely, a general one-factor model was solved and applied in the case of appreciation-rate reversing market dynamics. Here, we extend a general methodology to solve the stochastic control problem of optimal portfolio hedging under momentum market dynamics: the corresponding HJB PDE is transformed into the associated Monge–Ampère PDE, which is, utilizing the special structure of the problem, further reduced to a lower-dimensional Monge–Ampère PDE, which is then finally solved numerically. The present problem, in addition to being a two-factor model, has a substantive difficulty due to the degeneracy of the underlying Markov process, yielding the hypoellipticity of its infinitesimal generator, and the corresponding degeneracy of all the fully nonlinear PDEs derived. Furthermore, we solve the problem of optimal hedging and pricing of European and American options in momentum markets, derive a hypoelliptic Black–Scholes PDE/obstacle problem, and introduce a notion of options trading opportunity.

Key words. stochastic control, dynamic programming, optional hedging, Merton's problem, hypoellipticity

AMS subject classifications. 93E20, 60J60, 35H10, 35Q80, 62P05

DOI. 10.1137/S0363012903421170

1. Introduction.

Optimal portfolio hedging under Log-Normal price dynamics is introduced and solved in the seminal work in the early 1970s by Merton (see, e.g., [6, 7]). More precisely, suppose that the asset market prices are governed by the SDE

$$\boldsymbol{d}S(t) = S(t)a\boldsymbol{d}t + S(t)\sigma.\boldsymbol{d}B(t), \tag{1.1}$$

where, for any t, vector (valued random variable) $S(t) = \{S_1(t), \dots, S_m(t)\}$ is a vector, i.e., a one-dimensional array of asset prices, where vector $a = \{a_1, \dots, a_m\} \in \mathbb{R}^m$ is the vector of appreciation rates of the corresponding assets, $\sigma = \{\{\sigma_{1,1}, \dots, \sigma_{1,n}\}, \dots, \{\sigma_{m,1}, \dots, \sigma_{m,n}\}\} \in \mathbb{R}^m \times \mathbb{R}^n$ is the (volatility) matrix, i.e., a two-dimensional array, and $B(t) = \{B_1(t), \dots, B_n(t)\}$ is a vector of n independent Brownian motions. So, m is the number of tradable assets, while n is the number of independent sources of randomness. In (1.1) the multiplication of two vectors of the same length, such as $S(t)a$, is done componentwise, the multiplication of a vector by a matrix of the same length (a length of a matrix is the number of rows), such as $S(t)\sigma$, is done by multiplying a component of vector $S(t)$ by a row of matrix σ, while the "dot" product between a matrix and a vector, such as $\sigma.\boldsymbol{d}B(t)$, is the usual linear-algebra-matrix–vector multiplication; also one can check easily that in such a framework $(S(t)\sigma).\boldsymbol{d}B(t) = S(t)(\sigma.\boldsymbol{d}B(t))$.

What exactly a and σ are, whether they are constants or only t-dependent functions, or whether they depend possibly in a complicated way on other quantities or on other random events, will make a great deal of difference below. For example, if

*Received by the editors January 10, 2003; accepted for publication (in revised form) January 31, 2004; published electronically December 1, 2004.
http://www.siam.org/journals/sicon/43-4/42117.html
†Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221-0025 (srdjan@math.uc.edu, http://math.uc.edu/~srdjan/).

some of the assets are stock-options, then the evolution of their prices depends on the evolution of underlying stock-prices (in addition to time). Later in this work we shall be very precise about such dependencies.

An investor has a cash account, whose balance at time t, denoted by $C(t)$, is obeying the equation

$$\boldsymbol{d}C(t) = rC(t)\boldsymbol{d}t + \boldsymbol{d}c(t), \tag{1.2}$$

where r is the interest rate and where, much more importantly, the $\boldsymbol{d}c(t)$ are the cumulative cash transactions in ($\boldsymbol{d}c(t) > 0$), or outside ($\boldsymbol{d}c(t) < 0$) of the cash account as a consequence of selling and buying assets during the time interval $\boldsymbol{d}t$ (see [8]).

The total wealth, or designated wealth for this investment exercise, shall be denoted by $X(t)$. It represents the cash value of the whole assets-portfolio, plus the balance in the cash account. The problem is to design a trading, or more precisely, a portfolio hedging strategy. A trading or hedging strategy is going to be any vector-valued function $\Pi(t) = \{\Pi_1(t), \Pi_2(t), \dots, \Pi_m(t)\}$, where $\Pi_j(t) = \Pi_j(t, X, .)$ is the cash value, positive or negative, depending on whether it is a long or short position, of the investment in the jth asset considered for trading.

So, the total portfolio value, i.e., the total wealth, is equal to

$$X(t) = C(t) + \sum_{i=1}^{m} \Pi_i(t). \tag{1.3}$$

It can be derived (see section 7.3.1 of [8]) that

$$\boldsymbol{d}X^{\Pi}(t) = \boldsymbol{d}X(t) = (\Pi(t).(a - r) + rX(t))\boldsymbol{d}t + \Pi(t).\sigma.\boldsymbol{d}B(t), \tag{1.4}$$

where, by definition, $a - r = \{a_1, \dots, a_m\} - r = \{a_1 - r, \dots, a_m - r\}$. This equation is referred to as the wealth evolution SDE. Notice that unless a and σ (or even r) are either constants or functions of t only, the first-order SDE (1.4) is not closed, and its solution will depend on some other equations as well.

The objective of the investor/trader is to maximize the expected value of the utility ψ of the final total wealth:

$$v(t, X, .) = \sup_{\Pi = \Pi(t,X,.)} E_{t,X,.}\psi(X^{\Pi}(T)) = \max_{\Pi} E_{t,X,.}\psi(X^{\Pi}(T)) = E_{t,X,.}\psi(X^{\Pi^*}(T)), \tag{1.5}$$

where $E_{t,X,.}$ is the conditional expectation under the condition that at time t the total portfolio value was equal to X, and under some other conditions generically denoted as ".". Also, most often, the class of utility functions considered is HARA class: $\mathcal{U} = \{\psi_\gamma, 0 < \gamma < \infty\}$, where

$$\psi_\gamma(X) = \frac{X^{1-\gamma}}{1-\gamma} \tag{1.6}$$

for $0 < \gamma \neq 1$, while for $\gamma = 1$

$$\psi_1(X) = \log(X). \tag{1.7}$$

The value function v, corresponding to a HARA utility ψ_γ, will be denoted v_γ. Although we shall not do that in this work, one can solve problems when the class of

admissible portfolio rules Π in (1.5) is constrained in some of the convenient ways (see [8, Chapters 7 and 8]).

It is the celebrated result of Merton that, if a and σ (and r) are functions of t only (or constants), then for any $\gamma \in (0,\infty)$, the *value function* $v_\gamma(t,X,.) = v_\gamma(t,X)$ can be computed explicitly as the unique solution of the Monge–Ampère-type PDE

$$-\frac{1}{2}(a(t)-r(t)).(\sigma(t).\sigma(t)^T)^{-1}.(a(t)-r(t))\left(\frac{\partial v_\gamma(t,X)}{\partial X}\right)^2 + X\frac{\partial^2 v_\gamma(t,X)}{\partial X^2}r(t)\frac{v_\gamma(t,X)}{\partial X} + \frac{\partial v_\gamma(t,X)}{\partial t}\frac{\partial^2 v_\gamma(t,X)}{\partial X^2} = 0 \tag{1.8}$$

together with the terminal condition

$$v_\gamma(T,X) = \psi_\gamma(X) \tag{1.9}$$

and such that $\frac{\partial^2 v_\gamma(t,X)}{\partial X^2} < 0$. The solution is equal to

$$v_\gamma(t,X) = \frac{f(t)X^{1-\gamma}}{1-\gamma} = \frac{\boldsymbol{e}^{\frac{(\gamma-1)\int_t^T (a(\tau)-r(\tau)).(\sigma(\tau).\sigma(\tau)^T)^{-1}.(a(\tau)-r(\tau))+2\gamma r(\tau)\boldsymbol{d}\tau}{-2\gamma}}X^{1-\gamma}}{1-\gamma} \tag{1.10}$$

for $\gamma \neq 1$. (If $\gamma = 1$, another formula holds; see, e.g., [8] and also below.) Equation (1.8) is already quite interesting: notice that if $\gamma \geq 1$, the solution $v_\gamma(t,X) \to -\infty$, when $X \to 0$, while $v_\gamma(t,0) = 0$ for $0 < \gamma < 1$—either way, a boundary value cannot be imposed at $X = 0$ due to the degeneracy of the equation. Furthermore, the Merton optimal portfolio hedging strategy $\Pi^\star_\gamma$ is equal to

$$\begin{aligned}\Pi^\star_\gamma(t,X) &= -\frac{\frac{\partial v_\gamma(t,X)}{\partial X}(\sigma(t).\sigma(t)^T)^{-1}.(a(t)-r(t))}{\frac{\partial^2 v_\gamma(t,X)}{\partial X^2}} \\ &= \frac{X(\sigma(t).\sigma(t)^T)^{-1}.(a(t)-r(t))}{\gamma} = XP^\star_\gamma(t)\end{aligned} \tag{1.11}$$

for any $\gamma \in (0,\infty)$, where

$$P^\star_\gamma(t) = \frac{(\sigma(t).\sigma(t)^T)^{-1}.(a(t)-r(t))}{\gamma} \tag{1.12}$$

is the optimal investment per unit wealth. Of course, above it is assumed that the matrix $\sigma(t).\sigma(t)^T$ is invertible. There will be similar assumptions below which we shall consider self-evident without specifying them explicitly. In [8] the author extends Merton's theory to the case when the underlying asset price dynamics is much more general, yielding complicated Monge–Ampère-type PDEs. The general methodology introduced there will be applied here on a new asset price dynamics, i.e., to the case of *momentum* asset price dynamics. The Monge–Ampère-type PDEs considered here will have an additional *degeneracy* feature; as a matter of fact, we shall consider equations that can be called *hypoelliptic Monge–Ampère-type PDEs*. No theory for such equations exists today.

A model for momentum asset price dynamics was introduced in [8] as

$$\begin{aligned}\boldsymbol{d}Y(t) &= Z(t)\boldsymbol{d}t + \sigma_0\boldsymbol{d}b(t), \\ \boldsymbol{d}Z(t) &= \left(\frac{2\pi}{p}\right)^2(e - Y(t))\boldsymbol{d}t,\end{aligned} \tag{1.13}$$

where $Y(t)$ is the price of the considered security at time t, where $Z(t)$ is the *trend* (or "price velocity"), and where $b(t)$ is a scalar Brownian motion. A Markov process $\{Y(t), Z(t)\}$ is called the Price/Trend process. The parameters of the model are σ_0 (the price-*diffusion*), e (the price-*equilibrium*), and p (the price-dynamics *period*). Notice that in that model $Z(t)$ is not the appreciation rate—$Z(t)/Y(t)$ is—and that σ_0 is not the volatility—$\sigma_0/Y(t)$ is. Notice that if $\sigma_0 = 0$, then (1.13) is equivalent to the second-order ODE

$$Y''(t) + \left(\frac{2\pi}{p}\right)^2 Y(t) = \left(\frac{2\pi}{p}\right)^2 e, \tag{1.14}$$

which is the classical model for (undamped) oscillations around an equilibrium e, with a period of oscillations being equal to p.

Since in the optimal portfolio hedging it is much more convenient to work with appreciation rates (growth rates or maybe, more concisely, *momentum*) and volatilities, we shall introduce below a simple modification of the model (1.13); i.e., we shall introduce the Price/Momentum process.

2. SDE model: Price/Momentum process. We modify here the Price/Trend system (1.13) into a Price/Momentum system; i.e., we *postulate* that the value of a market-index, or an underlying stock-price if options hedging is attempted (see below), is governed (in the short-run) by

$$\begin{aligned} \boldsymbol{d}Y(t) &= Y(t)A(t)\boldsymbol{d}t + Y(t)\sigma\boldsymbol{d}b(t), \\ \boldsymbol{d}A(t) &= \frac{1}{Y(t)}\left(\frac{2\pi}{p}\right)^2 (e - Y(t))\boldsymbol{d}t \end{aligned} \tag{2.1}$$

for $0 < t < T$, with an initial condition

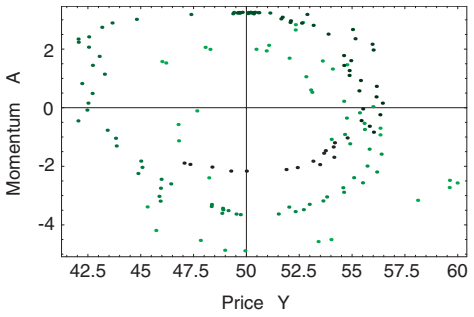
$$\begin{aligned} Y(0) &= Y_0, \\ A(0) &= A_0, \end{aligned} \tag{2.2}$$

where $Y(t)$ is the price or value of the asset at time t, while $A(t)$ is the growth rate or, more concisely, *momentum*. Notice that, unfortunately, in real trading the momentum $A(t)$ is not going to be observable—only the current price $Y(t)$ is. So, $A(t)$ would have to be estimated; i.e., one would have to compute $E[A(t)|Y(s), s \le t]$. Actually, it is not difficult, using the Liptser and Shiryaev theory of optimal filtering of conditionally Gaussian processes (see, e.g., [8] and references given there) that if $A(0) \sim N[m_0, g_0]$, then $m(t) = E[A(t)|Y(s), s \le t]$ can be computed from the equation

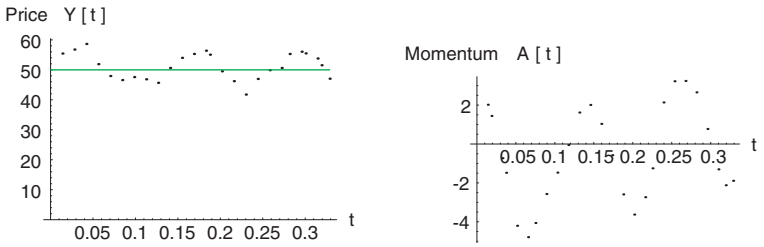
$$\begin{aligned} \boldsymbol{d}m(t) &= \left(\frac{2\pi}{p}\right)^2 \frac{(e - Y(t))}{Y(t)}\boldsymbol{d}t + \frac{g_0^2}{\sigma^2 + tg_0^2}\left(\frac{\boldsymbol{d}Y(t)}{Y(t)} - m(t)\boldsymbol{d}t\right), \\ m(0) &= m_0. \end{aligned} \tag{2.3}$$

Knowing $m(t)$ as opposed to knowing $A(t)$ would ideally have to be taken into account when computing an optimal hedging strategy. We shall not do that and we shall simply assume that $A(t)$, along with $Y(t)$, are observable or, equivalently, that $g_0 = 0$.

The model (2.1) is somewhat different from (1.13) and indeed may seem a bit arbitrary. Possibly the best defense are the Monte-Carlo experiments. For example, a single trajectory in the *phase plane*, for $e = 50$, $p = 60/365$, $Y_0 = 50, A_0 = 3$, and $0 < t < T = 2p$, looks like



The phase-plane plot does not refer to time. So, in order to see time dependence, we need two plots—one for the Price-trajectory and one for the Momentum-trajectory:



Notice that the momentum $A(t)$ is smooth, while the price $Y(t)$ is not.

It is not difficult to see that the infinitesimal generator of the Markov process $\{Y(t), A(t)\}_t$ is equal to

$$\mathcal{A} = \frac{1}{2}\sigma^2 Y^2 \frac{\partial^2}{\partial Y^2} + AY\frac{\partial}{\partial Y} + \frac{e-Y}{Y}\left(\frac{2\pi}{p}\right)^2 \frac{\partial}{\partial A}. \tag{2.4}$$

Operator $\mathcal{A}$ is quite interesting in $\{\{Y, A\}, Y > 0\}$. Indeed, $\frac{e-Y}{Y}$ changes sign at $Y = e$, and therefore $\mathcal{A}$ is backward parabolic in $Y < e$, and forward parabolic in $Y > e$. As a matter of fact, operator $\mathcal{A}$ is hypoelliptic in $\{\{Y, A\}, Y > 0\}$ (see [3]). Indeed, let a first-order differential operator X_1 be defined as

$$X_1 = \frac{1}{\sqrt{2}}\sigma Y \frac{\partial}{\partial Y}. \tag{2.5}$$

Then

$$X_1{}^2 = \left(\frac{1}{\sqrt{2}}\sigma Y\frac{\partial}{\partial Y}\right)\left(\frac{1}{\sqrt{2}}\sigma Y\frac{\partial}{\partial Y}\right) = \frac{1}{2}\sigma^2 Y\frac{\partial}{\partial Y} + \frac{1}{2}\sigma^2 Y^2\frac{\partial^2}{\partial Y^2}. \tag{2.6}$$

Therefore, if X_0 is a differential operator defined as

$$X_0 = \mathcal{A} - X_1{}^2 = \left(A - \frac{1}{2}\sigma^2\right) Y\frac{\partial}{\partial Y} + \frac{e-Y}{Y}\left(\frac{2\pi}{p}\right)^2\frac{\partial}{\partial A}, \tag{2.7}$$

then $\mathcal{A} = X_1{}^2 + X_0$. Notice that at $Y = e$, $\mathrm{Span}[X_0, X_1] \neq T_{(Y,A)}\mathbb{R}^2$. Nevertheless

the commutator can be computed to be

(2.8)
$$[X_0, X_1] = X_0X_1 - X_1X_0 = \left(\left(A - \frac{1}{2}\sigma^2\right) Y\frac{\partial}{\partial Y} + \frac{e-Y}{Y}\left(\frac{2\pi}{p}\right)^2 \frac{\partial}{\partial A}\right)\left(\frac{1}{\sqrt{2}}\sigma Y\frac{\partial}{\partial Y}\right)$$
$$- \left(\frac{1}{\sqrt{2}}\sigma Y\frac{\partial}{\partial Y}\right)\left(\left(A - \frac{1}{2}\sigma^2\right) Y\frac{\partial}{\partial Y} + \frac{e-Y}{Y}\left(\frac{2\pi}{p}\right)^2 \frac{\partial}{\partial A}\right) = \frac{1}{\sqrt{2}}\sigma\frac{e}{Y}\left(\frac{2\pi}{p}\right)^2 \frac{\partial}{\partial A},$$

and therefore, for any $Y \neq 0$ (see [3]),

(2.9)
$$\begin{aligned}\mathrm{Span}[\mathrm{Lie}[X_0, X_1]][Y, A] &= \mathrm{Span}[X_0, X_1, [X_0, X_1]][Y, A]\\ &= \mathrm{Span}\left[\frac{\partial}{\partial Y}, \frac{\partial}{\partial A}\right][Y, A] = T_{(Y,A)}\mathbb{R}^2,\end{aligned}$$

and consequently the Hörmander condition holds in $\{Y > 0\}$, and operator $\mathcal{A}$ is hypoelliptic there. From the applied point of view, the importance of hypoellipticity lies in the (numerical) solvability of the boundary value problems that will follow, as well as for the computational strategy. We notice, however, that in regard to the theoretical solvability of some of the problems below hypoelliptic Monge–Ampère-type PDEs are too difficult and still beyond the current theoretical reach. We hope that these applied considerations may provide some guidance to the theoretical investigations as well.

3. Optimal momentum portfolio hedging problem. Now consider an index (or underlying stock) with value (price) $Y(t)$ obeying the dynamics (2.1) or, a bit more generally, the dynamics

(3.1)
$$\begin{aligned}\boldsymbol{d}Y(t) &= Y(t)A(t)\boldsymbol{d}t + Y(t)\sigma_y(t, Y(t), A(t)).\boldsymbol{d}B(t),\\ \boldsymbol{d}A(t) &= \frac{1}{Y(t)}\left(\frac{2\pi}{p}\right)^2 (e - Y(t))\boldsymbol{d}t,\end{aligned}$$

where the vector-valued function $\sigma_y(t, Y, A) = \{\sigma_{y,1}(t, Y, A), \dots, \sigma_{y,n}(t, Y, A)\} \in \mathbb{R}^n$ is the vector-volatility of the considered index, and $B(t) = \{B_1(t), \dots, B_n(t)\}$ is the vector of n independent Brownian motions. Now in addition to the index, consider a set of *tradable assets* with prices $S(t) = \{S_1(t), \dots, S_m(t)\}$ obeying the dynamics

(3.2)
$$\boldsymbol{d}S(t) = S(t)a_s(t, Y(t), A(t))\boldsymbol{d}t + S(t)\sigma_s(t, Y(t), A(t)).\boldsymbol{d}B(t),$$

where the vector-valued function $a_s(t, Y, A) = \{a_{s,1}(t, Y, A), \dots, a_{s,m}(t, Y, A)\} \in \mathbb{R}^m$ is the vector of appreciation rates of the corresponding assets, and $\sigma_s(t, Y, A) \in \mathbb{R}^m \times \mathbb{R}^n$ is the (volatility) matrix. If the index is tradable (indeed there are securities that represent indices; for example, *qqq*), then it can be represented also as one of the equations in system (3.2). So, the above model attempts to describe a situation when a dominant security or an index, exhibiting a momentum price dynamics described by SDE system (3.1), is driving a number of additional securities considered for trading, whose price dynamics obeys the SDE system (3.2).

Also, we consider the cash account evolving according to (1.2). In such a framework the wealth evolution equation (1.4) still holds, now written more explicitly as

(3.3)
$$\begin{aligned}&\boldsymbol{d}X(t) = (\Pi(t, X(t), Y(t), A(t)).(a_s(t, Y(t), A(t)) - r) + rX(t))\boldsymbol{d}t\\ &+ \Pi(t, X(t), Y(t), A(t)).\sigma_s(t, Y(t), A(t)).\boldsymbol{d}B(t).\end{aligned}$$

In particular, the trading strategies are going to be vector-valued functions of time t, wealth X, index-value Y, and index-momentum A: $\Pi(t,X,Y,A)$. System (3.1), (3.3) is then a closed stochastic system to be controlled ((3.2) is hidden in (3.3)).

For any fixed trading strategy Π, the stochastic process $\mathcal{M}^{\Pi}=\{t,X(t),Y(t),A(t)\}_t$ is a Markov process. We compute its infinitesimal generator $\partial\mathcal{M}^{\Pi}$. To this end, using Itô's chain rule, we get after some calculus (we shall use an alternative notation for partial derivatives: for example, $\varphi^{(0,1,1,0)}=\varphi^{(0,1,1,0)}(t,X,Y,A)=\frac{\partial^2\varphi(t,X,Y,A)}{\partial X\,\partial Y}$)

(3.4)
$$\begin{aligned}
&\boldsymbol{d}\varphi=\varphi^{(1,0,0,0)}\boldsymbol{d}t+\varphi^{(0,1,0,0)}\boldsymbol{d}X(t)+\varphi^{(0,0,1,0)}\boldsymbol{d}Y(t)+\varphi^{(0,0,0,1)}\boldsymbol{d}A(t)+\frac{1}{2}\varphi^{(0,2,0,0)}(\boldsymbol{d}X(t))^2\\
&+\frac{1}{2}\varphi^{(0,0,2,0)}(\boldsymbol{d}Y(t))^2+\varphi^{(0,1,1,0)}\boldsymbol{d}X(t)\boldsymbol{d}Y(t)=\Bigg(\varphi^{(1,0,0,0)}+\varphi^{(0,1,0,0)}(\Pi.(a_s-r)\\
&+rX(t))+\varphi^{(0,0,1,0)}Y(t)A(t)+\varphi^{(0,0,0,1)}\frac{1}{Y(t)}\left(\frac{2\pi}{p}\right)^2(e-Y(t))+\frac{1}{2}\varphi^{(0,2,0,0)}\\
&\Pi.\sigma_s.\sigma_s{}^T.\Pi+\frac{1}{2}\varphi^{(0,0,2,0)}Y(t)^2\sigma_y.\sigma_y+\varphi^{(0,1,1,0)}Y(t)\Pi.\sigma_s.\sigma_y\Bigg)\boldsymbol{d}t+\varphi^{(0,1,0,0)}\Pi.\sigma_s.\boldsymbol{d}B(t)\\
&+\varphi^{(0,0,1,0)}Y(t)\sigma_y.\boldsymbol{d}B(t)=(\varphi^{(1,0,0,0)}+\mathcal{B}^{\Pi}(t)\varphi)\boldsymbol{d}t+\varphi^{(0,1,0,0)}\Pi.\sigma_s.\boldsymbol{d}B(t)\\
&+\varphi^{(0,0,1,0)}Y(t)\sigma_y.\boldsymbol{d}B(t),
\end{aligned}$$

and therefore, for any fixed hedging strategy Π, operator $\partial\mathcal{M}^{\Pi}$ is identified to be

(3.5)
$$\begin{aligned}
\partial\mathcal{M}^{\Pi}&=\frac{\partial}{\partial t}+\mathcal{B}^{\Pi}(t)=\frac{\partial}{\partial t}+(\Pi(t,X,Y,A).(a_s(t,Y,A)-r)+rX)\frac{\partial}{\partial X}+YA\frac{\partial}{\partial Y}\\
&+\frac{1}{Y}\left(\frac{2\pi}{p}\right)^2(e-Y)\frac{\partial}{\partial A}+\frac{1}{2}\Pi(t,X,Y,A).\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T.\Pi(t,X,Y,A)\frac{\partial^2}{\partial X^2}\\
&+\frac{1}{2}Y^2\sigma_y(t,Y,A).\sigma_y(t,Y,A)\frac{\partial^2}{\partial Y^2}+Y\Pi(t,X,Y,A).\sigma_s(t,Y,A).\sigma_y(t,Y,A)\frac{\partial^2}{\partial X\partial Y}.
\end{aligned}$$

Under *appropriate conditions* on Π, similarly as in the case of operator $\mathcal{A}$ above, operators $\mathcal{B}^{\Pi}(t)$ are hypoelliptic in the semi-infinite domain $\{\{X,Y,A\},X>0,Y>0\}$, and so is $\partial\mathcal{M}^{\Pi}$ in

$$Q_T=\{\{t,X,Y,A\},t<T,X>0,Y>0\}. \tag{3.6}$$

This means, in particular, that for such a fixed hedging strategy Π, if φ is the solution of the *linear* hypoelliptic PDE

$$\partial\mathcal{M}^{\Pi}\varphi=0 \tag{3.7}$$

in Q_T, together with the terminal condition

$$\varphi(T,X,Y,A)=\psi(X,Y,A) \tag{3.8}$$

for some given function ψ, then, due to the hypoellipticity of $\partial\mathcal{M}^{\Pi}$, we have that $\varphi\in C^{\infty}(Q_T)$, and moreover

$$\varphi(t,X,Y,A)=E_{t,X,Y,A}\psi(X(T),Y(T),A(T)). \tag{3.9}$$

The objective of the investor/trader is to, for a given utility function ψ, maximize the expected value of the utility of the final total wealth, i.e., to find an optimal hedging strategy $\Pi^\star(t,X,Y,A)$ such that

(3.10)
$$v(t,X,Y,A)=\sup_{\Pi} E_{t,X,Y,A}\psi(X^{\Pi}(T))=\max_{\Pi} E_{t,X,Y,A}\psi(X^{\Pi}(T))=E_{t,X,Y,A}\psi(X^{\Pi^\star}(T)).$$

Above $E_{t,X,Y,A}$ is the conditional expectation under the condition that at time t the total portfolio value was equal to X, the index had a value Y, and the index-momentum was equal to A. Function $v(t,X,Y,A)$ is the value function.

4. From the HJB to the first Monge–Ampère PDE. The standard formalism for solving the stochastic control problem (3.10) is to attempt to solve the associated Hamilton–Jacobi–Bellman (HJB) PDE characterizing the value function $v=\varphi=\varphi(t,X,Y,A)$:

(4.1)
$$\begin{aligned}\mathrm{Max}[\partial\mathcal{M}^{\Pi}\varphi]=\mathrm{Max}\Big[&\varphi^{(1,0,0,0)}+\varphi^{(0,1,0,0)}(\Pi.(a_s-r)+rX)+\varphi^{(0,0,1,0)}YA\\&+\varphi^{(0,0,0,1)}\frac{1}{Y}\left(\frac{2\pi}{p}\right)^2(e-Y)+\frac{1}{2}\varphi^{(0,2,0,0)}\Pi.\sigma_s.{\sigma_s}^T.\Pi+\frac{1}{2}\varphi^{(0,0,2,0)}Y^2\sigma_y.\sigma_y\\&+\varphi^{(0,1,1,0)}Y\Pi.\sigma_s.\sigma_y\Big]=0\end{aligned}$$

in the domain Q_T (see (3.6)), with the terminal condition

$$\varphi(T,X,Y,A)=\psi(X). \tag{4.2}$$

We notice that (4.1) is very much degenerate, if considered in the usual way, as a backward "parabolic" equation. Indeed, the linear operator under the max above is only hypoelliptic (due to the term $\varphi^{(0,0,0,1)}\frac{1}{Y}(\frac{2\pi}{p})^2(e-Y)$, whose coefficient changes sign at $Y=e$, and due to the absence of the second-order term $\varphi^{(0,0,0,2)}$). There are additional, less consequential, sources of degeneracies in (4.1), such as the fact that max is taken over all the strategies—think of $\Pi=0$ or, less dramatically, think of any Π proportional to X (as a matter of fact the optimal Π, the one for which the max in (4.1) is realized, and therefore the consequential one, is going to be such). Yet another already explicit degeneracy is at $Y=0$, due to the term $\frac{1}{2}\varphi^{(0,0,2,0)}(t,X,Y,A)Y^2\sigma_y(t,Y,A).\sigma_y(t,Y,A)$. These degeneracies account for lack of the boundary condition there. All of those degeneracies, and especially the first one, yield that (4.1) is beyond what is understood in the *theory* of HJB PDEs today. Our understanding of such equations is based on their equivalence to some simpler equations to be derived below, on their numerical solutions, and on stochastic representations. We rewrite (4.1) by extracting from the max all the terms independent of Π:

$$\begin{aligned}0&=\mathrm{Max}[\partial\mathcal{M}^{\Pi}\varphi]\\&=\varphi^{(1,0,0,0)}+rX\varphi^{(0,1,0,0)}+\varphi^{(0,0,1,0)}YA+\varphi^{(0,0,0,1)}\frac{1}{Y}\left(\frac{2\pi}{p}\right)^2(e-Y)\\&\quad+\frac{1}{2}\varphi^{(0,0,2,0)}Y^2\sigma_y.\sigma_y+\mathrm{Max}F(\Pi),\end{aligned} \tag{4.3}$$

where

$$F(\Pi) = \varphi^{(0,1,0,0)}\Pi.(a_s - r) + \frac{1}{2}\varphi^{(0,2,0,0)}\Pi.\sigma_s.{\sigma_s}^T.\Pi + \varphi^{(0,1,1,0)}Y\Pi.\sigma_s.\sigma_y. \tag{4.4}$$

Since F is quadratic in Π, to find $\Pi^\star$ such that $\mathrm{Max}_\Pi F(\Pi) = F(\Pi^\star)$, we just need to solve the equation $\nabla F(\Pi) = 0$, yielding

(4.5)

$$\begin{aligned}\Pi^\star(t,X,Y,A) &= -\frac{\varphi^{(0,1,0,0)}}{\varphi^{(0,2,0,0)}}(\sigma_s.{\sigma_s}^T)^{-1}.(a_s - r) - \frac{\varphi^{(0,1,1,0)}}{\varphi^{(0,2,0,0)}}Y(\sigma_s.{\sigma_s}^T)^{-1}.\sigma_s.\sigma_y \\ &= -\frac{\varphi^{(0,1,0,0)}}{\varphi^{(0,2,0,0)}}(a_s - r).(\sigma_s.{\sigma_s}^T)^{-1} - \frac{\varphi^{(0,1,1,0)}}{\varphi^{(0,2,0,0)}}Y\sigma_y.{\sigma_s}^T.(\sigma_s.{\sigma_s}^T)^{-1}.\end{aligned}$$

We are going to use both expressions in (4.5). Going back to the HJB PDE (4.3), we get (after multiplying by $\varphi^{(0,2,0,0)}$)

$$\begin{aligned}&\varphi^{(0,2,0,0)}\varphi^{(1,0,0,0)} + rX\varphi^{(0,2,0,0)}\varphi^{(0,1,0,0)} + \varphi^{(0,2,0,0)}\varphi^{(0,0,1,0)}YA \\ &+ \varphi^{(0,2,0,0)}\varphi^{(0,0,0,1)}\frac{1}{Y}\left(\frac{2\pi}{p}\right)^2(e - Y) + \frac{1}{2}\varphi^{(0,2,0,0)}\varphi^{(0,0,2,0)}Y^2\sigma_y.\sigma_y \\ &+ \varphi^{(0,2,0,0)}\varphi^{(0,1,0,0)}\Pi^\star.(a_s - r) + \frac{1}{2}{\varphi^{(0,2,0,0)}}^2\Pi^\star.\sigma_s.{\sigma_s}^T.\Pi^\star \\ &+ \varphi^{(0,2,0,0)}\varphi^{(0,1,1,0)}Y\Pi^\star.\sigma_s.\sigma_y = 0.\end{aligned} \tag{4.6}$$

We also want to plug (4.5) back into (4.6). To this end, we prepare the formulas

$$\begin{aligned}&\varphi^{(0,2,0,0)}\varphi^{(0,1,0,0)}\Pi^\star.(a_s - r) \\ &= \varphi^{(0,1,0,0)}\big(-\varphi^{(0,1,0,0)}(a_s - r).(\sigma_s.{\sigma_s}^T)^{-1} \\ &\quad - \varphi^{(0,1,1,0)}Y\sigma_y.{\sigma_s}^T.(\sigma_s.{\sigma_s}^T)^{-1}\big).(a_s - r) \\ &= -{\varphi^{(0,1,0,0)}}^2(a_s - r).(\sigma_s.{\sigma_s}^T)^{-1}.(a_s - r) \\ &\quad - \varphi^{(0,1,0,0)}\varphi^{(0,1,1,0)}Y\sigma_y.{\sigma_s}^T.(\sigma_s.{\sigma_s}^T)^{-1}.(a_s - r)\end{aligned} \tag{4.7}$$

as well as

(4.8)

$$\begin{aligned}&\frac{1}{2}{\varphi^{(0,2,0,0)}}^2\Pi^\star.\sigma_s.{\sigma_s}^T.\Pi^\star \\ &= \frac{1}{2}{\varphi^{(0,2,0,0)}}^2\left(-\frac{\varphi^{(0,1,0,0)}}{\varphi^{(0,2,0,0)}}(a_s - r).(\sigma_s.{\sigma_s}^T)^{-1} - \frac{\varphi^{(0,1,1,0)}}{\varphi^{(0,2,0,0)}}Y\sigma_y.{\sigma_s}^T.(\sigma_s.{\sigma_s}^T)^{-1}\right).\sigma_s.{\sigma_s}^T \\ &.\left(-\frac{\varphi^{(0,1,0,0)}}{\varphi^{(0,2,0,0)}}(\sigma_s.{\sigma_s}^T)^{-1}.(a_s - r) - \frac{\varphi^{(0,1,1,0)}}{\varphi^{(0,2,0,0)}}Y(\sigma_s.{\sigma_s}^T)^{-1}.\sigma_s.\sigma_y\right) \\ &= \frac{1}{2}{\varphi^{(0,1,0,0)}}^2(a_s - r).(\sigma_s.{\sigma_s}^T)^{-1}.(a_s - r) + \varphi^{(0,1,0,0)}\varphi^{(0,1,1,0)}Y(a_s - r).(\sigma_s.{\sigma_s}^T)^{-1}.\sigma_s.\sigma_y \\ &+ \frac{1}{2}{\varphi^{(0,1,1,0)}}^2Y^2\sigma_y.{\sigma_s}^T.(\sigma_s.{\sigma_s}^T)^{-1}.\sigma_s.\sigma_y\end{aligned}$$

and

(4.9)

$$\begin{aligned}&\varphi^{(0,2,0,0)}\varphi^{(0,1,1,0)}Y\Pi^\star.\sigma_s.\sigma_y \\ &= \varphi^{(0,1,1,0)}Y\big(-\varphi^{(0,1,0,0)}(a_s - r).(\sigma_s.{\sigma_s}^T)^{-1} - \varphi^{(0,1,1,0)}Y\sigma_y.{\sigma_s}^T.(\sigma_s.{\sigma_s}^T)^{-1}\big).\sigma_s.\sigma_y \\ &= -Y\varphi^{(0,1,1,0)}\varphi^{(0,1,0,0)}(a_s - r).(\sigma_s.{\sigma_s}^T)^{-1}.\sigma_s.\sigma_y - {\varphi^{(0,1,1,0)}}^2Y^2\sigma_y.{\sigma_s}^T.(\sigma_s.{\sigma_s}^T)^{-1}.\sigma_s.\sigma_y.\end{aligned}$$

Using (4.7)–(4.9), (4.6) becomes

$$\begin{aligned}
&\varphi^{(0,2,0,0)}\varphi^{(1,0,0,0)}+rX\varphi^{(0,2,0,0)}\varphi^{(0,1,0,0)}+\varphi^{(0,2,0,0)}\varphi^{(0,0,1,0)}YA\\
&+\varphi^{(0,2,0,0)}\varphi^{(0,0,0,1)}\frac{1}{Y}\left(\frac{2\pi}{p}\right)^2(e-Y)\\
&+\frac{1}{2}\varphi^{(0,2,0,0)}\varphi^{(0,0,2,0)}Y^2\sigma_y.\sigma_y-\varphi^{(0,1,0,0)^2}(a_s-r).(\sigma_s.\sigma_s{}^T)^{-1}.(a_s-r)\\
&-\varphi^{(0,1,0,0)}\varphi^{(0,1,1,0)}Y\sigma_y.\sigma_s{}^T.(\sigma_s.\sigma_s{}^T)^{-1}.(a_s-r)+\frac{1}{2}\varphi^{(0,1,0,0)^2}(a_s-r).(\sigma_s.\sigma_s{}^T)^{-1}.(a_s-r)\\
&+\varphi^{(0,1,0,0)}\varphi^{(0,1,1,0)}Y(a_s-r).(\sigma_s.\sigma_s{}^T)^{-1}.\sigma_s.\sigma_y+\frac{1}{2}\varphi^{(0,1,1,0)^2}Y^2\sigma_y.\sigma_s{}^T.(\sigma_s.\sigma_s{}^T)^{-1}.\sigma_s.\sigma_y\\
&-Y\varphi^{(0,1,1,0)}\varphi^{(0,1,0,0)}(a_s-r).(\sigma_s.\sigma_s{}^T)^{-1}.\sigma_s.\sigma_y\\
&-\varphi^{(0,1,1,0)^2}Y^2\sigma_y.\sigma_s{}^T.(\sigma_s.\sigma_s{}^T)^{-1}.\sigma_s.\sigma_y=0,
\end{aligned}$$

and consequently

$$\begin{aligned}
&\varphi^{(0,2,0,0)}\varphi^{(1,0,0,0)}+rX\varphi^{(0,2,0,0)}\varphi^{(0,1,0,0)}+\varphi^{(0,2,0,0)}\varphi^{(0,0,1,0)}YA\\
&+\varphi^{(0,2,0,0)}\varphi^{(0,0,0,1)}\frac{1}{Y}\left(\frac{2\pi}{p}\right)^2(e-Y)\\
&+\frac{1}{2}\varphi^{(0,2,0,0)}\varphi^{(0,0,2,0)}Y^2\sigma_y.\sigma_y-\frac{1}{2}\varphi^{(0,1,0,0)^2}(a_s-r).(\sigma_s.\sigma_s{}^T)^{-1}.(a_s-r)\\
&-\varphi^{(0,1,0,0)}\varphi^{(0,1,1,0)}Y\sigma_y.\sigma_s{}^T.(\sigma_s.\sigma_s{}^T)^{-1}.(a_s-r)\\
&-\frac{1}{2}\varphi^{(0,1,1,0)^2}Y^2\sigma_y.\sigma_s{}^T.(\sigma_s.\sigma_s{}^T)^{-1}.\sigma_s.\sigma_y=0
\end{aligned}\tag{4.10}$$

in the domain Q_T (see (3.6)), together with the terminal condition (4.2), and such that

$$\varphi^{(0,2,0,0)}(t,X,Y,A)<0.\tag{4.11}$$

So, the solution of the HJB PDE (4.1) solves the Monge–Ampère-type PDE (4.10). Equation (4.10) is simpler than (4.1), but it is still fully nonlinear, and strongly degenerate, if considered as backward parabolic. In much the same way as in (4.1), the most interesting term in (4.10) is $\varphi^{(0,2,0,0)}\varphi^{(0,0,0,1)}\frac{1}{Y}(\frac{2\pi}{p})^2(e-Y)$—the coefficient changes sign at $Y=e$, and (4.10) has no second-order term $\varphi^{(0,0,0,2)}$. Indeed, it yields a sort of a "hypoellipticity" of the Monge–Ampère equation (4.10). Also, other degeneracies are explicit now; at $X=0$ the degeneracy is due to the term $rX\varphi^{(0,1,0,0)}(t,X,Y,A)\varphi^{(0,2,0,0)}(t,X,Y,A)$, and similarly at $Y=0$. Consequently, the boundary condition is not imposed at $X=0$ and $Y=0$. Only the terminal condition (4.2) is imposed. One can compare (4.10) and (1.8)—they are analogous, (4.10) being so much more complicated.

5. From the first Monge–Ampère PDE to the reduced Monge–Ampère PDE. So far, we could have worked with pretty much any wealth-utility function $\psi(X)$. For example, the nonconcave or even the discontinuous utility function $\psi(X)$ (such a problem arises, for example, if one attempts to maximize a probability of reaching a certain wealth by the given deadline) would contradict (4.11), which would cause the terminal condition (4.2) to not be taken in a continuous way. Nevertheless, everything would work just fine, at least from the point of view of numerical solutions

and stochastic control, i.e., financial implications. On the other hand, from now on, the most fruitful study will be very much dependent on restricting our attention to the HARA class $\mathcal{U} = \{\psi_\gamma(X), \gamma > 0\}$ (see (1.6) and (1.7)). Furthermore, we shall have to restrict ourselves to the subclass $\mathcal{U}_1 = \{\psi_\gamma(X), \gamma \geq 1\}$. This is not a serious disadvantage considering the importance of $\mathcal{U}_1$—for the most reasonable investing one indeed has to use utilities ψ_γ for large γ's, since otherwise the investment decisions would be too aggressive and consequently too risky.

How do we solve (4.10), (4.2)? Similarly as was already introduced in [8] and inspired by (1.10), in the case of HARA utility, we seek the solution of (4.10), (4.2) in the form

$$\varphi(t, X, Y, A) = \frac{X^{1-\gamma} f(t, Y, A)}{1-\gamma} \tag{5.1}$$

if $\gamma \neq 1$, or

$$\varphi(t, X, Y, A) = f(t, Y, A) + \log(X) \tag{5.2}$$

in the case $\gamma = 1$. φ being the value function, we call f the *reduced value function.* Plugging (5.1) into (4.10), after symbolic simplifications (see [8] for a similar computation), we arrive at the Monge–Ampère-type PDE for $f(t, Y, A)$:

(5.3)

$$\begin{aligned}
\mathfrak{MA}[f] = {} & -p^2(\gamma-1)\sigma_y(t,Y,A).\sigma_s(t,Y,A)^T.\big(\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T\big)^{-1}.\sigma_s(t,Y,A).\sigma_y(t,Y,A) \\
& f^{(0,1,0)}(t,Y,A)^2 Y^3 - p^2(\gamma-1)\big(2r\gamma + (a_s(t,Y,A)-r).\big(\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T\big)^{-1}.(a_s(t,Y,A)-r)\big) \\
& f(t,Y,A)^2 Y + f(t,Y,A)\big(Y\big(2Y\big(A\gamma - (\gamma-1)\sigma_y(t,Y,A).\sigma_s(t,Y,A)^T \\
& .\big(\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T\big)^{-1}.(a_s(t,Y,A)-r)\big) f^{(0,1,0)}(t,Y,A) \\
& + \gamma(\sigma_y(t,Y,A).\sigma_y(t,Y,A) f^{(0,2,0)}(t,Y,A) Y^2 + 2f^{(1,0,0)}(t,Y,A))\big)p^2 \\
& + 8\pi^2(e-Y)\gamma f^{(0,0,1)}(t,Y,A)\big) = 0
\end{aligned}$$

in the semi-infinite domain $\Omega_{T,\infty} = \{\{t, Y, A\}, t < T, Y > 0\}$, together with the terminal condition

$$f(T, Y, A) = 1 \tag{5.4}$$

in the case $0 < \gamma \neq 1$, or in the case $\gamma = 1$

$$\begin{aligned}
& Y(a_s(t,Y,A)-r).\big(\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T\big)^{-1}.(a_s(t,Y,A)-r)p^2 \\
& + Y(\sigma_y(t,Y,A).\sigma_y(t,Y,A) f^{(0,2,0)}(t,Y,A) Y^2 + 2A f^{(0,1,0)}(t,Y,A) Y \\
& + 2(r + f^{(1,0,0)}(t,Y,A)))p^2 + 8\pi^2(e-Y) f^{(0,0,1)}(t,Y,A) = 0
\end{aligned} \tag{5.5}$$

in the same domain $\Omega_{T,\infty}$, together with the terminal condition

$$f(T, Y, A) = 0. \tag{5.6}$$

So, as in Merton's case, the variable X is eliminated. We refer to (5.3)–(5.4) as the *reduced Monge–Ampère PDE of optimal portfolio hedging.* Equations (5.5)–(5.6), on the other hand, are linear.

Due to the term $8\pi^2(e-Y)\gamma f^{(0,0,1)}(t,Y,A) f(t,Y,A)$, which changes sign, as before, at $Y = e$, we can describe (5.3) as the "hypoelliptic Monge–Ampère PDE." Equation (5.3) is also degenerate at $Y = 0$, but this is much less interesting, if not for

other reasons then because this degeneracy can and will be eliminated by truncation of the considered domain.

Also, plugging (5.1) as well as (5.2) into (4.5), we get the optimal portfolio hedging rule

(5.7)
$$\Pi^{\star}_{\gamma}(t,X,Y,A)=\frac{X}{\gamma}\Bigg((\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T)^{-1}.(a_s(t,Y,A)-r) + \frac{f_{\gamma}{}^{(0,1,0)}(t,Y,A)}{f_{\gamma}(t,Y,A)}Y\sigma_y(t,Y,A).\sigma_s(t,Y,A)^T.(\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T)^{-1}\Bigg)$$

in the case $\gamma\neq 1$, and

$$\Pi^{\star}_{1}(t,X,Y,A)=X(\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T)^{-1}.(a_s(t,Y,A)-r) \tag{5.8}$$

in the case $\gamma=1$. So, the case $\gamma=1$ (as always) is trivial: not only is the "reduced Monge–Ampère PDE" (5.5) linear but it actually does not need to be solved—(5.8) does not depend on f_1. On the other hand, if $\gamma\neq 1$, the optimal hedging strategy (5.7) is given in terms of the solution $f_\gamma=f$ of the reduced Monge–Ampère PDE (5.3). Nevertheless, we see that not only was one of the variables (X) eliminated, but also the optimal portfolio rule $\Pi^\star$ was computed in terms of f and its *first derivative* $f^{(0,1,0)}$, while before, if (4.10) was used, the optimal portfolio rule $\Pi^\star$ (see (4.5)) was computed using first derivative $\varphi^{(0,1,0,0)}$, and second derivatives $\varphi^{(0,2,0,0)}$ and $\varphi^{(0,1,1,0)}$, of the value function φ. Also, since $\varphi\sim X^{1-\gamma}$, the value function φ is extremely flat for large values of X when $\gamma\gg 1$. In practice, those are the values of interest for γ, since otherwise very aggressive trading strategies are inferred (see [8]). By studying the reduced value function f instead of the value function φ, one of the practical consequences is that we can compute the solution of the optimal portfolio hedging problem for large values of γ.

We further emphasize that the optimal portfolio rule $\Pi^\star(t,X,Y,A)$ is again proportional to the wealth:

$$\Pi^{\star}_{\gamma}(t,X,Y,A)=XP^{\star}_{\gamma}(t,Y,A), \tag{5.9}$$

where

(5.10)
$$P^{\star}_{\gamma}(t,Y,A)=\frac{1}{\gamma}\Bigg((\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T)^{-1}.(a_s(t,Y,A)-r) + \frac{f_{\gamma}{}^{(0,1,0)}(t,Y,A)}{f_{\gamma}(t,Y,A)}Y\sigma_y(t,Y,A).\sigma_s(t,Y,A)^T.(\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T)^{-1}\Bigg)$$

is the optimal investment per unit of wealth. It is interesting to compare the computed strategy $P^\star_\gamma(t,Y,A)$ with the Merton strategy (1.12). To that end write $P^\star_\gamma(t,Y,A)=M^\star_\gamma(t,Y,A)+S_\gamma(t,Y,A)$, where

$$M^{\star}_{\gamma}(t,Y,A)=\frac{(\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T)^{-1}.(a_s(t,Y,A)-r)}{\gamma} \tag{5.11}$$

is Merton's strategy (1.12) as applied directly in the present problem, while

(5.12)
$$S_{\gamma}(t,Y,A)=\frac{1}{\gamma}\frac{f_{\gamma}{}^{(0,1,0)}(t,Y,A)}{f_{\gamma}(t,Y,A)}Y\sigma_y(t,Y,A).\sigma_s(t,Y,A)^T.(\sigma_s(t,Y,A).\sigma_s(t,Y,A)^T)^{-1}$$

for $\gamma \neq 1$ and $S_1(t, Y, A) = 0$ is the correction provided by this methodology, assuming we can compute f_γ.

It turns out that (5.3)–(5.4) is (numerically) solvable in the case $\gamma > 1$. Of course, if numerical solution is attempted, then the semi-infinite domain $\Omega_{T,\infty}$ has to be truncated. So define

$$\Omega_T = \{\{t, Y, A\}, t_{\min} < t < T, y_{\min} < Y < y_{\max}, a_{\min} < A < a_{\max}\} \tag{5.13}$$

for some $0 < y_{\min} < e < y_{\max} < \infty$ and $-\infty < a_{\min} < 0 < a_{\max} < \infty$. The good understanding of the hypoelliptic operator $\mathcal{A}$ above is now quite important, since its properties are propagated into (5.3). We need to identify the $\left(\mathcal{A} + \frac{\partial}{\partial t}\right)$-hypoelliptic boundary of Ω_T—the set of all points on the boundary $\partial\Omega_T$ that *can* possibly be reached by the process (recall (3.1)) $\{t, Y(t), A(t)\}$ starting inside Ω_T. This is where the terminal/boundary condition is imposed. Since the time is running forward, the process obviously *cannot* exit at $t = t_{\min}$, but furthermore, more interestingly, due to the degeneracy, the process also *cannot* exit Ω_T at $\{\{t, Y, a_{\min}\}, t_{\min} < t < T, y_{\min} < Y < e\}$, just as well as it cannot exit at $\{\{t, Y, a_{\max}\}, t_{\min} < t < T, e < Y < y_{\max}\}$. Therefore, the $\left(\mathcal{A} + \frac{\partial}{\partial t}\right)$-hypoelliptic boundary of Ω_T, or $\partial_{\mathcal{A}+\frac{\partial}{\partial t}}\Omega_T$, which is the same $\partial_{\mathfrak{MA}}\Omega_T$, has two components: $\{t < T\} \cap \partial_{\mathfrak{MA}}\Omega_T$ (the complement of which was described above), where the lateral boundary condition is imposed, and $\{t = T\} \cap \partial_{\mathfrak{MA}}\Omega_T$, where the terminal condition $f = 1$ is imposed.

What would be a good (artificial) lateral boundary condition for f on $\{t < T\} \cap \partial_{\mathfrak{MA}}\Omega_T$? The simple choice is the Dirichlet condition $f = 1$. Nevertheless such a choice is shown to be a poor choice (see [8]). It turns out that a very useful property of the solution f is that it satisfies the inequality $0 < f < 1$ in $\Omega_{T,\infty}$. Consequently a nonlinear ("noninterfering") boundary condition

$$\frac{\partial f}{\partial \nu} = -\alpha f(1 - f) \tag{5.14}$$

on $\{t < T\} \cap \partial_{\mathfrak{MA}}\Omega_T$, where ν is the exterior unit normal, and where $\alpha > 0$ is a constant chosen experimentally in such a way to minimize the interference of the boundary condition with a solution (see [8]), is introduced. The condition (5.14) makes sense since, for example, the condition $0 < f < 1$ in $\Omega_{T,\infty}$ (not just Ω_T) implies that f is flat whenever taking values close to 0 or 1. The nonlinear boundary condition (5.14) was not studied in the context of Monge–Ampère PDEs as of yet (see, e.g., [2], [5]). Nevertheless the numerical calculations, some of which are showcased below, suggest very convincingly that the problem (5.3), (5.4), (5.14) is well posed.

Of course, by solving problem (5.3), (5.4), (5.14) in the truncated domain Ω_T, as opposed to solving problem (5.3), (5.4) in $\Omega_{T,\infty}$, we introduce a small error—it is an approximate solution to the original problem. In the rest of the paper we shall ignore the difference.

Thereby the extension of Merton's theory to momentum trading is complete: the optimal trading strategy is given by (5.7), where $f = f_\gamma$, the reduced value function, is computed as an appropriate solution of (5.3), (5.4), (5.14).

6. The simplest example: A single-stock portfolio. Consider the simplest case: a single stock whose price obeys (3.1) is considered for trading. More precisely, let it be in (3.1)–(3.2), $m = n = 1$,

$$\sigma_y(t, Y, A) = \{s\}, \sigma_s(t, Y, A) = \{\{s\}\}, a_s(t, Y, A) = \{A\}. \tag{6.1}$$

In other words, the stock-price dynamics is described by the system

$$\boldsymbol{d}Y(t) = Y(t)A(t)\boldsymbol{d}t + Y(t)\{s\}.\boldsymbol{d}B(t), \tag{6.2}$$

$$\boldsymbol{d}A(t) = \frac{1}{Y(t)}\left(\frac{2\pi}{p}\right)^2 (e - Y(t))\boldsymbol{d}t \tag{6.3}$$

and then again, as a tradable asset,

$$\boldsymbol{d}S(t) = S(t)\{A(t)\}\boldsymbol{d}t + S(t)\{\{s\}\}.\boldsymbol{d}B(t), \tag{6.4}$$

where $s > 0$ is constant volatility, $e > 0$ is constant price-equilibrium, and $p > 0$ is constant price-period. Also, $Y(t)$ and $A(t)$ are scalar-valued, while $S(t)$ and $B(t)$ are 1-vector-valued. $B(t) = \{B_1(t)\}$ is the one-dimensional "vector" Brownian motion. Of course, (6.2) and (6.4) are redundant; the purpose of having them both is to use the general framework above.

Then, from (5.7), the optimal portfolio rule is given by (we shall allow a bit of abuse of notation here: $P^\star_\gamma = \{P^\star_\gamma\}$, $M^\star_\gamma = \{M^\star_\gamma\}$, $S_\gamma = \{S_\gamma\}$)

$$\Pi^\star_\gamma(t, X, Y, A) = X\{P^\star_\gamma(t, Y, A)\} = \frac{X}{\gamma}\left\{\frac{A - r}{s^2} + \frac{Y f_\gamma{}^{(0,1,0)}(t, Y, A)}{f_\gamma(t, Y, A)}\right\} \tag{6.5}$$

for $\gamma \neq 1$, and $\Pi^\star_1(t, X, Y, A) = X\{P^\star_1(t, Y, A)\} = X\{\frac{A-r}{s^2}\}$, where the reduced value function $f = f_\gamma$ is the solution of the reduced Monge–Ampère PDE (5.3), which now simplifies to

$$\begin{aligned} &-p^2(\gamma - 1)s^2 f^{(0,1,0)}(t, Y, A)^2 Y^3 - p^2(\gamma - 1)\left(\frac{(A - r)^2}{s^2} + 2r\gamma\right) f(t, Y, A)^2 Y \\ &+ f(t, Y, A)(p^2 s^2 \gamma f^{(0,2,0)}(t, Y, A)Y^3 + 2p^2(A + r(\gamma - 1))f^{(0,1,0)}(t, Y, A)Y^2 + 2p^2\gamma \\ &\qquad f^{(1,0,0)}(t, Y, A)Y + 8\pi^2(e - Y)\gamma f^{(0,0,1)}(t, Y, A)) = 0 \end{aligned} \tag{6.6}$$

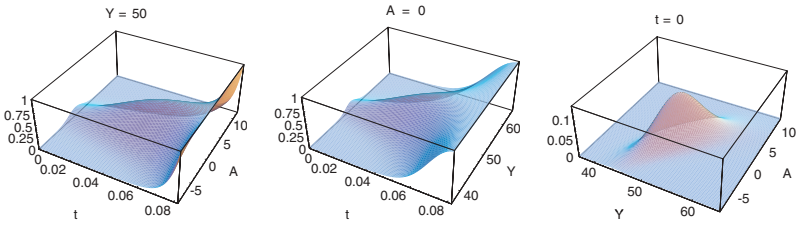
in Ω_T, together with boundary conditions (5.4) and (5.14). As before, we can write $P^\star_\gamma = M^\star_\gamma + S_\gamma$, where

$$M^\star_\gamma(t, Y, A) = \frac{A - r}{\gamma s^2} \tag{6.7}$$

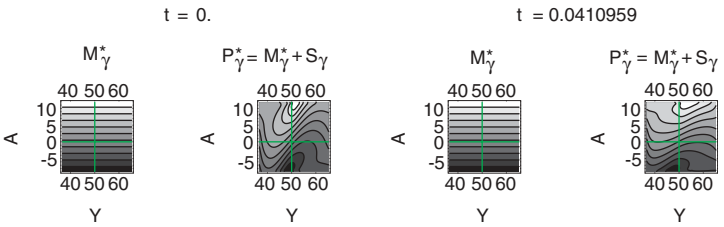
is the classical Merton's strategy, while

$$S_\gamma(t, Y, A) = \frac{Y f_\gamma{}^{(0,1,0)}(t, Y, A)}{\gamma f_\gamma(t, Y, A)} \tag{6.8}$$

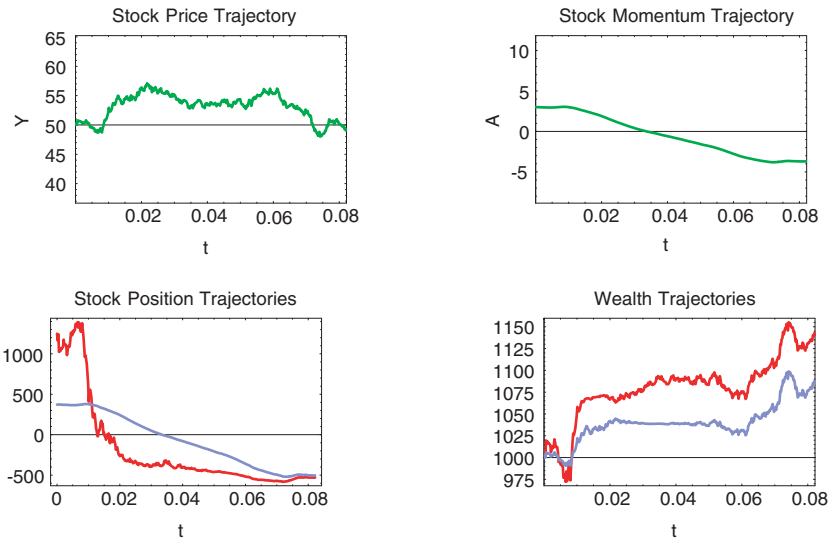
is the momentum correction provided by this methodology. Many numerical experiments were performed. Here are some results. For example, if $s = 0.4$, $e = 50$, $p = 2/12$, $r = 0.025$, $\gamma = 50$, $T = 30/365$, $y_{\min} = 36.6709$, $y_{\max} = 65.2239$, $a_{\min} = -8.85934$, $a_{\max} = 11.8334$, $\alpha = 0.05$, then the reduced value function $f = f_\gamma(t, Y, A)$, for fixed stock-price Y, for fixed momentum A, and for fixed time t, looks like



In particular, one can use the previous three figures to examine the validity of the artificial boundary condition (5.14); one can also try to examine the expected (hypoelliptic) C^∞-regularity of the solution. No PDE theory exists today to support such an expectation for equations such as (6.6). Next, we compare the strategy $M^\star_\gamma(t, Y, A)$ with the new optimal momentum strategy $P^\star_\gamma(t, Y, A)$ for a couple of fixed times (these are contour plots; dark means low relative value, and in this case short-selling, while light means a long position; see (6.7)):



We can conclude that the two strategies are quite a bit different long before the time of investment-success evaluation (T), and then, as time goes by, those two strategies converge. Finally, how much better is the new momentum strategy? We attempted to answer this question by means of Monte-Carlo experiments. For example, with initial condition $\{Y(0), A(0)\} = \{50, 3\}$, a very typical outcome looked like this:



where plotted are the wealth and investment-positions evolutions corresponding to strategy $M^\star_\gamma$ (blue-light), and corresponding to the strategy $P^\star_\gamma$ (red-dark); price and

momentum trajectories. One can notice that the investment position corresponding to $P^\star_\gamma$ is more aggressive in the beginning of the investment period and as time goes by converges to the one corresponding to the strategy $M^\star_\gamma$.

We postpone the more precise comparison based on extensive Monte-Carlo experiments until the end of the final section, when we compare these two strategies with three additional ones, involving also options.

7. Optimal hedging and pricing of European and American options in momentum markets. Now we consider a portfolio of an option, European or American, and the corresponding underlying stock. Let T_0 be the expiration time for the option, while $T \leq T_0$ is the time when, as before, we stop the trading and measure a level of success or failure. Let k be the option strike price, and let r be the interest rate. For simplicity, we assume no dividends are paid, although everything that precedes and follows can be modified for either continuous or instantaneous dividend-paying underlying. One tracks the stock-price $Y(t)$ and the price-momentum $A(t)$ while trading stocks and options. We assume that the underlying stock-price has volatility $s(t, Y, A)$, equilibrium e, and period p. More precisely, in (3.1), let $m = n = 2$, and let

$$\boldsymbol{d}Y(t) = Y(t)A(t)\boldsymbol{d}t + Y(t)\{s(t, Y(t), A(t)), 0\}.\boldsymbol{d}B(t), \tag{7.1}$$

$$\boldsymbol{d}A(t) = \frac{1}{Y(t)}\left(\frac{2\pi}{p}\right)^2 (e - Y(t))\boldsymbol{d}t, \tag{7.2}$$

identifying

$$\sigma_y(t, Y, A) = \{s(t, Y, A), 0\}. \tag{7.3}$$

Now, since we allow volatility $s = s(t, Y, A)$ to depend also on the momentum A, there is no reason to expect that the option price V does not depend on A. So, denote the option price as $V = V(t, Y, A)$. Under such a scenario, the tradable assets—the stock and the option—have a vector-price $S(t) = \{S_1(t), S_2(t)\} = \{Y(t), V(t, Y(t), A(t))\}$ with the dynamics

(7.4)

$$\begin{aligned}
\boldsymbol{d}S(t) = \boldsymbol{d}\{Y(t), V(t, Y(t), A(t))\} = \{Y(t), V(t, Y(t), A(t))\}\Bigg\{A(t), \frac{1}{V(t, Y(t), A(t))} \\
\left(\frac{1}{2}Y(t)^2 V^{(0,2,0)}(t, Y(t), A(t))s(t, Y(t), A(t))^2 + \frac{4\pi^2(e - Y(t))V^{(0,0,1)}(t, Y(t), A(t))}{p^2 Y(t)}\right. \\
\left. + A(t)Y(t)V^{(0,1,0)}(t, Y(t), A(t)) + V^{(1,0,0)}(t, Y(t), A(t))\right)\Bigg\}\boldsymbol{d}t + \{Y(t), V(t, Y(t), A(t))\} \\
\begin{pmatrix} s(t, Y(t), A(t)) & 0 \\ \dfrac{Y(t)s(t, Y(t), A(t))V^{(0,1,0)}(t, Y(t), A(t))}{V(t, Y(t), A(t))} & \epsilon \end{pmatrix}.\boldsymbol{d}B(t) = S(t)a_s(t, Y(t), A(t))\boldsymbol{d}t \\
+ S(t)\sigma_s(t, Y(t), A(t)).\boldsymbol{d}B(t)
\end{aligned}$$

for $\epsilon = 0$. This follows directly from Itô's chain rule. On the other hand, we shall assume that $\epsilon > 0$; i.e., we shall augment the option-price dynamics with an additional layer of randomness. This is motivated/justified by the well-known fact that, due to the reduced liquidity of options compared to the underlying stock, the relative difference between bid and ask prices for options is much more significant than for

the stocks. The assumption $\epsilon > 0$ is also of technical significance, and it will be further justified below by the beauty and by the consistency with the usual theory of the results obtained.

So, we identify the coefficients in (3.2) as

$$
\begin{aligned}
a_s(t,Y,A) = \Bigg\{ A, \frac{1}{V(t,Y,A)} \Bigg(& \frac{1}{2} Y^2 V^{(0,2,0)}(t,Y,A) s(t,Y,A)^2 \\
& + \frac{4\pi^2 (e-Y) V^{(0,0,1)}(t,Y,A)}{p^2 Y} + AY V^{(0,1,0)}(t,Y,A) + V^{(1,0,0)}(t,Y,A) \Bigg) \Bigg\}
\end{aligned} \tag{7.5}
$$

and

$$
\sigma_s(t,Y,A) = \begin{pmatrix} s(t,Y,A) & 0 \\ \dfrac{Y s(t,Y,A) V^{(0,1,0)}(t,Y,A)}{V(t,Y,A)} & \epsilon(t,Y,A) \end{pmatrix}. \tag{7.6}
$$

We are in a position now to apply the general theory, and in particular formulas (5.7) and (5.8), which now read as

(7.7)

$$
\begin{aligned}
\Pi^\star_\gamma(t,X,Y,A) = \{ & (X(2p^2 Y s(t,Y,A)^2 V(t,Y,A)^2 f^{(0,1,0)}(t,Y,A) \epsilon(t,Y,A)^2 \\
& + f(t,Y,A)(2p^2 r Y V(t,Y,A) V^{(0,1,0)}(t,Y,A) s(t,Y,A)^2 - V^{(0,1,0)}(t,Y,A) \\
& \quad (Y(Y^2 V^{(0,2,0)}(t,Y,A) s(t,Y,A)^2 + 2rY V^{(0,1,0)}(t,Y,A) \\
& + 2V^{(1,0,0)}(t,Y,A)) p^2 + 8\pi^2 (e-Y) V^{(0,0,1)}(t,Y,A)) s(t,Y,A)^2 \\
& + 2p^2 (A-r) V(t,Y,A)^2 \epsilon(t,Y,A)^2))) \\
& \quad / (2p^2 \gamma f(t,Y,A) s(t,Y,A)^2 V(t,Y,A)^2 \epsilon(t,Y,A)^2), \\
& \quad (X(-2rYV(t,Y,A)p^2 + Y(Y^2 V^{(0,2,0)}(t,Y,A) s(t,Y,A)^2 \\
& + 2rY V^{(0,1,0)}(t,Y,A) + 2V^{(1,0,0)}(t,Y,A)) p^2 \\
& + 8\pi^2 (e-Y) V^{(0,0,1)}(t,Y,A))) / (2p^2 Y \gamma V(t,Y,A) \epsilon(t,Y,A)^2) \}
\end{aligned}
$$

for $\gamma \neq 1$, where $f = f_\gamma$, the reduced value function, is the solution of the reduced Monge–Ampère PDE (5.3), which now becomes

(7.8)

$$
\begin{aligned}
f(t,Y,A) (& Y(2Y(A + r(\gamma-1)) f^{(0,1,0)}(t,Y,A) + \gamma(Y^2 f^{(0,2,0)}(t,Y,A) s(t,Y,A)^2 \\
& + 2f^{(1,0,0)}(t,Y,A))) p^2 + 8\pi^2 (e-Y) \gamma f^{(0,0,1)}(t,Y,A)) - \big((\gamma-1) \\
& \big(4p^4 Y^4 f^{(0,1,0)}(t,Y,A)^2 s(t,Y,A)^4 + \big(f(t,Y,A)^2 \big(p^4 Y^6 V^{(0,2,0)}(t,Y,A)^2 s(t,Y,A)^6 \\
& + 4p^2 Y^3 V^{(0,2,0)}(t,Y,A) (Y(rYV^{(0,1,0)}(t,Y,A) + V^{(1,0,0)}(t,Y,A)) p^2 \\
& - p^2 rYV(t,Y,A) + 4\pi^2 (e-Y) V^{(0,0,1)}(t,Y,A)) s(t,Y,A)^4 \\
& + 4\big(rY^2 V(t,Y,A)^2 (2\gamma \epsilon(t,Y,A)^2 + r) p^4 - 2rYV(t,Y,A) (Y(rY \\
& V^{(0,1,0)}(t,Y,A) + V^{(1,0,0)}(t,Y,A)) p^2 + 4\pi^2 (e-Y) V^{(0,0,1)}(t,Y,A)) p^2 \\
& + (Y(rYV^{(0,1,0)}(t,Y,A) + V^{(1,0,0)}(t,Y,A)) p^2 + 4\pi^2 (e-Y) V^{(0,0,1)}(t,Y,A))^2 \big) \\
& s(t,Y,A)^2 + 4p^4 (A-r)^2 Y^2 V(t,Y,A)^2 \epsilon(t,Y,A)^2 \big) \big) / (V(t,Y,A)^2 \epsilon(t,Y,A)^2) \big) \big) \\
& / (4p^2 Y s(t,Y,A)^2) = 0
\end{aligned}
$$

in Ω_T, together with boundary conditions (5.4) and (5.14), while in the case $\gamma = 1$,

the formula is complete—nothing is left to compute:

(7.9)
$$\begin{aligned}\Pi_1^\star(t,X,Y,A)=\{&(X(2p^2rYV(t,Y,A)V^{(0,1,0)}(t,Y,A)s(t,Y,A)^2-V^{(0,1,0)}(t,Y,A)\\&(Y(Y^2V^{(0,2,0)}(t,Y,A)s(t,Y,A)^2+2rYV^{(0,1,0)}(t,Y,A)\\&+2V^{(1,0,0)}(t,Y,A))p^2+8\pi^2(e-Y)V^{(0,0,1)}(t,Y,A))s(t,Y,A)^2\\&+2p^2(A-r)V(t,Y,A)^2\epsilon(t,Y,A)^2))/(2p^2s(t,Y,A)^2V(t,Y,A)^2\epsilon(t,Y,A)^2),\\&(X(-2rYV(t,Y,A)p^2+Y(Y^2V^{(0,2,0)}(t,Y,A)\\&s(t,Y,A)^2+2rYV^{(0,1,0)}(t,Y,A)+2V^{(1,0,0)}(t,Y,A))p^2\\&+8\pi^2(e-Y)V^{(0,0,1)}(t,Y,A)))/(2p^2YV(t,Y,A)\epsilon(t,Y,A)^2)\}.\end{aligned}$$

In various places in (7.7)–(7.9) a linear (hypoelliptic) differential operator of fundamental importance appears. Indeed, let operator $\mathcal{H}$ be defined as

(7.10)
$$\begin{aligned}\mathcal{H}[V](t,Y,A)&=V^{(1,0,0)}(t,Y,A)+\frac{1}{2}Y^2V^{(0,2,0)}(t,Y,A)s(t,Y,A)^2+rYV^{(0,1,0)}(t,Y,A)\\&-rV(t,Y,A)+\frac{4\pi^2(e-Y)V^{(0,0,1)}(t,Y,A)}{Yp^2}.\end{aligned}$$

The importance of $\mathcal{H}$ can be deduced immediately by its comparison with the classical Black–Scholes operator $\mathcal{L}$:

(7.11)
$$\mathcal{L}[V](t,Y)=V^{(1,0)}(t,Y)+\frac{1}{2}Y^2V^{(0,2)}(t,Y)s(t,Y)^2+rYV^{(0,1)}(t,Y)-rV(t,Y).$$

$\mathcal{H}$ is a momentum version of $\mathcal{L}$—if there is no momentum-dependence, $\mathcal{L}=\mathcal{H}$, since $V^{(0,0,1)}=0$. $\mathcal{H}$ is not backward-parabolic like $\mathcal{L}$ but only hypoelliptic, due to the term $4\pi^2(e-Y)V^{(0,0,1)}(t,Y,A)/(Yp^2)$ and due to the absence of $V^{(0,0,2)}$ in (7.10). Now using (7.10), after some work (7.7)–(7.9) simplify quite significantly: the optimal portfolio rule is equal to

(7.12)
$$\begin{aligned}\Pi_\gamma^\star(t,X,Y,A)&=X\left\{Q_\gamma^\star(t,Y,A),\frac{\mathcal{H}[V](t,Y,A)}{\gamma\epsilon(t,Y,A)^2V(t,Y,A)}\right\}=\frac{X}{\gamma}\left\{\frac{A-r}{s(t,Y,A)^2}\right.\\&\left.+\frac{Yf^{(0,1,0)}(t,Y,A)}{f(t,Y,A)}-\frac{\mathcal{H}[V](t,Y,A)}{\epsilon(t,Y,A)^2V(t,Y,A)}\frac{YV^{(0,1,0)}(t,Y,A)}{V(t,Y,A)},\frac{\mathcal{H}[V](t,Y,A)}{\epsilon(t,Y,A)^2V(t,Y,A)}\right\}\end{aligned}$$

for $\gamma\neq1$, where $f=f_\gamma$, the reduced value function, is the solution of the reduced Monge–Ampère PDE

(7.13)
$$\begin{aligned}&-p^2(\gamma-1)s(t,Y,A)^2f^{(0,1,0)}(t,Y,A)^2Y^3\\&-p^2(\gamma-1)\left(\frac{(A-r)^2}{s(t,Y,A)^2}+2r\gamma+\frac{(\mathcal{H}[V](t,Y,A))^2}{\epsilon(t,Y,A)^2V(t,Y,A)^2}\right)\\&\quad f(t,Y,A)^2Y+f(t,Y,A)(p^2s(t,Y,A)^2\gamma f^{(0,2,0)}(t,Y,A)Y^3+2p^2(A+r(\gamma-1))\\&\quad f^{(0,1,0)}(t,Y,A)Y^2+2p^2\gamma f^{(1,0,0)}(t,Y,A)Y+8\pi^2(e-Y)\gamma f^{(0,0,1)}(t,Y,A))=0\end{aligned}$$

in Ω_T, together with boundary conditions (5.4) and (5.14), while in the case $\gamma = 1$

(7.14)
$$\Pi_1^\star(t,X,Y,A) = X\left\{Q_1^\star(t,Y,A), \frac{\mathcal{H}[V](t,Y,A)}{\epsilon(t,Y,A)^2 V(t,Y,A)}\right\}$$
$$= X\left\{\frac{A-r}{s(t,Y,A)^2} - \frac{\mathcal{H}[V](t,Y,A)}{\epsilon(t,Y,A)^2 V(t,Y,A)}\frac{YV^{(0,1,0)}(t,Y,A)}{V(t,Y,A)}, \frac{\mathcal{H}[V](t,Y,A)}{\epsilon(t,Y,A)^2 V(t,Y,A)}\right\}.$$

Notice that in the above, for comparison purposes below, we have introduced the notation $Q_\gamma^\star(t,Y,A)$, as opposed to $P_\gamma^\star(t,Y,A)$ from the previous section.

In (7.12) and (7.14) the quantity $\mathcal{H}[V]/(\epsilon^2 V)$ has a very prominent role: it is, modulo X/γ, the optimal investment in the option. Therefore, one can think of it as an *option trading opportunity* or, more precisely, *option-underpricing* (if positive, it implies a long position in the option, as a consequence of the option being cheap, and vice versa). So, let us define the *option trading opportunity* ω as

$$\omega = \frac{\mathcal{L}[V]}{\epsilon^2 V} \tag{7.15}$$

if there is no momentum-dependence, or $\omega = \mathcal{H}[V]/(\epsilon^2 V)$ in momentum markets.

This definition now makes it very natural to define a *fair price of a European option* (cf. [4] and references given there) as a function V for which option-underpricing ω is equal to zero, i.e., for which the Black–Scholes PDE

$$\mathcal{L}[V] = 0 \tag{7.16}$$

(or $\mathcal{H}[V] = 0$, a hypoelliptic Black–Scholes PDE) holds.

Furthermore (let $\eta = \eta_k(Y)$ denote the option-payoff), a *fair price of an American option* in a way does not always exist; it is a function $V \geq \eta$, for which option-underpricing ω is either equal to zero (the fair price does exist; implying $L[V] = 0$), or it is less than zero (option-underpricing ω is negative, i.e., the option is overpriced and therefore it is not rational to trade in such an option—it is rational to exercise it; implying $L[V] \leq 0$), or put all together, for which

$$\operatorname{Max}[\mathcal{L}[V], \eta - V] = 0 \tag{7.17}$$

or $\operatorname{Max}[\mathcal{H}[V], \eta - V] = 0$ in momentum markets. Equation (7.17) is the Black–Scholes obstacle problem (see [8] for many other equivalent formulations of obstacle problems). In the rest of the paper, for simplicity, we shall consider European options only.

Consider the simplest possible example. Let all the data be the same as in the previous section. In particular, that means that the assumed underlying volatility is constant, and therefore $\mathcal{H} = \mathcal{L}$. Additionally, let the considered option have a strike price $k = 50$, and the expiration date $T_0 = 60/365$. Let also the above-discussed bid/ask indeterminacy be modeled by $\epsilon = 1/2$. We shall also need an option-pricing formula $V(t,Y,A) = V(t,Y)$. We choose

(7.18)
$$V(t,Y) = V_{\gamma,\kappa,\epsilon}(t,Y) = \frac{1}{2}\boldsymbol{e}^{\left(\gamma\kappa\epsilon^2+r\right)(t-T_0)}\left(\left(\operatorname{erfc}\left(\frac{2\log\left(\frac{Y}{k}\right)-(2r-s^2)(t-T_0)}{2\sqrt{2}\sqrt{s^2(T_0-t)}}\right)-2\right)k\right.$$
$$\left.+\left(\operatorname{erf}\left(\frac{2\log\left(\frac{Y}{k}\right)-(s^2+2r)(t-T_0)}{2\sqrt{2}\sqrt{s^2(T_0-t)}}\right)+1\right)\boldsymbol{e}^{r(T_0-t)}Y\right).$$

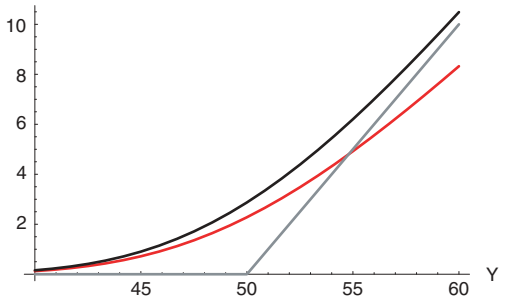
The pricing formula $V_{\gamma,\kappa,\epsilon}(t,Y)$ was computed as a solution of

$$\frac{\mathcal{L}[V](t,Y)}{\gamma\epsilon^2 V(t,Y)} = \kappa \tag{7.19}$$

with the usual Black–Scholes terminal condition

$$V(T_0,Y) = \eta_k(Y) = \mathrm{Max}[0, Y-k]. \tag{7.20}$$

Notice the difference between k, the strike price, and κ (Kappa) appearing in (7.19) and throughout below. Of course, for $\kappa = 0$ the solution $V_{\gamma,0,\epsilon}(t,Y)$ of (7.19)–(7.20) is the Black–Scholes fair option price, and according to the above, it would yield an optimal option investment of $\kappa = 0$. We choose $\kappa = 0.15$, which means we do a case study when an optimal option investment is *constant* and equal to 15% of the available wealth, and we are interested in seeing how such an optimal option investment, in such a badly priced option market, is hedged. It might be interesting to compare $V_{\gamma,0,\epsilon}$ and $V_{\gamma,\kappa,\epsilon}$, i.e., to see how badly the option is priced. For example, for $t = T/2 = 15/365$,



So, in order that under the risk-avoidance parameter $\gamma = 50$ it is optimal to invest 15% of the available wealth in the considered option, and under all of the other above conditions, the option needs to be quite a bit underpriced.

We present the results of numerical computations. The optimal trading strategy (7.12) is now equal to

$$\Pi^\star_\gamma(t,X,Y,A) = X\left\{Q^\star_\gamma(t,Y,A), \frac{\kappa}{\gamma}\right\} = \frac{X}{\gamma}\left\{\frac{A-r}{s^2} + \frac{Yf^{(0,1,0)}(t,Y,A)}{f(t,Y,A)} - \kappa\frac{YV^{(0,1)}(t,Y)}{V(t,Y)}, \kappa\right\}, \tag{7.21}$$

where the reduced value function f is the solution of (7.13), which now reduces to

$$\begin{aligned} &-p^2(\gamma-1)s^2 f^{(0,1,0)}(t,Y,A)^2 Y^3 - p^2(\gamma-1)\left(\frac{(A-r)^2}{s^2} + 2r\gamma + (\gamma\epsilon\kappa)^2\right) f(t,Y,A)^2 Y \\ &+ f(t,Y,A)(p^2s^2\gamma f^{(0,2,0)}(t,Y,A)Y^3 + 2p^2(A + r(\gamma-1))f^{(0,1,0)}(t,Y,A)Y^2 \\ &\qquad + 2p^2\gamma f^{(1,0,0)}(t,Y,A)Y + 8\pi^2(e-Y)\gamma f^{(0,0,1)}(t,Y,A)) = 0. \end{aligned} \tag{7.22}$$

For the purpose of comparison of various strategies later, let us introduce also the notation

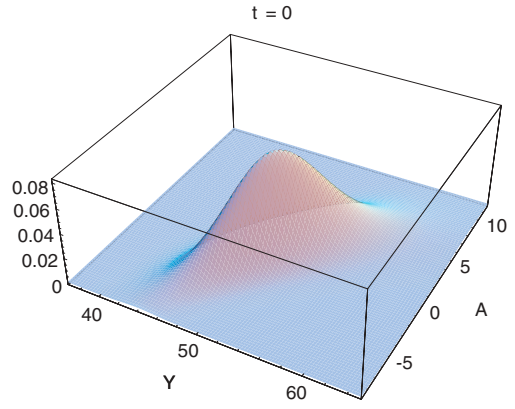
$$J^\star_\gamma(t,Y,A) = \frac{1}{\gamma}\left(\frac{A-r}{s^2} - \kappa\frac{YV^{(0,1)}(t,Y)}{V(t,Y)}\right) \tag{7.23}$$

and

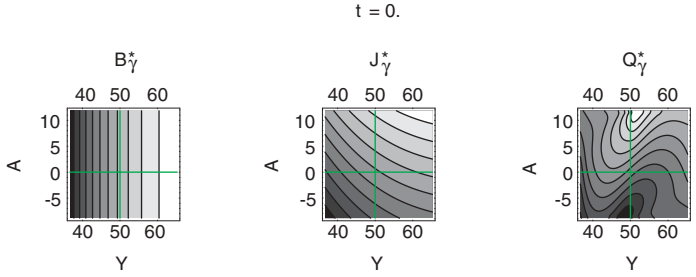
$$B^\star_\gamma(t,Y,A) = -\frac{\kappa}{\gamma}\frac{YV^{(0,1)}(t,Y)}{V(t,Y)}. \tag{7.24}$$

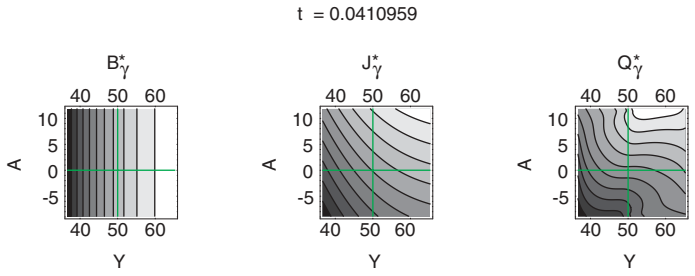
Notice that $B^\star_\gamma$ is the "Black–Scholes hedging strategy." Also notice, as it is going to be transparent on a sample trajectory below, that $B^\star_\gamma$ is not a riskless strategy, since the options pricing formula $V(t,Y)$ is not correct; i.e., it does not satisfy the Black–Scholes PDE (7.16). Therefore, strictly speaking, $B^\star_\gamma$ is *not* the Black–Scholes hedging strategy—it is just given by the same formula as the right one, when the correct options pricing formula is used. So, eventually we shall compare strategies $\{Q^\star_\gamma, \frac{\kappa}{\gamma}\}$, $\{P^\star_\gamma, 0\}$, $\{J^\star_\gamma, \frac{\kappa}{\gamma}\}$, $\{M^\star_\gamma, 0\}$, and $\{B^\star_\gamma, \frac{\kappa}{\gamma}\}$. Notice also that (7.22) and (6.6) are somewhat different when $\kappa\epsilon \neq 0$.

The solution of (7.22) looks like (for fixed time)

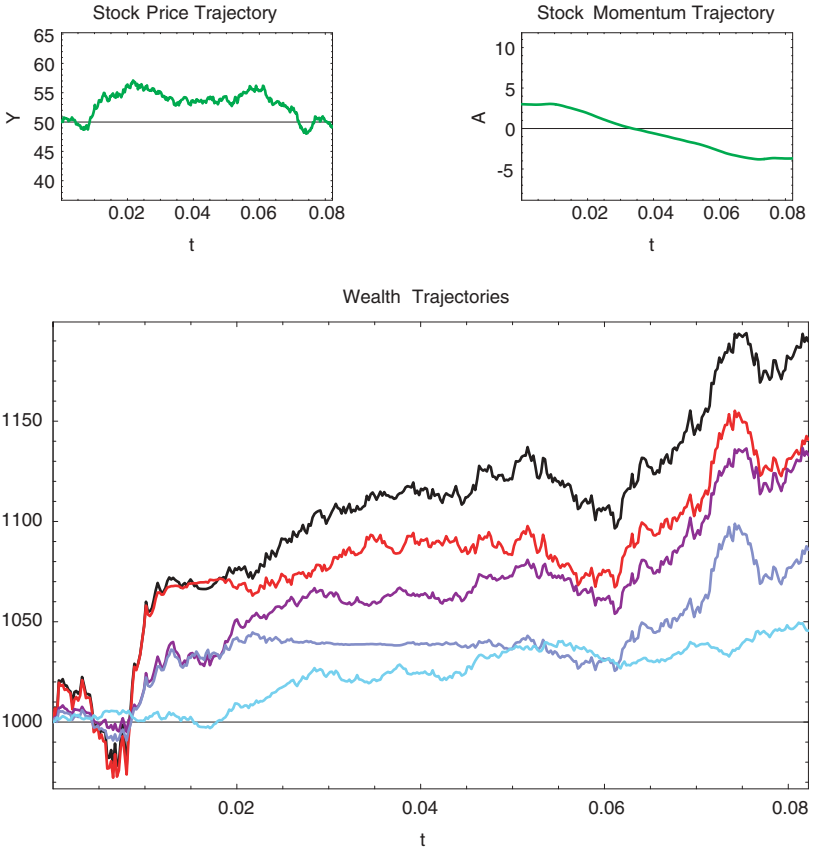


Compare this with the similar figure in the previous section: the present reduced value function is smaller than the one there, which implies that the current *value function* is bigger. Of course, the expected payoff is bigger now, since we have an additional opportunity—we have the opportunity to trade in options, as well. Next, we compare strategies $B^\star_\gamma(t,Y,A)$ and $J^\star_\gamma(t,Y,A)$, which do not depend on any numerical computations, and $Q^\star_\gamma(t,Y,A)$, which depends on the reduced value function above, for several fixed times (compare also with the similar plots in the previous section):





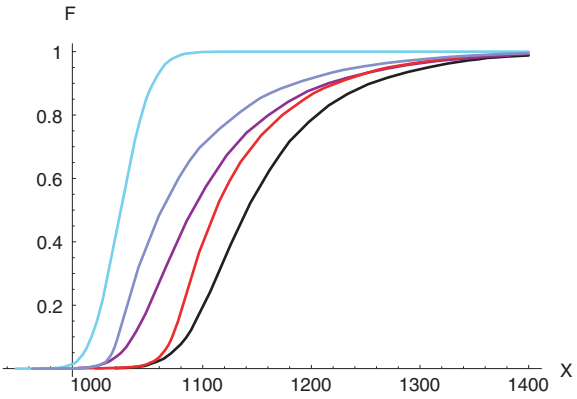
Finally, those three strategies $\{Q^\star_\gamma, \frac{\kappa}{\gamma}\}$, $\{J^\star_\gamma, \frac{\kappa}{\gamma}\}$, and $\{B^\star_\gamma, \frac{\kappa}{\gamma}\}$ were applied (in addition to two strategies already applied: $\{P^\star_\gamma, 0\}$ and $\{M^\star_\gamma, 0\}$) on some 10000 market trajectories (actually, 10452 trajectories were needed to select 10000 of them which stay inside Ω_T). A very typical one looks like (the same one is in the previous section)



where, looking on the right from the top, the wealth trajectories correspond to $\{Q^\star_\gamma, \frac{\kappa}{\gamma}\}$, $\{P^\star_\gamma, 0\}$, $\{J^\star_\gamma, \frac{\kappa}{\gamma}\}$, $\{M^\star_\gamma, 0\}$, and $\{B^\star_\gamma, \frac{\kappa}{\gamma}\}$. The average profit returns by these strategies were 15.767%, 13.176%, 11.138%, 8.605%, and 2.545%, respectively (a profit of 15.767% over 30 days yields a profit of 493.765% over a year).

Finally, as the most precise strategy-comparison gauge, the empirical cumulative distribution functions $F^{\{Q^\star_\gamma, \frac{\kappa}{\gamma}\}}(X)$, $F^{\{P^\star_\gamma, 0\}}(X)$, $F^{\{J^\star_\gamma, \frac{\kappa}{\gamma}\}}(X)$, $F^{\{M^\star_\gamma, 0\}}(X)$, $F^{\{B^\star_\gamma, \frac{\kappa}{\gamma}\}}(X)$, where $F(X) = \frac{1}{N}\sum_{i=1}^{N} \chi_{\{X \geq X_i(T)\}}$, and where χ_A is the characteristic function of an

event A, look like (recall that $N = 10000$, out of 10452; from the right to the left)



REFERENCES

[1] W. H. Fleming and S. J. Sheu, *Risk-sensitive control and an optimal investment model*, Math. Finance, 10 (2000), pp. 197–213.
[2] C. E. Gutiérrez, *The Monge–Ampère Equations*, Birkhäuser Boston, Boston, 2001.
[3] L. Hörmander, *Hypoelliptic second order differential equations*, Acta Math., 119 (1967), pp. 147–171.
[4] J. Kallsen, *Utility-based derivative pricing in incomplete markets*, in Mathematical Finance—Bachelier Congress 2000, H. Geman, D. Madan, S. R. Pliska, and T. Vorst, eds., Springer-Verlag, Berlin, 2002, pp. 313–338.
[5] N. V. Krylov, *Nonlinear Elliptic and Parabolic Equations of the Second Order*, D. Reidel, Kluwer, Dordrecht, The Netherlands, 1987.
[6] R. C. Merton, *Optimum consumption and portfolio rules in a continuous-time model*, J. Econom. Theory, 3 (1971), pp. 373–413.
[7] R. C. Merton, *Continuous-Time Finance*, Blackwell, Cambridge, UK, 1992.
[8] S. Stojanovic, *Computational Financial Mathematics Using Mathematica®: Optimal Trading in Stocks and Options*, Birkhäuser Boston, Boston, 2003.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1174–1195

EFFICIENT HEDGING WHEN ASSET PRICES FOLLOW A GEOMETRIC POISSON PROCESS WITH UNKNOWN INTENSITIES*

MICHAEL KIRCH† AND WOLFGANG J. RUNGGALDIER‡

Abstract. We consider the problem of determining a strategy that is efficient in the sense that it minimizes the expectation of a convex loss function of the hedging error for the case when prices change at discrete random points in time according to a geometric Poisson process. The intensities of the jump process need not be fully known by the investor. The solution algorithm is based on dynamic programming for piecewise deterministic control problems, and its implementation is discussed as well.

Key words. geometric Poisson process, piecewise deterministic control problems, incomplete information, incomplete markets, efficient hedging, Bayesian approach

AMS subject classifications. Primary, 91B28, 93E20; Secondary, 91B70, 90C39, 60G55

DOI. 10.1137/S0363012903423168

1. Introduction.

This paper concerns the problem of hedging a future liability. Depending on the hedging criterion, various approaches have been considered in the literature. A mathematically attractive criterion, related to mean-variance hedging, is the quadratic criterion. This criterion leads to the mathematical problem of approximating an L^2-random variable by stochastic integrals. Starting from the work by Föllmer and Sondermann [9] and Schweizer [22], much research has been dedicated to this criterion so that this topic can by now be considered as a well-studied one (for a survey with extensive literature see [24]). For more general criteria the approaches are in a sense in common with utility maximization. They are mainly the so-called *martingale approach* (see, e.g., [3], [16], [13], [21]) and approaches based on *optimal stochastic control*.

One of the main goals in the present paper is to study the hedging problem in the context of incomplete/partial information on the underlying price evolution model. The various approaches mentioned above have, to some extent, also been applied to the case of incomplete information. For the quadratic/mean variance approach and under a martingale measure see, e.g., [23] and [10]. The martingale approach for the case of incomplete information was first considered in [18]. For more recent studies see, e.g., [4], [14], [26]; they concern mainly diffusion-type models where the uncertainty is in the stock appreciation rates that are supposed to be unknown constants and treated, from the Bayesian point of view, as random variables with a given prior distribution. For a more typically stochastic control approach see [20]. In the present paper we concentrate on the stochastic control approach that can indeed be viewed as a rather general approach for problems with partial information.

Stochastic control methods, in particular the method of dynamic programming (DP), have been applied mainly to diffusion-type models, and here DP leads to HJB-

*Received by the editors February 24, 2003; accepted for publication (in revised form) March 2, 2004; published electronically December 1, 2004.

http://www.siam.org/journals/sicon/43-4/42316.html

†Department of Financial and Actuarial Mathematics, 8-10 Wiedner Hauptstrasse, 1040 Vienna, Austria (michael.kirch@gmx.net). Current address: Goldman Sachs, Peterborough Court, 133 Fleet Street, London EC4A 2BB (michael.kirch@gs.com). The research of this author was partially supported by the EU RTM-project Dynstoch (IMP, Fifth Framework Programme).

‡Dipartimento di Matematica Pura ed Applicata, Universitá di Padova, 7 Via Belzoni, 35131 - Padova, Italy (runggal@math.unipd.it).

type equations; much current research is going on concerning analytical solutions to these equations. Diffusion-type models lead to continuous trajectories for the prices. In reality, prices change at discrete random points in time, and, in addition, they move at fixed increments (multiples of tick size). In this paper we consider a price evolution model that possesses these features, namely a geometric Poisson process (see also [15]). In order to concentrate better on the main issues, we consider a rather simple such model with a single risky asset and assume that the liability to be hedged is adapted to the filtration generated by the underlying price evolution. The approach can, however, be extended to markets with many assets, not all of them available for hedging, and with a liability that may be adapted to the filtration generated by all these assets. Furthermore, the simple geometric Poisson process can be extended to include compound Poisson processes. This simple model is also a natural generalization of the successful binomial (or Cox–Ross–Rubinstein) model, in which prices change at fixed proportions either up or down. While in the binomial model the changes occur at fixed points in time; here they occur more realistically at random points in time. In any case, an important test for a model is to check whether it is able to reproduce option prices observed in the market. The simple model (1) passes this test surprisingly well: it is able to reproduce both the smile observed in a foreign exchange market and the skew observed in equity markets.

A geometric Poisson process is driven by random jump processes that are characterized by their intensities. We assume the intensities to be constant in time, but, since one of our purposes is to highlight the problem of model uncertainty, we allow for the possibility that these intensities are not fully known to the investor. Taking the Bayesian point of view, they are considered as random variables with a distribution that is continuously updated on the basis of the information coming from observing the actual price evolution. This allows us to capture some of the most important features while keeping complexity low.

In the setting as described above, the market is incomplete, and so it is not possible to obtain for any given claim a self-financing and perfect hedging strategy. As a hedging criterion we consider here the expected value of a convex loss function applied to the hedging error and call a hedging strategy that minimizes this criterion *efficient (or optimal)*. Notice that this includes the quadratic criterion and also the well-known *shortfall risk criterion* (see, e.g., [2], [8]).

For our geometric Poisson models, the DP approach of stochastic control leads to Bellman equations for piecewise deterministic processes (see, e.g., [5], [25], [6], [1]) that are studied here both under full as well as partial information on the underlying price evolution model.

In summary, we consider the following optimization problem where, letting all the prices be discounted with respect to the nonrisky asset, X_t denotes the (for simplicity scalar) price process, V_t is the value process corresponding to a self-financing and predictable investment strategy ξ_t, and $F(X_S)$ is the claim for a fixed maturity S. The processes N_t^+ and N_t^- are jump processes where the intensity need not be fully known, a, b are given positive constants, and $l(\cdot)$ is an increasing and convex loss function:

$$\begin{cases} dX_t = X_{t-}\left[(e^a-1)dN_t^+ + (e^{-b}-1)dN_t^-\right], \\ dV_t = \xi_t X_{t-}\left[(e^a-1)dN_t^+ + (e^{-b}-1)dN_t^-\right], \\ E\left\{l\left(F(X_S)-V_S\right)\right\} \quad \longrightarrow \quad \min. \end{cases}$$

The outline of the paper is as follows. In section 2 we describe more precisely the problem setup and obtain some preliminary results. The DP approach for piecewise deterministic processes is studied, in the context of our problem, in section 3 and is extended to the case of incomplete information in section 4, more precisely in subsection 4.2. Since, in the context of our model, the incomplete information concerns uncertainty about the intensities of the driving Poisson processes, in subsection 4.1 we recall some facts about the Bayesian approach to uncertain intensities. The computation of real-size problems needs some approximations to be introduced. These are discussed in section 5, where also an example is given to better illustrate the approximation procedure itself.

2. Problem setup. We examine efficient strategies in the situation where, considering for simplicity only a single risky asset, its price follows a *geometric Poisson process*, i.e.,

$$X_t = x_0 e^{aN_t^+ - bN_t^-} \tag{1}$$

for two independent Poisson processes N^+, N^- with intensities λ^+, λ^- defined on a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ and constants $a, b > 0$. To keep the presentation as simple as possible, we shall assume that all processes/values are already discounted; i.e. we implicitly assume the short rate r_t to be equal to zero. The more realistic case of $r_t > 0$ would not change the essence of the results but would considerably complicate the presentation.

For a hedging strategy ξ, where ξ_t denotes the number of units of the risky asset held in the portfolio at time t, and an initial capital V_0, we define the associated wealth process by

$$V_t = V_0 + \int_0^t \xi_s dX_s, \tag{2}$$

thereby enforcing ξ to be self-financing. Although we shall not pursue this here, transaction costs may be easily incorporated by subtracting from the right-hand side in (2) the total amount of transaction costs incurred up to time t. In order to keep the results below as general as possible, we shall consider ξ_t to be real valued. We say that a strategy ξ is admissible for initial capital V_0 if it is predictable and the associated wealth process satisfies

$$V_t \geq -c, \quad t \in [0, S], \quad P\text{-a.s.} \tag{3}$$

for some fixed $c \geq 0$ and a given time horizon S. Let $\mathcal{A}_{V_0}$ denote the class of all admissible strategies for initial capital $V_0 \geq -c$.

We denote by τ_n the time of the nth jump

$$\begin{aligned} \tau_n &= \inf\{t \geq 0 \mid N_t^+ + N_t^- = n\}, \\ \hat{\tau}_n &:= \tau_n \wedge S. \end{aligned}$$

PROPOSITION 2.1. *We have $\xi \in \mathcal{A}_{V_0}$ if and only if ξ is predictable and satisfies*

$$\xi_t \in \left[-\frac{c + V_{\hat{\tau}_n}}{X_{\hat{\tau}_n}(e^a - 1)}, \frac{c + V_{\hat{\tau}_n}}{X_{\hat{\tau}_n}(1 - e^{-b})} \right], \quad t \in (\hat{\tau}_n, \hat{\tau}_{n+1}], \quad n = 0, 1, \ldots, \quad P\text{-a.s.} \tag{4}$$

Proof. First, observe that

$$(0, S] = \bigcup_{n=0,1,\ldots} (\hat{\tau}_n, \hat{\tau}_{n+1}], \quad P\text{-a.s.}$$

holds; i.e., condition (3) is satisfied if and only if it is satisfied on every interval $(\hat{\tau}_n, \hat{\tau}_{n+1}]$.

Since X is a pure jump process, we obtain from (2)

$$\begin{aligned} V_t &= V_{\hat{\tau}_n} + \xi_t(X_{\hat{\tau}_{n+1}\wedge t} - X_{\hat{\tau}_n}), \quad t \in (\hat{\tau}_n, \hat{\tau}_{n+1}] \\ &= \begin{cases} V_{\hat{\tau}_n} & \text{for } t \in (\hat{\tau}_n, \hat{\tau}_{n+1}), \\ V_{\hat{\tau}_n} + \xi_t X_{\hat{\tau}_n}(e^a - 1) & \text{for } t = \hat{\tau}_{n+1},\ N_t^+ - N_{t-}^+ = 1, \\ V_{\hat{\tau}_n} + \xi_t X_{\hat{\tau}_n}(e^{-b} - 1) & \text{for } t = \hat{\tau}_{n+1},\ N_t^- - N_{t-}^- = 1. \end{cases} \end{aligned}$$

Hence condition (3) is satisfied if and only if (4) holds. □

Next, we examine *efficient hedging strategies* for a given European claim with maturity S and payoff $F(X_S) \geq -c$, assuming $F(\cdot)$ is a continuous function.

Let $\mathcal{M}$ denote the family of all equivalent martingale measures for X. The superhedge price for the option F at time $t < S$ admits the representation

$$F_{u,d} := \sup_{Q \in \mathcal{M}} E_Q[F(X_S) | N_t^+ = u, N_t^- = d]; \tag{5}$$

see [7] and [17]. It does not depend on t since, given t, we can find a measure P' equivalent to P such that N^i has constant intensity $\lambda^i \frac{S}{S-t}$, $i = +, -$. The name *superhedge price* is derived from the fact that, if the capital/wealth available to an investor at time t is $V_t \geq F_{u,d}$, then there exists a self-financing strategy (see (2)) such that the capital/wealth at maturity satisfies $V_S \geq F(X_S)$ a.s. Clearly, the model specified by (1) is incomplete. Especially, it can be shown that the superhedge price for a European call option with payoff $F(X_S) = (X_S - K)^+$ is given by x_0 for any maturity S and strike K (see, e.g., [11]). This price allows for arbitrage for the seller of the option. This example illustrates that superhedging may not be appropriate in model (1).

A feasible efficient strategy depends on the investors attitude towards risk. For our purposes, this attitude is incorporated in the choice of a loss function l such that l is increasing, convex, and $l(z) = 0$ for $z \leq 0$. A typical choice is $l(z) = z^p$ for $z \geq 0$. Here the parameter $p \geq 1$ corresponds directly to the investor's degree of risk aversion. We shall assume that

$$E[l(F(X_S) + c)] < \infty \tag{6}$$

holds.

As introduced in [8], an efficient hedging strategy ξ^* is a solution to the optimization problem

$$J_0^* := \min_{\xi \in \mathcal{A}_{V_0}} E\left[l\left(F(X_S) - V_0 - \int_0^S \xi_s dX_s \right) \right]. \tag{7}$$

The main reason for this criterion is that, since the market is incomplete, no perfect replication is possible. On the other hand, superreplication is neither appropriate nor economical, and so one tolerates some risk that one wants to minimize while taking into account the investor's attitude towards risk. The minimal risk J^* is called the value function (or optimal cost-to-go function). More generally, the value function at time t is given by

$$J^*(v,u,d,t) = \min_{\xi \in \mathcal{A}_{v,u,d,t}} E\left[l\left(F(X_S) - v - \int_t^S \xi_s dX_s \right) | \, N_t^+ = u,\ N_t^- = d \right], \tag{8}$$

where $\mathcal{A}_{v,u,d,t}$ denotes the class of admissible strategies on the interval $[t, S]$ given $V_t = v$, $N_t^+ = u$, and $N_t^- = d$. The value $J^*(v, u, d, t)$ represents the optimal minimal value at time t over the remaining period when the current information corresponds to $V_t = v$, $N_t^+ = u$, and $N_t^- = d$. In line with the comment after (5) notice also that, if at any time t one has $v \geq F_{u,d}$ with $F_{u,d}$ as in (5), then $J^* \equiv 0$. The optimization problem as such is therefore meaningful only if $v < F_{u,d}$. In what follows we shall consider for v the entire closed interval $[-c, F_{u,d}]$ since we want to obtain the hedging strategy also for $v = F_{u,d}$. In this situation, and up to the first jump after t, the interval in (4) is given by

$$I_{v,u,d} := \left[-\frac{c+v}{x_0 e^{au-bd}(e^a - 1)}, \frac{c+v}{x_0 e^{au-bd}(1 - e^{-b})} \right]. \tag{9}$$

We gather two facts about the structure of the model in the next lemma.

LEMMA 2.2.

(i) *The number $N_t := N_t^+ + N_t^-$ of jumps up to time t is a Poisson process with intensity $\lambda = \lambda^+ + \lambda^-$.*

(ii) *For any nonnegative function f we have*

$$\begin{aligned} &E[f(N^+_{\hat{\tau}_{u+d+1}}, N^-_{\hat{\tau}_{u+d+1}}, \hat{\tau}_{u+d+1}) |\ N_t^+ = u,\ N_t^- = d] \\ &\quad = \int_0^{S-t} \Big\{ \lambda^+ f(u+1, d, t+s) + \lambda^- f(u, d+1, t+s) \Big\} e^{-\lambda s} ds \\ &\qquad + e^{-\lambda(S-t)} f(u, d, S). \end{aligned} \tag{10}$$

Remark 1. The expression in (10) represents the expected value of f, taking into account that there may be a next jump either upwards or downwards or no further jump at all and given that at the current time t one has observed u jumps upwards and d downwards.

Proof. Only item (ii) requires a proof. To simplify notation, we assume, without loss of generality, $u = d = t = 0$. Let τ_1^+ (respectively, τ_1^-) denote the time of the first jump up (respectively, down). We then have

$$\begin{aligned} E[f(N^+_{\hat{\tau}_1}, N^-_{\hat{\tau}_1}, \hat{\tau}_1), \tau_1 \leq S] &= E[f(1,0,\tau_1^+),\ \tau_1^+ < \tau_1^-,\ \tau_1^+ \leq S] \\ &\quad + E[f(0,1,\tau_1^-),\ \tau_1^- < \tau_1^+,\ \tau_1^- \leq S] \\ &= E[f(1,0,\tau_1^+)\ P[\tau_1^+ < \tau_1^-,\ \tau_1^+ \leq S |\, \tau_1^+]\,] \\ &\quad + E[f(0,1,\tau_1^-)\ P[\tau_1^- < \tau_1^+,\ \tau_1^- \leq S |\, \tau_1^-]\,] \\ &= \int_0^S f(1,0,t)\ P[t < \tau_1^-] P[\tau_1^+ \in dt] \\ &\quad + \int_0^S f(0,1,t)\ P[t < \tau_1^+] P[\tau_1^- \in dt] \\ &= \int_0^S f(1,0,t)\ e^{-\lambda^- t} \lambda^+ e^{-\lambda^+ t} dt \\ &\quad + \int_0^S f(0,1,t)\ e^{-\lambda^+ t} \lambda^- e^{-\lambda^- t} dt. \end{aligned}$$

Adding to this expression the term

$$E[f(N^+_{\hat{\tau}_1}, N^-_{\hat{\tau}_1}, \hat{\tau}_1), \tau_1 > S] = e^{-\lambda S} f(0,0,S),$$

we arrive at (10). ∎

3. Known intensities. We first examine the structure of the efficient strategy in the case where the investor has no doubt regarding the true values of the jump intensities λ^+, λ^-.

3.1. The PD-DP equation. In our situation, the state space is given by

$$E = \{(v,u,d,t) \mid v \geq -c,\ u,d \in \mathbb{N},\ t \in [0,S]\}.$$

Let $\mathcal{C}(E)$ denote the class of all functionals J on E such that $(v,t) \mapsto J(v,u,d,t)$ is continuous for all $u,d \in \mathbb{N}$. We endow $\mathcal{C}(E)$ with the supremum norm

$$||J|| = \sup_{(v,u,d,t)\in E} |J(v,u,d,t)|.$$

The process X is piecewise deterministic. Hence problem (8) is a piecewise deterministic control problem. We only remark that, although the price process X is piecewise constant, the optimal control will, in general, not be piecewise constant. This is due to the fact that if X remains constant from time t to time $t+s$, the time horizon changes, and thus the optimal strategy needs to be adapted.

For $J : E \to \mathbb{R}^+$, we define the operator T mapping J to $TJ :\ E \to \mathbb{R}^+$ by

$$\begin{aligned}(11)\quad (TJ)(v,u,d,t) &= \int_0^{S-t} e^{-\lambda s} \cdot \min_{\zeta \in I_{v,u,d}} \left\{ \begin{matrix} \lambda^+ J(v+\zeta x_0 e^{au-bd}(e^a - 1), u+1, d, t+s) \\ +\lambda^- J(v+\zeta x_0 e^{au-bd}(e^{-b}-1), u, d+1, t+s) \end{matrix} \right\} ds \\ &\quad + e^{-\lambda(S-t)} l(F(x_0 e^{au-bd}) - v),\end{aligned}$$

where $I_{v,u,d}$ is as in (9). In (11) one takes the min over the present control actions of the expectation of the optimal cost-to-go function at the next jump time, where one of the two possibilities may occur: a jump upwards for which the value process changes to $v+\zeta x_0 e^{au-bd}(e^a - 1)$ or a jump downwards (see also the proof of Lemma 3.2 and the explicit expressions (20) and (21) of J^1 and J^2 after its proof). From the proof of the next lemma it follows that the integral in (11) is well defined: indeed, the integrand is given by the continuous function $\hat{g}$ defined in (12).

LEMMA 3.1. *The operator $T : \mathcal{C}(E) \to \mathcal{C}(E)$ is a contraction with contraction constant $1 - e^{-\lambda S}$.*

Proof. 1. We first demonstrate that, for $J \in \mathcal{C}(E)$, the integral in (11) is well defined and $TJ \in \mathcal{C}(E)$. Let

$$\begin{aligned} g(\zeta,v,s) = \lambda^+ J(v+\zeta x_0 e^{au-bd}(e^a - 1), u+1, d, t+s) \\ +\lambda^- J(v+\zeta x_0 e^{au-bd}(e^{-b}-1), u, d+1, t+s)\end{aligned}$$

denote the function inside the curly brackets in (11). This function is continuous on

$$\left[-\frac{c+F_{u,d}}{x_0 e^{au-bd}(e^a-1)}, \frac{c+F_{u,d}}{x_0 e^{au-bd}(1-e^{-b})}\right] \times [-c, F_{u,d}] \times [0,T];$$

hence it is also bounded on this domain. The multifunction $v \mapsto I_{v,u,d}$ is continuous. Applying Proposition D.3 (c) of [12], we obtain that the function

$$(12)\qquad \hat{g}(v,s) = \min_{z \in I_{v,u,d}} g(z,v,s)$$

is continuous.

2. The proof that T is a contraction is inspired by Theorem 3.2 of [1]. For $\xi : [0, S-t] \to I_{v,u,d}$, let

$$\begin{aligned}(T_\xi J)(v,u,d,t) = e^{-\lambda(S-t)} l(F(x_0 e^{au-bd}) - v) \\ + \int_0^{S-t} e^{-\lambda s} \Big\{ \lambda^+ J(v + \xi(s) x_0 e^{au-bd}(e^a - 1), u+1, d, t+s) \\ + \lambda^- J(v + \xi(s) x_0 e^{au-bd}(e^{-b} - 1), u, d+1, t+s) \Big\} ds.\end{aligned}$$

We then have

$$||T_\xi J - T_\xi J'|| \leq (1 - e^{-\lambda S}) ||J - J'||.$$

Due to the continuity of J, there exists $\xi : [0, S-t] \to I_{v,u,d}$ such that

$$T_\xi J = TJ.$$

Hence we obtain

$$\begin{aligned}TJ' - TJ &\leq T_\xi J' - T_\xi J \\ &\leq (1 - e^{-\lambda S}) ||J' - J||.\end{aligned}$$

By symmetry we can conclude that

$$|TJ' - TJ| \leq (1 - e^{-\lambda S}) ||J' - J||. \qquad \square$$

Given $n \in \mathbb{N}$, let

$$J^0 = 0, \quad \text{and, for } h \leq n, \quad J^h = TJ^{h-1}, \tag{13}$$

and let $(\xi^n_s)_{s \in [0,S]}$ be the strategy induced by computing $J^n(V_0, 0, 0, 0)$ via (13) and (11). More precisely, this strategy is defined as follows. By induction over $h \leq n-2$ we define ξ^n on each interval $(\hat\tau_h, \hat\tau_{h+1}]$: For $s \in [0, \hat\tau_1]$ let $v = V_0$ and

$$\xi^n_s := \arg \min_{\zeta \in I_{v,0,0}} \Big\{ \lambda^+ J^{n-1}(v + \zeta x_0 (e^a - 1), 1, 0, s) + \lambda^- J^{n-1}(v + \zeta x_0 (e^{-b} - 1), 0, 1, s) \Big\}. \tag{14}$$

It was shown in [8] that the mapping $v \mapsto J^n(v, u, d, t)$ is decreasing and strictly convex on $(-c, F_{u,d})$. Hence the minimum in (14) and (15) below is assumed at a unique point ζ.

Suppose we defined the strategy ξ^n_s for all $s \leq \hat\tau_h$. At time $t = \hat\tau_h$, we observed $u = N^+_t$ jumps up and $d = N^-_t$ jumps down with $u + d = h$, and we have some capital $v = V_{\hat\tau_h}$ available. For $s \in (\hat\tau_h, \hat\tau_{h+1}]$, we define

$$\begin{aligned}\xi^n_s := \arg \min_{\zeta \in I_{v,u,d}} \Big\{ \lambda^+ J^{n-u-d-1}(v + \zeta x_0 e^{au-bd}(e^a - 1), u+1, d, s) \\ + \lambda^- J^{n-u-d-1}(v + \zeta x_0 e^{au-bd}(e^{-b} - 1), u, d+1, s) \Big\}.\end{aligned} \tag{15}$$

We have thus defined the strategy ξ^n_t for $t \in [0, \hat\tau_{n-1}]$. On the event $\{\tau_{n-1} > S\}$, this is sufficient. For the general case, we set $\xi^n_t = 0$ for $t \in (\hat\tau_{n-1}, S]$; i.e., we transfer all funds to the cash account after the $(n-1)$st jump happened prior to S.

Clearly, the strategy ξ^n thus defined satisfies $\xi^n \in \mathcal{A}_{V_0}$. The associated wealth process

$$V_t^n := V_0 + \int_0^t \xi_s^n dX_s \tag{16}$$

is constant from time $\hat{\tau}_{n-1}$ on.

By analogy to (8), we define a new value function

$$J^{*,n}(v,u,d,t) = \min_{\xi \in \mathcal{A}_{v,u,d,t}} E[l(F(X_S) - V_S^\xi),\ \tau_{u+d+n} > S |\, N_t^+ = u,\ N_t^- = d], \tag{17}$$

where

$$V_S^\xi = v + \int_t^S \xi_s dX_s.$$

The difference is that for $J^{*,n}(v,u,d,t)$ we take the expectation on the event that less than n jumps occur on the interval $(t,S]$.

We have now defined three functions J^*, J^n, and $J^{*,n}$ via (8), (13), and (17). Of these, J^* is the value function associated with problem (7). The function J^n is defined as the nth iteration of the dynamic programming operator associated with problem (7). Hence, a priori, J^n is not a value function. However, $J^{*,n}$ is the value function for the problem obtained from (7) by replacing Ω by $\Omega \cap \{\tau_n > S\}$. In the next lemma, we demonstrate that J^n and $J^{*,n}$ coincide. Consequently, we obtain for $t = 0$ the equality

$$J_0^n := J^n(V_0,0,0,0) = E[l(F(X_S) - V_S^n),\ \tau_n > S] = \min_{\xi \in \mathcal{A}_{V_0}} E[l(F(X_S) - V_S^\xi),\ \tau_n > S] \tag{18}$$

with V^n defined as in (16). This is remarkable in that it gives a new interpretation for J^n as a value function to problem (17) which is related to the original problem in a simple intuitive way.

LEMMA 3.2. *We have* $J^n = J^{*,n}$.

Proof. Essentially, the assertion is an application of the DP principle, Proposition 2.1, and (10).

For $n = 0, 1$, the assertion is immediate, see also (20).

Assume $J^{*,n-1} = J^{n-1}$ holds for $n - 1 \in \mathbb{N}$. We now demonstrate the assertion for n. Conditioning the right-hand side of (17) on $\hat{\tau}_{u+d+1}$, we obtain from the DP principle that

$$\begin{aligned} &J^{*,n}(v,u,d,t) \\ &\quad = \min_{\xi \in \mathcal{A}_{v,u,d,t}} E[J^{*,n-1}(V_{\hat{\tau}_{u+d+1}}, N^+_{\hat{\tau}_{u+d+1}}, N^-_{\hat{\tau}_{u+d+1}}, \hat{\tau}_{u+d+1}) |\, N_t^+ = u,\ N_t^- = d] \\ &\quad = \min_{\xi \in \mathcal{A}_{v,u,d,t}} E[J^{n-1}(V_{\hat{\tau}_{u+d+1}}, N^+_{\hat{\tau}_{u+d+1}}, N^-_{\hat{\tau}_{u+d+1}}, \hat{\tau}_{u+d+1}) |\, N_t^+ = u,\ N_t^- = d]. \end{aligned}$$

Due to (10) and Proposition 2.1, the right-hand side of the previous equation evaluates to

$$\begin{aligned} \min_{\xi \in \mathcal{A}_{v,u,d,t}} \int_0^{S-t} e^{-\lambda s} &\left\{ \begin{array}{c} \lambda^+ J^{n-1}(v + \xi_s x_0 e^{au-bd}(e^a - 1), u+1, d, t+s) \\ + \lambda^- J^{n-1}(v + \xi_s x_0 e^{au-bd}(e^{-b} - 1), u, d+1, t+s) \end{array} \right\} ds \\ &+ e^{-\lambda(S-t)} l(F(x_0 e^{au-bd}) - v) \end{aligned}$$

$$= \int_0^{S-t} e^{-\lambda s} \min_{\zeta \in I_{v,u,d}} \left\{ \begin{array}{c} \lambda^+ J^{n-1}(v + \zeta x_0 e^{au-bd}(e^a - 1), u+1, d, t+s) \\ +\lambda^- J^{n-1}(v + \zeta x_0 e^{au-bd}(e^{-b} - 1), u, d+1, t+s) \end{array} \right\} ds$$

$$+ e^{-\lambda(S-t)} l(F(x_0 e^{au-bd}) - v),$$

$$= (TJ^{n-1})(v,u,d,t) = J^n(v,u,d,t) \tag{19}$$

which proves the assertion for n. □

We have explicit expressions for the first two iterations of T:

$$J^1(v,u,d,t) = e^{-\lambda(S-t)} l(F(x_0 e^{au-bd}) - v), \tag{20}$$

$$J^2(v,u,d,t) = J^1(v,u,d,t) \tag{21}$$

$$+ (S-t) \min_{\zeta \in I_{v,u,d}} \left\{ \begin{array}{c} \lambda^+ J^1(v + \zeta x_0 e^{au-bd}(e^a - 1), u+1, d, t) \\ +\lambda^- J^1(v + \zeta x_0 e^{au-bd}(e^{-b} - 1), u, d+1, t) \end{array} \right\},$$

where we have used the fact that J^1 depends on t only through the factor given by the exponential function. Obtaining an explicit expression of $J^n(v,u,d,t)$ becomes difficult, if not impossible, for $n > 2$, unless one makes some simplifying assumptions that would, however, lead to a suboptimal solution. In section 5 we shall therefore describe a computable approximation approach and show its convergence.

THEOREM 3.3.

(i) *The value function J^* is the unique fixed point of T, i.e.,*

$$J^* = TJ^*, \tag{22}$$

and we have

$$||J^n - J^*|| \le e^{\lambda S}(1 - e^{-\lambda S})^n ||J^1||. \tag{23}$$

(ii) *The following strategy ξ^* is efficient: For $s \in (\hat\tau_{u+d}, \hat\tau_{u+d+1}]$ and $v = V_{\tau_{u+d}}$, let ξ_s^* be given by the unique solution to the deterministic optimization problem embedded in the computation of $(TJ^*)(u,d,v,t)$ according to (11), i.e.,*

$$\xi_s^* := \arg \min_{\zeta \in I_{v,u,d}} \Big\{ \lambda^+ J^*(v + \zeta x_0 e^{au-bd}(e^a - 1), u+1, d, s)$$

$$+\lambda^- J^*(v + \zeta x_0 e^{au-bd}(e^{-b} - 1), u, d+1, s) \Big\}.$$

Proof. From (18) we obtain

$$J_0^n \le J_0^* \le J_0^n + E[l(F(X_S) + c),\ \tau_n \le S]. \tag{24}$$

Due to estimates (24) and (6), we have

$$\lim_{n \uparrow \infty} J^n(v,0,0,t) = J^*(v,0,0,t) \tag{25}$$

uniformly in v and t. Since N^+, N^- are stationary Markov processes, we obtain $J^n(v,u,d,t) \to J^*(v,u,d,t)$ uniformly in v and t for all u,d. This implies $J^* \in \mathcal{C}(E)$; i.e., J^* is in the domain of T. Application of the DP principle implies that J^* is a fixed point of T (alternatively, this follows from (25) and Lemma 3.1).

For item (ii), we obtain from the DP principle

$$J^*(v,u,d,t) = \min_{\xi \in \mathcal{A}_{v,u,d,t}} E\Bigg[J^* \left(v + \int_t^{\tau_{u+d+1}} \xi_s dX_s, N^+_{\tau_{u+d+1}}, N^-_{\tau_{u+d+1}}, \tau_{u+d+1} \right) \Bigg|$$

$$N_t^+ = u,\ N_t^- = d \Bigg].$$

Due to (10), the right-hand side of the previous equation evaluates to

$$\min_{\xi\in\mathcal{A}_{v,u,d,t}} \int_0^{S-t} e^{-\lambda s} \begin{Bmatrix} \lambda^+ J^*(v+\xi_s x_0 e^{au-bd}(e^a - 1), u+1, d, t+s) \\ +\lambda^- J^*(v+\xi_s x_0 e^{au-bd}(e^{-b}-1), u, d+1, t+s) \end{Bmatrix} ds$$
$$+ e^{-\lambda(S-t)} l(F(x_0 e^{au-bd}) - v).$$

The minimum is achieved by the strategy ξ^* prescribed in item (ii) (see also (19)). This proves optimality of the strategy ξ^*, and we obtain again the fixed-point equation

$$J^*(v,u,d,t) = (TJ^*)(v,u,d,t). \qquad \square$$

Remark 2. In computing the approximating strategy ξ^n, (23) gives a handle on the distance between the value of the nth iteration J_0^n and the optimal value J_0^*. However, from an economic point of view, one is equally interested in the expected loss incurred from implementing the strategy ξ^n associated with this iteration, i.e., in the term $E[l(F(X_S) - V_S^n)]$, where, we recall from (16), V_t^n is the wealth process associated with the strategy ξ^n. This can be estimated via

$$J_0^n \le E[l(F(X_S) - V_S^n)] \le J_0^n + E[l(F(X_S)+c),\ \tau_n \le S] \tag{26}$$

which follows from (18). In section 5, we examine the numerical implementation of the problem in more detail.

4. Uncertain intensities.

4.1. Bayesian updating. In this section, we assume λ^+ and λ^- are constant but unknown to the investor; i.e., the true probability distribution $P = P_{\lambda^+,\lambda^-}$ is unknown to the investor. In modeling this situation we take the Bayesian point of view, thereby assuming that λ^+ and λ^- are random variables to which the investor assigns some (prior) distributions $\pi_0^+(d\lambda^+)$ and $\pi_0^-(d\lambda^-)$. Let $(\Omega, \mathcal{F}, P)$ denote the given probability space where $P = P_{\lambda^+,\lambda^-}$. On $\overline{\Omega} := \mathbb{R}_+^2 \times \Omega$, the subjective probability measure of the investor is thus given by

$$P(d\overline{\omega}) = \pi_0^+(d\lambda^+)\pi_0^-(d\lambda^-)P_{\lambda^+,\lambda^-}(d\omega).$$

We consider two filtrations on $\overline{\Omega}$:

(i) The filtration $(\mathcal{F}_t)$ generated by the processes N^+ and N^-. This is the information available to the investor.

(ii) Let $\mathcal{G}_0$ denote the σ-algebra generated by λ^+ and λ^-. We then define the filtration $\mathcal{G}_t = \mathcal{F}_t \vee \mathcal{G}_0$ as the filtration corresponding to "full information."

We have that, with respect to the filtration $(\mathcal{F}_t)$, N is a Cox process, whereas, with respect to the filtration $(\mathcal{G}_t)$, N is a Poisson process with intensity $\lambda^+ + \lambda^-$.

We define P_0 via

$$\left.\frac{dP}{dP_0}\right|_{\mathcal{G}_t} = e^{-(\lambda^+ + \lambda^- - 2)t}(\lambda^+)^{N_t^+}(\lambda^-)^{N_t^-} =: L_t.$$

Under P_0, the random quantities λ^+, λ^-, N^+, and N^- are independent: N^+ and N^- are standard Poisson processes; i.e., both have known intensity one. The random variables λ^i have distribution π_0^i under P_0 for $i = +, -$. Let

$$\mathcal{L}_t(f) := E_0[L_t\, f(\lambda^+,\lambda^-)|\,\mathcal{F}_t]$$
$$= \int_0^\infty \int_0^\infty f(\lambda^+,\lambda^-) e^{-(\lambda^+ + \lambda^- - 2)t}(\lambda^+)^{N_t^+}(\lambda^-)^{N_t^-}\pi_0^+(d\lambda^+)\pi_0^-(d\lambda^-)$$

and put

$$\pi_t^i(d\lambda) := \frac{e^{-\lambda t}\lambda^{N_t^i}\pi^i(d\lambda)}{\int_0^\infty e^{-\lambda t}\lambda^{N_t^i}\pi^i(d\lambda)}, \quad i = +, -. \tag{27}$$

We then have the relation (filter equation)

$$E[f(\lambda^+, \lambda^-)|\,\mathcal{F}_t] = \frac{\mathcal{L}_t(f)}{\mathcal{L}_t(1)} = \int\int f(\lambda^+, \lambda^-)\pi_t^+(d\lambda^+)\pi_t^-(d\lambda^-), \tag{28}$$

and $\pi_t^i(d\lambda^i)$ is the posterior distribution of λ^i, $i = +, -$.

Recall that the density of the gamma distribution with shape parameter α and scale parameter β can be given in the form

$$\gamma(\lambda\,|\,\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}.$$

We conclude from (27) that, if the prior distribution π_0^i for λ^i is a gamma distribution with parameters α_0^i and β_0, then the posterior distribution π_t at time t is again a gamma distribution with parameters

$$\alpha_t^i = \alpha_0^i + N_t^i, \quad \beta_t = \beta_0 + t \tag{29}$$

for $i = +, -$. Furthermore, the distribution of $\lambda = \lambda^+ + \lambda^-$ at time 0 is gamma with parameters $\alpha_0 = \alpha_0^+ + \alpha_0^-$ and β_0. At time t, the posterior distribution of λ is again gamma with β_t as in (29) and

$$\alpha_t = \alpha_t^+ + \alpha_t^- = \alpha_0 + N_t. \tag{30}$$

We can calculate by means of (28)

$$\overline{\lambda_t^i} := E[\lambda^i|\,\mathcal{F}_t] = \frac{\alpha_t^i}{\beta_t}, \quad i = +, -. \tag{31}$$

4.2. The PD-DP equation. The observed process X is still piecewise deterministic. Hence problem (8) is still a piecewise deterministic control problem. In the case of uncertain intensities, if X remains constant from time t to time $t+s$, this not only changes the time to maturity, but it also reveals additional information regarding the true intensities.

Let

$$p^+(u,d,t,s) := \left(\frac{\beta_0+t}{\beta_0+t+s}\right)^{\alpha_0+u+d}\frac{\alpha_0^+ + u}{\beta_0+t+s},$$
$$p^-(u,d,t,s) := \left(\frac{\beta_0+t}{\beta_0+t+s}\right)^{\alpha_0+u+d}\frac{\alpha_0^- + d}{\beta_0+t+s},$$
$$p^0(n,t) := \left(\frac{\beta_0+t}{\beta_0+S}\right)^{\alpha_0+n}.$$

LEMMA 4.1. *For any nonnegative function f, we have*

$$\begin{aligned} &E[f(N^+_{\hat\tau_{u+d+1}}, N^-_{\hat\tau_{u+d+1}}, \hat\tau_{u+d+1})|\ N_t^+ = u,\ N_t^- = d] \\ &= \int_0^{S-t}\Big\{p^+(u,d,t,s)f(u+1,d,t+s) + p^-(u,d,t,s)f(u,d+1,t+s)\Big\}ds \\ &\quad + p^0(u+d,t)f(u,d,S). \end{aligned} \tag{32}$$

Proof. Since N is a Poisson process with intensity $\lambda^+ + \lambda^-$ with respect to the filtration $(\mathcal{G}_t)$, we obtain from (10)

$$\begin{aligned}
&E[f(N^+_{\hat\tau_{u+d+1}}, N^-_{\hat\tau_{u+d+1}}, \hat\tau_{u+d+1})|\ N_t^+ = u,\ N_t^- = d] \\
&\quad = E[E[f(N^+_{\hat\tau_{u+d+1}}, N^-_{\hat\tau_{u+d+1}}, \hat\tau_{u+d+1})|\ \mathcal{G}_t]|\ N_t^+ = u,\ N_t^- = d] \\
&\quad = E\bigg[\int_0^{S-t} \Big\{\lambda^+ f(u+1,d,t+s) + \lambda^- f(u,d+1,t+s)\Big\} e^{-\lambda s} ds \\
&\qquad\quad + e^{-\lambda(S-t)} f(u,d,S)|\ N_t^+ = u,\ N_t^- = d\bigg].
\end{aligned}$$

Due to (28) and (29), the last expression is given by

$$\begin{aligned}
&\int\int \gamma(\lambda^+ \,|\, \alpha_0^+ + u, \beta_0 + t)\gamma(\lambda^- \,|\, \alpha_0^- + d, \beta_0 + t) d\lambda^+ d\lambda^- \\
&\cdot\Big[\int_0^{S-t} \Big\{\lambda^+ f(u+1,d,t+s) + \lambda^- f(u,d+1,t+s)\Big\} e^{-\lambda s} ds + e^{-\lambda(S-t)} f(u,d,S)\Big].
\end{aligned}$$

Applying Fubini's theorem to integrate first over λ^+, λ^- and then over s, we arrive at the right-hand side of (32). ∎

In the case of uncertain intensities, the DP operator T maps $J : E \to \mathbb{R}^+$ to $TJ :\ E \to \mathbb{R}^+$ defined by

$$\begin{aligned}
(TJ)(v,u,d,t) &= \int_0^{S-t} \min_{\zeta \in I_{v,u,d}} \left\{ \begin{matrix} p^+(u,d,t,s) J(v + \zeta x_0 e^{au-bd}(e^a - 1), u+1, d, t+s) \\ + p^-(u,d,t,s) J(v + \zeta x_0 e^{au-bd}(e^{-b} - 1), u, d+1, t+s) \end{matrix} \right\} ds \\
&\quad + p^0(u+d,t) l(F(x_0 e^{au-bd}) - v).
\end{aligned} \tag{33}$$

Let

$$E^k = \{(v,u,d,t) \,|\, v \geq -c,\ u,d \in \mathbb{N},\ u+d \leq k,\ t \in [0,S]\} \tag{34}$$

and

$$(T^k J)(v,u,d,t) := \begin{cases} (TJ)(v,u,d,t), & u+d \leq k-1, \\ 0, & u+d = k. \end{cases} \tag{35}$$

LEMMA 4.2. *For every $k \in \mathbb{N}$, the operator $T^k : \mathcal{C}(E^k) \to \mathcal{C}(E^k)$ is a contraction with contraction constant*

$$1 - p^0(k,0) = 1 - \left(\frac{\beta_0}{\beta_0 + S}\right)^{\alpha_0 + k},$$

and T^k has a unique fixed point in $\mathcal{C}(E^k)$.

Proof. 1. It follows as in Lemma 3.1 that $T^k J \in \mathcal{C}(E^k)$ for $J \in \mathcal{C}(E^k)$.

2. For $\xi : [0, S-t] \to I_{v,u,d}$ with values denoted, as previously, by ζ, let

$$\begin{aligned}
(T_\xi J)(v,u,d,t) &= p^0(u+d,t) l(F(x_0 e^{au-bd}) - v) \\
&\quad + \int_0^{S-t} \Big\{p^+(u,d,t,s) J(v + \zeta(s) x_0 e^{au-bd}(e^a - 1), u+1, d, t+s) \\
&\quad + p^-(u,d,t,s) J(v + \zeta(s) x_0 e^{au-bd}(e^{-b} - 1), u, d+1, t+s)\Big\} ds.
\end{aligned}$$

We then have

$$\begin{aligned}||T_\xi J - T_\xi J'||_{E^k} &\leq \left|\left|\int_0^{S-t} \{p^+(u,d,t,s) + p^-(u,d,t,s)\} ds\right|\right|_{E^k} ||J - J'||_{E^k} \\ &= ||1 - p^0(u+d,t)||_{E^k}\ ||J - J'||_{E^k} \\ &= \left|\left|1 - \left(\frac{\beta_0 + t}{\beta_0 + S}\right)^{\alpha_0+u+d}\right|\right|_{E^k} ||J - J'||_{E^k} \\ &\leq \left(1 - \left(\frac{\beta_0}{\beta_0 + S}\right)^{\alpha_0+k}\right) ||J - J'||_{E^k}.\end{aligned}$$

Due to the continuity of J, there exists $\xi : [0, S-t] \to I_{v,u,d}$ such that, for $u+d \leq k-1$,

$$T_\xi J = T^k J.$$

Hence we obtain

$$T^k J' - T^k J \leq T_\xi J' - T_\xi J \leq \left(1 - \left(\frac{\beta_0}{\beta_0 + S}\right)^{\alpha_0+k}\right) ||J' - J||_{E^k}.$$

By symmetry we can conclude that

$$||T^k J' - T^k J||_{E^k} \leq \left(1 - \left(\frac{\beta_0}{\beta_0 + S}\right)^{\alpha_0+k}\right) ||J' - J||_{E^k}. \qquad \square$$

COROLLARY 4.3. *For $k \in \mathbb{N}$, a fixed point J_k^* of T^k coincides with a fixed point J^* of T on the set E^k.*

Proof. Let $G = J^*$ on E^{k-1} and $G = 0$ on $E^k \setminus E^{k-1}$. It follows from (35) that G is a fixed point of T^k. Due to Lemma 4.2, this implies $G = J_k^*$. $\square$

Remark 3. While the case of uncertain intensities follows in large part along the lines of known intensities, Lemma 4.2 highlights a major difference: The operator T is no longer a contraction on $\mathcal{C}(E)$. This is due to the fact that, as the number of observed jumps increases, the estimated probability that no more jumps occur prior to S tends to zero. But it is exactly this probability that makes T a contraction; see also [1] for the case without model uncertainty.

For practical purposes, however, this poses no difficulties since (see Corollary 4.3) we can instead work with the contraction T^k, and for this see section 5, especially Lemma 5.1.

We conclude this section by showing Theorem 4.5 below by which it follows that, despite the fact that in the present setting the operator T is no longer a contraction, this operator has a unique fixed point. Furthermore, an efficient strategy exists in complete analogy to the case of known intensities.

As in (13), we define

$$J^0 = 0, \quad \text{and, for } h \leq n, \quad J^h = TJ^{h-1}. \tag{36}$$

Again, we obtain explicit expressions for the first two iterations of T (compare with (20) and (21)):

$$J^1(v,u,d,t) = p^0(u+d,t)\, l(F(x_0 e^{au-bd}) - v), \tag{37}$$

$$\begin{aligned}J^2(v,u,d,t) = J^1(v,u,d,t) + \frac{S-t}{\beta_0 + S} \\ \cdot \min_{\zeta \in I_{v,u,d}} \left\{ \begin{array}{c} (\alpha_0^+ + u)\, J^1(v + \zeta x_0 e^{au-bd}(e^a - 1), u+1, d, t) \\ + (\alpha_0^- + d)\, J^1(v + \zeta x_0 e^{au-bd}(e^{-b} - 1), u, d+1, t) \end{array} \right\}.\end{aligned} \tag{38}$$

In (38), we see directly how the observed jumps affect the optimal strategy: The larger the number u of upward jumps in comparison to the number of downward jumps d, the higher the a posteriori probability that the next jump will be upwards. Concerning the explicit computation of the exact values of $J^n(v,u,d,t)$ for $n>2$ the same comments apply as after (21).

We define the strategy ξ^n analogous to (14) and (15) in which λ^i is replaced by p^i $(i=+,-)$.

LEMMA 4.4. *We have*

$$J^n(v,u,d,t)=\min_{\xi\in\mathcal{A}_{v,u,d,t}} E[l(F(X_S)-V_S^{\xi}),\ \tau_{u+d+n}>S|\,N_t^+=u,\ N_t^-=d].$$

Proof. The proof proceeds exactly as in Lemma 3.2, with the exception that (10) is replaced by (32). □

THEOREM 4.5.

(i) *The value function J^* is the unique fixed point of T, i.e.,*

$$J^*(v,u,d,t)=(TJ^*)(v,u,d,t),\quad (v,u,d,t)\in E. \tag{39}$$

(ii) *The following strategy ξ^* is efficient: For $s\in(\hat\tau_{u+d},\hat\tau_{u+d+1}]$ and $v=V_{\tau_{u+d}}$, let $\xi_s^*=\zeta^{u,d,v,t,*}(s)$, where the latter is given by the deterministic optimization problem embedded in the computation of $(TJ^*)(u,d,v,t)$, i.e.,*

$$\zeta_s^*:=\arg\min_{\zeta\in I_{v,u,d}}\Big\{p^+(u,d,t,s-t)J^*(v+\zeta x_0e^{au-bd}(e^a-1),u+1,d,s)$$
$$+p^-(u,d,t,s-t)J^*(v+\zeta x_0e^{au-bd}(e^{-b}-1),u,d+1,s)\Big\}.$$

Proof. We first prove (i), i.e., the uniqueness of the fixed point. Consider two fixed points J^a and J^b of T. Due to Corollary 4.3 and Lemma 4.2, we obtain

$$J^a(v,u,d,t)=J^b(v,u,d,t),\quad (v,u,d,t)\in E^{k-1}.$$

Since k is arbitrary, this implies $J^a=J^b$ on E.

As in (24), we obtain

$$J_0^n\le J_0^*\le J_0^n+E[l(F(X_S)+c),\ \tau_n\le S]. \tag{40}$$

From (40) and the uniqueness of the fixed point we obtain item (i) as in the proof of Theorem 3.3.

Similarly, (ii) follows from the DP principle and (32). Especially, this implies that J^* is a fixed point of T. □

For a version of the estimate (23) in the case of uncertain intensities, we refer to Lemma 5.1.

We also have the following analogue to the estimate (26):

$$J_0^n\le E[l(F(X_S)-V_S^n)]\le J_0^n+E[l(F(X_S)+c),\ \tau_n\le S]. \tag{41}$$

5. Algorithmic implementation: Interpolation of the value function. Due to Theorem 3.3 (respectively, 4.5), the value J_0^n converges to the optimal value and the strategy ξ^n to the efficient strategy in the sense of the estimate (26), respectively (41). In order to compute J_0^n and ξ^n for reasonably large values of n, we need to discretize the problem in the dimensions wealth and time (v,t) and then to interpolate it in the same variables. In this section, we provide Lemma 5.1 to

control the error incurred from both this interpolation and stopping after the nth iteration. To this effect, in order to include also the case of uncertain intensities, we shall consider the operator T^k in (35) for some fixed $k \leq n$ instead of T. This will lead to a fixed contraction constant $1 - p^0(k,0)$ (see Lemma 4.2) also for the case of uncertain intensities and thus also for this case to the same upper bound in Lemma 5.1 below that goes to zero for $n \to \infty$.

For the case of unknown intensities we shall examine an approximation J_k^n of J^n defined recursively via

$$J_k^0 = 0, \quad \text{and, for } \ h \leq n, \quad J_k^h = T^k J_k^{h-1}, \tag{42}$$

and restrict the arguments v and t to $(v,t) \in \mathcal{S} := [-c, F_k] \times [0, S]$ with

$$F_k := \max\{F_{u,d} |\ u + d \leq k - 1\}; \tag{43}$$

see (5). Notice in fact that, if at any time $v \geq F_k$, then superhedging is possible, and this leads to an optimal value $J^* = 0$. Since for $v = F_k$ we are interested also in the optimal strategy, in what follows we shall thus require that $v \leq F_k$, i.e., $v \in [-c, F_k]$.

For the case of known intensities we shall use the same approximations as in (42), and all that follows holds in the same way also for this case by simply putting $k = n$.

Consider then some finite grid $G \subset \mathcal{S}$ containing the extremal points of $\mathcal{S}$ and denote by $\mathcal{D}(E^k)$ the Banach space of cadlag functions on E^k endowed with the Skorokhod norm $||\cdot||_{E^k}$. Define the operator $T_G^k :\ \mathcal{D}(E^k) \to \mathcal{D}(E^k)$ via

$$(T_G^k H)(v,u,d,t) := \begin{cases} (TH)(v,d,u,t) & \text{if } (v,t) \in G,\ u + d \leq k - 1, \\ 0 & \text{if } u + d = k \ \text{ or } \ v > F_k, \\ \text{cadlag interpolation} & \text{else.} \end{cases}$$

Notice that, although for $v = F_k$ we have the possibility of superhedging, in our calculations we need to consider also this value in order to obtain the corresponding (superhedging) strategy. Due to our interpolation approximations, this strategy will, however, turn out to be only approximately superhedging.

Due to Corollary 4.3, we have

$$T_G^k J^* = \begin{cases} J^*(v,d,u,t) & \text{if } (v,t) \in G,\ u + d \leq k - 1, \\ 0 & \text{if } u + d = k \ \text{ or } \ v > F_k, \\ \text{cadlag interpolation} & \text{else.} \end{cases}$$

Due to the continuity of J^*, we have

$$\epsilon(G) := ||J^* - T_G^k J^*||_{E^{k-1}} \to 0 \tag{44}$$

for

$$\sup_{(v,t) \in \mathcal{S}} \min_{(v',t') \in G} (|v - v'| + |t - t'|) \to 0.$$

We approximate the optimal value J^* by the value H_k^n defined recursively via

$$H_k^1 = J_k^1, \quad H_k^n = T_G^k H_k^{n-1} \quad \text{for} \quad n \geq k. \tag{45}$$

This value can be achieved by some strategy $\tilde{\xi}^n$ defined in analogy to (14)–(16) and, similarly, for the case of uncertain intensities.

With relation (44), the next lemma provides a handle on the error.

LEMMA 5.1. *We have*

$$||J^* - H_k^n||_{E^{k-1}} \leq \frac{1}{1-\kappa}\Big(\kappa^n ||J_k^1||_{E^{k-1}} + \epsilon(G)\Big), \tag{46}$$

where κ *is given by*

(i) $\kappa = 1 - e^{-\lambda S}$ *in the case of known intensities,*

(ii) $\kappa = 1 - p^0(k,0)$ *in the case of uncertain intensities.*

Proof. 1. It follows by an immediate extension to the case of cadlag interpolations of the proof of Lemma 3.1 (respectively, Lemma 4.2) that the operator $T_G^k : \mathcal{D}(E^k) \to \mathcal{D}(E^k)$ is a contraction with contraction constant κ. Especially, T_G^k has a fixed point $H_k^* \in \mathcal{D}(E^k)$. Noticing that $H_k^1(v,u,d,t) = p^0(u+d,t)l(F(x_0e^{au-bd}) - v)$ and that, consequently,

$$\begin{aligned} ||H_k^2(v,u,d,t) - H_k^1(v,u,d,t)||_{E^{k-1}} &= ||T_G^k H_k^1(v,u,d,t) - H_k^1(v,u,d,t)||_{E^{k-1}} \\ &\leq \left| \int_0^{S-t} \{p^+(u,d,t,s) + p^-(u,d,t,s)\} ds \right| ||H_k^1||_{E^{k-1}}, \end{aligned}$$

one obtains

$$||H_k^n - H_k^*||_{E^{k-1}} \leq \frac{\kappa^n}{1-\kappa} ||J_k^1||_{E^{k-1}}.$$

2. We have

$$\begin{aligned} ||J^* - H_k^*||_{E^{k-1}} &\leq ||J^* - T_G^k J^*||_{E^{k-1}} + ||T_G^k J^* - T_G^k H_k^*||_{E^{k-1}} \\ &\leq \epsilon(G) + \kappa ||J^* - H_k^*||_{E^{k-1}}, \end{aligned}$$

namely,

$$||J^* - H_k^*||_{E^{k-1}} \leq \frac{1}{1-\kappa}\epsilon(G).$$

3. Due to

$$||J^* - H_k^n||_{E^{k-1}} \leq ||J^* - H_k^*||_{E^{k-1}} + ||H_k^n - H_k^*||_{E^{k-1}},$$

the estimate (46) follows from 1 and 2. ∎

By analogy to (26) and (41) we now have

$$H_k^n \leq E[l(F(X_S) - \tilde{V}_S^n)] \leq H_k^n + E[l(F(X_S) + c),\ \tau_k \leq S],$$

where $\tilde{V}_t^n$ is the wealth process associated with the strategy $\tilde{\xi}^n$.

Remark 4. The just-described algorithm hinges upon the specific form given to the operator T^k in (35) for the case $u + d = k$. Variants are possible, and they imply slight variants also for the algorithm with advantages and disadvantages.

5.1. Example. Here we consider a simple example to illustrate the interpolation algorithm described above in this section. The example is as follows.

Consider the geometric Poisson price model (1) with $x_0 = 1$ and suppose that a, b are such that $e^a = 2$, $e^{-b} = \frac{1}{2}$. For the case of known intensities of the driving Poisson processes we let them be given by $\lambda^+ = \lambda^- = 1$, so that $\lambda = 2$. For the case of unknown intensities we choose $\alpha_0^+ = \alpha_0^- = \beta_0 = 1$ so that $\overline{\lambda}_0^+ = \overline{\lambda}_0^- = 1$.

Take a claim of the form $F(X_S) = (X_S - 1)^+$, a time horizon of $S = 2$, and put $c = 0.5$. The superhedge price for F is (see the comment after (5)) $F_{u,d} \equiv x_0 = 1$. Let the loss function be of the form $l(z) = [\max(z, 0)]^p$ for $p = 2$. The optimization criterion is then

$$J_0^* := \inf_{\xi \in \mathcal{A}_{V_0}} E\left\{ \left[\max\left(\left(x_0 e^{au - bd} - 1 \right)^+ - V_S, 0 \right) \right]^2 \right\}, \tag{47}$$

and it is of the type of *shortfall risk minimization*. The interval of admissible values for ξ on $[0, \hat{\tau}_1]$ is given by (see (9))

$$I_{v,u,d} = \left[-\frac{c+v}{2^{u-d}},\ 2\,\frac{c+v}{2^{u-d}} \right]. \tag{48}$$

Since we have $F_{u,d} \equiv x_0 = 1$, also for F_k in (43) one has $F_k \equiv 1$. This implies that, for the given data, $(v, t) \in \mathcal{S} := [-\frac{1}{2}, 1] \times [0, 2]$. Consider then the following grid that, for the purpose of illustrating the procedure in the simplest possible way, is chosen to be very coarse. More precisely, $G \subset \mathcal{S}$ is obtained as follows: partition the time interval $[0, 2]$ into $L = 2$ subintervals $T_0 := [t_0 = 0, t_1 = 1]$, $T_1 := [t_1 = 1, t_2 = 2]$ and the wealth-value interval $[-\frac{1}{2}, 1]$ into $M = 3$ subintervals $V_0 := [v_0 = -\frac{1}{2}, v_1 = 0]$, $V_1 := [v_1 = 0, v_2 = \frac{1}{4}]$, $V_2 := [v_2 = \frac{1}{4}, v_3 = 1]$. The reason why we consider $v_2 = \frac{1}{4}$ and not, say, $v_2 = \frac{1}{2}$ is that at $v = \frac{1}{2}$ the optimal values are already equal to zero. Even though we are still far from the superhedge value of $v = 1$, this happens because we restrict ourselves to the event that no more than k jumps can occur.

5.1.1. Known intensities. We shall compute the values for H_k^n according to (45) (as well as (42)) for tuples (v_j, u, d, t_i) with $i \in \{0, 1\}$, $j \in \{0, 1, 2\}$, and $u + d \leq k - 1$. In fact, this last restriction for u and d and the sufficiency of considering only values of $v \leq 1$ come from the definition of the operator T_G^k. On the other hand, since we perform a cadlag interpolation, we do not need to consider the upper end point $t_i = 2$ of the time interval; i.e., it suffices to have $i \in \{0, 1\}$.

First we recall from (42) that, for our example,

$$J_k^1(v_j, u, d, t_i) = e^{-2(2-t_i)} \left[\max\left\{ \left(2^{u-d} - 1 \right)^+ - v_j, 0 \right\} \right]^2, \quad i \in \{0, 1\},\ j \in \{0, 1, 2\}, \tag{49}$$

and notice that (see by analogy the motivation for (21))

$$J_k^1(v_j, u, d, t_i + s) = e^{2s} J_k^1(v_j, u, d, t_i). \tag{50}$$

The recursions (45) now become, always for the case of our example and recalling

that we consider cadlag interpolations,

$$
\begin{aligned}
(51)\qquad & H_k^n(v_j,u,d,t_i) = \left(T_G^k H_k^{n-1}\right)(v_j,u,d,t_i) \\
& = J_k^1(v_j,u,d,t_i) + \int_0^{S-t_i} e^{-2s} \\
& \min_{\zeta\in I_{v_j,u,d}} \left\{ \begin{array}{l} \displaystyle\sum_{l=0}^{L-1}\sum_{m=0}^{M-1} 1_{\{T_l\}}(t_i+s)\, 1_{\{V_m\}}(v_j+\zeta x_0 e^{au-bd}(e^a-1)) \\ \quad H_k^{n-1}(v_m,u+1,d,t_l) \\ \displaystyle +\sum_{l=0}^{L-1}\sum_{m=0}^{M-1} 1_{\{T_l\}}(t_i+s)\, 1_{\{V_m\}}(v_j+\zeta x_0 e^{au-bd}(e^{-b}-1)) \\ \quad H_k^{n-1}(v_m,u,d+1,t_l) \end{array} \right\} ds \\
& = J_k^1(v_j,u,d,t_i) + \sum_{l=0}^{L-1} 1_{\{t_i\le t_l\}}\gamma_{i,l} \\
& \min_{\zeta\in I_{v_j,u,d}} \left\{ \begin{array}{l} \displaystyle\sum_{m=0}^{M-1} 1_{\{V_m\}}(v_j+\zeta x_0 e^{au-bd}(e^a-1))\, H_k^{n-1}(v_m,u+1,d,t_l) \\ \displaystyle +\sum_{m=0}^{M-1} 1_{\{V_m\}}(v_j+\zeta x_0 e^{au-bd}(e^{-b}-1))\, H_k^{n-1}(v_m,u,d+1,t_l) \end{array} \right\},
\end{aligned}
$$

where we have used the fact that

$$
\int_0^{S-t_i} e^{-2s} 1_{[t_l,t_{l+1})}(t_i+s)\, ds = \sum_{l=0}^{L-1} 1_{\{t_i\le t_l\}} \frac{e^{2t_i}}{2}\left[e^{-2t_l}-e^{-2t_{l+1}}\right]
$$

so that $\gamma_{i,l} := \frac{e^{2t_i}}{2}\left[e^{-2t_l}-e^{-2t_{l+1}}\right]$, in particular $\gamma_{0,0} = \frac{1-e^{-2}}{2}$ $\gamma_{0,1} = e^{-2}\gamma_{0,0}$. Since J_k^1 satisfies property (50), we have that for $n=2$ the above recursions (51) simplify to become

(52)

$$
\begin{aligned}
& H_k^2(v_j,u,d,t_i) = J_k^1(v_j,u,d,t_i) \\
& \quad +(S-t_i)\cdot \min_{\zeta\in I_{v_j,u,d}} \left\{ \begin{array}{l} \displaystyle\sum_{m=0}^{M-1} 1_{\{V_m\}}(v_j+\zeta x_0 e^{au-bd}(e^a-1))\, J_k^1(v_m,u+1,d,t_i) \\ \displaystyle +\sum_{m=0}^{M-1} 1_{\{V_m\}}(v_j+\zeta x_0 e^{au-bd}(e^{-b}-1))\, J_k^1(v_m,u,d+1,t_i) \end{array} \right\}.
\end{aligned}
$$

Always in order to illustrate the procedure in the simplest possible way, we shall now consider the case of $k=n=3$ with the objective of computing the values for $H_3^3(0,0,0,0)$ and $H_3^3(\frac{1}{4},0,0,0)$ as well as the corresponding minimizing strategy ξ_s^3 for $s\in[0,2]$.

For this purpose we first compute $J_3^1(v_j,u,d,t_i)$ for $v_j\in\{-\frac{1}{2},0,\frac{1}{4},1\}, t_i\in\{0,1\}$, and $u+d\le 2$. This can be easily done on the basis of (49) noticing that these values are zero for all tuples (v_j,u,d,t_i) with $u\le d$ and $v_j\ge 0$ and this is, thanks to (50), independently of t_i.

One then proceeds to compute $H_3^2(v_j, u, d, t_i)$ for $v_j \in \{-\frac{1}{2}, 0, \frac{1}{4}, 1\}, t_i \in \{0, 1\}$, and $u + d = 1$. Notice that the computation of these values leads then also to the minimizing strategy for $s \in (\hat{\tau}_1, \hat{\tau}_2]$. Since by (48) we have $I_{-\frac{1}{2},u,d} = \{0\}$, no minimization is required for the computation of $H_3^2(-\frac{1}{2}, u, d, t_i)$, and so, for this case of $v_j = -\frac{1}{2}$, the minimizing strategy is $\xi_s \equiv 0$ when $s \in (\hat{\tau}_1, \hat{\tau}_2]$. From (52) we now have

$$\begin{aligned} H_3^2\left(-\frac{1}{2}, 1, 0, 0\right) &= J_3^1\left(-\frac{1}{2}, 1, 0, 0\right) + 2\left\{J_3^1\left(-\frac{1}{2}, 2, 0, 0\right) + J_3^1\left(-\frac{1}{2}, 1, 1, 0\right)\right\} \\ &= \frac{109}{4}e^{-4}. \end{aligned}$$

With similar calculations one obtains $H_3^2(-\frac{1}{2}, 0, 1, 0) = \frac{5}{4}e^{-4}$, and, from here, by (50) $H_3^2(-\frac{1}{2}, 1, 0, 1) = \frac{59}{4}e^{-2}$ and $H_3^2(-\frac{1}{2}, 0, 1, 1) = \frac{3}{4}e^{-2}$.

Moving on to $v_j = 0$, we have that $I_{0,u,d} = \left[-\frac{c}{2^{u-d}}, 2\frac{c}{2^{u-d}}\right] \neq \{0\}$ so that this time one has to perform also the minimization according to (52). For this purpose notice that the partition $\{V_m\}_{m=1}^M$ of the wealth-value interval induces a partition of $I_{0,u,d}$ characterized by the fact that the right-hand side in (52) remains constant for all values of ζ in a same subinterval of the partition. We thus have

$$\begin{aligned} H_3^2(0,1,0,0) &= J_3^1(0,1,0,0) \\ &\quad + 2\min\Big\{1_{\{\zeta=-\frac{1}{4}\}}\left[J_3^1(-\tfrac{1}{2},2,0,0) + J_3^1(\tfrac{1}{4},1,1,0)\right], \\ &\qquad 1_{\{\zeta\in(-\frac{1}{4},0)\}}\left[J_3^1(-\tfrac{1}{2},2,0,0) + J_3^1(0,1,1,0)\right], \\ &\qquad 1_{\{\zeta=0\}}\left[J_3^1(0,2,0,0) + J_3^1(0,1,1,0)\right], \\ &\qquad 1_{\{\zeta\in(0,\frac{1}{8})\}}\left[J_3^1(0,2,0,0) + J_3^1(-\tfrac{1}{2},1,1,0)\right], \\ &\qquad 1_{\{\zeta\in(\frac{1}{8},\frac{1}{8})\}}\left[J_3^1(\tfrac{1}{4},2,0,0) + J_3^1(-\tfrac{1}{2},1,1,0)\right], \\ &\qquad 1_{\{\zeta=\frac{1}{2}\}}\left[J_3^1(1,2,0,0) + J_3^1(-\tfrac{1}{2},1,1,0)\right]\Big\} \\ &= e^{-4} + 2\min\left\{\tfrac{49}{4}e^{-4}, \tfrac{49}{4}e^{-4}, 9e^{-4}, \tfrac{37}{4}e^{-4}, \tfrac{125}{16}e^{-4}, \tfrac{17}{4}e^{-4}\right\} = \tfrac{19}{2}e^{-4}, \end{aligned}$$

where the min is achieved for $\zeta = \frac{1}{2}$ so that, for this case of $v_j = 0$, the minimizing strategy is $\xi_s^3 = \frac{1}{2}$ when $s \in (\hat{\tau}_1, \hat{\tau}_2] \cap [0, 1]$. Analogously, one obtains $H_3^2(0,1,0,1) = \frac{21}{4}e^{-2}$ where, thanks to (50), the minimizing value of ζ is the same as before so that we have again $\xi_s^3 = \frac{1}{2}$ for $v_j = 0$ also when $s \in (\hat{\tau}_1, \hat{\tau}_2] \cap [(1, 2]$. By similar calculations one then obtains

$$\begin{aligned} &H_3^2(0,0,1,0) = H_3^2(0,0,1,1) = 0 && \text{with minimizing } \zeta = 0, \\ &H_3^2(\tfrac{1}{4},1,0,0) = \tfrac{145}{16}e^{-4} && \text{with minimizing } \zeta = \tfrac{3}{8}, \\ &H_3^2(\tfrac{1}{4},1,0,1) = \tfrac{77}{16}e^{-2} && \text{with minimizing } \zeta = \tfrac{3}{8}, \\ &H_3^2(\tfrac{1}{4},0,1,0) = H_3^2(\tfrac{1}{4},0,1,1) = 0 && \text{with minimizing } \zeta \text{ any } \zeta \in \left[-\tfrac{1}{2}, 1\right], \\ &H_3^2(1,1,0,0) = 8e^{-4} && \text{with minimizing } \zeta = 0, \\ &H_3^2(1,1,0,1) = 4e^{-2} && \text{with minimizing } \zeta = 0. \end{aligned}$$

At this point one can finally compute

$$
\begin{aligned}
H_3^3(0,0,0,0) &= J_3^1(0,0,0,0) \\
&\quad + \textstyle\sum_{l=0}^{1} \gamma_{0,l} \min \Big\{ 1_{\{\zeta=-\frac{1}{2}\}} \left[H_3^2(-\tfrac{1}{2},1,0,t_l) + H_3^2(\tfrac{1}{4},0,1,t_l) \right], \\
&\qquad\qquad 1_{\{\zeta\in(-\frac{1}{2},0)\}} \left[H_3^2(-\tfrac{1}{2},1,0,t_l) + H_3^2(0,0,1,t_l) \right], \\
&\qquad\qquad 1_{\{\zeta=0\}} \left[H_3^2(0,1,0,t_l) + H_3^2(0,0,1,t_l) \right], \\
&\qquad\qquad 1_{\{\zeta\in(0,\frac{1}{4})\}} \left[H_3^2(0,1,0,t_l) + H_3^2(-\tfrac{1}{2},0,1,t_l) \right], \\
&\qquad\qquad 1_{\{\zeta\in[\frac{1}{4},1)\}} \left[H_3^2(\tfrac{1}{4},1,0,t_l) + H_3^2(-\tfrac{1}{2},0,1,t_l) \right], \\
&\qquad\qquad 1_{\{\zeta=1\}} \left[H_3^2(1,1,0,t_l) + H_3^2(-\tfrac{1}{2},0,1,t_l) \right] \Big\} \\
&= 7e^{-4}(1-e^{-2}),
\end{aligned}
$$

where the min is achieved for $\zeta = 1$ both when $t_l = 0$ and when $t_l = 1$ so that, for $v_j = 0$, the minimizing strategy is $\xi_s^3 = 1$ when $s \in [0, \hat{\tau}_1] \cap [0, 2]$. Peforming analogous calculations one obtains that $H_3^3(\frac{1}{4}, 0, 0, 0) = \frac{221}{32} e^{-4}(1 - e^{-2})$, which is slightly less than the value for $H_3^3(0, 0, 0, 0)$, and the min here is achieved for any $\zeta \in [0, \frac{1}{2}]$ when $t_l = 0$ and for $\zeta = \frac{3}{4}$ when $t_l = 1$. This implies that, for $v_j = \frac{1}{4}$, the minimizing strategy is $\xi_s^3 \in [0, \frac{1}{2}]$ when $s \in (0, \hat{\tau}_1] \cap [0, 1]$ and $\xi_s^3 = \frac{3}{4}$ when $s \in (0, \hat{\tau}_1] \cap (1, 2]$.

Summarizing, we have obtained for the (approximating) minimizing strategy the following expression where, due to the right continuous interpolation, the strategy is the same for all values v of the wealth belonging to a same subinterval of the partition $\{V_m\}$ and given by the value computed in its lower end point. We have in fact

$$
\xi_s^3 = \begin{cases}
1 & \text{if } s \in [0, \hat{\tau}_1] \cap [0, 2] \text{ and } V_0 = 0, \\
\text{any } \zeta \in [0, \frac{1}{2}] & \text{if } s \in [0, \hat{\tau}_1] \cap [0, 1] \text{ and } V_0 = \frac{1}{4}, \\
\zeta = \frac{3}{4} & \text{if } s \in [0, \hat{\tau}_1] \cap (1, 2] \text{ and } V_0 = \frac{1}{4}, \\
\zeta = \frac{3}{8} & \text{if } s \in (\hat{\tau}_1, \hat{\tau}_2] \cap [0, 2],\ u = 1, d = 0,\ V_{\hat{\tau}_1} \in [\frac{1}{4}, 1), \\
\zeta = 0 & \text{if } s \in (\hat{\tau}_1, \hat{\tau}_2] \cap [0, 2],\ u = 1, d = 0,\ V_{\hat{\tau}_1} \geq 1, \\
0 & \text{if } s \in (\hat{\tau}_1, \hat{\tau}_2] \cap [0, 2],\ u = 0, d = 1,\ V_{\hat{\tau}_1} < 0, \\
0 & \text{if } s \in (\hat{\tau}_1, \hat{\tau}_2] \cap [0, 2],\ u = 0, d = 1,\ V_{\hat{\tau}_1} \in [0, \frac{1}{4}), \\
\text{any } \zeta \in [-\frac{1}{2}, 1] & \text{if } s \in (\hat{\tau}_1, \hat{\tau}_2] \cap [0, 2],\ u = 0, d = 1,\ V_{\hat{\tau}_1} = \frac{1}{4}, \\
0 & \text{if } s > \hat{\tau}_2.
\end{cases}
$$

5.1.2. Uncertain intensities. We choose $\alpha_0^+ = \alpha_0^- = \beta_0 = 1$ so that $\overline{\lambda}_0^+ = \overline{\lambda}_0^- = 1$. We start from the expression for J_k^1 that is given here by (see (42))

(53)
$$J_k^1(v_j, u, d, t_i)$$

$$= \left(\frac{t_i+1}{3}\right)^{u+d+2} \left[\max\left\{\left(2^{u-d}-1\right)^+ - v_j, 0\right\}\right]^2,\ i \in \{0,1\},\ j \in \{0,1,2\},$$

for which, analogously to (50), we have

$$J_k^1(v_j,u,d,t+s) = \left(\frac{t+s+1}{t+1}\right)^{u+d+2} J_k^1(v_j,u,d,t). \tag{54}$$

On the other hand, the recursions (45) become

(55)

$$\begin{aligned}
H_k^n(v_j,u,d,t_i) &= J_k^1(v_j,u,d,t_i) \\
&\quad + \int_0^{S-t_i} \min_{\zeta\in I_{v_j,u,d}} \left\{ \sum_{l,m=0}^{L-1} 1_{\{T_l\}}(t_i+s) \left(\frac{1+t_i}{1+t_l}\right)^{1+u+d} \frac{1+u}{1+t_l} \right. \\
&\qquad \cdot 1_{\{V_m\}}(v_j+\zeta x_0 e^{au-bd}(e^a-1)) H_k^{n-1}(v_m,u+1,d,t_l) \\
&\quad + \sum_{l,m=0}^{L-1} 1_{\{T_l\}}(t_i+s) \left(\frac{1+t_i}{1+t_l}\right)^{1+u+d} \frac{1+d}{1+t_l} \\
&\qquad \left. \cdot 1_{\{V_m\}}(v_j+\zeta x_0 e^{au-bd}(e^{-b}-1)) H_k^{n-1}(v_m,u,d+1,t_l) \right\} \\
&= J_k^1(v_j,u,d,t_i) + \sum_{l=0}^{L-1} 1_{\{t_i\le t_l\}} \\
&\quad \cdot \min_{\zeta\in I_{v_j,u,d}} \left\{ \gamma_{i,l}^u \sum_{m=0}^{L-1} 1_{\{V_m\}}(v_j+\zeta x_0 e^{au-bd}(e^a-1)) H_k^{n-1}(v_m,u+1,d,t_l) \right. \\
&\quad \left. + \gamma_{i,l}^d \sum_{m=0}^{L-1} 1_{\{V_m\}}(v_j+\zeta x_0 e^{au-bd}(e^{-b}-1)) H_k^{n-1}(v_m,u,d+1,t_l) \right\}
\end{aligned}$$

having put

$$\begin{cases} \gamma_{i,l}^u &= (t_{l+1}-t_l)\left(\frac{1+t_i}{1+t_l}\right)^{1+u+d} \frac{1+u}{1+t_l}, \\ \gamma_{i,l}^d &= (t_{l+1}-t_l)\left(\frac{1+t_i}{1+t_l}\right)^{1+u+d} \frac{1+d}{1+t_l}. \end{cases}$$

Due to (54) the relations (55) simplify slightly for $n=2$ but not anymore as much as in (52).

At this point the procedure parallels mostly the one described in the previous subsection for known values of the intensities and the calculations are similar.

Note added in proof. Improvements and extensions to the contents of section 5 can be found in [19].

REFERENCES

[1] A. Almudevar, *A dynamic programming algorithm for the optimal control of piecewise deterministic Markov processes*, SIAM J. Control Optim., 40 (2001), pp. 525–539.

[2] J. Cvitanić, *Minimizing expected loss of hedging in incomplete and constrained markets*, SIAM J. Control Optim., 38 (2000), pp. 1050–1066.
[3] J. Cvitanić and I. Karatzas, *Hedging and portfolio optimization under transaction costs: A martingale approach*, Math. Finance, 6 (1996), pp. 133–165.
[4] J. Cvitanić and I. Karatzas, *On dynamic measures of risk*, Finance Stoch., 4 (1999), pp. 451–482.
[5] M. H. A Davis, *Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models*, J. R. Stat. Soc. Ser. B Stat. Methodol., 46 (1984), pp. 353–388.
[6] M. A. H. Dempster and J. J. Ye, *Necessary and sufficient optimality conditions for control of piecewise deterministic Markov processes*, Stochastics Stochastics Rep., 40 (1992), pp. 125–145.
[7] N. El Karoui and M.-C. Quenez, *Dynamic programming and pricing of contingent claims in an incomplete market*, SIAM J. Control Optim., 33 (1995), pp. 29–66.
[8] H. Föllmer and P. Leukert, *Efficient hedging: Cost versus shortfall risk*, Finance Stoch., 4 (2000), pp. 117–146.
[9] H. Föllmer and D. Sondermann, *Hedging of non-redundant contingemt claims*, in Contributions to Mathematical Economics in Honor of Géerard Debreu, W. Hildenbrand and A. Mas-Colell, eds., North–Holland, Amsterdam, 1986, pp. 205–223.
[10] R. Frey and W. J. Runggaldier, *Risk-minimizing hedging strategies under restricted information: The case of stochastic volatility models observed only at discrete random times*, Math. Methods Oper. Res., 50 (1999), pp. 339–350.
[11] A. A. Gushchin and E. Mordecki, *Bounds on option prices for semimartingale market models*, Proc. Steklov Inst. Math., 237 (2002), pp. 73–113.
[12] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, Berlin, 1989.
[13] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.
[14] I. Karatzas and X. Zhao, *Bayesian adaptive portfolio optimization*, in Handbook of Mathematical Finance, Cambridge University Press, Cambridge, UK, 2001, pp. 632–670.
[15] M. Kirch, *Efficient Hedging in Incomplete Markets under Model Uncertainty*, Ph.D. thesis, Humboldt University, Berlin, Germany, 2002.
[16] R. Korn, *Optimal Portfolios. Stochastic Models for Optimal Investment and Risk Management in Continuous Time*, World Scientific, Singapore, 1997.
[17] D. Kramkov, *Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets*, Probab. Theory Related Fields, 105 (1996), pp. 459–479.
[18] P. Lakner, *Utility maximization with partial information*, Stochastic Process. Appl., 56 (1995), pp. 247–273.
[19] W. J. Runggaldier and S. Di Emidio, *Computing efficient hedging strategies in discontinuous market models*, in Proceedings of International Conference "Stochastic Finance 2004," Lisbon, Portugal, 2004.
[20] R. Rishel, *Optimal portfolio management with partial observations and power utility function*, in Stochastic Analysis, Control, Optimization and Applications, Volume in Honor of W. H. Fleming, W. M. McEneany, G. Yin, and Q. Zhang, eds., Birkhäuser, Boston, 1999, pp. 605–619.
[21] W. Schachermayer, *Optimal investment in incomplete financial markets*, in Mathematical Finance—Bachelier Congress 2000, H. Geman, D. Madan, S. Pliska, and T. Vorst, eds., Springer-Verlag, Berlin, 2002, pp. 427–462.
[22] M. Schweizer, *Mean variance hedging for general claims*, Ann. Appl. Probab., 2 (1992), pp. 171–179.
[23] M. Schweizer, *Risk minimizing hedging strategies under restricted information*, Math. Finance, 4 (1994), pp. 327–342.
[24] M. Schweizer, *A guided tour through quadratic hedging approaches*, in Option Pricing, Interest Rates and Risk Management, E. Jouini, J. Cvitanić, and M. Musiela, eds., Cambridge University Press, Cambridge, UK, 2000, pp. 538–574.
[25] D. Vermes, *Optimal control of piecewise deterministic Markov processes*, Stochastics, 14 (1985), pp. 165–208.
[26] G. Zohar, *A generalized Cameron-Martin formula with applications to partially observed dynamic portfolio optimization*, Math. Finance, 11 (2001), pp. 475–494.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1196–1221

LINEAR QUADRATIC GAUSSIAN BALANCING FOR DISCRETE-TIME INFINITE-DIMENSIONAL LINEAR SYSTEMS*

MARK R. OPMEER† AND RUTH F. CURTAIN†

Abstract. In this paper, we study the existence of linear quadratic Gaussian (LQG)–balanced realizations for discrete-time infinite-dimensional systems. LQG-balanced realizations are those for which the smallest nonnegative self-adjoint solutions of the control and filter Riccati equations are equal. We show that the control (filter) Riccati equation has a nonnegative self-adjoint solution if and only if the system is output (input) stabilizable. Our main result is that the transfer function of a discrete-time linear system has an approximately controllable and observable LQG-balanced realization iff it has an input and output stabilizable realization. The corresponding control and filter Riccati equations have unique nonnegative self-adjoint solutions. Moreover, approximately controllable and observable LQG-balanced realizations are unique up to a unitary state-space transformation. Finally, we show that the spectrum of the product of the smallest nonnegative self-adjoint solutions of the control and filter Riccati equations is independent of the particular realization.

Key words. balanced realization, discrete-time system, infinite-dimensional system, LQG-balanced realization, normalized factorization, Riccati equations

AMS subject classifications. 47A48, 47N70, 93B28, 93C55

DOI. 10.1137/S0363012903431189

1. Introduction.

Simple models are normally preferred over complex ones in control systems design. Sometimes it is obvious how to construct a simple model for a physical system, but sometimes the characteristics essential to the controller design of a physical system are not obvious. One way to obtain a simple model in this last case is to first obtain a sophisticated model that takes into account every aspect that could be of interest and then perform model reduction on this sophisticated model. A simple model reduction procedure was introduced by Moore [7] and is now a textbook subject (see, e.g., Zhou and Doyle [17, Chapter 7]). The method proposed by Moore consists of truncating a balanced realization. A balanced realization (also called Lyapunov-balanced or internally balanced) is a realization for which the controllability and observability gramians are equal and diagonal. This procedure is applicable only to stable systems. Alternatively for unstable systems one can use truncations of a linear quadratic Gaussian (LQG)–balanced realization, which for rational transfer functions always exists. A LQG-balanced realization is a realization for which the smallest nonnegative self-adjoint solutions of the standard LQG control and filter Riccati equations are equal and diagonal. This method was proposed by Verriest [13], [14] and further developed by Jonckheere and Silverman [5]. For an alternative treatment see Mustafa and Glover [8]. The discrete-time case was considered by Hoffmann, Prätzel-Wolters, and Zerz [4].

In the case that the system is infinite-dimensional, the model/controller approximation becomes essential. One would like to use the methods of balanced truncation and LQG-balanced truncation in this case, too. The existence of Lyapunov-balanced and LQG-balanced realizations for irrational transfer functions, however, is nontrivial. Sufficient conditions for the existence of Lyapunov-balanced realizations were proved

*Received by the editors July 10, 2003; accepted for publication (in revised form) March 29, 2004; published electronically December 1, 2004.
http://www.siam.org/journals/sicon/43-4/43118.html
†Mathematics Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands (opmeer@math.rug.nl, curtain@math.rug.nl).

by Young [15], [16]. The purpose of this article is to give necessary and sufficient conditions for the existence of LQG-balanced realizations for discrete-time infinite-dimensional systems.

The proof is based on the correspondence between the Riccati equations of the plant and the Lyapunov equations of a certain closed-loop system and the result of Young on the existence of Lyapunov-balanced realizations. Although discrete-time systems have bounded operators, a number of features make the infinite-dimensional case more complicated than the finite-dimensional case. One is that input and output stabilizability, the natural infinite-dimensional generalizations of stabilizability and detectability, are not sufficient to obtain unique solutions of the control Riccati equation. Another is that it is not a priori clear that the natural factorization generated by the closed-loop system is coprime. These uniqueness and coprime properties were key features in the finite-dimensional proofs. Consequently, we have been forced to develop different proofs, predominantly algebraic, to get around these complications. We exploit the factorization idea previously used by Meyer and Franklin [6] (see also Ober and McFarlane [9]) for the finite-dimensional continuous-time case.

This article is organized as follows. We begin by reviewing the known finite-dimensional theory on discrete-time LQG-balancing in section 2. In section 3 we review the relevant theory of discrete-time infinite-dimensional systems and, in particular, the linear quadratic regulator problem for this class of systems. Some of our results appear to be new. The previous standard results (e.g., Halanay and Ionescu [3]) assume a type of "exponential" stabilizability that is too strong for our purposes. In section 4 we review the relevant theory on Lyapunov-balanced realizations for discrete-time infinite-dimensional systems. The key result on the connection between normalized factorizations and the linear quadratic regulator theory is developed in section 5. In section 6 we define the LQG-characteristic values and show that they are system invariants. In section 7 we derive many algebraic relations between the solutions of the control and filter Riccati equations and the Lyapunov equations of the closed-loop system. The often tedious algebraic proofs are relegated to the appendix. Finally, all the results from the previous sections are linked up in section 8 to prove our main result: an input and output stabilizable discrete-time system possesses an approximately controllable and observable LQG-balanced realization. These realizations are unique up to a unitary state-space transformation. This represents an elegant generalization of the finite-dimensional theory under minimal assumptions.

2. LQG-balanced realizations: The finite-dimensional case. In this section we review some of the results on finite-dimensional LQG-balanced realizations. We consider systems of the form

$$x_{n+1} = Ax_n + Bu_n, \quad x(0) = x_0, \quad y_n = Cx_n + Du_n, \tag{2.1}$$

where A, B, C, D are matrices of compatible dimensions. For simplicity we consider the linear quadratic regulator (LQR) problem for the cost functional

$$J(x_0, u) := \sum_{n=0}^{\infty} \|u_n\|^2 + \|y_n\|^2,$$

where y is given in terms of x_0 and u by (2.1). The LQR problem consists of finding for a given x_0 that u for which $J(x_0, u)$ is minimal. As is well known, this problem has a unique solution when (A, B, C, D) is minimal: the optimal input $u^{\min}$ is given

by the state feedback $u_n^{\min} = -(I + D^*D + B^*QB)^{-1}(D^*C + B^*QA)x_n$, where Q is the unique nonnegative solution of the Riccati equation

$$A^*QA - Q + C^*C = (C^*D + A^*QB)(I + D^*D + B^*QB)^{-1}(D^*C + B^*QA),$$

and the optimal cost is given by $J(x_0, u^{\min}) = \langle x_0, Qx_0\rangle$. By duality the optimal filter cost is given by $\langle x_0, Px_0\rangle$, where P is the unique nonnegative solution of the Riccati equation

$$APA^* - P + BB^* = (BD^* + APC^*)(I + DD^* + CPC^*)^{-1}(DB^* + CPA^*).$$

The quantity $\langle x_0, Px_0\rangle$ can be interpreted as a measure of the difficulty of reconstructing the initial state x_0 from noisy measurements. The eigenvalues of the product PQ are similarity invariants. It was shown by Fuhrmann and Ober [2] that the square roots of the eigenvalues of PQ (called the LQG-characteristic values) are the singular values of a certain Hankel operator associated with the system. These invariants can be interpreted as a measure of how important the subspace generated by the eigenvector is for the compensator design. This can be seen from the LQG-balanced realization. An LQG-balanced realization is a realization of the plant such that $P = Q = \Lambda$, where Λ is the diagonal matrix containing the LQG-characteristic values. Let λ_i be the square root of an eigenvalue of PQ with eigenvector x_i of length one. Then, in the LQG-balanced realization, the optimal cost with initial condition x_i is λ_i and the difficulty of reconstructing this initial state from noisy measurements is also λ_i. The idea behind LQG-balanced truncation is to restrict the system to the subspace generated by the eigenvectors corresponding to the largest eigenvalues. Since this subspace is most important for compensator design, the system obtained by LQG-balanced truncation seems to be a reasonable approximation. There is a bound on the distance between a plant and an LQG-balanced truncation of the plant in terms of the discarded LQG-characteristic values; see Mustafa and Glover [8, section 8.4.5].

The existence of LQG-balanced realizations in the finite-dimensional case is easily proved as follows:

1. Start with a minimal realization (A, B, C, D) and compute the solutions Q and P of the Riccati equations.
2. Write $P^{1/2}QP^{1/2} = U\Lambda^2U^*$ with Λ diagonal.
3. Let $T := Q^{1/2}U\Lambda^{-1/2}$.

Then it is easily seen that $(TAT^{-1}, TB, CT^{-1}, D)$ is an LQG-balanced realization.

In the infinite-dimensional case this proof no longer works. The main problem is that the singular value decomposition performed at Step 2 cannot always be made in the infinite-dimensional case. (One has to assume a compactness condition.) Even if it can, then usually the singular values form a sequence with zero as limit point and the operator $\Lambda^{-1/2}$ mentioned at the third step is unbounded. Because of this unboundedness, it is unclear whether the expressions TAT^{-1} and TB make sense. To avoid these problems we take a different approach, mentioned in the introduction: we use the known result for the existence of Lyapunov-balanced realizations. In this article we consider only the existence of LQG-balanced realizations. The study of the properties of truncated LQG-balanced realizations will be done elsewhere. In slight contrast with the definition above we will call a realization LQG-balanced if $P = Q$. We do not require that they are diagonal, since this is not always possible in infinite dimensions.

3. Discrete-time infinite-dimensional systems. In this section we review that part of the theory of discrete-time infinite-dimensional systems that we need in

this article. Discrete-time infinite-dimensional systems have been treated in a number of texts (e.g., [1], [3], [12]). However, the standard treatments of the linear quadratic theory assume the strong concept of power stabilizability, i.e., the existence of an F such that $\|(A+BF)^n\| \leq M\lambda^n$ for some constants $M > 0, 0 < \lambda < 1$ and all positive integers n. Unfortunately, this concept is not suitable for a nice theory of LQG-balanced realizations. Thus in this section we reexamine the basic concepts under weaker stabilizability assumptions.

A discrete-time infinite-dimensional system or simply a system is a quadruple of bounded operators $(A, B, C, D) \in \mathcal{L}(X) \times \mathcal{L}(U, X) \times \mathcal{L}(X, Y) \times \mathcal{L}(U, Y)$, where X, U, Y are separable Hilbert spaces. For an input u and initial condition x_0, the state x and output y of the system are defined by

$$x_{n+1} = Ax_n + Bu_n \quad x(0) = x_0 \quad y_n = Cx_n + Du_n. \tag{3.1}$$

The observability map $\mathcal{C}$ of a discrete-time system (A, B, C, D) is defined by

$$(\mathcal{C}x)_i := CA^i x, \quad i \in \mathbb{N}. \tag{3.2}$$

The discrete-time system (A, B, C, D) is said to be approximately observable if $\ker \mathcal{C} = \{0\}$ (see Curtain and Zwart [1]). There are many generalizations of the finite-dimensional concept of observability to an infinite-dimensional setting, and the concept of approximate observability is one. Our main use of approximate observability is that realizations may, without loss of generality, be assumed to be approximately controllable and observable. This may not always be the case for other generalizations of observability. Here, approximately controllable and observable plays the role that minimal plays in finite dimensions.

The discrete-time system (A, B, C, D) is said to be output stable if the image of $\mathcal{C}$ is contained in $l^2(\mathbb{N}; Y)$. The observability gramian L_C of an output stable system is defined as $L_C := \mathcal{C}^*\mathcal{C}$. We now give an alternative characterization of output stability in terms of solutions of a certain Lyapunov equation.

LEMMA 3.1. *Let (A, B, C, D) be a discrete-time system. The following are equivalent statements:*

1. *The system is output stable.*
2. *The* observation Lyapunov equation

$$A^*LA - L + C^*C = 0 \tag{3.3}$$

has a nonnegative self-adjoint solution.

If one (and hence both) of these hold, then the observability gramian is the smallest nonnegative self-adjoint solution of the observation Lyapunov equation.

Proof. We first show that 1 implies 2. It is easily seen that the observability gramian is given by the formula $L_C = \sum_{i=0}^{\infty} A^{*i}C^*CA^i$, and substituting this into the observation Lyapunov equation shows that it is a solution.

That 2 implies 1 is proved as follows. Suppose that the observation Lyapunov equation has a nonnegative self-adjoint solution L. Then multiplying (3.3) from the left with A^{*n} and from the right with A^n and summing from $n = 0$ to N gives

$$\begin{aligned}\sum_{n=0}^{N} A^{*n}C^*CA^n &= \sum_{n=0}^{N} A^{*n}LA^n - \sum_{n=0}^{N} A^{*n+1}LA^{n+1} \\ &= L - A^{*N+1}LA^{N+1} \leq L.\end{aligned}$$

Letting $N \to \infty$ shows that L_C is a bounded map and so $\mathcal{C}$ is bounded. That L_C is smaller than any other nonnegative self-adjoint solution of the observation Lyapunov equation is obvious from the above inequality. ∎

The following result shows that strong stability implies the uniqueness of solutions of Lyapunov equations. We remind the reader that an operator A is called strongly (or asymptotically) stable if for all $x \in X$ we have $A^n x \to 0$ as $n \to \infty$.

LEMMA 3.2. *Let (A, B, C, D) be output stable and let A be strongly stable. Then the observability gramian is the unique nonnegative self-adjoint solution of the observation Lyapunov equation* (3.3).

Proof. According to Lemma 3.1 the observability gramian is a nonnegative self-adjoint solution of the observation Lyapunov equation, so we only have to show that it is the unique nonnegative self-adjoint solution. Let L be a nonnegative self-adjoint solution of the observation Lyapunov equation. Then, as in the proof of Lemma 3.1, we have for all $N \in \mathbb{N}$

$$\sum_{n=0}^{N} A^{*n} C^* C A^n = \sum_{n=0}^{N} A^{*n} L A^n - \sum_{n=0}^{N} A^{*n+1} L A^{n+1} = L - A^{*N+1} L A^{N+1}.$$

We then have for all $x, y \in X$

$$\left\langle \sum_{n=0}^{N} A^{*n} C^* C A^n x, y \right\rangle = \langle Lx, y\rangle - \langle LA^{N+1}x, A^{N+1}y\rangle.$$

Letting $N \to \infty$ and using that A is strongly stable, we have for all $x, y \in X$

$$\langle L_C x, y\rangle = \langle Lx, y\rangle.$$

This implies that $L = L_C$. Since L was an arbitrary nonnegative self-adjoint solution, this implies that L_C is the unique nonnegative self-adjoint solution of the observation Lyapunov equation. ∎

A discrete-time system (A, B, C, D) is called output stabilizable if there exists an $F \in \mathcal{L}(X, U)$ such that $(A + BF, 0, [F; C + DF], 0)$ is output stable. (Note that we use the notation $[-; -]$ for a block column vector and $[-, -]$ for a block row vector.) Output stabilizability is a necessary and sufficient condition for the solvability of the LQR problem. To show this we first review some well-known results on the LQR problem in infinite dimensions. For a system (A, B, C, D) with input, state, and output related by (3.1), we consider the cost functional

$$J(x_0, u) := \sum_{n=0}^{\infty} \|u_n\|^2 + \|y_n\|^2.$$

The well-known linear quadratic regulator problem is as follows. Find a sequence $u^{\min}$ such that $J(x_0, u^{\min}) \le J(x_0, u)$ for all sequences u. A system is said to satisfy the finite cost condition if for every initial state x_0 there exists a u such that $J(x_0, u) < \infty$. Just as in the finite-dimensional case one can prove that if the finite cost condition is satisfied, then there exists a unique optimal control $u^{\min}$; we actually have $J(x_0, u^{\min}) < J(x_0, u)$ for all other sequences u and $J(x_0, u^{\min}) = \langle x_0, Qx_0\rangle$, where Q is the smallest nonnegative self-adjoint solution of the control algebraic Riccati equation (CARE) associated with the system (A, B, C, D),

$$A^*QA - Q + C^*C = (C^*D + A^*QB)(S + B^*QB)^{-1}(D^*C + B^*QA), \tag{3.4}$$

where $S := I + D^*D$. Moreover, $u^{\min}$ can be given by a state feedback. All of this can be found, for example, in Curtain and Zwart [1, Exercise 6.34]. The operator Q is called the optimal cost operator. The following lemma gives conditions that are equivalent to the finite cost condition.

LEMMA 3.3. *The following statements about a discrete-time system* (A, B, C, D) *are equivalent:*

1. *The discrete-time system is output stabilizable.*
2. *The discrete-time system satisfies the finite cost condition.*
3. *The CARE* (3.4) *of the discrete-time system has a nonnegative self-adjoint solution.*

Proof. Suppose the discrete-time system is output stabilizable. Then there exists an F such that $(A + BF, 0, [F; C + DF], 0)$ is output stable; denote the observability map of this system by $\mathcal{C}_F$. Add the equation $u := Fx$ to (3.1). Call the solution (u, x, y) of this set of equations $\bar{u}, \bar{x}, \bar{y}$. Then $[\bar{u}, \bar{y}] = \mathcal{C}_F x_0$ and since $\mathcal{C}_F$ is bounded we see that $[\bar{u}, \bar{y}]$ has finite l^2 norm. That is, $J(x_0, \bar{u}) < \infty$ and the system satisfies the finite cost condition.

An outline of the proof that 2 implies 3 can be found in Curtain and Zwart [1, Exercise 6.34].

For 3 implies 1, we will show that the feedback $F := -(B^*QB+I+D^*D)^{-1}(D^*C+B^*QA)$, where Q is a solution of the CARE (3.4), is output stabilizing. We will do this by showing that Q is a solution of the observation Lyapunov equation of the system $(A + BF, 0, [F; C + DF], 0)$. We want to show that Q satisfies

$$(A + BF)^*Q(A + BF) - Q + [F^*, C^* + F^*D^*] \begin{bmatrix} F \\ C + DF \end{bmatrix} = 0.$$

This is equivalent to

$$\begin{aligned} A^*QA - Q + C^*C + F^*(B^*QB + I + D^*D)F \\ + F^*(B^*QA + D^*C) + (A^*QB + C^*D)F = 0. \end{aligned} \tag{3.5}$$

Substituting for F in (3.5) we obtain

$$A^*QA - Q + C^*C = (C^*D + A^*QB)(S + B^*QB)^{-1}(D^*C + B^*QA),$$

which is precisely CARE (3.4). ∎

The controllability map $\mathcal{B}$ of a discrete-time system (A, B, C, D) is defined for finitely nonzero U-valued sequences u by

$$\mathcal{B}u := \sum_{i=0}^{\infty} A^i B u_{-i-1}. \tag{3.6}$$

The discrete-time system (A, B, C, D) is said to be approximately controllable if $\ker \mathcal{B}^* = \{0\}$. The discrete-time system (A, B, C, D) is said to be input stable if $\mathcal{B}$ extends to a bounded map from $l^2(\mathbb{Z}^-; U)$ to X. The controllability gramian L_B of an input stable system is defined as $L_B := \mathcal{B}\mathcal{B}^*$. A discrete-time system (A, B, C, D) is called input stabilizable if there exists an $L \in \mathcal{L}(Y, X)$ such that $(A + LC, [L, B + LD], 0, 0)$ is input stable. The following dual results of the results proven earlier hold.

LEMMA 3.4. *Let* (A, B, C, D) *be a discrete-time system. The following are equivalent statements:*

1. *The system is input stable.*
2. *The* control Lyapunov equation

$$ALA^* - L + BB^* = 0 \tag{3.7}$$

has a nonnegative self-adjoint solution.

If one (and hence both) of the above holds, then the controllability gramian is the smallest nonnegative self-adjoint solution of the control Lyapunov equation.

LEMMA 3.5. *Let (A, B, C, D) be input stable and let A^* be strongly stable. Then the controllability gramian is the unique nonnegative self-adjoint solution of the control Lyapunov equation* (3.7).

LEMMA 3.6. *The following statements about a discrete-time system (A, B, C, D) are equivalent:*

1. *The discrete-time system is input stabilizable.*
2. *The dual system (A^*, C^*, B^*, D^*) satisfies the finite cost condition.*
3. *The filter algebraic Riccati equation (FARE)*

$$APA^* - P + BB^* = (BD^* + APC^*)(R + CPC^*)^{-1}(DB^* + CPA^*), \tag{3.8}$$

where $R := I + DD^*$, of the discrete-time system has a nonnegative self-adjoint solution.

We now give a condition under which CARE (3.4) has a unique nonnegative self-adjoint solution.

LEMMA 3.7. *Let (A, B, C, D) be an input and output stabilizable discrete-time system. Let Q be a nonnegative self-adjoint solution of the CARE* (3.4)*, and assume that $A_Q := A - B(S + B^*QB)^{-1}(D^*C + B^*QA)$ is strongly stable. Then Q is the unique nonnegative self-adjoint solution of CARE* (3.4).

Proof. For the proof we need the following algebraic relations, which are proved in the appendix (Lemmas 10.3 and 10.4). Suppose Q_1 and P_1 are nonnegative self-adjoint solutions of the CARE and FARE, respectively, and define A_{Q_1} similar to A_Q above and $A_{P_1} := A - (BD^* + AP_1C^*)(R + CP_1C^*)^{-1}C$. Then the following relation holds:

$$(I + P_1Q_1)A_{Q_1} = A_{P_1}(I + P_1Q_1). \tag{3.9}$$

The following algebraic relation is also proven in the appendix (Lemma 10.4). If Q_1 and Q_2 are nonnegative self-adjoint solutions of the CARE, and if A_{Q_1} and A_{Q_2} are defined similarly as A_Q above, then

$$Q_1 - Q_2 = A_{Q_2}^*(Q_1 - Q_2)A_{Q_1}. \tag{3.10}$$

With induction it follows that for all $n \in \mathbb{N}$ we have

$$Q_1 - Q_2 = A_{Q_2}^{*n}(Q_1 - Q_2)A_{Q_1}^n. \tag{3.11}$$

Using these facts we now prove the statement. Since (A, B, C, D) is input stabilizable, there exists a nonnegative self-adjoint solution P of the FARE (3.8). Since A_Q is assumed to be strongly stable and (3.9) shows that A_P is similar to A_Q, we have that A_P is strongly stable. Now let $\tilde{Q}$ be an arbitrary nonnegative self-adjoint solution of the CARE. According to (3.9), $A_{\tilde{Q}}$ is similar to the strongly stable operator A_P and hence is strongly stable. Since $A_{\tilde{Q}}$ is strongly stable there exists for every $x \in X$ a real number c_x such that for every $n \in \mathbb{N}$ we have $\|A_{\tilde{Q}}^n x\| \leq c_x$. By the uniform

boundedness theorem this implies that there exists a real number c such that for every $n \in \mathbb{N}$ we have $\|A_{\tilde{Q}}^n\| \leq c$. Using (3.11) with $Q_1 = Q$ and $Q_2 = \tilde{Q}$ we have for all $x \in X$ and $n \in \mathbb{N}$

$$\|(Q - \tilde{Q})x\| = \|A_{\tilde{Q}}^{*n}(Q - \tilde{Q})A_Q^n x\| \leq \|A_{\tilde{Q}}^{*n}\| \, \|Q - \tilde{Q}\| \, \|A_Q^n x\| \leq c \, \|Q - \tilde{Q}\| \, \|A_Q^n x\|.$$

Since A_Q is strongly stable, the right-hand side converges to zero as $n \to \infty$. This implies that the left-hand side is zero and so $\tilde{Q} = Q$. □

The input-output map $\mathcal{D}$ of a discrete-time system (A, B, C, D) is defined for finitely nonzero U-valued sequences u by

$$(\mathcal{D}u)_k := \sum_{i=0}^{\infty} CA^i Bu_{k-i-1} + Du_k. \tag{3.12}$$

The discrete-time system (A, B, C, D) is said to be input-output stable if $\mathcal{D}$ extends to a bounded map from $l^2(\mathbb{Z}; U)$ to $l^2(\mathbb{Z}; Y)$.

We define the transfer function G of a system (A, B, C, D) by

$$G(z) = D + \sum_{i=0}^{\infty} CA^i Bz^{i+1} \tag{3.13}$$

for those z for which the sum converges absolutely. Note that it converges absolutely for $|z| < 1/r(A)$ ($r(A)$ denotes the spectral radius of A) and it is equal to $D + Cz(I - zA)^{-1}B$ for those z. It is obvious that the transfer function of a system can be constructed from the input-output map and vice versa; in this sense transfer functions and input-output maps are equivalent notions. Given a transfer function G we call any system (A, B, C, D) such that (3.13) holds a realization of the transfer function. We note that the functions that appear as transfer functions of discrete-time infinite-dimensional systems are exactly the operator-valued functions that are analytic on some disc centered at the origin.

The (time-domain) Hankel operator Γ of a system is defined as $\Gamma := \mathcal{CB}$, where $\mathcal{C}$ and $\mathcal{B}$ are the observability and controllability maps of the system, respectively. It is easily seen that the Hankel operator does not depend on the particular realization but only on the input-output map.

4. Lyapunov-balanced realizations. In this section we review some results from Young [15], [16] and Ober and Wu [10] and translate them in terms more suitable for our purposes. The following result on the existence of Lyapunov-balanced realizations was proved by Young [15], [16]. We recall that an input and output stable system is called Lyapunov balanced if its controllability and observability gramians are equal (again, we do not require them to be diagonal).

LEMMA 4.1. *Every transfer function that has a bounded Hankel operator has an approximately controllable and observable Lyapunov-balanced realization. Moreover, approximately controllable and observable Lyapunov-balanced realizations are unique up to a unitary transformation.*

The next corollary gives an alternative condition for the existence of Lyapunov-balanced realizations.

COROLLARY 4.2. *A transfer function has a Lyapunov-balanced realization if and only if it has a realization that is both input and output stable.*

Proof. Since a Lyapunov-balanced realization is input and output stable, one implication is immediate. If the transfer function has a realization such that both

its controllability map $\mathcal{B}$ and observability map $\mathcal{C}$ are bounded, then its Hankel operator $\Gamma = \mathcal{C}\mathcal{B}$ is bounded and Lemma 4.1 shows that it has a Lyapunov-balanced realization. ∎

The following result was proved by Ober and Wu [10].

LEMMA 4.3. *Let (A, B, C, D) be Lyapunov-balanced and approximately controllable and approximately observable. Then A and A^* are strongly stable.*

Combining Lemmas 4.3, 3.2, and 3.5 we have the following corollary.

COROLLARY 4.4. *The gramian of an approximately controllable and approximately observable Lyapunov-balanced realization is the unique nonnegative self-adjoint solution of both the control and the observation Lyapunov equation.*

5. Normalized factorizations. In this section we generalize a result of Meyer and Franklin [6] on the connection between normalized factorizations and linear quadratic regulator theory to the infinite-dimensional case. This result will allow us to relate LQG-balanced realizations to Lyapunov-balanced realizations of a normalized factorization of the given transfer function.

Given an output stabilizable discrete-time system (A, B, C, D) with optimal cost operator Q we form the optimal closed-loop system

$$\check{A} := A + BF, \quad \check{B} := BW^{-1/2}, \quad \check{C} := [F; C + DF], \quad \check{D} := [I; D]W^{-1/2}, \tag{5.1}$$

where

$$W := S + B^*QB, \quad S := I + D^*D, \quad F := -W^{-1}(D^*C + B^*QA).$$

We first remark that the F above is the optimal state feedback operator for the LQR problem. We obtain the optimal closed-loop system from the system (A, B, C, D) by choosing $u = Fx + W^{-1/2}\check{u}$ and considering $\check{u}$ as the input of this new system and $[u; y]$ as the output. This amounts to closing the loop by the optimal state feedback operator, considering the input and output of the plant as the new output and prefiltering the new input.

Our first result in this section states that the optimal cost operator of the plant equals the observability gramian of the optimal closed-loop system.

LEMMA 5.1. *Let (A, B, C, D) be an output stabilizable discrete-time system. Denote its optimal cost operator by Q and define its optimal closed-loop system by* (5.1). *Denote the observability gramian of the optimal closed-loop system by L_C. Then $Q = L_C$.*

Proof. From the discussion above it is obvious that if $\check{u} = 0$, then the output of the optimal closed-loop system is $[u^{\min}; y^{\min}]$, the optimal input and output of the plant. From this it follows that

$$\langle L_C x_0, x_0\rangle = \langle \mathcal{C}x_0, \mathcal{C}x_0\rangle = \| \begin{bmatrix} u^{\min} \\ y^{\min} \end{bmatrix} \|^2 = J(x_0, u^{\min}) = \langle Qx_0, x_0\rangle.$$

Since this holds for all x_0 in the state space we have $Q = L_C$. ∎

The next lemma shows that the observability gramian of the optimal closed-loop system satisfies two additional equations.

LEMMA 5.2. *Let (A, B, C, D) be an output stabilizable discrete-time system. Denote its optimal cost operator by Q and define its optimal closed-loop system $(\check{A}, \check{B}, \check{C}, \check{D})$ by* (5.1). *Denote the observability gramian of the optimal closed-loop system by L_C. Then*

$$\check{B}^* L_C \check{B} + \check{D}^* \check{D} = I, \tag{5.2}$$

$$\check{B}^* L_C \check{A} + \check{D}^* \check{C} = 0. \tag{5.3}$$

Proof. The equations (5.2) and (5.3) are readily verified using (5.1) and the fact that $Q = L_C$ from Lemma 5.1. □

An input-output map $\mathcal{D}$ is called inner if it maps $l^2(\mathbb{Z}; U)$ into $l^2(\mathbb{Z}; Y)$ and satisfies $\mathcal{D}^*\mathcal{D} = I$. In the next lemma we give necessary and sufficient conditions on a realization for the input-output map to be inner.

LEMMA 5.3. *Let (A, B, C, D) be an output stable realization of the input-output map $\mathcal{D}$. Denote the observability gramian of this system by L_C. If*

$$B^* L_C B + D^* D = I$$

and

$$B^* L_C A + D^* C = 0,$$

then $\mathcal{D}$ is inner. If the system (A, B, C, D) is approximately controllable, then these conditions are also necessary.

Proof. We take u_k equal to u at the kth position and zero elsewhere and compute

$$(\mathcal{D}u_k)_m = \begin{cases} 0, & m < k, \\ Du, & m = k, \\ CA^{m-k-1}Bu, & m > k. \end{cases}$$

We define v_i similar to u_k above and compute for $k > i$

$$\langle (\mathcal{D}u_k)_n, (\mathcal{D}v_i)_n \rangle = \begin{cases} 0, & n < k, \\ \langle Du, CA^{k-i-1}Bv \rangle, & n = k, \\ \langle CA^{n-k-1}Bu, CA^{n-i-1}Bv \rangle, & n > k. \end{cases}$$

We then compute

$$\begin{aligned} \langle \mathcal{D}u_k, \mathcal{D}v_i \rangle &= \sum_{n=-\infty}^{\infty} \langle (\mathcal{D}u_k)_n, (\mathcal{D}v_i)_n \rangle \\ &= \langle Du, CA^{k-i-1}Bv \rangle + \sum_{n=k+1}^{\infty} \langle CA^{n-k-1}Bu, CA^{n-i-1}Bv \rangle \\ &= \langle u, D^*CA^{k-i-1}Bv \rangle + \sum_{n=k+1}^{\infty} \langle u, B^*A^{*n-k-1}C^*CA^{n-k-1}AA^{k-i-1}Bv \rangle \\ &= \langle u, (D^*C + B^*L_CA)A^{k-i-1}Bv \rangle. \end{aligned}$$

By the assumptions we thus have $\langle \mathcal{D}u_k, \mathcal{D}v_i \rangle = 0$ for $k > i$. Then obviously $\langle \mathcal{D}u_k, \mathcal{D}v_i \rangle = 0$ for $k \neq i$.

For $k = i$ we have

$$\begin{aligned} \langle \mathcal{D}u_k, \mathcal{D}v_i \rangle &= \sum_{n=-\infty}^{\infty} \langle (\mathcal{D}u_k)_n, (\mathcal{D}v_i)_n \rangle \\ &= \langle Du, Dv \rangle + \sum_{n=k+1}^{\infty} \langle CA^{n-k-1}Bu, CA^{n-k-1}Bv \rangle \\ &= \langle D^*Du, v \rangle + \sum_{j=0}^{\infty} \langle B^*A^{*j}C^*CA^jBu, v \rangle = \langle (D^*D + B^*L_CB)u, v \rangle. \end{aligned}$$

By the assumptions we thus have $\langle \mathcal{D}u_k, \mathcal{D}v_i\rangle = \langle u, v\rangle$ for $k = i$.

Let u and v be finitely nonzero sequences. Then

$$u = \sum_{k=-n}^{n} u^k e_k, \quad v = \sum_{i=-n}^{n} v^i e_i,$$

where $u^k, v^i \in U$ and e_j is the element of $l^2(\mathbb{Z}; U)$ with a 1 at the jth position and zeros elsewhere. Then

$$\begin{aligned}\langle \mathcal{D}u, \mathcal{D}v\rangle &= \sum_{k=-n}^{n}\sum_{i=-n}^{n} \langle \mathcal{D}(u^k e_k), \mathcal{D}(v^i e_i)\rangle = \sum_{j=-n}^{n} \langle \mathcal{D}(u^j e_j), \mathcal{D}(v^j e_j)\rangle \\ &= \sum_{j=-n}^{n} \langle u^j, v^j\rangle = \langle u, v\rangle.\end{aligned}$$

From the above we have for every finitely nonzero sequence u that $\|\mathcal{D}u\| = \|u\|$. Since the set of finitely nonzero sequences is dense in $l^2(\mathbb{Z}; U)$ this implies that $\mathcal{D}$ has a continuous extension to a map from $l^2(\mathbb{Z}; U)$ to $l^2(\mathbb{Z}; Y)$, and so $\mathcal{D}$ is stable. Further, since this extension satisfies $\langle \mathcal{D}u, \mathcal{D}v\rangle = \langle u, v\rangle$, we must have $\mathcal{D}^*\mathcal{D} = I$, and so $\mathcal{D}$ is inner.

Suppose that $\mathcal{D}$ is inner and the realization is approximately controllable. Since $\mathcal{D}$ is inner we have for all $i < 0$ that $\langle \mathcal{D}u_0, \mathcal{D}v_i\rangle = \langle u_0, v_i\rangle = 0$, where u_0 and v_i are defined as above. From the above we see that this implies that $\langle u, (D^*C + B^*L_CA)A^{-i-1}Bv\rangle = 0$. Since this holds for all $u \in U$ we must have $(D^*C + B^*L_CA)A^{-i-1}Bv = 0$ for all $v \in U$ and $i < 0$. This implies that for finitely nonzero U-valued sequences z we have $(D^*C + B^*L_CA)\mathcal{B}z = 0$. Since the system is approximately controllable, the set of elements of the form $\mathcal{B}z$ is dense in the state space, and so $D^*C + B^*L_CA = 0$ on a dense set and by continuity on the whole state space.

Since $\mathcal{D}$ is inner we have $\langle \mathcal{D}u_0, \mathcal{D}v_0\rangle = \langle u_0, v_0\rangle$, where u_0 and v_0 are as above. This implies that $\langle (D^*D + B^*L_CB)u, v\rangle = \langle u, v\rangle$ for all $u, v \in U$ and hence $D^*D + B^*L_CB = I$. ∎

Combining Lemmas 5.2 and 5.3 we have the following.

COROLLARY 5.4. *Let (A, B, C, D) be an output stabilizable discrete-time system. Then the input-output map of its optimal closed-loop system is inner.*

We can recover the system (A, B, C, D) from its optimal closed-loop system as follows.

LEMMA 5.5. *Let (A, B, C, D) be an output stabilizable discrete-time system. Let $(\check{A}, \check{B}, \check{C}, \check{D})$ be its optimal closed-loop system defined by* (5.1). *Partition $\check{C}$ and $\check{D}$ in the obvious way as $\check{C} = [\check{C}_1; \check{C}_2]$ and $\check{D} = [\check{D}_1; \check{D}_2]$. Then $\check{D}_1$ is boundedly invertible and*

$$(5.4)\quad A := \check{A} - \check{B}\check{D}_1^{-1}\check{C}_1, \quad B := \check{B}\check{D}_1^{-1}, \quad C := \check{C}_2 - \check{D}_2\check{D}_1^{-1}\check{C}_1, \quad D := \check{D}_2\check{D}_1^{-1}.$$

Proof. That $\check{D}_1$ is boundedly invertible is obvious from its definition. The identities (5.4) follow from simple algebraic manipulations. ∎

To relate the input-output maps of the plant and its optimal closed-loop system we first study the series interconnection of two systems.

Consider two systems (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) such that the output space of the first system and the input space of the second system are equal. Define the series interconnection of these two systems as the system we obtain by choosing

the input of the second system equal to the output of the first system. Obviously, the input-output map of the series interconnection is $\mathcal{D}_2\mathcal{D}_1$, the composition of the input-output maps of the first and second systems. A realization of this series interconnection is the following:

$$A = \begin{bmatrix} A_1 & 0 \\ B_2C_1 & A_2 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2D_1 \end{bmatrix}, \quad C = \begin{bmatrix} D_2C_1, & C_2 \end{bmatrix}, \quad D = D_2D_1.$$

If we apply the invertible state space transformation

$$\begin{bmatrix} I & 0 \\ I & I \end{bmatrix}$$

to this realization we obtain another realization of the series interconnection, namely,

$$A_s = \begin{bmatrix} A_1 & 0 \\ A_2 + B_2C_1 - A_1 & A_2 \end{bmatrix}, \quad B_s = \begin{bmatrix} B_1 \\ B_2D_1 - B_1 \end{bmatrix}, \tag{5.5}$$

$$C_s = \begin{bmatrix} D_2C_1 + C_2, & C_2 \end{bmatrix}, \quad D_s = D_2D_1.$$

We first use the series interconnection to obtain a result about the invertibility of input-output maps.

LEMMA 5.6. *Let (A, B, C, D) be a system with input-output map $\mathcal{D}$. If D is boundedly invertible, then the input-output map $\tilde{\mathcal{D}}$ of the system*

$$(A - BD^{-1}C, BD^{-1}, -D^{-1}C, D^{-1})$$

satisfies $\tilde{\mathcal{D}}\mathcal{D} = I = \mathcal{D}\tilde{\mathcal{D}}$. Thus in this case the input-output map $\mathcal{D}$ has an inverse.

Proof. Using (5.5) we see that the series interconnection of the two given systems has a realization

$$A_s = \begin{bmatrix} A & 0 \\ 0 & A - BD^{-1}C \end{bmatrix}, \quad B_s = \begin{bmatrix} B \\ 0 \end{bmatrix}, \quad C_s = \begin{bmatrix} 0, & -D^{-1}C \end{bmatrix}, \quad D_s = I.$$

From this we see that $C_sA_s^kB_s = 0$ for all $k \geq 0$ and so the input-output map of the series interconnection is the identity. This implies that $\tilde{\mathcal{D}}\mathcal{D} = I$. The other equality mentioned follows from interconnecting the systems in the opposite order. ∎

We now state the relation between the input-output map of a plant and its optimal closed-loop system.

LEMMA 5.7. *Let $(\check{A}, \check{B}, [\check{C}_1; \check{C}_2], [\check{D}_1; \check{D}_2])$ be a system such that $\check{D}_1$ is boundedly invertible. Denote its input-output map by $[\mathcal{M}; \mathcal{N}]$. Define the system (A, B, C, D) by* (5.4) *and denote its input-output map by $\mathcal{D}$. Then $\mathcal{D} = \mathcal{N}\mathcal{M}^{-1}$.*

Proof. The realization $(\check{A}, \check{B}, [\check{C}_1; \check{C}_2], [\check{D}_1; \check{D}_2])$ of $[\mathcal{M}; \mathcal{N}]$ gives us (by Lemma 5.6) the following realization (A_1, B_1, C_1, D_1) of $\mathcal{M}^{-1}$:

$$A_1 = \check{A} - \check{B}\check{D}_1^{-1}\check{C}_1, \quad B_1 = \check{B}\check{D}_1^{-1}, \quad C_1 = -\check{D}_1^{-1}\check{C}_1, \quad D_1 = \check{D}_1^{-1}.$$

It also gives us the realization (A_2, B_2, C_2, D_2) of $\mathcal{N}$:

$$A_2 = \check{A}, \quad B_2 = \check{B}, \quad C_2 = \check{C}_2, \quad D_2 = \check{D}_2.$$

Using (5.5) we obtain the following realization of the series interconnection which has input-output map $\mathcal{N}\mathcal{M}^{-1}$:

$$A_s = \begin{bmatrix} \check{A} - \check{B}\check{D}_1^{-1}\check{C}_1 & 0 \\ 0 & \check{A} \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & \check{A} \end{bmatrix}, \quad B_s = \begin{bmatrix} \check{B}\check{D}_1^{-1} \\ 0 \end{bmatrix} = \begin{bmatrix} B \\ 0 \end{bmatrix},$$

$$C_s = \begin{bmatrix} \check{C}_2 - \check{D}_2\check{D}_1^{-1}\check{C}_1, & \check{C}_2 \end{bmatrix} = \begin{bmatrix} C, & \check{C}_2 \end{bmatrix}, \quad D_s = \check{D}_2\check{D}_1^{-1} = D,$$

where we have used (5.4). It follows that $D_s = D$ and $C_s A_s^k B_s = CA^k B$ for all $k \geq 0$. This implies that $\mathcal{N}\mathcal{M}^{-1} = \mathcal{D}$. ∎

We call an input-output map $[\mathcal{M};\mathcal{N}]$ a right factor of the input-output map $\mathcal{D}$ if $\mathcal{M}^{-1}$ is the input-output map of a system, $\mathcal{D} = \mathcal{N}\mathcal{M}^{-1}$ and $\mathcal{M}$ and $\mathcal{N}$ are stable. We call the factor normalized if $[\mathcal{M};\mathcal{N}]$ is inner. From Lemmas 5.5 and 5.7 and Corollary 5.4 we have the following.

COROLLARY 5.8. *Let (A, B, C, D) be an output stabilizable discrete-time system. Then the input-output map of its optimal closed-loop system is a normalized right factor of the input-output map of the plant.*

We next state a result about the uniqueness of a normalized right factor. We remark that we can interpret an operator V in $\mathcal{L}(U)$ as a map $\mathcal{V}$ from $l^2(\mathbb{Z};U)$ into itself by $(\mathcal{V}u)_k = Vu_k$.

LEMMA 5.9. *If the input-output map of a system has a normalized right factor $[\mathcal{M};\mathcal{N}]$, then all normalized right factors of this input-output map are $\{[\mathcal{M}V;\mathcal{N}V] : V \in \mathcal{L}(U)$ unitary$\}$.*

Proof. Let $[\mathcal{M};\mathcal{N}]$ be an arbitrary normalized right factor of $\mathcal{D}$. Since $[\mathcal{M};\mathcal{N}]$ is normalized we have

$$\mathcal{M}^*\mathcal{M} + \mathcal{N}^*\mathcal{N} = I.$$

Multiplying this equality with $\mathcal{M}^{-*}$ from the left and $\mathcal{M}^{-1}$ from the right we obtain

$$I + \mathcal{D}^*\mathcal{D} = \mathcal{M}^{-*}\mathcal{M}^{-1}.$$

Since the left-hand side of this equation does not depend on the particular factor, we have for two normalized factors $[\mathcal{M}_1;\mathcal{N}_1]$ and $[\mathcal{M}_2;\mathcal{N}_2]$ of $\mathcal{D}$ that $\mathcal{M}_1^{-*}\mathcal{M}_1^{-1} = \mathcal{M}_2^{-*}\mathcal{M}_2^{-1}$, which implies

$$\mathcal{M}_2^*\mathcal{M}_1^{-*} = \mathcal{M}_2^{-1}\mathcal{M}_1. \tag{5.6}$$

The operator on the right-hand side of (5.6) is the input-output map of some system (namely, the series interconnection of the systems corresponding to $\mathcal{M}_1$ and $\mathcal{M}_2^{-1}$). Define u_k to be the sequence equal to u at the kth position and zero elsewhere. Then $\mathcal{M}_2^{-1}\mathcal{M}_1 u_k$ is equal to zero at the positions i with $i < k$. The operator on the left-hand side of (5.6) is the adjoint of the input-output map of some system (namely, the series interconnection of the systems corresponding to $\mathcal{M}_2$ and $\mathcal{M}_1^{-1}$). This implies that $\mathcal{M}_2^*\mathcal{M}_1^{-*}u_k$ is zero at the positions i with $i > k$.

Since $\mathcal{M}_2^*\mathcal{M}_1^{-*} = \mathcal{M}_2^{-1}\mathcal{M}_1$ we must have that $\mathcal{M}_2^{-1}\mathcal{M}_1 u_k$ is equal to zero at all positions except possibly the kth one. Thus $\mathcal{M}_2^{-1}\mathcal{M}_1$ is a constant operator V, and so $\mathcal{M}_1 = \mathcal{M}_2 V$.

We have $\mathcal{N}_1 = \mathcal{D}\mathcal{M}_1 = \mathcal{D}\mathcal{M}_2 V = \mathcal{N}_2 V$.

It remains to be proved that V is unitary. We have that $V = \mathcal{M}_2^{-1}\mathcal{M}_1$ and so $V^* = \mathcal{M}_1^*\mathcal{M}_2^{-*} = \mathcal{M}_1^{-1}\mathcal{M}_2 = V^{-1}$ by (5.6). This proves that V is unitary. ∎

In the finite-dimensional case the transfer function of the optimal closed-loop system is known to be a normalized coprime factorization. In the infinite-dimensional case this is an open problem.

6. Some algebraic relations. In section 5 we proved that the optimal cost operator equals the observability gramian of the optimal closed-loop system. This can also be interpreted as follows. The smallest nonnegative self-adjoint solution of

the control algebraic Riccati equation equals the smallest nonnegative self-adjoint solution of the observation Lyapunov equation of a certain closed-loop system. In this section we show that a similar result holds for all nonnegative self-adjoint solutions of the control algebraic Riccati equation. We also study the relation between nonnegative self-adjoint solutions of the filter algebraic Riccati equation of the plant and nonnegative self-adjoint solutions of the control Lyapunov equation of the closed-loop system.

The CARE closed-loop system $(\check{A}, \check{B}, \check{C}, \check{D})$ associated with an output stabilizable discrete-time system (A, B, C, D) and a nonnegative self-adjoint solution Q of the CARE (3.4) is defined by (5.1). For the special case that Q is the smallest nonnegative self-adjoint solution of the CARE (3.4) the CARE closed-loop system is equal to the optimal closed-loop system defined earlier.

Lemma 5.5 holds in this more general case, which is obvious from its proof.

LEMMA 6.1. *Let (A, B, C, D) be a discrete-time system such that its CARE* (3.4) *has a nonnegative self-adjoint solution Q. Let $(\check{A}, \check{B}, \check{C}, \check{D})$ be its CARE closed-loop system defined by* (5.1). *Partition $\check{C}$ and $\check{D}$ in the obvious way as $\check{C} = [\check{C}_1; \check{C}_2]$ and $\check{D} = [\check{D}_1; \check{D}_2]$. Then $\check{D}_1$ is boundedly invertible and* (5.4) *holds.*

This lemma implies that the input-output map of the CARE closed-loop system is a factor of the input-output map of the plant by Lemma 5.7. If Q is not the smallest nonnegative self-adjoint solution of the CARE, the factorization need not be normalized (see, e.g., [11, Example 3.1.2]).

We now show that if a system is approximately controllable or observable, then its CARE closed-loop system is too.

LEMMA 6.2. *Let (A, B, C, D) be an output stabilizable discrete-time system and let Q be a nonnegative self-adjoint solution of its CARE* (3.4). *If (A, B, C, D) is approximately controllable, then its CARE closed-loop system is too. If (A, B, C, D) is approximately observable, then its CARE closed-loop system is too.*

Proof. As is well known, a system $(\check{A}, \check{B}, \check{C}, \check{D})$ is approximately controllable (observable) iff $(\check{A} + \check{B}K\check{C}, \check{B}, \check{C}, \check{D})$ is approximately controllable (observable). With $K = -W^{1/2}[I, 0]$ we see that the CARE closed-loop system is approximately controllable (observable) iff $(A, \check{B}, \check{C}, \check{D})$ is approximately controllable (observable). Obviously $(A, \check{B}) = (A, BW^{-1/2})$ is approximately controllable iff (A, B) is approximately controllable and $(A, \check{C}) = (A, [F; C + DF])$ is approximately observable if (A, C) is (but not only if). □

In the next four lemmas we prove a correspondence between the Riccati equations of a system and the Lyapunov equations of its CARE closed-loop system. Because this requires some extensive algebraic manipulations we have relegated some of the proofs to the appendix.

LEMMA 6.3. *Let (A, B, C, D) be an output stabilizable discrete-time system and let Q be a nonnegative self-adjoint solution of its CARE* (3.4). *Let $(\check{A}, \check{B}, \check{C}, \check{D})$ be its CARE closed-loop system defined by* (5.1). *Then*

$$\check{B}^*Q\check{B} + \check{D}^*\check{D} = I, \tag{6.1}$$

$$\check{B}^*Q\check{A} + \check{D}^*\check{C} = 0, \tag{6.2}$$

and Q is a solution of the observation Lyapunov equation of $(\check{A}, \check{B}, \check{C}, \check{D})$:

$$\check{A}^*Q\check{A} - Q + \check{C}^*\check{C} = 0. \tag{6.3}$$

Proof. The equations (6.1) and (6.2) are readily verified using (5.1). Equation (6.3) is more complicated and the proof is given in the appendix. ∎

LEMMA 6.4. *Let* $(\check{A}, \check{B}, [\check{C}_1; \check{C}_2], [\check{D}_1, \check{D}_2])$ *be a discrete-time system such that* $\check{D}_1$ *is boundedly invertible and define the discrete-time system* (A, B, C, D) *by* (5.4). *If the nonnegative self-adjoint operator* Q *satisfies* (6.2) *and* (6.3), *then* Q *is a solution of the CARE* (3.4) *of* (A, B, C, D).

Proof. See the appendix. ∎

LEMMA 6.5. *Let* (A, B, C, D) *be an input and output stabilizable discrete-time system and let* Q *be a nonnegative self-adjoint solution of its CARE* (3.4) *and* P *be a nonnegative self-adjoint solution of its FARE* (3.8). *Let* $(\check{A}, \check{B}, \check{C}, \check{D})$ *be its CARE closed-loop system defined by* (5.1). *Define* $L := (I + PQ)^{-1}P$. *Then* L *is a solution of the control Lyapunov equation of* $(\check{A}, \check{B}, \check{C}, \check{D})$:

$$\check{A}L\check{A}^* - L + \check{B}\check{B}^* = 0. \tag{6.4}$$

Proof. See the appendix for the proof. ∎

LEMMA 6.6. *Let* $(\check{A}, \check{B}, [\check{C}_1; \check{C}_2], [\check{D}_1, \check{D}_2])$ *be a discrete-time system such that* $\check{D}_1$ *is boundedly invertible and define the discrete-time system* (A, B, C, D) *by* (5.4). *If the nonnegative self-adjoint operators* Q *and* L *satisfy* (6.1), (6.2), (6.3), *and* (6.4) *and* $I - QL$ *is boundedly invertible, then* $L(I - QL)^{-1}$ *is a solution of the FARE* (3.8) *of* (A, B, C, D).

Proof. See the appendix for the proof. ∎

If Q and P are the smallest nonnegative self-adjoint solutions of their respective Riccati equations, then the operator $L := (I + PQ)^{-1}P$ defined in Lemma 6.5 is actually the smallest nonnegative self-adjoint solution of the Lyapunov equation (6.4). To prove this we first prove the following lemma. We note that the Hankel norm of an input-output map is the norm of the associated Hankel operator.

LEMMA 6.7. *Let* (A, B, C, D) *be an input and output stabilizable discrete-time system. Then the Hankel norm of the input-output map of the optimal closed-loop system* (5.1) *is strictly smaller than one.*

Proof. Denote the optimal cost operators by P and Q, the gramians of the optimal closed-loop system by L_B and L_C, and the Hankel operator of the optimal closed-loop system by Γ. From Lemma 5.1 it follows that $Q = L_C$. From Lemma 6.5 we know that $(I + PQ)^{-1}P$ is a solution of the control Lyapunov equation and hence by Lemma 3.4 we have $L_B \leq (I + PQ)^{-1}P$. This implies that $L_C^{1/2} L_B L_C^{1/2} \leq Q^{1/2}(I + PQ)^{-1}PQ^{1/2} = (I + Q^{1/2}PQ^{1/2})^{-1}Q^{1/2}PQ^{1/2}$. Now $\|\Gamma\| = \|\Gamma\Gamma^*\| = r(\Gamma\Gamma^*) = r(\mathcal{C}\mathcal{B}\mathcal{B}^*\mathcal{C}^*) = r(\mathcal{C}^*\mathcal{C}\mathcal{B}\mathcal{B}^*) = r(L_C L_B) = r(L_C^{1/2} L_B L_C^{1/2}) \leq r((I + Q^{1/2}PQ^{1/2})^{-1}Q^{1/2}PQ^{1/2})$. Next we show that if X is a nonnegative self-adjoint operator, then $(I + X)^{-1}X < I$. Since $X < I + X$ we have $(I + X)^{-1}X = (I + X)^{-1/2}X(I + X)^{-1/2} < (I + X)^{-1/2}(I + X)(I + X)^{-1/2} = I$. Finally, we apply this last result with $X := Q^{1/2}PQ^{1/2}$ to obtain $\|\Gamma\| \leq r((I + X)^{-1}X) < 1$. ∎

Lemma 6.7 has the following corollary.

COROLLARY 6.8. *Let* (A, B, C, D) *be an input and output stabilizable discrete-time system. Let* L_B *and* L_C *denote the gramians of an input and output stable realization of the input-output map of the optimal closed-loop system* (5.1). *Then* $r(L_B L_C) < 1$.

We now show for an approximately observable system that if Q and P are the smallest nonnegative self-adjoint solutions of their respective Riccati equations then the operator $L := (I + PQ)^{-1}P$ defined in Lemma 6.5 is actually the smallest nonnegative self-adjoint solution of the Lyapunov equation (6.4).

LEMMA 6.9. *Let (A, B, C, D) be an approximately observable input and output stabilizable discrete-time system. Let Q and P denote the optimal cost operators of the system and its dual, respectively. Define $L := (I + PQ)^{-1}P$. Then L is the controllability gramian of the optimal closed-loop system* (5.1).

Proof. We have by Lemma 5.1 that $Q = L_C$, the observability gramian of the optimal closed-loop system, and by Lemma 6.5 that $L = (I + PQ)^{-1}P$ is a solution of the control Lyapunov equation of the optimal closed-loop system. Since the control Lyapunov equation of the optimal closed-loop system has a nonnegative self-adjoint solution, the optimal closed-loop system is input stable and has a controllability gramian L_B which satisfies $L_B \leq L$. From Lemma 6.8 we know that $I - L_C L_B$ is boundedly invertible and hence by Lemma 6.6 we have that $\tilde{P} := L_B(I - L_C L_B)^{-1}$ is a solution of the FARE (3.8) of the system (A, B, C, D). We thus have $P \leq \tilde{P}$. Note that approximate observability ensures that $Q = L_C > 0$. Since Q is positive and thus has dense range we have $\tilde{P} \geq P$ iff $Q^{1/2}\tilde{P}Q^{1/2} \geq Q^{1/2}PQ^{1/2}$. This is true iff

$$\begin{aligned} I - Q^{1/2}(I + \tilde{P}Q)^{-1}\tilde{P}Q^{1/2} &= (I + Q^{1/2}\tilde{P}Q^{1/2})^{-1} \\ &\leq (I + Q^{1/2}PQ^{1/2})^{-1} = I - Q^{1/2}(I + PQ)^{-1}PQ^{1/2}. \end{aligned}$$

This is true iff $Q^{1/2}(I + \tilde{P}Q)^{-1}\tilde{P}Q^{1/2} \geq Q^{1/2}(I + PQ)^{-1}PQ^{1/2}$ and again using that Q has dense range this is true iff $(I + \tilde{P}Q)^{-1}\tilde{P} \geq (I + PQ)^{-1}P$. We conclude that $L = (I + PQ)^{-1}P \leq (I + \tilde{P}Q)^{-1}\tilde{P} = L_B$. Since we already had $L_B \leq L$, we must have $L = L_B$. ∎

7. LQG-characteristic values. In this section we show that the spectrum of the product of the optimal cost operators of a system and its dual does not depend on the realization but only on the input-output map. This generalizes the result from finite-dimensional theory that the eigenvalues of PQ are similarity invariants. We define the LQG-characteristic values of an input and output stabilizable discrete-time system to be the square roots of the spectral values of the product of the optimal cost operators P and Q. We first prove the following lemma on the spectrum of PQ.

LEMMA 7.1. *Let (A, B, C, D) be an approximately observable input and output stabilizable discrete-time system. Let Q and P denote the optimal cost operators of the system and its dual, respectively, and let L_B and L_C denote the gramians of the optimal closed-loop system* (5.1). *Then $\lambda \in \sigma(PQ)$ iff $\lambda/(1 + \lambda) \in \sigma(L_B L_C)$.*

Proof. From Lemmas 5.1 and 6.9 we have that $L_B L_C = (I + PQ)^{-1}PQ$, from which it follows that $PQ = (I - L_B L_C)^{-1}L_B L_C$. Let $\lambda \in \mathbb{C} - \{-1\}$ and define $\mu := \lambda/(1 + \lambda)$, then $\lambda = \mu/(1 - \mu)$. We have

$$\begin{aligned} \lambda I - PQ &= \frac{\mu}{1-\mu} I - L_B L_C (I - L_B L_C)^{-1} \\ &= \frac{1}{1-\mu}\left[\mu I - (1-\mu) L_B L_C (I - L_B L_C)^{-1}\right] \\ &= \frac{1}{1-\mu}\left[\mu(I - L_B L_C) - (1-\mu) L_B L_C\right](I - L_B L_C)^{-1} \\ &= \frac{1}{1-\mu}(\mu I - L_B L_C)(I - L_B L_C)^{-1}. \end{aligned}$$

This shows that $\lambda \in \sigma(PQ)$ iff $\mu = \lambda/(1 + \lambda) \in \sigma(L_B L_C)$. ∎

The following lemma gives the desired result.

LEMMA 7.2. *Let (A_i, B_i, C_i, D) with $i = 1, 2$ be two approximately observable input and output stabilizable discrete-time systems. Let Q_i and P_i denote the optimal cost operators of the system and its dual, respectively. If the two systems have the same input-output map, then the spectra of P_1Q_1 and P_2Q_2 are equal, with the possible exception of zero.*

Proof. Denote the gramians of the optimal closed-loop system of (A_i, B_i, C_i, D) by L_{B_i} and L_{C_i}. Then according to Lemma 7.1, the lemma would be proved if the nonzero elements in the spectrum of $L_{B_1}L_{C_1}$ equal the nonzero elements in the spectrum of $L_{B_2}L_{C_2}$. Since the input-output map of both optimal closed-loop systems is a normalized factor of the input-output map of the plant by Lemma 5.8, there exists a unitary V such that $[\mathcal{M}_2; \mathcal{N}_2] = [\mathcal{M}_1; \mathcal{N}_1]V$ by Lemma 5.9. For the Hankel operators of the optimal closed-loop systems this implies $\Gamma_2 = \Gamma_1 V$, which implies that $\Gamma_2\Gamma_2^* = \Gamma_1\Gamma_1^*$. Since for arbitrary bounded operators S and T we have that the nonzero elements in the spectrum of ST equal the nonzero elements in the spectrum of TS, we have that the nonzero elements in the spectrum of $L_BL_C = \mathcal{B}\mathcal{B}^*\mathcal{C}^*\mathcal{C}$ equal the nonzero elements in the spectrum of $\Gamma\Gamma^* = \mathcal{C}\mathcal{B}\mathcal{B}^*\mathcal{C}^*$. This shows that the nonzero elements in the spectrum of $L_{B_1}L_{C_1}$ equal the nonzero elements in the spectrum of $L_{B_2}L_{C_2}$. ∎

8. LQG-balanced realizations. In this section we prove the existence and some properties of LQG-balanced realizations. We first show that the input-output map of the optimal closed-loop system has a Lyapunov-balanced realization.

LEMMA 8.1. *Let (A, B, C, D) be an output stabilizable discrete-time system. Then the input-output map of the optimal closed-loop system has an approximately controllable and approximately observable Lyapunov-balanced realization.*

Proof. Lemma 5.8 shows that the optimal closed-loop system is input-output stable. This implies that the Hankel operator of the optimal closed-loop system is bounded and Lemma 4.1 then shows that the input-output map of the optimal closed-loop system has a Lyapunov-balanced realization. ∎

From this Lyapunov-balanced realization we can construct a LQG-balanced realization of the plant.

THEOREM 8.2. *Let (A, B, C, D) be an input and output stabilizable discrete-time system. Then the input-output map of the system (A, B, C, D) has a LQG-balanced realization.*

Proof. Denote the approximately controllable and approximately observable Lyapunov-balanced realization of the input-output map of the optimal closed-loop system by $(\check{A}, \check{B}, \check{C}, \check{D})$. Denote the equal controllability and observability gramians of this realization by L. Define

$$A_s := \check{A} - \check{B}\check{D}_1^{-1}\check{C}_1, \quad B_s := \check{B}\check{D}_1^{-1}, \quad C_s := \check{C}_2 - \check{D}_2\check{D}_1^{-1}\check{C}_1, \quad D_s := \check{D}_2\check{D}_1^{-1}.$$

Then (A_s, B_s, C_s, D_s) is a realization of the input-output map of (A, B, C, D) according to Lemma 5.7. Since the input-output map of the optimal closed-loop system is inner (Corollary 5.8) and $(\check{A}, \check{B}, \check{C}, \check{D})$ is approximately controllable we know from Lemma 5.3 that $\check{B}^*L\check{B} + \check{D}^*\check{D} = I$ and $\check{B}^*L\check{A} + \check{D}^*\check{C} = 0$. From Lemma 6.4 we see that L is a solution of the CARE of the system (A_s, B_s, C_s, D_s). From Corollary 6.8 we know that $I - L^2$ is boundedly invertible, and Lemma 6.6 now tells us that $L(I-L^2)^{-1}$ is a solution of the FARE of the system (A_s, B_s, C_s, D_s). From Lemma 4.3 we know that $\check{A}$ is strongly stable. Using Lemma 3.7 we see that this implies that L is the unique nonnegative self-adjoint solution of the CARE of the system (A_s, B_s, C_s, D_s). Assume that the FARE of the system (A_s, B_s, C_s, D_s) has two nonnegative self-adjoint

solutions, P_1 and P_2. From Lemma 6.5 we see that $(I+P_1L)^{-1}P_1$ and $(I+P_2L)^{-1}P_2$ are solutions of the control Lyapunov equation of the optimal closed-loop system of (A_s, B_s, C_s, D_s), which is the balanced realization $(\check{A}, \check{B}, \check{C}, \check{D})$. From Corollary 4.4 we see that $(I + P_1L)^{-1}P_1 = (I + P_2L)^{-1}P_2$, which implies that $P_1 = P_2$. We recall that if a system (A_1, B_1, C_1, D_1) has a solution Q to its CARE and P to its FARE and S is a boundedly invertible operator, then the system $(SAS^{-1}, SB, CS^{-1}, D)$ has a solution $S^{-*}QS^{-1}$ to its CARE and SPS^* to its FARE. It is easily seen that $I - L^2$ is a nonnegative operator (for example, using that the Hankel operator has norm smaller than one), from which it follows that $S := (I - L^2)^{-1/4}$ is well defined. Define $(A_l, B_l, C_l, D_l) = (SA_sS^{-1}, SB_s, C_sS^{-1}, D_s)$. Then the system (A_l, B_l, C_l, D_l) has $L(I - L^2)^{-1/2}$ as the unique solution to both its CARE and its FARE. □

We now prove that LQG-balanced realizations are essentially unique.

LEMMA 8.3. *Let (A, B, C, D) be an approximately controllable and approximately observable LQG-balanced realization. Then all approximately controllable and approximately observable LQG-balanced realizations of the same input-output map are given by $\{(TAT^{-1}, TB, CT^{-1}, D) : T \in \mathcal{L}(X)$ unitary$\}$.*

Proof. It is obvious that the given realizations are indeed LQG balanced; it remains to be proved that these are all approximately controllable and approximately observable LQG-balanced realizations. Let (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) be two approximately controllable and approximately observable LQG-balanced realizations of the same input-output map. Let Q_1 and Q_2 be the optimal cost operators of the respective systems. Define $S_i := (I + Q_i^2)^{1/4}$ and $(A_i^b, B_i^b, C_i^b, D_i^b) = (S_iA_iS_i^{-1}, S_iB_i, C_iS_i^{-1}, D_i)$ for $i = 1, 2$. Then $(A_i^b, B_i^b, C_i^b, D_i^b)$ has $Q_i^b := Q_i(I + Q_i^2)^{-1/2}$ as its optimal cost operator and $P_i^b := Q_i(I + Q_i^2)^{1/2}$ as the optimal cost operator of its dual system. Let $(\check{A}_i^b, \check{B}_i^b, \check{C}_i^b, \check{D}_i^b)$ be the optimal closed-loop system of $(A_i^b, B_i^b, C_i^b, D_i^b)$. Using Lemmas 5.1 and 6.9 we see that they are Lyapunov-balanced realizations with gramians $L_i := Q_i(I + Q_i^2)^{-1/2}$. According to Corollary 5.8 the input-output maps of $(\check{A}_1^b, \check{B}_1^b, \check{C}_1^b, \check{D}_1^b)$ and $(\check{A}_2^b, \check{B}_2^b, \check{C}_2^b, \check{D}_2^b)$ are normalized factors of the input-output map of (A_1, B_1, C_1, D_1) (which equals the input-output map of (A_2, B_2, C_2, D_2)). By Lemma 5.9 there exists an operator V such that $[\mathcal{M}_2; \mathcal{N}_2] = [\mathcal{M}_1; \mathcal{N}_1]V$. It is easily seen that $(\check{A}_1^b, \check{B}_1^bV, \check{C}_1^b, \check{D}_1^bV)$ is a Lyapunov-balanced realization of $[\mathcal{M}_1; \mathcal{N}_1]V = [\mathcal{M}_2; \mathcal{N}_2]$. From Lemma 6.2 it follows that both $(\check{A}_1^b, \check{B}_1^bV, \check{C}_1^b, \check{D}_1^bV)$ and $(\check{A}_2^b, \check{B}_2^b, \check{C}_2^b, \check{D}_2^b)$ are approximately controllable and approximately observable Lyapunov-balanced realizations of the input-output map $[\mathcal{M}_2; \mathcal{N}_2]$. From Lemma 4.1 it follows that there exists a unitary T such that

$$(T\check{A}_1^bT^{-1}, T\check{B}_1^bV, \check{C}_1^bT^{-1}, \check{D}_1^bV) = (\check{A}_2^b, \check{B}_2^b, \check{C}_2^b, \check{D}_2^b).$$

Using (5.4) for $i = 1, 2$ we see that

$$(TA_1^bT^{-1}, TB_1^b, C_1^bT^{-1}, D_1^b) = (A_2^b, B_2^b, C_2^b, D_2^b).$$

It now follows that

$$(S_2^{-1}TS_1A_1S_1^{-1}T^{-1}S_2, S_2^{-1}TS_1B_1, C_1S_1^{-1}T^{-1}S_2, D_1) = (A_2, B_2, C_2, D_2).$$

To complete the proof that (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) are unitarily equivalent we prove that $S_2^{-1}TS_1 = T$ or equivalently $TS_1T^{-1} = S_2$. Since $(I + Q_i^2)^{-1} = I - [Q_i(I + Q_i^2)^{-1/2}]^2 = I - L_i^2$ we have $S_i = (I + Q_i^2)^{-1/4} = (I - L_i^2)^{1/4}$ and since $L_2 = TL_1T^{-1}$ we then have

$$S_2 = (I - L_2^2)^{1/4} = T(I - L_1^2)^{1/4}T^{-1} = TS_1T^{-1}.$$

This proves that all approximately controllable and approximately observable LQG-balanced realizations of the same input-output map are unitarily equivalent. ∎

The following lemma states what the proof of Theorem 8.2 already indicated, namely, that the filter and control Riccati equations of an approximately controllable and approximately observable LQG-balanced realization have unique nonnegative self-adjoint solutions.

LEMMA 8.4. *Let (A, B, C, D) be an approximately controllable and approximately observable LQG-balanced realization. Then both the CARE* (3.4) *and the FARE* (3.8) *of (A, B, C, D) have a unique nonnegative self-adjoint solution.*

Proof. As in the proof of Lemma 8.3 we construct the approximately controllable and approximately observable realization (A^b, B^b, C^b, D^b) and the approximately controllable and approximately observable Lyapunov-balanced realization $(\check{A}^b, \check{B}^b, \check{C}^b, \check{D}^b)$. From Lemma 4.3 we know that $\check{A}^b$ is strongly stable. Lemma 3.7 shows that the CARE of (A^b, B^b, C^b, D^b) has a unique nonnegative self-adjoint solution. Obviously this implies that the CARE of (A, B, C, D) has a unique nonnegative self-adjoint solution. From Lemma 4.4 we know that the control Lyapunov equation of $(\check{A}^b, \check{B}^b, \check{C}^b, \check{D}^b)$ has a unique nonnegative self-adjoint solution. As in the proof of Theorem 8.2, it follows that the FARE of (A^b, B^b, C^b, D^b) and hence the FARE of (A, B, C, D) has a unique nonnegative self-adjoint solution. ∎

9. Conclusions. We have proved the existence of LQG-balanced realizations for the class of transfer functions that are analytic on some disc centered at the origin and have a (infinite-dimensional) realization that is both input and output stabilizable. We have also proved that approximately controllable and approximately observable LQG-balanced realizations are essentially unique and that the Riccati equations of approximately controllable and approximately observable LQG-balanced realizations have unique nonnegative self-adjoint solutions. Analogous continuous-time results and error bounds for truncations of LQG-balanced realizations will be given elsewhere.

10. Appendix.

10.1. Miscellaneous results on Riccati operators.

LEMMA 10.1. *Let P and Q be nonnegative self-adjoint operators. Define*

$$A_P := A - (BD^* + APC^*)(R + CPC^*)^{-1}C, \tag{10.1}$$

$$A_Q := A - B(S + B^*QB)^{-1}(D^*C + B^*QA), \tag{10.2}$$

$$\underline{A} := A - BS^{-1}D^*C. \tag{10.3}$$

where $S := I + D^*D$ and $R := I + DD^*$. Then

$$A_P(I + PC^*R^{-1}C) = \underline{A} = (I + BS^{-1}B^*Q)A_Q, \tag{10.4}$$

$$A_Q = (I + BS^{-1}B^*Q)^{-1}A_P(I + PC^*R^{-1}C), \tag{10.5}$$

$$A_P = (I + BS^{-1}B^*Q)A_Q(I + PC^*R^{-1}C)^{-1}. \tag{10.6}$$

Proof. We prove that $A_P(I+PC^*R^{-1}C) = \underline{A}$. The equality $\underline{A} = (I+BS^{-1}B^*Q)A_Q$ is proved similarly. By writing out A_P in full we have

$$\begin{aligned}
&A_P(I + PC^*R^{-1}C)\\
&\quad = A(I + PC^*R^{-1}C) - (BD^* + APC^*)(R + CPC^*)^{-1}C(I + PC^*R^{-1}C)\\
&\quad = A(I + PC^*R^{-1}C) - (BD^* + APC^*)(R + CPC^*)^{-1}(R + CPC^*)R^{-1}C\\
&\quad = A + APC^*R^{-1}C - (BD^* + APC^*)R^{-1}C = A - BD^*R^{-1}C\\
&\quad = A - BS^{-1}D^*C,
\end{aligned}$$

since $D^*R^{-1} = S^{-1}D^*$. This completes the proof of (10.4). Equations (10.5) and (10.6) easily follow from (10.4). □

Note that in the above lemma we have not assumed that P and Q are solutions of the Riccati equations.

We now prove that the Riccati equations can be written in several different but equivalent versions.

LEMMA 10.2.

1. *P is a nonnegative self-adjoint solution of*

$$A_P P(I + C^*R^{-1}CP)A_P^* - P + BS^{-1}B^* = 0, \tag{10.7}$$

where A_P is defined by (10.1)*, iff it is a nonnegative self-adjoint solution of*

$$\underline{A}P(I + C^*R^{-1}CP)^{-1}\underline{A}^* - P + BS^{-1}B^* = 0, \tag{10.8}$$

where $\underline{A}$ is defined by (10.3)*.*

2. *P is a nonnegative self-adjoint solution of* (10.7) *iff it is a nonnegative self-adjoint solution of the FARE* (3.8)*.*
3. *Q is a nonnegative self-adjoint solution of*

$$A_Q^*(I + QBS^{-1}B^*)QA_Q - Q + C^*R^{-1}C = 0, \tag{10.9}$$

where A_Q is defined by (10.2)*, iff it is a nonnegative self-adjoint solution of*

$$\underline{A}^*Q(I + BS^{-1}B^*Q)^{-1}\underline{A} - Q + C^*R^{-1}C = 0, \tag{10.10}$$

where $\underline{A}$ is defined by (10.3)*.*

4. *Q is a nonnegative self-adjoint solution of* (10.9) *iff it is a nonnegative self-adjoint solution of the CARE* (3.4)*.*

Proof. We shall prove the equivalence of the filter equations; the equivalence of the control equations is similar.

1. The equations (10.7) and (10.8) are equivalent iff the following holds:

$$\underline{A}P(I + C^*R^{-1}CP)^{-1}\underline{A}^* = A_P P(I + C^*R^{-1}CP)A_P^*. \tag{10.11}$$

We use Lemma 10.1 (which tells us that $\underline{A} = A_P(I + PC^*R^{-1}C)$) to write the left-hand side of (10.11) as

$$A_P(I + PC^*R^{-1}C)P(I + C^*R^{-1}CP)^{-1}(I + C^*R^{-1}CP)A_P^*,$$

which is indeed equal to the right-hand side of (10.11).

2. To prove the equivalence of (10.7) and (3.8) we substitute in (10.7) for A_P from (10.1) and for $(I + C^*R^{-1}CP)A_P^*$, we substitute $\underline{A}^*$ (using (10.4)) and then substitute (10.3) for $\underline{A}$. We then get

$$(A - (BD^* + APC^*)(R + CPC^*)^{-1}C)P(A^* - C^*DS^{-1}B^*) - P + BS^{-1}B^* = 0.$$

Rewriting this gives

$$\begin{aligned} APA^* - P + BB^* = {} & (BD^* + APC^*)(R + CPC^*)^{-1}CPA^* \\ & - (BD^* + APC^*)(R + CPC^*)^{-1}CPC^*DS^{-1}B^* + APC^*DS^{-1}B^* \\ & - BS^{-1}B^* + BB^*. \end{aligned}$$

We now focus on the last two lines of this last equation. We note that $I - S^{-1} = D^*DS^{-1}$ and we can thus rewrite these last two lines as

$$-(BD^* + APC^*)(R + CPC^*)^{-1}CPC^*DS^{-1}B^* + APC^*DS^{-1}B^* + BD^*DS^{-1}B^*,$$

and this can be rewritten as

$$(BD^* + APC^*)(R + CPC^*)^{-1}[-CPC^* + R + CPC^*]DS^{-1}B^*.$$

Noting that $RDS^{-1} = D$, we see that this is equal to

$$(BD^* + APC^*)(R + CPC^*)^{-1}DB^*.$$

This completes the proof of the equivalence of (10.7) and (3.8). ∎

We now show that the known relationship between A_Q and A_P extends to the infinite-dimensional case.

LEMMA 10.3. *Let (A, B, C, D) be an input and output stabilizable discrete-time system, let Q be a nonnegative self-adjoint solution of its CARE* (3.4)*, and let P be a nonnegative self-adjoint solution of its FARE* (3.8)*. Define A_P and A_Q by* (10.1) *and* (10.2)*, respectively. Then*

$$(I + PQ)A_Q = A_P(I + PQ).$$

Proof. We use FARE (10.7) to write

$$P = A_PP(I + C^*R^{-1}CP)A_P^* + BS^{-1}B^*,$$

which leads to

$$I + PQ = I + A_PP(I + C^*R^{-1}CP)A_P^*Q + BS^{-1}B^*Q$$

and so

$$(I + PQ)A_Q = (I + BS^{-1}B^*Q)A_Q + A_PP(I + C^*R^{-1}CP)A_P^*QA_Q.$$

We use (10.5) to write the right-hand side as

$$A_P(I + PC^*R^{-1}C) + A_PP(I + C^*R^{-1}CP)A_P^*QA_Q.$$

Rearranging gives

$$A_P + A_PP[C^*R^{-1}C + (I + C^*R^{-1}CP)A_P^*QA_Q],$$

and using (10.5) again we obtain

$$A_P + A_P P[C^* R^{-1} C + A_Q^*(I + QBS^{-1}B^*)QA_Q].$$

According to CARE (10.9), the term in square brackets equals Q. So the above is equal to $A_P(I + PQ)$. □

We now prove a relation concerning the difference of two solutions of a Riccati equation.

LEMMA 10.4. *Let (A, B, C, D) be an output stabilizable discrete-time system and let Q_1 and Q_2 be nonnegative self-adjoint solutions of its CARE* (3.4). *Define A_{Q_1} and A_{Q_2} similarly to* (10.2). *Then*

$$Q_1 - Q_2 = A_{Q_2}^*(Q_1 - Q_2)A_{Q_1}.$$

Proof. Subtract the form (10.9) of the CARE for Q_1 and Q_2 to obtain

$$Q_1 - Q_2 = A_{Q_1}^*(I + Q_1 BS^{-1}B^*)Q_1 A_{Q_1} - A_{Q_2}^*(I + Q_2 BS^{-1}B^*)Q_2 A_{Q_2}. \tag{10.12}$$

According to Lemma 10.1 (say with $P = I$) we have

$$\begin{aligned} A_{Q_2} &= (I + BS^{-1}B^*Q_2)^{-1} A_P (I + PC^* R^{-1} C) \\ &= (I + BS^{-1}B^*Q_2)^{-1}(I + BS^{-1}B^*Q_1)A_{Q_1}. \end{aligned} \tag{10.13}$$

Combining (10.12) and (10.13) we obtain

$$\begin{aligned} Q_1 - Q_2 &= A_{Q_2}^*(I + Q_2 BS^{-1}B^*)Q_1 A_{Q_1} - A_{Q_2}^* Q_2 (I + BS^{-1}B^*Q_1)A_{Q_1} \\ &= A_{Q_2}^*(Q_1 - Q_2)A_{Q_1}. \end{aligned}$$ □

10.2. Proofs for section 6. In this section we prove the relationships between the CARE and the FARE of a system and the control and observation Lyapunov equations of its CARE closed-loop system (Lemmas 6.3, 6.4, 6.5, and 6.6).

LEMMA 10.5. *Suppose that the system $(\check{A}, \check{B}, [\check{C}_1; \check{C}_2], [\check{D}_1; \check{D}_2])$ is such that $\check{D}_1$ is boundedly invertible and that there exists a nonnegative self-adjoint operator V such that*

$$\check{B}^* V \check{A} + \check{D}^* \check{C} = 0. \tag{10.14}$$

Define the system (A, B, C, D) by (5.4) *and $S := I + D^*D$ and $R := I + DD^*$. Then*

1. $\check{A} = A - B(S + B^*VB)^{-1}(D^*C + B^*VA)$ *and*
2. $\check{A}^* V \check{A} - V + \check{C}^* \check{C} = \check{A}^*(I + VBS^{-1}B^*)V\check{A} - V + C^* R^{-1} C$.

Proof. We first prove the equality

$$S\check{C}_1 = -(B^* V \check{A} + D^* C). \tag{10.15}$$

From (10.14) and (5.4) we obtain

$$\check{D}_1^* B^* V \check{A} + \check{D}_1^* \check{C}_1 + \check{D}_2^* \check{C}_2 = 0.$$

Thus

$$\check{C}_1 = -B^* V \check{A} - D^* \check{C}_2 = -B^* V \check{A} - D^*(C + D\check{C}_1)$$

and this yields (10.15):

$$S\check{C}_1 = (I + D^*D)\check{C}_1 = -B^* V \check{A} - D^* C.$$

We now prove the first equality stated in the lemma. We take the equality just proved (10.15) and substitute $\check{A} = A + B\check{C}_1$ to obtain

$$S\check{C}_1 = -(B^*V(A + B\check{C}_1) + D^*C).$$

Thus

$$(S + B^*VB)\check{C}_1 = -(B^*VA + D^*C).$$

We now solve for $\check{C}_1$ and substitute to obtain

$$\check{A} = A + B\check{C}_1 = A - B(S + B^*VB)^{-1}(B^*VA + D^*C).$$

We now prove the equality

$$\check{C}^*\check{C} = \check{A}^*VBS^{-1}B^*V\check{A} + C^*R^{-1}C. \tag{10.16}$$

We have

$$\check{C}^*\check{C} = \check{C}_1^*\check{C}_1 + \check{C}_2^*\check{C}_2$$

and substituting for $\check{C}_2$ from (5.4) gives

$$\check{C}^*\check{C} = \check{C}_1^*\check{C}_1 + (C + D\check{C}_1)^*(C + D\check{C}_1).$$

Finally, substituting for $\check{C}_1$ from (10.15) and simplifying gives the result.
The second equality stated in the lemma follows easily from (10.16). □

Proof of Lemma 6.3. We only have to prove (6.3). Noting (6.2), we apply Lemma 10.5 with $V := Q$ to the CARE closed-loop system. Since $\check{A}$ defined by (5.1) equals A_Q defined by (10.2), we have

$$\check{A}^*Q\check{A} - Q + \check{C}^*\check{C} = A_Q^*(I + QBS^{-1}B^*)QA_Q - Q + C^*R^{-1}C.$$

Since Q is a solution of the CARE (3.4) of (A, B, C, D) the right-hand side of this equation is zero by Lemma 10.2 and we have shown (6.3). □

Proof of Lemma 6.4. By assumption, there exists a nonnegative self-adjoint Q such that $\check{B}^*Q\check{A} + \check{D}^*\check{C} = 0$. We apply Lemma 10.5 with $V := Q$ to obtain $\check{A} = A_Q$ given by (10.2) and

$$\check{A}^*Q\check{A} - Q + \check{C}^*\check{C} = A_Q^*(I + QBS^{-1}B^*)QA_Q - Q + C^*R^{-1}C.$$

The left-hand side of this equation is zero since by assumption Q satisfies (6.3). This proves that the right-hand side is zero, and Lemma 10.2 now shows that Q is a solution of the CARE of the system (A, B, C, D). □

LEMMA 10.6. *Let a system* $(\check{A}, \check{B}, [\check{C}_1; \check{C}_2], [\check{D}_1; \check{D}_2])$ *with* $\check{D}_1$ *boundedly invertible be given and assume that a nonnegative self-adjoint operator* V *exists such that*

$$\check{B}^*V\check{B} + \check{D}^*\check{D} = I.$$

Define the system (A, B, C, D) *by* (5.4). *Then we have*

1. $B^*VB + S = \check{D}_1^{-*}\check{D}_1^{-1}$ *and*
2. $\check{B}\check{B}^*(I + VBS^{-1}B^*) = BS^{-1}B^*$.

Proof. 1. The given equation for V translates to

$$\check{D}_1^* B^* V B \check{D}_1 + \check{D}_1^* \check{D}_1 + \check{D}_1^* D^* D \check{D}_1 = I,$$

and multiplying from the left with $\check{D}_1^{-*}$ and from the right with $\check{D}_1^{-1}$ gives the result.

2. The first equality implies that $(S+B^*VB)^{-1} = \check{D}_1\check{D}_1^*$ and so $B(S+B^*VB)^{-1}B^* = \check{B}\check{B}^*$. Hence

$$\begin{aligned}\check{B}\check{B}^*(I + VBS^{-1}B^*) &= B(S + B^*VB)^{-1}B^*(I + VBS^{-1}B^*) \\ &= B(S + B^*VB)^{-1}(S + B^*VB)S^{-1}B^* = BS^{-1}B^*,\end{aligned}$$

which proves the second equality. □

We now prove Lemma 6.5.

Proof of Lemma 6.5. We remark that $L = (I + PQ)^{-1}P = P(I + QP)^{-1}$ and so we have to show that

$$\check{A}P(I + QP)^{-1}\check{A}^* - P(I + QP)^{-1} + \check{B}\check{B}^* = 0. \tag{10.17}$$

Recalling that $\check{A} = A_Q$ from (5.1) and (10.2) and using Lemmas 10.1 (10.5) and 10.3 we can substitute $\check{A} = (I + BS^{-1}B^*Q)^{-1}A_P(I + PC^*R^{-1}C)$ and $(I + QP)^{-1}\check{A}^* = A_P^*(I + QP)^{-1}$ to obtain for the left-hand side of (10.17)

$$\begin{aligned}&(I + BS^{-1}B^*Q)^{-1}A_P(I + PC^*R^{-1}C)PA_P^*(I + QP)^{-1} - P(I + QP)^{-1} + \check{B}\check{B}^* \\ &\quad = (I + BS^{-1}B^*Q)^{-1}[A_P(I + PC^*R^{-1}C)PA_P^* - (I + BS^{-1}B^*Q)P \\ &\quad + (I + BS^{-1}B^*Q)^{-1}\check{B}\check{B}^*(I + QP)](I + QP)^{-1}.\end{aligned}$$

We now apply Lemma 10.6 with $V := Q$ to obtain $(I+BS^{-1}B^*Q)^{-1}\check{B}\check{B}^* = BS^{-1}B^*$, and so we obtain for the left-hand side of (10.17)

$$\begin{aligned}(I + BS^{-1}B^*Q)^{-1}[&A_P(I + PC^*R^{-1}C)PA_P^* - (I + BS^{-1}B^*Q)P \\ &+ BS^{-1}B^*(I + QP)](I + QP)^{-1}.\end{aligned}$$

The term in square brackets is zero according to the FARE (10.7) and we have proved (10.17). □

Proof of Lemma 6.6. We first recall from the proof of Lemma 6.4 that $\check{A} = A_Q$, where A_Q is defined by (10.2), and that (10.6) holds with $A_Q = \check{A}$. Define $P := L(I - QL)^{-1}$ and define A_P by (10.1). The second step in the proof is establishing the identity

$$(I - LQ)A_P = \check{A}(I - LQ) \tag{10.18}$$

or by (10.6) the equivalent identity

$$(I - LQ)(I + BS^{-1}B^*Q)\check{A} = \check{A}(I - LQ)(I + PC^*R^{-1}C). \tag{10.19}$$

Since $P = L(I - QL)^{-1} = (I - LQ)^{-1}L$ the right-hand side of (10.19) is equal to

$$\check{A} - \check{A}LQ + \check{A}LC^*R^{-1}C.$$

We substitute $Q - C^*R^{-1}C = \check{A}^*(I + QBS^{-1}B^*)Q\check{A}$ (this identity holds because Q is a solution of the CARE; see (10.9)) to obtain for the right-hand side of (10.19)

$$\check{A} - \check{A}L\check{A}^*(I + QBS^{-1}B^*)Q\check{A}.$$

The control Lyapunov equation tells us that $\check{A}L\check{A}^* = L - \check{B}\check{B}^*$ and so the right-hand side of (10.19) is equal to

$$\check{A} - L(I + QBS^{-1}B^*)Q\check{A} + \check{B}\check{B}^*(I + QBS^{-1}B^*)Q\check{A}.$$

Substituting $\check{B}\check{B}^*(I + QBS^{-1}B^*) = BS^{-1}B^*$ from Lemma 10.6 with $V = Q$ we obtain for the right-hand side of (10.19)

$$\check{A} - L(I + QBS^{-1}B^*)Q\check{A} + BS^{-1}B^*Q\check{A},$$

which is equal to the left-hand side of (10.19). This proves (10.18). We now make the third and last step of the proof that P is a solution of the FARE. We start with the control Lyapunov equation

$$\check{A}L\check{A}^* - L + \check{B}\check{B}^* = 0$$

and substitute $\check{A} = (I-LQ)A_P(I-LQ)^{-1}$ from (10.18) and $\check{A}^* = (I+C^*R^{-1}CP)A_P^*(I+QBS^{-1}B^*)^{-1}$ from (10.6) to obtain

$$(I - LQ)A_P(I - LQ)^{-1}L(I + C^*R^{-1}CP)A_P^*(I + QBS^{-1}B^*)^{-1} - L + \check{B}\check{B}^* = 0.$$

We multiply by $(I - LQ)^{-1}$ from the left and by $(I + QBS^{-1}B^*)$ from the right to obtain

$$A_PP(I + C^*R^{-1}CP)A_P^* - P(I + QBS^{-1}B^*) + (I - LQ)^{-1}\check{B}\check{B}^*(I + QBS^{-1}B^*) = 0.$$

We again use the fact that $\check{B}\check{B}^*(I + QBS^{-1}B^*) = BS^{-1}B^*$ to obtain

$$A_PP(I + C^*R^{-1}CP)A_P^* - P - PQBS^{-1}B^* + (I - LQ)^{-1}BS^{-1}B^* = 0. \tag{10.20}$$

Using that $P = (I - LQ)^{-1}L$ we see that the sum of the two last terms of the left-hand side of (10.20) equals $BS^{-1}B^*$. This proves that P is a solution of the equivalent version (10.7) of the FARE. ∎

REFERENCES

[1] R.F. Curtain and H.J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[2] P.A. Fuhrmann and R. Ober, *A functional approach to LQG balancing*, Internat. J. Control, 57 (1993), pp. 627–741.

[3] A. Halanay and V. Ionescu, *Time Varying Discrete Linear Systems*, Birkhäuser, Basel, 1994.

[4] J. Hoffmann, D. Prätzel-Wolters, and E. Zerz, *A balanced canonical form for discrete-time minimal systems using characteristic maps*, Linear Algebra Appl., 277 (1998), pp. 63–81.

[5] E.A. Jonckheere and L.M. Silverman, *A new set of invariants for linear systems–application to reduced order compensator design*, IEEE Trans. Automat. Control., 28 (1983), pp. 953–964.

[6] D.G. Meyer and G.F. Franklin, *A connection between normalized coprime factorizations and linear quadratic regulator theory*, IEEE Trans. Automat. Control, 32 (1987), pp. 227–228.

[7] B.C. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–32.

[8] D. Mustafa and K. Glover, *Minimum Entropy Control*, Springer-Verlag, Berlin, 1990.

[9] R.J. Ober and D.C. McFarlane, *Balanced canonical forms for minimal systems: A normalized coprime factor approach*, Linear Algebra Appl., 122 (1989), pp. 23–64.

[10] R. Ober and Y. Wu, *Asymptotic stability of infinite-dimensional discrete-time balanced realizations*, SIAM J. Control Optim., 31 (1993), pp. 1321–1339.

[11] A.J. Sasane, *Hankel Norm Approximation for Infinite-Dimensional Systems*, Springer-Verlag, Berlin, 2002.

[12] O.J. Staffans, *Well-Posed Linear Systems*, Encyclopedia Math. Appl. 103, Cambridge University Press, Cambridge, UK, to appear.

[13] E.I. Verriest, *Low sensitivity design and optimal order reduction for the LQG-problem*, in Proceedings of the 24th Midwest Symposium on Circuit Systems, Albuquerque, NM, 1981, pp. 365–369.

[14] E.I. Verriest, *Suboptimal LQG-design via balanced realizations*, in Proceedings of the 20th IEEE Conference on Decision and Control, San Diego, CA, 1981, pp. 686–687.

[15] N.J. Young, *Balanced realizations in infinite dimensions*, Oper. Theory Adv. Appl., 19 (1986), pp. 449–471.

[16] N.J. Young, *Balanced, normal, and intermediate realizations of nonrational transfer functions*, IMA J. Math. Control Inform., 3 (1986), pp. 43–58.

[17] K. Zhou and J.C. Doyle, *Essentials of Robust Control*, Prentice–Hall, Upper Saddle River, NJ, 1998.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1222–1233

NONZERO-SUM STOCHASTIC DIFFERENTIAL GAMES WITH DISCONTINUOUS FEEDBACK*

PAOLA MANNUCCI†

Abstract. The existence of a Nash equilibrium feedback is established for a two-player nonzero-sum stochastic differential game with discontinuous feedback. This is obtained by studying a parabolic system strongly coupled by discontinuous terms.

Key words. nonzero-sum stochastic games, Nash point, strongly coupled parabolic systems, discontinuous feedbacks

AMS subject classifications. 35K50, 49K30, 49N35, 91A10, 91A15

DOI. 10.1137/S0363012903423715

1. Introduction. The aim of this paper is to study the existence of Nash equilibrium points for a two-player nonzero-sum stochastic differential game. The game is governed by a stochastic differential equation with two controls and two payoffs.

This problem can be found, for instance, in Friedman [7] and in a series of papers by Bensoussan and Frehse [2], [3], [4]. All these papers make the assumption that feedback is continuous.

We are interested in studying the problem assuming that the controls take values in compact sets. In this case one cannot expect a Nash equilibrium among continuous feedback, and the Hamiltonian functions associated with the game are nonsmooth.

We consider a simple multidimensional model problem taking two players, affine dynamics, affine payoff functions, and compact control sets.

The loss of continuity of the feedback, due to the hard constraints, leads us to consider a parabolic system strongly coupled by discontinuous terms. In fact, from the usual necessary condition satisfied by the value of the Nash equilibrium feedback in terms of the Hamilton–Jacobi theory, we reduce ourselves to studying the existence of a sufficiently regular solution to a system of nonlinear parabolic equations which contains the Heaviside graph. By this regularity result, we are able to construct Nash equilibrium feedback whose optimality is proved by using the verification approach in the sense of [2], [3], [4].

The motivation for studying games in compact control sets comes from standard nonlinear control theory; this seems a natural assumption in many applications. In particular, Nash equilibria for nonzero-sum deterministic differential games were recently studied by Olsder [12] and Cardaliaguet and Plaskacz [5].

2. Statement of the problem. Let Ω be a bounded smooth domain in $\mathbf{R}^N$. Let X be a process which satisfies the following stochastic differential equation

$$dX(s) = f(s, X(s), u_1(s, X(s)), u_2(s, X(s)))ds + \sigma(s, X(s))dw, \tag{2.1}$$

$$X(t) = x, \tag{2.2}$$

$$s \in [t, T],\ x \in \Omega \subset \mathbf{R}^N.$$

*Received by the editors February 26, 2003; accepted for publication (in revised form) March 8, 2004; published electronically December 1, 2004. This work was partially supported by the Italian MURST national project "Analisi e controllo di equazioni di evoluzione deterministiche e stocastiche."
http://www.siam.org/journals/sicon/43-4/42371.html

†Dipartimento di Matematica Pura e Applicata, Università degli Studi di Padova, 35135 Padova, Italy (mannucci@math.unipd.it).

For each s, $X(s)$ represents the state evolution of a system controlled by two players. The ith player acts by means of a feedback control function $u_i : (t,T) \times \mathbf{R}^N \to U_i \subset \mathbf{R}^{k_i}$, where $i = 1, 2$.

Let $\mathcal{U}_i :\equiv \{u_i$ Borel-measurable applications $(t,T)\times\mathbf{R}^N \to U_i \subset \mathbf{R}^{k_i}\}$, $i = 1, 2$, be the set of the control functions u_i with values $u_i(s, X)$ in U_i. The term $\sigma(s, X(s))dw$ represents the "noise," where w is an N-dimensional standard Brownian motion and σ is an $N \times N$ matrix. We assume that σ does not depend on the control variables u_1, u_2 and that σ and σ^{-1} are bounded and Lipschitz on X. The function $f(s, X, u_1, u_2)$ is called the dynamic of the game (2.1).

We refer to [7] for the definitions about stochastic processes, stochastic differential equations and functional spaces.

A control function $u_i \in \mathcal{U}_i$ will be called *admissible* if it is adapted to the filtration defined on the probability space.

It is possible to prove, using Girsanov's theorem, that, under convenient assumptions on f, for all $(u_1, u_2) \in \mathcal{U}_1 \times \mathcal{U}_2$ admissible controls, there exists a unique weak solution to the problem (2.1), (2.2) (see, for example, [4], [9], [6, Chapter 4]).

For any choice of admissible controls u_1, u_2 we have the following *payoff* functions:

$$J_i(t,x,u_1,u_2) = E_{tx}\left\{ \int_t^\tau l_i(s, X(s), u_1(s, X(s)), u_2(s, X(s)))ds + g_i(T, X(T)) \right\}, \quad i = 1, 2, \tag{2.3}$$

where $\tau \equiv T \wedge \inf\{s \geq t, X(s) \notin \Omega\}$, E_{tx} is the expectation under the probability P_{tx}, l_i and g_i are prescribed functions (the assumptions will be specified later), and $X = X(s)$ is the unique weak solution of (2.1)–(2.2) corresponding to $(u_1, u_2) \in \mathcal{U}_1 \times \mathcal{U}_2$ admissible controls.

Each player wants to maximize his own payoff.

DEFINITION 2.1. *A pair of admissible controls $(\overline{u}_1, \overline{u}_2) \in \mathcal{U}_1 \times \mathcal{U}_2$ is called the* Nash equilibrium point *of the differential game* (2.1)–(2.2)*, with payoff* (2.3)*, if*

$$J_1(t,x,\overline{u}_1,\overline{u}_2) \geq J_1(t,x,u_1,\overline{u}_2), \tag{2.4}$$

$$J_2(t,x,\overline{u}_1,\overline{u}_2) \geq J_2(t,x,\overline{u}_1,u_2), \tag{2.5}$$

for all $(t,x) \in (0,T) \times \Omega$ and for all $(u_1, u_2) \in \mathcal{U}_1 \times \mathcal{U}_2$ admissible controls.

The functions

$$V_1(t,x) :\equiv J_1(t,x,\overline{u}_1,\overline{u}_2), \quad V_2(t,x) :\equiv J_2(t,x,\overline{u}_1,\overline{u}_2) \tag{2.6}$$

are a *value* of the Nash equilibrium point $(\overline{u}_1, \overline{u}_2)$.

We define the pre-Hamiltonians $H_i(t,x,p,u_1,u_2) : (0,T)\times\mathbf{R}^N\times\mathbf{R}^N\times U_1\times U_2 \to \mathbf{R}$, $i = 1, 2$:

$$\begin{aligned} H_1(t,x,p,u_1(t,x),u_2(t,x)) :\equiv\ & p \cdot f(t,x,u_1(t,x),u_2(t,x)) \\ & + l_1(t,x,u_1(t,x),u_2(t,x)), \\ H_2(t,x,p,u_1(t,x),u_2(t,x)) :\equiv\ & p \cdot f(t,x,u_1(t,x),u_2(t,x)) \\ & + l_2(t,x,u_1(t,x),u_2(t,x)). \end{aligned} \tag{2.7}$$

We set

$$a = \frac{1}{2}\sigma\sigma^*$$

(σ^* is the transpose of σ) to be the matrix with elements $a_{h,k}$, $h, k = 1, \ldots, N$.

If the value functions $V_1,\ V_2 \in C^{1,2}$, we can apply Itô's formula; changing the time variable $(T-t \to t)$, we get that $V_1,\ V_2$ solve, in $\Omega_T :\equiv (0,T)\times\Omega$, the following nonlinear parabolic system coupled by the Nash equilibrium problem:

$$\begin{aligned}(2.8)\qquad &\frac{\partial V_1(t,x)}{\partial t} - \sum_{h,k=1}^{N} a_{hk}(t,x)\frac{\partial^2 V_1(t,x)}{\partial x_h \partial x_k}\\ &= \max_{\{u_1\in U_1\}} H_1(t,x,\nabla_x V_1(t,x), u_1(t,x), \overline{u}_2(t,x))\\ &= H_1(t,x,\nabla_x V_1(t,x), \overline{u}_1(t,x), \overline{u}_2(t,x)),\end{aligned}$$

$$\begin{aligned}(2.9)\qquad &\frac{\partial V_2(t,x)}{\partial t} - \sum_{h,k=1}^{N} a_{hk}(t,x)\frac{\partial^2 V_2(t,x)}{\partial x_h \partial x_k}\\ &= \max_{\{u_2\in U_2\}} H_2(t,x,\nabla_x V_2(t,x), \overline{u}_1(t,x), u_2(t,x))\\ &= H_2(t,x,\nabla_x V_2(t,x), \overline{u}_1(t,x), \overline{u}_2(t,x)),\end{aligned}$$

$$(2.10)\qquad \overline{u}_1(t,x) \in \operatorname{argmax}_{\{u_1\in U_1\}} H_1(t,x,\nabla_x V_1(t,x), u_1(t,x), \overline{u}_2(t,x)),$$

$$(2.11)\qquad \overline{u}_2(t,x) \in \operatorname{argmax}_{\{u_2\in U_2\}} H_2(t,x,\nabla_x V_2(t,x), \overline{u}_1(t,x), u_2(t,x)),$$

$$(2.12)\qquad V_1 = g_1(t,x),\ V_2 = g_2(t,x) \quad \text{on } \partial_p\Omega_T,$$

where $\partial_p\Omega_T :\equiv \big((0,T)\times\partial\Omega\big) \cup \big(\{t=0\}\times\Omega\big)$. (Here and in the following we write, for the sake of brevity, $\operatorname{argmax}_{u_i\in U_i} H_i$, which means $\operatorname{argmax}_{u_i(t,x)\in U_i} H_i$).

The functions

$$\begin{aligned}&H_1(t,x,\nabla_x V_1(t,x), \overline{u}_1(t,x), \overline{u}_2(t,x))\\ &\quad = \max_{\{u_1\in U_1\}} H_1(t,x,\nabla_x V_1(t,x), u_1(t,x), \overline{u}_2(t,x)),\\ &H_2(t,x,\nabla_x V_1(t,x), \overline{u}_1(t,x), \overline{u}_2(t,x))\\ &\quad = \max_{\{u_2\in U_2\}} H_2(t,x,\nabla_x V_2(t,x), \overline{u}_1(t,x), u_2(t,x))\end{aligned}$$

are called the Hamiltonian functions associated with the game (2.1)–(2.3).

We want to outline here the classical procedure used in Friedman's book [7] to prove the existence of a Nash equilibrium point $\overline{u}_1,\ \overline{u}_2$.

1. Suppose that, for any fixed $p\in\mathbf{R}^N$, there exist $u_1^*,\ u_2^*$ such that

$$\begin{aligned}&u_1^*(t,x,p) \in \operatorname{argmax}_{\{u_1\in U_1\}} H_1(t,x,p,u_1(t,x),u_2^*(t,x)),\\ &u_2^*(t,x,p) \in \operatorname{argmax}_{\{u_2\in U_2\}} H_2(t,x,p,u_1^*(t,x),u_2(t,x))\\ &\text{are measurable in } (t,x)\in\Omega_T \text{ and continuous in } p.\end{aligned}\qquad (2.13)$$

2. Solve the parabolic system

$$\begin{aligned}(2.14)\qquad &\frac{\partial V_1(t,x)}{\partial t} - \sum_{h,k=1}^{N} a_{hk}(t,x)\frac{\partial^2 V_1(t,x)}{\partial x_h \partial x_k}\\ &= H_1(t,x,\nabla_x V_1, u_1^*(t,x,\nabla_x V_1), u_2^*(t,x,\nabla_x V_2)),\end{aligned}$$

$$\begin{aligned}&\frac{\partial V_2(t,x)}{\partial t} - \sum_{h,k=1}^{N} a_{hk}(t,x)\frac{\partial^2 V_2(t,x)}{\partial x_h \partial x_k}\\ &= H_2(t,x,\nabla_x V_2, u_1^*(t,x,\nabla_x V_1), u_2^*(t,x,\nabla_x V_2)),\\ &V_1 = g_1(t,x),\ V_2 = g_2(t,x) \quad \text{on } \partial_p\Omega_T.\end{aligned}$$

3. Prove that the pair of functions $(\overline{u}_1, \overline{u}_2)$ with values

$$\overline{u}_1(t,x) :\equiv u_1^*(t,x,\nabla_x V_1(t,x)),$$

$$\overline{u}_2(t,x) :\equiv u_2^*(t,x,\nabla_x V_2(t,x))$$

is a Nash equilibrium point (see Definition 2.1).

Therefore to obtain Nash equilibrium points for the associated stochastic differential game, we look for a "sufficiently regular" solution of system (2.14).

A similar procedure is used, in the elliptic case, by Bensoussan and J. Frehse [2], [3], [4] to study systems of Bellman equations.

We want to emphasize that the results of Friedman and Bensoussan and J. Frehse on the existence of classical solutions and of Nash equilibrium points are obtained under the assumption that there exist some feedback

$$u_1^* \in \operatorname{argmax}_{\{u_1 \in U_1\}} H_1(t,x,p,u_1,u_2^*), \qquad u_2^* \in \operatorname{argmax}_{\{u_2 \in U_2\}} H_2(t,x,p,u_1^*,u_2)$$

that are continuous in p (see, for example, assumption (D) [7, section 17, p. 497]). If we assume that the sets U_i, $i = 1, 2$, are compact, the assumption on the continuity of the feedback can be too restrictive.

Weaker assumptions on the regularity of the feedback can be found in [8] and [9].

In this paper we consider a model problem with U_1, U_2 compact sets in $\mathbf{R}$, affine dynamics of the game, and affine payoff.

Let us list the assumptions:

$$U_1 = U_2 = [0,1], \tag{2.15}$$

$$\begin{aligned} &f(x,u_1(t,x),u_2(t,x)) :\ \Omega \times U_1 \times U_2 \to \mathbf{R}^N, \\ &f(x,u_1(t,x),u_2(t,x)) = f_1(x)u_1(t,x) + f_2(x)u_2(t,x), \\ &f_i(x) :\ \Omega \to \mathbf{R}^N, f_i(x) \in C^1(\overline{\Omega}),\ i = 1,2, \end{aligned} \tag{2.16}$$

$$\begin{aligned} &l_i(x,u_1(t,x),u_2(t,x)) :\ \Omega \times U_1 \times U_2 \to \mathbf{R}, \\ &l_i(x,u_1(t,x),u_2(t,x)) = l_i(x)u_i(t,x), \quad i = 1,2, \\ &l_i(x) :\ \Omega \to \mathbf{R}^N,\ l_i(x) \in C^1(\overline{\Omega}), \quad i = 1,2, \end{aligned} \tag{2.17}$$

$$g_i(t,x) \in H^{1+\alpha}(\overline{\Omega}_T),\ \alpha \in (0,1), \quad i = 1,2, \tag{2.18}$$

$$a_{hk}(t,x) \in C^2(\overline{\Omega}_T),$$

$$\nu|\xi|^2 \le \sum_{h,k=1}^{N} a_{hk}(t,x)\xi_h\xi_k \le \mu|\xi|^2, \quad \nu,\mu > 0, \tag{2.19}$$

$$\text{for all } (t,x) \in \Omega_T \text{ and for all } \xi \in \mathbf{R}^N.$$

Taking into account the affine structure of f and l in (2.16)–(2.17), the functions H_1 and H_2 in (2.7) become

$$\begin{aligned} &H_1(x,p,u_1(t,x),u_2(t,x)) = \big(p \cdot f_1(x) + l_1(x)\big)u_1(t,x) + p \cdot f_2(x)u_2(t,x), \\ &H_2(x,p,u_1(t,x),u_2(t,x)) = \big(p \cdot f_2(x) + l_2(x)\big)u_2(t,x) + p \cdot f_1(x)u_1(t,x). \end{aligned} \tag{2.20}$$

From (2.15) and (2.20), for any fixed p, we have

$$\begin{aligned}
&\overline{u}_1(t,x) \in \operatorname{argmax}_{\{u_1\in U_1\}} H_1(x,p,u_1(t,x),\overline{u}_2(t,x))\\
&= \operatorname{argmax}_{\{u_1(t,x)\in[0,1]\}} \big(p\cdot f_1(x)+l_1(x)\big)u_1(t,x) + p\cdot f_2(x)\overline{u}_2(t,x)\\
&= \operatorname{Heav}\big(p\cdot f_1(x)+l_1(x)\big), \qquad (2.21)
\end{aligned}$$

where $\operatorname{Heav}(\eta)$ is the Heaviside graph defined as $\operatorname{Heav}(\eta)=1$ if $\eta>0$, $\operatorname{Heav}(\eta)=0$ if $\eta<0$, and $\operatorname{Heav}(0)=[0,1]$.

Analogously

$$\overline{u}_2(t,x) \in \operatorname{Heav}\big(p\cdot f_2(x)+l_2(x)\big). \qquad (2.22)$$

From (2.20), (2.21), (2.22), the Hamiltonian functions assume the form

$$\begin{aligned}
H_1(x,p,\overline{u}_1(t,x),\overline{u}_2(t,x)) &\equiv \max_{\{u_1\in U_1\}} H_1(x,p,u_1(t,x),\overline{u}_2(t,x))\\
&= \big(p\cdot f_1(x)+l_1(x)\big)_+ + p\cdot f_2(x)\overline{u}_2(t,x), \qquad (2.23)\\
H_2(x,p,\overline{u}_1,\overline{u}_2) &\equiv \max_{\{u_2\in U_2\}} H_2(x,p,\overline{u}_1(t,x),u_2(t,x))\\
&= \big(p\cdot f_2(x)+l_2(x)\big)_+ + p\cdot f_1(x)\overline{u}_1(t,x), \qquad (2.24)
\end{aligned}$$

where we denoted by $(h)_+$ the positive part of the function h.

Taking, for any fixed p,

$$u_1^*(t,x,p) \in \operatorname{Heav}\big(p\cdot f_1(x)+l_1(x)\big), \qquad (2.25)$$

$$u_2^*(t,x,p) \in \operatorname{Heav}\big(p\cdot f_2(x)+l_2(x)\big), \qquad (2.26)$$

we have that u_1^*, u_2^* do not satisfy (2.13) because they are not continuous in p. Hence, in this case, we cannot use the results of [7].

From (2.25), (2.26), the parabolic system (2.8), (2.9) becomes

$$\begin{aligned}
\frac{\partial V_1}{\partial t} - \sum_{h,k=1}^{N} a_{hk}\frac{\partial^2 V_1}{\partial x_h \partial x_k} \in{}& \Big(\nabla_x V_1\cdot f_1(x)+l_1(x)\Big)\operatorname{Heav}(\nabla_x V_1\cdot f_1(x)+l_1(x))\\
&+ \nabla_x V_1\cdot f_2(x)\operatorname{Heav}(\nabla_x V_2\cdot f_2(x)+l_2(x)) \text{ in } \Omega_T, \qquad (2.27)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial V_2}{\partial t} - \sum_{h,k=1}^{N} a_{hk}\frac{\partial^2 V_2}{\partial x_h \partial x_k} \in{}& \Big(\nabla_x V_2\cdot f_2(x)+l_2(x)\Big)\operatorname{Heav}(\nabla_x V_2\cdot f_2(x)+l_2(x))\\
&+ \nabla_x V_2\cdot f_1(x)\operatorname{Heav}(\nabla_x V_1\cdot f_1(x)+l_1(x)) \text{ in } \Omega_T, \qquad (2.28)
\end{aligned}$$

$$V_1(t,x)=g_1(t,x),\ V_2(t,x)=g_2(t,x) \quad \text{on } \partial_p\Omega_T. \qquad (2.29)$$

This is a uniformly parabolic system strongly coupled by the Heaviside graph containing the first order derivatives of the unknown functions.

Equations (2.27) and (2.28) are to be interpreted in the following way:

$$\frac{\partial V_1}{\partial t} - \sum_{h,k=1}^{N} a_{hk}\frac{\partial^2 V_1}{\partial x_h \partial x_k} = \Big(\nabla_x V_1\cdot f_1+l_1\Big)h_1(t,x) + \nabla_x V_1\cdot f_2 h_2(t,x),$$

$$\frac{\partial V_2}{\partial t} - \sum_{h,k=1}^{N} a_{hk}\frac{\partial^2 V_2}{\partial x_h \partial x_k} = \Big(\nabla_x V_2\cdot f_2+l_2\Big)h_2(t,x) + \nabla_x V_2\cdot f_1 h_1(t,x),$$

$$h_1(t,x)\in\operatorname{Heav}(\nabla_x V_1\cdot f_1(x)+l_1(x)),\ h_2(t,x)\in\operatorname{Heav}(\nabla_x V_2\cdot f_2(x)+l_2(x)).$$

Following the previous scheme, first we investigate the existence of a solution V_1, V_2 of (2.27)–(2.29). Next, if we find sufficient regularity, it will be possible to prove the existence of a Nash equilibrium point.

In section 3 we provide an existence result for a solution V_1, V_2 in $H^{1+\alpha}(\overline{\Omega}_T) \cap W_q^{1,2}(\Omega_T)$ of the system (2.27)–(2.29), and in section 4 we prove the existence of a Nash equilibrium point.

3. Existence of a solution to the parabolic system. We give the following definition.

DEFINITION 3.1. (V_1, V_2) *is a* strong *solution of the system* (2.27)–(2.29) *if*

(a) $V_1(t,x),\ V_2(t,x) \in H^{1+\alpha}(\overline{\Omega}_T) \cap W_q^{1,2}(\Omega_T)$ *for some* $\alpha \in (0,1),\ q > N+2$*;*

(b) *equations* (2.27)–(2.28) *hold almost everywhere and* (2.29) *holds.*

THEOREM 3.2. *Under assumptions* (2.15)–(2.19)*, taking* $\mathrm{Heav}(\eta) = 1$ *if* $\eta > 0$*,* $\mathrm{Heav}(\eta) = 0$ *if* $\eta < 0$*, and* $\mathrm{Heav}(0) = [0,1]$*, there exists at least a strong solution* $(V_1,\ V_2)$ *of the parabolic system* (2.27)–(2.29)*.*

Proof. Let us consider the approximating problems obtained by replacing the Heaviside graph $\mathrm{Heav}(\eta)$ with smooth functions H_n:

$$
\begin{aligned}
&H_n(\eta) \in C^\infty(\mathbf{R}),\ H_n(\eta) \in L_\infty(\mathbf{R}),\\
&H_n(\eta) = 0 \text{ if } \eta \le 0, \quad H_n(\eta) = 1 \text{ if } \eta \ge \frac{1}{n},\\
&H_n' \ge 0,\\
&H_n(\eta) \to \mathrm{Heav}(\eta) \text{ in } L_p(K),\ p > 1,\ K \subset \mathbf{R} \text{ is any compact of } \mathbf{R},\\
&H_n(\eta) \to \mathrm{Heav}(\eta) \text{ in } C^0 \text{ outside a neighbourhood of } \eta = 0.
\end{aligned}
\tag{3.1}
$$

We denote by V_{1n}, V_{2n} the solution of the problem

$$
\begin{aligned}
&\frac{\partial V_{1n}}{\partial t} - \sum_{h,k=1}^{N} a_{hk} \frac{\partial^2 V_{1n}}{\partial x_h \partial x_k} = \Big(\nabla_x V_{1n} \cdot f_1 + l_1\Big) H_n(\nabla_x V_{1n} \cdot f_1 + l_1)\\
&+ \nabla_x V_{1n} \cdot f_2 H_n(\nabla_x V_{2n} \cdot f_2 + l_2) \text{ in } \Omega_T,
\end{aligned}
\tag{3.2}
$$

$$
\begin{aligned}
&\frac{\partial V_{2n}}{\partial t} - \sum_{h,k=1}^{N} a_{hk} \frac{\partial^2 V_{2n}}{\partial x_h \partial x_k} = \Big(\nabla_x V_{2n} \cdot f_2 + l_2\Big) H_n(\nabla_x V_{2n} \cdot f_2 + l_2)\\
&+ \nabla_x V_{2n} \cdot f_1 H_n(\nabla_x V_{1n} \cdot f_1 + l_1) \text{ in } \Omega_T,
\end{aligned}
\tag{3.3}
$$

$$
V_{1n} = g_1,\ V_{2n} = g_2 \ \text{ in } \partial_p \Omega_T. \tag{3.4}
$$

From [11, Theorem 7.1, p. 596] on quasi-linear parabolic systems with smooth coefficients, there exists an unique solution of problem (3.2)–(3.4), $V_{1n}(t,x)$, $V_{2n}(t,x) \in H^{2+\alpha}(\Omega_T)$.

At this point, regarding the terms $H_n(\nabla_x V_{1n} \cdot f_1 + l_1) + f_2 H_n(\nabla_x V_{2n} \cdot f_2 + l_2)$ and $H_n(\nabla_x V_{2n} \cdot f_2 + l_2) + f_1 H_n(\nabla_x V_{1n} \cdot f_1 + l_1)$ in (3.2)–(3.3) as bounded uniformly on n, from [11, Theorem 9.1, p. 341], we find an uniform estimate in $W_q^{1,2}$:

$$
\|V_{1n}\|_{q,\Omega_T}^{(2)} + \|V_{2n}\|_{q,\Omega_T}^{(2)} \le C(\|H_n\|_{q,\Omega_T}, \|g_1\|_{q,\partial_p\Omega_T}^{(2-1/q)}, \|g_2\|_{q,\partial_p\Omega_T}^{(2-1/q)}) \le C, \tag{3.5}
$$

where C is independent of n and $q > 1$.

By means of an embedding theorem (see, for example, [11, Chapter 2, Lemma 3.3]), taking $q > N+2$, we obtain

$$
|V_{1n}|_{\Omega_T}^{(1+\alpha)} + |V_{2n}|_{\Omega_T}^{(1+\alpha)} \le C, \qquad \alpha = 1 - \frac{N+2}{q}, \tag{3.6}
$$

where C is independent of n.

We can now extract two subsequences, which we denote again by V_{1n}, V_{2n} such that

$$\begin{aligned} &V_{in} \to V_i \text{ in } C^0(\Omega_T), \quad i=1,2, \\ &\frac{\partial V_{in}}{\partial x_h} \to \frac{\partial V_i}{\partial x_h} \text{ in } C^0(\Omega_T), \quad i=1,2,\ h=1,\dots N. \end{aligned} \tag{3.7}$$

From the weak precompactness of the unit ball of $W^{2,1}_q$, we have

$$\begin{aligned} &\frac{\partial V_{in}}{\partial t} \rightharpoonup \frac{\partial V_i}{\partial t} \text{ weakly in } L_2(\Omega_T), \quad i=1,2, \\ &\frac{\partial^2 V_{in}}{\partial x_h \partial x_k} \rightharpoonup \frac{\partial^2 V_i}{\partial x_h \partial x_k} \text{ weakly in } L_2(\Omega_T), \quad i=1,2,\ h,k=1,\dots N. \end{aligned} \tag{3.8}$$

From (3.7), (3.8)

$$V_1(t,x),\ V_2(t,x) \in H^{1+\alpha}(\Omega_T) \cap W^{1,2}_q(\Omega_T), \text{ with } \alpha = 1 - \frac{N+2}{q}. \tag{3.9}$$

Now we have to prove that V_1, V_2 solve (2.27)–(2.28) almost everywhere in Ω_T.

From assumptions (3.1) and (3.8), the two sequences $H_n(\nabla_x V_{1n} \cdot f_1 + l_1)$, $H_n(\nabla_x V_{2n} \cdot f_2 + l_2)$ are uniformly bounded in $L_2(\Omega_T)$, and hence we can extract two subsequences such that

$$\begin{aligned} &H_n(\nabla_x V_{1n} \cdot f_1 + l_1) \rightharpoonup h_1(t,x) \text{ weakly in } L_2(\Omega_T), \\ &H_n(\nabla_x V_{2n} \cdot f_2 + l_2) \rightharpoonup h_2(t,x) \text{ weakly in } L_2(\Omega_T). \end{aligned} \tag{3.10}$$

We now show that $h_i(t,x) \in \mathrm{Heav}(\nabla_x V_i \cdot f_i + l_i)$ almost everywhere $i=1,2$. To do this let us consider the following sets:

$$\mathcal{P}_i :\equiv \{(t,x) \in \Omega_T : \nabla_x V_i(t,x) \cdot f_i(x) + l_i(x) > 0\},$$

$$\mathcal{N}_i :\equiv \{(t,x) \in \Omega_T : \nabla_x V_i(t,x) \cdot f_i(x) + l_i(x) < 0\},$$

$$\mathcal{Z}_i :\equiv \{(t,x) \in \Omega_T : \nabla_x V_i(t,x) \cdot f_i(x) + l_i(x) = 0\},$$

$$i = 1,2.$$

From (3.7), we have that, for a sufficiently large n, $\nabla_x V_{in}(t,x) \cdot f_i(x) + l_i(x) > 0$ for all $(t,x) \in \mathcal{P}_i$. Hence, from (3.1), we obtain that

$$H_n(\nabla_x V_{in}(t,x) \cdot f_i(x) + l_i(x)) \to 1 = \mathrm{Heav}(\nabla_x V_i(t,x) \cdot f_i(x) + l_i(x))$$

for all $(t,x) \in \mathcal{P}_i$, $i=1,2$.

Analogously, in $\mathcal{N}_i$, for a sufficiently large n, $\nabla_x V_{in}(x,t) \cdot f_i(x) + l_i(x) < 0$, and hence

$$H_n(\nabla_x V_{in}(x,t) \cdot f_i(x) + l_i(x)) \equiv 0 = \mathrm{Heav}(\nabla_x V_i(x,t) \cdot f_i(x) + l_i(x))$$

for all $(t,x) \in \mathcal{N}_i$, $i=1,2$.

In $\mathcal{Z}_i$, from the assumptions on H_n (see (3.1)), we have that $0 \le h_i \le 1$ almost everywhere, and hence

$$h_1(t,x) \in \mathrm{Heav}(\nabla_x V_1(t,x)\cdot f_1(x) + l_1(x)) \text{ almost everywhere in } \Omega_T, \tag{3.11}$$

$$h_2(t,x) \in \mathrm{Heav}(\nabla_x V_2(t,x)\cdot f_2(x) + l_2(x)) \text{ almost everywhere in } \Omega_T. \tag{3.12}$$

At this point, from (3.7), (3.8), (3.10), (3.11), (3.12), we obtain that V_1, V_2 satisfy (2.27)–(2.28) almost everywhere in Ω_T and from the regularity of the functions V_1, V_2, we have that $V_1(t,x) = g_1(t,x)$, $V_2(t,x) = g_2(t,x)$ for all $(t,x) \in \partial_p\Omega_T$. □

REMARK 3.1. *If we choose $W(\eta) \in \mathrm{Heav}(\eta)$, we are not able to solve the problem*

$$\begin{aligned} \frac{\partial V_1}{\partial t} - \sum_{h,k=1}^{N} a_{hk}\frac{\partial^2 V_1}{\partial x_h \partial x_k} &= \big(\nabla_x V_1\cdot f_1 + l_1\big)W(\nabla_x V_1\cdot f_1 + l_1) \\ &\quad + \nabla_x V_1\cdot f_2\, W(\nabla_x V_2\cdot f_2 + l_2), \end{aligned} \tag{3.13}$$

$$\begin{aligned} \frac{\partial V_2}{\partial t} - \sum_{h,k=1}^{N} a_{hk}\frac{\partial^2 V_2}{\partial x_h \partial x_k} &= \big(\nabla_x V_2\cdot f_2 + l_2\big)W(\nabla_x V_2\cdot f_2 + l_2) \\ &\quad + \nabla_x V_2\cdot f_1\, W(\nabla_x V_1\cdot f_1 + l_1), \end{aligned} \tag{3.14}$$

because we cannot exclude that $\mathrm{meas}\{(t,x)\in\Omega_T : \nabla_x V_i\cdot f_i(x) + l_i(x) = 0\} > 0$, $i = 1,2$. *Hence we cannot prove that*

$$H_n(\nabla_x V_{in}\cdot f_i + l_i) \rightharpoonup W(\nabla_x V_i\cdot f_i + l_i), \quad i = 1,2,$$

but only that

$$H_n(\nabla_x V_{in}\cdot f_i + l_i) \rightharpoonup h_i \in \mathrm{Heav}(\nabla_x V_i\cdot f_i + l_i), \quad i = 1,2.$$

4. Existence of a Nash equilibrium point. We now prove the following.

THEOREM 4.1. *Suppose that the assumptions of Theorem* 3.2 *hold. Let* $(V_1,\ V_2)$ *be a strong solution of the parabolic system* (2.27)–(2.29); *then any admissible control* $(\overline{u}_1, \overline{u}_2)$ *such that*

$$\overline{u}_1(t,x) \in \mathrm{Heav}(\nabla_x V_1(t,x)\cdot f_1(x) + l_1(x)), \tag{4.1}$$

$$\overline{u}_2(t,x) \in \mathrm{Heav}(\nabla_x V_2(t,x)\cdot f_2(x) + l_2(x)) \tag{4.2}$$

is a Nash equilibrium point for the stochastic differential game (2.1)–(2.2) *with payoff* (2.3).

Proof. The existence of a strong solution $(V_1,\ V_2)$ of the parabolic system (2.27)–(2.29) is stated by Theorem 3.2.

To prove that $\overline{u}_i(t,x) \in \mathrm{Heav}(\nabla_x V_i(t,x)\cdot f_i(x) + l_i(x))$, $i = 1,2$, are the values of a Nash equilibrium point, as in Definition 2.1, we have to show that

$$\begin{aligned} &J_1(t,x,\overline{u}_1,\overline{u}_2) \ge J_1(t,x,u_1,\overline{u}_2), \\ &J_2(t,x,\overline{u}_1,\overline{u}_2) \ge J_2(t,x,\overline{u}_1,u_2) \end{aligned} \tag{4.3}$$

for all $(u_1,u_2) \in \mathcal{U}_1 \times \mathcal{U}_2$ admissible controls.

Let us denote

$$\begin{aligned} &v_1(t,x) :\equiv J_1(t,x,\overline{u}_1,\overline{u}_2), \\ &v_2(t,x) :\equiv J_2(t,x,\overline{u}_1,\overline{u}_2). \end{aligned} \tag{4.4}$$

Using a generalization of Itô's formula applied to functions in $W_q^{1,2}$ (see, for example, [1, Theorem 4.1, p. 126]), we have that the couple (v_1, v_2) solves the following parabolic system (here and in the following we omit, for the sake of brevity, the dependence on the variables (t, x)):

$$\frac{\partial v_1}{\partial t} - \sum_{h,k=1}^{N} a_{hk} \frac{\partial^2 v_1}{\partial x_h \partial x_k} = H_1(x, t, \nabla_x v_1, \overline{u}_1, \overline{u}_2) \tag{4.5}$$
$$= (\nabla_x v_1 f_1 + l_1)\overline{u}_1 + \nabla_x v_1 f_2 \overline{u}_2 \ \text{ in } \ \Omega_T,$$

$$\frac{\partial v_2}{\partial t} - \sum_{h,k=1}^{N} a_{hk} \frac{\partial^2 v_2}{\partial x_h \partial x_k} = H_2(x, t, \nabla_x v_2, \overline{u}_1, \overline{u}_2) \tag{4.6}$$
$$= (\nabla_x v_2 \cdot f_2 + l_2)\overline{u}_2 + \nabla_x v_2 f_1 \overline{u}_1 \ \text{ in } \ \Omega_T,$$

$$v_1 = g_1(t, x), \ v_2 = g_2(t, x) \ \text{ on } \ \partial_p \Omega_T. \tag{4.7}$$

From (4.1), (4.2), we have

$$\frac{\partial v_1}{\partial t} - \sum_{h,k=1}^{N} a_{hk} \frac{\partial^2 v_1}{\partial x_h \partial x_k} \in (\nabla_x v_1 \cdot f_1 + l_1)\text{Heav}(\nabla_x V_1 \cdot f_1 + l_1)$$
$$+ \nabla_x v_1 \cdot f_2 \text{Heav}(\nabla_x V_2 \cdot f_2 + l_2) \ \text{ in } \ \Omega_T, \tag{4.8}$$
$$\frac{\partial v_2}{\partial t} - \sum_{h,k=1}^{N} a_{hk} \frac{\partial^2 v_2}{\partial x_h \partial x_k} \in (\nabla_x v_2 \cdot f_2 + l_2)\text{Heav}(\nabla_x V_2 \cdot f_2 + l_2)$$
$$+ \nabla_x v_2 \cdot f_1 \text{Heav}(\nabla_x V_1 \cdot f_1 + l_1) \ \text{ in } \ \Omega_T,$$

$$v_1 = g_1(t, x), \ v_2 = g_2(t, x) \ \text{ on } \ \partial_p \Omega_T. \tag{4.9}$$

Let us now fix $(u_1, u_2) \in \mathcal{U}_1 \times \mathcal{U}_2$ admissible controls and denote

$$w_1(t, x) :\equiv J_1(t, x, u_1, \overline{u}_2), \tag{4.10}$$
$$w_2(t, x) :\equiv J_2(t, x, \overline{u}_1, u_2).$$

The couple (w_1, w_2) solves the following parabolic system:

$$\frac{\partial w_1}{\partial t} - \sum_{h,k=1}^{N} a_{hk} \frac{\partial^2 w_1}{\partial x_h \partial x_k} = H_1(x, t, \nabla_x w_1, u_1, \overline{u}_2) \tag{4.11}$$
$$= (\nabla_x w_1 f_1 + l_1)u_1 + \nabla_x w_1 f_2 \overline{u}_2 \ \text{ in } \ \Omega_T,$$

$$\frac{\partial w_2}{\partial t} - \sum_{h,k=1}^{N} a_{hk} \frac{\partial^2 w_2}{\partial x_h \partial x_k} = H_2(x, t, \nabla_x w_2, \overline{u}_1, u_2) \tag{4.12}$$
$$= (\nabla_x w_2 \cdot f_2 + l_2)u_2 + \nabla_x w_2 f_1 \overline{u}_1 \ \text{ in } \ \Omega_T,$$

$$w_1 = g_1(t, x), \ w_2 = g_2(t, x) \ \text{ on } \ \partial_p \Omega_T, \tag{4.13}$$
$$w_1, w_2 \in W_q^{1,2}(\Omega_T).$$

From the expressions (2.20) of H_1, H_2, taking into account (2.10), (2.11), we have that, for any p fixed,

$$(pf_1 + l_1)u_1(t, x) \le (pf_1 + l_1)\overline{u}_1(t, x), \tag{4.14}$$
$$(pf_2 + l_2)u_2(t, x) \le (pf_2 + l_2)\overline{u}_2(t, x).$$

Consider now the functions $z_1 :\equiv v_1 - w_1$, $z_2 :\equiv v_2 - w_2$. From systems (4.5)–(4.7) and (4.11)–(4.13) we have

$$\begin{aligned}
\frac{\partial z_1}{\partial t} - \sum_{h,k=1}^{N} a_{hk}\frac{\partial^2 z_1}{\partial x_h \partial x_k} &= (\nabla_x v_1 \cdot f_1 + l_1)\overline{u}_1 + \nabla_x v_1 \cdot f_2 \overline{u}_2 \\
&\quad - (\nabla_x w_1 \cdot f_1 + l_1)u_1 - \nabla_x w_1 \cdot f_2 \overline{u}_2 \quad \text{in } \Omega_T, \\
\frac{\partial z_2}{\partial t} - \sum_{h,k=1}^{N} a_{hk}\frac{\partial^2 z_2}{\partial x_h \partial x_k} &= (\nabla_x v_2 \cdot f_2 + l_2)\overline{u}_2 + \nabla_x v_2 \cdot f_1 \overline{u}_1 \\
&\quad - (\nabla_x w_2 \cdot f_2 + l_2)u_2 - \nabla_x w_2 \cdot f_1 \overline{u}_1 \quad \text{in } \Omega_T, \\
z_1 = z_2 &= 0 \quad \text{on } \partial_p \Omega_T.
\end{aligned} \tag{4.15}$$

Taking into account (4.14), we obtain

$$\begin{aligned}
\frac{\partial z_1}{\partial t} - \sum_{h,k=1}^{N} a_{hk}\frac{\partial^2 z_1}{\partial x_h \partial x_k} &\geq (\nabla_x v_1 \cdot f_1 + l_1)u_1 + \nabla_x v_1 \cdot f_2 \overline{u}_2 \\
&\quad - (\nabla_x w_1 \cdot f_1 + l_1)u_1 - \nabla_x w_1 \cdot f_2 \overline{u}_2 \\
&= \nabla_x z_1 \cdot \Big(f_1 u_1 + f_2 \overline{u}_2 \Big) \quad \text{in } \Omega_T,
\end{aligned} \tag{4.16}$$

$$\begin{aligned}
\frac{\partial z_2}{\partial t} - \sum_{h,k=1}^{N} a_{hk}\frac{\partial^2 z_2}{\partial x_h \partial x_k} &\geq (\nabla_x v_2 \cdot f_2 + l_2)u_2 + \nabla_x v_2 \cdot f_1 \overline{u}_1 \\
&\quad - (\nabla_x w_2 \cdot f_2 + l_2)u_2 - \nabla_x w_2 \cdot f_1 \overline{u}_1 \\
&= \nabla_x z_2 \cdot \Big(f_2 u_2 + f_1 \overline{u}_1 \Big) \quad \text{in } \Omega_T, \\
z_1 = z_2 &= 0 \quad \text{on } \partial_p \Omega_T.
\end{aligned} \tag{4.17}$$

Equations (4.16), (4.17) are no longer coupled and the terms $f_1 u_1 + f_2 \overline{u}_2$, $f_2 u_2 + f_1 \overline{u}_1$ are known and bounded. Hence we can apply an extension of the maximum principle to parabolic equations whose coefficients are in L^∞ (see, for example, [10, Chapter 7]), obtaining

$$z_1(t,x) \geq 0, \quad z_2(t,x) \geq 0 \quad \text{in } \Omega_T. \tag{4.18}$$

Taking into account (4.4) and (4.10), from (4.18) we obtain (4.3), i.e., the result. □

REMARK 4.1. *The results proved in section* 3 *and* 4 *hold true even if we take* $M > 2$ *players and if we take the functions* f *and* l *linear and dependent explicitly on* t*, i.e.,*

$$\begin{aligned}
f(t,x,u_1,u_2) &= f_1(t,x)u_1 + f_2(t,x)u_2 + f_3(t,x), \\
l_1(t,x,u_1,u_2) &= l_1(t,x)u_1 + h_1(t,x), \\
l_2(t,x,u_1,u_2) &= l_2(t,x)u_2 + h_2(t,x).
\end{aligned}$$

The only difference is the appearance in (2.27)–(2.28), *as source terms, of the functions* $f_3(t,x) + h_1(t,x)$ *and* $f_3(t,x) + h_2(t,x)$*, respectively.*

REMARK 4.2. *In* [12], *Olsder studied the Nash equilibria of the following nonzero-sum deterministic differential game with open-loop bang-bang controls:*

$$\dot{x} = (1-x)u_1 - xu_2,$$

with payoff

$$J_1 = \int_t^T (c_1 x - u_1)\, ds, \quad J_2 = \int_t^T (c_2(1-x) - u_2)\, ds,$$

and controls subject to

$$0 \le u_i(t) \le 1.$$

Because of the hard constraints, also in this case, the optimal controls contain Heaviside functions.

REMARK 4.3. *If we change the control sets taking* $U_1 = U_2 = [-1, 1]$, *as done in* [5] *in the deterministic case, the optimal feedback equilibria are*

$$\overline{u}_1 \in \operatorname{sign}\big(p \cdot f_1 + l_1\big),$$
$$\overline{u}_2 \in \operatorname{sign}\big(p \cdot f_2 + l_2\big),$$

where $\operatorname{sign}(\eta) = 1$ *if* $\eta > 0$, $\operatorname{sign}(\eta) = -1$ *if* $\eta < 0$, *and* $\operatorname{sign}(0) = [-1, 1]$.

The Hamiltonian functions take the form

$$H_1(x, p, \overline{u}_1, \overline{u}_2) = \big|p \cdot f_1(x) + l_1(x)\big| + p \cdot f_2(x)\overline{u}_2,$$
$$H_2(x, p, \overline{u}_1, \overline{u}_2) = \big|p \cdot f_2(x) + l_2(x)\big| + p \cdot f_1(x)\overline{u}_1,$$

and also in this case our method can be applied.

Acknowledgments. The author would like to thank Martino Bardi, who stated the problem, for his helpful comments and suggestions. The author thanks the referees for their useful remarks.

REFERENCES

[1] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, New York, 1982.

[2] A. BENSOUSSAN AND J. FREHSE, *Regularity Results for Nonlinear Elliptic Systems and Applications*, Appl. Math. Sci. 151, Springer-Verlag, Berlin, 2002.

[3] A. BENSOUSSAN AND J. FREHSE, *Stochastic games for N players*, J. Optim. Theory Appl., 105 (2000), pp. 543–565.

[4] A. BENSOUSSAN AND J. FREHSE, *Nonlinear elliptic systems in stochastic game theory*, J. Reine Angew. Math., 350 (1984), pp. 23–67.

[5] P. CARDALIAGUET AND S. PLASKACZ, *Existence and uniqueness of a Nash equilibrium feedback for a simple nonzero-sum differential game*, Internat. J. Game Theory, 32 (2003), pp. 33–71.

[6] W. H. FLEMING AND H. METE SONER, *Controlled Markov Processes and Viscosity Solutions* Springer-Verlag, New York, 1993.

[7] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, Academic Press, New York, 1976.

[8] S. HAMEDÈNE AND J. P. LEPELTIER, *Points d'equilibre dans le jeux stochastiques de somme non nulle*, C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 251–256.

[9] S. HAMEDÈNE, J. P. LEPELTIER, AND S. PENG, *BSDEs with continuous coefficients and stochastic differential games*, in Backward Stochastic Differential Equations (Paris, 1995–1996), El Karoui et al., ed., Pitman Res. Notes Math. Ser. 364, Longman, Harlow, 1997, pp. 115–128.

[10] G. M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, River Edge, NJ, 1996.
[11] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.
[12] G. J. Olsder, *On open- and closed-loop bang-bang control in nonzero-sum differential games*, SIAM J. Control Optim., 40 (2001), pp. 1087–1106.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1234–1252

A SUFFICIENT CONDITION ON RIESZ BASIS WITH PARENTHESES OF NON–SELF-ADJOINT OPERATOR AND APPLICATION TO A SERIALLY CONNECTED STRING SYSTEM UNDER JOINT FEEDBACKS*

BAO-ZHU GUO† AND YU XIE‡

Abstract. This paper gives an abstract sufficient condition on Riesz basis with parentheses property for the generators of C_0-groups in Hilbert spaces whose eigenvalues are comprised of some finite unification of separable sets after taking the algebraic multiplicities into account. The condition is then applied to the closed-loop system of a serially connected string system under joint damping feedbacks to show that there is a family of generalized eigenfunctions that form a Riesz basis with parentheses in the state space. The spectrum-determined growth condition is concluded as a consequence.

Key words. C_0-group, string equation, completeness, Riesz basis, function of exponentials

AMS subject classifications. 93C20, 93D15, 35B35, 35P10

DOI. 10.1137/S0363012902420352

1. Introduction.

The Riesz basis property of linear infinite-dimensional systems $\dot{x}(t) = Ax(t)$ in Hilbert spaces is usually studied in the context of stability. The property is also closely related to the moment method, a powerful method in the study of controllability of hyperbolic systems [1], [19], [25]. Being the basis of frequency analysis, Riesz basis generation is one of the fundamental properties for vibrating systems. The verification of Riesz basis generation usually solves another difficult problem, the so-called spectrum-determined growth condition. The Riesz basis property also offers insight into the solution under the expansion of nonharmonic Fourier series. A recent interesting result [10] says that when A generates a C_0-group and the one-dimensional control operator b is unbounded but admissible [23], then under conditions that the eigenvalues are simple and separable, system $\dot{x}(t) = Ax(t) + bu(t)$ is exactly controllable if and only if $\dot{x}(t) = Ax(t)$ is a Riesz spectral system [4].

Let us recall that a sequence $\{Y_i\}_{i=1}^{\infty}$ in H is called a basis if any element $Y \in H$ has a unique representation

$$Y = \sum_{i=1}^{\infty} \alpha_i Y_i \tag{1}$$

with respect to the norm of H. It is called a Riesz basis for H if

(a) $\overline{\text{span}}\{Y_i\} = H$ and

*Received by the editors December 22, 2002; accepted for publication (in revised form) March 27, 2004; published electronically December 1, 2004. This research was supported by the National Natural Science Foundation of China.

http://www.siam.org/journals/sicon/43-4/42035.html

†Corresponding author. Institute of Systems Science, Academy of Mathematics and System Sciences, Academia Sinica, Beijing 100080, China (bzguo@iss03.iss.ac.cn), and Department of Computational and Applied Mathematics, University of the Witwatersrand, Private Bag-3, Wits-2050, Johannesburg, South Africa.

‡Institute of Systems Science, Academy of Mathematics and System Sciences, Academia Sinica, Beijing 100080, China (xieyu_cn@hotmail.com).

(b) there exist positive constants m and M such that for an arbitrary positive integer n and arbitrary scalars $\alpha_i, i = 1, 2, \ldots, n$, one has

$$m \sum_{i=1}^{n} \mid \alpha_i \mid^2 \leq \left\| \sum_{i=1}^{n} \alpha_i Y_i \right\|^2 \leq M \sum_{i=1}^{n} \mid \alpha_i \mid^2 .$$

Furthermore, a basis $\{Y_i\}_{i=1}^{\infty}$ up to normalization is called a *Riesz basis with parentheses* if (1) converges unconditionally in H only after putting some of its items in parentheses in a way that is independent of Y (see [20]). We refer the reader to [25] for more details on Riesz bases. The Riesz basis property was confirmed for Euler–Bernoulli equations in [8], [5] for one beam and in [6] for two connected beams by the methods of perturbation. General results by perturbation methods to general collocated systems can be found in [7]. The perturbation methods, however, are not applicable to string equations [18]. For the Riesz basis property of one string equations under boundary feedbacks, we refer to [21], [22] and the references therein. Two connected strings under joint feedbacks were discussed recently in [24], where a sufficient condition was developed with the help of the basis property of functions of exponentials in L^2 space under the basic condition that the eigenvalues are algebraically simple and separable.

In this article, we develop an abstract sufficient condition on the Riesz basis with parentheses property for generators of C_0-groups in Hilbert spaces. The distinct contributions of this paper are that (a) the functions of exponentials formulated by eigenvalues are not necessarily Riesz basis in $L^2(0, T)$ but are Riesz basis on their closed span in $L^2(0, T)$ for some $T > 0$; (b) the eigenvalues are allowed to be algebraically multiple; (c) the eigenvalues are comprised of some finite unification of separable sets after taking the algebraic multiplicities into account; (d) the result can be used to deal with the Riesz basis (with parentheses) property of N-connected string equations under joint linear feedback controls. The main tool used in this paper, which is different from Levin and Golovin's theorem [1] used in [24], is the notion of generalized divided difference (GDD) introduced very recently in [2] and [3].

The paper is organized as follows. In section 2, we give an abstract sufficient condition on the Riesz basis with parentheses property for generators of C_0-groups in Hilbert spaces. Section 3 is devoted to the application of a serially connected string equation under joint damping feedbacks. It is shown that the associated system operator generates a C_0-group, and its root subspace is complete in the state space. Moreover, there is a family of generalized eigenfunctions that forms a Riesz basis with parentheses in the state space. The spectrum-determined growth condition is concluded as a consequence.

2. A sufficient condition on the Riesz basis with parentheses. Assume that B is a discrete operator in a separable Hilbert space $\mathbf{H}$ (that is, there is a $\mu \in \sigma(B)$ such that $(\mu - B)^{-1}$ is compact in $\mathbf{H}$). Let us recall that a nonzero $Y \in \mathbf{H}$ is called a generalized eigenvector of B, corresponding to an eigenvalue μ, if there is a positive integer n such that $(\mu - B)^n Y = 0$. The number (denoted by $n_{a\mu}$) of all linearly independent generalized eigenvectors is called the algebraic multiplicity of μ. μ is said to be algebraically simple if $n_{a\mu} = 1$. It is well known in functional analysis that each eigenvalue of a discrete operator must have finite algebraic multiplicity. A nonzero $Y \in \mathbf{H}$ is called an eigenvector of B if $(\mu - B)Y = 0$. The number (denoted by $n_{g\mu}$) of all linearly independent eigenvectors is called the geometric multiplicity of μ. μ is said to be geometrically simple if $n_{g\mu} = 1$. Suppose that $\sigma(B) = \{\mu_n\}_{n\in\mathbb{Z}}$,

where (and henceforth) $\mathbb{Z}$ stands for some set of integers and each μ_n is of algebraic multiplicity m_n. Thus we have a set of complex exponentials associated with μ_n:

$$E_n(t) = \{e^{\mu_n t}, te^{\mu_n t}, \dots, t^{m_n-1}e^{\mu_n t}\}.$$

The Riesz basis property of $\{E_n(t)\}$ in $L^2(0,T)$ has been studied extensively by former Soviet mathematicians (Levin, Pavlov, Nikolskiĭ, and many others), and necessary and sufficient conditions are already available in the literature [1], [9], [25] for the case that the $\{\mu_n\}$ are separable (a set Ω of complex plane is called separable if $\inf_{a,b\in\Omega,a\neq b}|a-b|>0$). It was shown recently in [24] that the Riesz basis property of $\{E_n(t)\}$ in $L^2(0,T)$ is closely related to the Riesz basis generation of the root subspace of A, provided that each eigenvalue is algebraically simple ($m_n = 1$). Since it is difficult to check the algebraic multiplicity and the separability of the eigenvalues in applications, we have to generalize the results of [24] in order to deal with the cases where the eigenvalues are not simple and separable. For this purpose, we need the help of the concept of generalized divided difference (GDD) introduced in [2] and [3].

DEFINITION 2.1. *Let $\mu_k, k = 1,2,\dots,m$, be arbitrary complex numbers (not necessarily different). The GDD of order zero of the function $e^{\mu t}$ corresponding to the point μ_1 is defined as $[\mu_1](t) = e^{\mu_1 t}$. GDD of the order $n-1, n \le m$ of $e^{\mu t}$ corresponding to $\{\mu_k, k = 1,2,\dots,n\}$ is defined by*

$$[\mu_1,\mu_2,\dots,\mu_n](t) = \begin{cases} \dfrac{[\mu_1,\mu_2,\dots,\mu_{n-1}](t) - [\mu_2,\mu_3,\dots,\mu_n](t)}{\mu_1-\mu_n}, & \mu_1 \neq \mu_n, \\ \dfrac{\partial}{\partial\mu}[\mu,\mu_2,\dots,\mu_{n-1}](t)|_{\mu=\mu_1}, & \mu_1 = \mu_n. \end{cases}$$

The following formula is valid for any $\{\mu_k\}_{k=1}^n$:

$$[\mu_1,\mu_2,\dots,\mu_n](t) = \int_0^1 d\tau_1 \int_0^{\tau_1} d\tau_2 \dots \int_0^{\tau_{n-2}} d\tau_{n-1} t^{n-1} e^{t[\mu_1+\tau_1(\mu_2-\mu_1)+\dots+\tau_{n-1}(\mu_n-\mu_{n-1})]}. \tag{2}$$

And hence if $\mathrm{Re}\mu_n \le \mathrm{Re}\mu_{n-1} \le \cdots \le \mathrm{Re}\mu_1$, then

$$|[\mu_1,\mu_2,\dots,\mu_n](t)| \le t^{n-1}e^{\mathrm{Re}\mu_1 t} \qquad \forall\, t \ge 0. \tag{3}$$

Note that if $\mu_i = \mu, i = 1,2,\dots,n$, then

$$[\mu_1,\mu_2,\dots,\mu_i](t) = t^{i-1}e^{\mu t}, \qquad 1 \le i \le n.$$

Generally, if there are m number of different elements in $\{\mu_1,\mu_2,\dots,\mu_n\}$, $\{\mu_1,\mu_2,\dots,\mu_n\} = \{\nu_1,\nu_2,\dots,\nu_m\}$, each ν_k repeats n_k times, $\sum_{k=1}^m n_k = n$. Then by Lemma 3.1 of [3], the GDD $[\mu_1,\mu_2,\dots,\mu_n](t)$ is the linear combination of functions $t^{j-1}e^{\nu_k t}, 1 \le j \le n_k, 1 \le k \le m$, and the coefficients of the leading terms $t^{n_k-1}e^{\nu_k t}$ are not equal to zero. The latter implies conversely that for any $1 \le k \le n_m$, $t^{k-1}e^{\nu_m t}$ is also the linear combination of $[\mu_1](t), [\mu_1,\mu_2](t), \dots, [\mu_1,\mu_2,\dots,\mu_n](t)$. In particular, we have the following proposition.

PROPOSITION 2.2. *Let $\{\mu_1,\mu_2,\dots,\mu_n\} = \{\nu_1,\nu_2,\dots,\nu_m\}$, $\nu_i \neq \nu_j$, $i \neq j$, $1 \le i, j \le m$, and each ν_j repeat n_j times, $\sum_{j=1}^m n_j = n$. Then any $\phi(t) = \sum_{j=1}^m e^{\nu_j t}\sum_{i=1}^{n_j} a_{ij}t^{i-1}$ can be represented as*

$$\phi(t) = \sum_{i=1}^n G_i[\mu_1,\mu_2,\dots,\mu_i](t),$$

where $G_1 = \sum_{j=1}^m a_{1j}$.

Let $\Omega = \{\mu_k\}_{k\in\mathbb{Z}}$ be a sequence in $\mathbb{C}$. Suppose Ω is a union of N separable sets Ω_j (each μ_k is assigned according to its multiplicity): $\Omega = \bigcup_{j=1}^{N} \Omega_j$ and $\sup_k |\mathrm{Re}\mu_k| < \infty$. Suppose that $\{\mu_k\}$ have been ordered in such a way that $\{\mathrm{Im}\mu_k\}$ forms a nondecreasing sequence [2]. Define

$$D^+(\Omega) = \lim_{r\to\infty} \frac{n^+(r)}{r},$$

where

$$n^+(r) = \sup_{x\in\mathbb{R}} \#\{\mathrm{Im}(\Omega) \cap [x, x+r)\}.$$

Let $\delta = \min_j[\inf_{a,b\in\Omega_j, a\neq b} |a-b|]$. For any $x \in \mathbb{R}$, suppose that there are M balls with radius $\delta/3$, which covers the compact region $\Omega(x) = \{\mu || \mathrm{Re}\mu| \le \sup_k |\mathrm{Re}\mu_k|,\ \mathrm{Im}\mu \in [x, x+1]\}$ of $\mathbb{C}$. Note that M is independent of x by unit shift. Then there are at most NM number of μ_k inside $\Omega(x)$. Hence, for any $r > 0$, we have

$$n^+(r) = \sup_{x\in\mathbb{R}} \#\{\mathrm{Im}(\Omega)\cap[x,x+r)\} \le \sup_{x\in\mathbb{R}} \#\{\mathrm{Im}(\Omega)\cap[x,x+([r]+1))\} \le ([r]+1)NM,$$

where $[r]$ denotes the maximal integer not exceeding r. Therefore, $D^+(\Omega) \le NM$. In particular,

$$D^+(\Omega) < \infty. \tag{4}$$

For any $\mu \in \mathbb{C}$, denote by $D_\mu(r)$ a disk with center μ and radius r. Let $G^p(r)$, $p = 1, 2, \ldots$, be the connected components of the union $\cup_{\mu\in\Omega} D_\mu(r)$ and write $\Omega^p(r) = \{\mu_{j,p}\}$ for the subsequence of Ω in $G^p(r)$, $\Omega^p(r) = \Omega \cap G^p(r)$. Then Lemma 1 of [2] says that, for any $r < r_0 = \delta/(2N)$, the number of $\Omega^p(r)$ is less than or equal to N. Set

$$\Omega^p(r) = \{\mu_{j,p}\}, \qquad j = 1, 2, \ldots, M^p(r) \le N. \tag{5}$$

Denote by

$$\varepsilon^p(\Omega, r) = \{[\mu_{1,p}], [\mu_{1,p}, \mu_{2,p}], \ldots, [\mu_{1,p}, \mu_{2,p}, \ldots, \mu_{M^p,p}]\}$$

the family of GDD corresponding to $\Omega^p(r)$. Then the following result, which is Theorem 3 of [2], characterizes the basis property of $\varepsilon^p(\Omega, r)$ in L^2 space.

PROPOSITION 2.3. *Assume that $\Omega = \{\mu_k\}_{k\in\mathbb{Z}}$ is defined as above. Then for any $2\pi D^+(\Omega) < T < \infty$ the family $\varepsilon^p(\Omega, r)$ forms a Riesz basis in the closed subspace of $L^2(0, T)$ spanned by itself.*

With these preliminary results, we come to the proof of the abstract result which links the Riesz basis property of the corresponding family of GDD in L^2 space and that of the root subspace of associated operators in cases that the eigenvalues are comprised of some finite unification of separable sets after taking the algebraic multiplicities into account.

LEMMA 2.4. *Let $\mathbf{H}$ be a separable Hilbert space. Suppose $\{e_n(t)\}_{n\in\mathbb{Z}}$ forms a Riesz basis in the closed subspace spanned by itself in $L^2(0, T)$, $T > 0$. Then for any $\phi \in L^2(0, T; \mathbf{H})$, $\phi(t) = \sum_{n\in\mathbb{Z}} e_n(t)\phi_n$, there exist constants $C_1, C_2 > 0$ such that*

$$C_1 \sum_{n\in\mathbb{Z}} \|\phi_n\|^2_{\mathbf{H}} \le \|\phi\|^2_{L^2(0,T;\mathbf{H})} \le C_2 \sum_{n\in\mathbb{Z}} \|\phi_n\|^2_{\mathbf{H}}. \tag{6}$$

Proof. Take some orthonormal basis $\{\psi_n\}$ of $\mathbf{H}$, and for almost all $t \in [0,T]$ expand ϕ as

$$\phi(t) = \sum_{n\in\mathbb{Z}} \langle \phi(t), \psi_n\rangle_{\mathbf{H}} \psi_n, \qquad t \in [0,T] \text{ a.e.}$$

Then

$$\|\phi(t)\|^2_{\mathbf{H}} = \sum_{n\in\mathbb{Z}} |\ \langle \phi(t), \psi_n\rangle_{\mathbf{H}}\ |^2 \qquad \forall t \in [0,T] \text{ a.e.} \tag{7}$$

Since for any $m \in \mathbb{Z}$, $\langle \phi(t), \psi_m\rangle_{\mathbf{H}} \in \text{span}\{e_n(t)\}_{n\in\mathbb{Z}}$ and

$$\langle \phi(t), \psi_m\rangle_{\mathbf{H}} = \sum_{n\in\mathbb{Z}} \langle \phi_n, \psi_m\rangle_H e_n(t) \qquad \forall\ m \in \mathbb{Z} \text{ in } L^2(0,T). \tag{8}$$

By assumption,

$$C_1 \sum_{n\in\mathbb{Z}} |\ \langle \phi_n, \psi_m\rangle_{\mathbf{H}}\ |^2 \le \int_0^T |\ \langle \phi(t), \psi_m\rangle_{\mathbf{H}}\ |^2\, dt \le C_2 \sum_{n\in\mathbb{Z}} |\ \langle \phi_n, \psi_m\rangle_{\mathbf{H}}\ |^2 \tag{9}$$

for some constants $C_1, C_2 > 0$ that depend on $\{e_n(t)\}$. Hence it follows from (7) that

$$\begin{aligned} C_1 \sum_{m\in\mathbb{Z}} \sum_{n\in\mathbb{Z}} |\ \langle \phi_n, \psi_m\rangle_{\mathbf{H}}\ |^2 &\le \sum_{m\in\mathbb{Z}} \int_0^T |\ \langle \phi(t), \psi_m\rangle_{\mathbf{H}}\ |^2\, dt \\ &= \int_0^T \|\phi(t)\|^2_{\mathbf{H}} dt \le C_2 \sum_{m\in\mathbb{Z}} \sum_{n\in\mathbb{Z}} |\ \langle \phi_n, \psi_m\rangle_{\mathbf{H}}\ |^2 . \end{aligned} \tag{10}$$

Note that

$$\phi_n = \sum_{m\in\mathbb{Z}} \langle \phi_n, \psi_m\rangle_{\mathbf{H}} \psi_m, \qquad \|\phi_n\|^2 = \sum_{m\in\mathbb{Z}} |\langle \phi_n, \psi_m\rangle_{\mathbf{H}}|^2. \tag{11}$$

Then (6) follows from (10). □

Assume further that B generates a C_0-group on $\mathbf{H}$ and that its root subspace (the closed subspace spanned by all generalized eigenvectors of B) is complete in $\mathbf{H}$: $Sp(B) = \mathbf{H}$. Suppose

$$\sigma(B) = \bigcup_{p\in\mathbb{Z}} \Omega(p), \quad \Omega(p) = \{\nu_j^p\}_{j=1}^{N^p}, \quad \text{Re}\nu_1^p \ge \text{Re}\nu_2^p \ge \cdots \ge \text{Re}\nu_{N^p}^p, \quad \nu_i^n \ne \nu_j^m,$$

unless $m = n, i = j$. Assume that each ν_j^p has algebraic multiplicity m_j^p and that both the number N^p of $\Omega(p)$ and m_j^p have uniform upper bound; i.e., there exists an $N > 0$ such that $\sup_p[N^p \max_{1\le j\le N^p} m_j^p] \le N$. Set $\tilde{m}_0^p = 0$, $\tilde{m}_l^p = \sum_{q=1}^l m_q^p, l = 1, \ldots, N^p$. Arranging $\Omega(p)$ again according to multiplicity, we obtain a new set $\Lambda^p = \{\{\mu^p_{i+\tilde{m}^p_{j-1}}\}_{i=1}^{m_j^p}\}_{j=1}^{N^p}$,

$$\mu^p_{i+\tilde{m}^p_{j-1}} = \nu_j^p, \quad 1 \le i \le m_j^p, \quad 1 \le j \le N^p. \tag{12}$$

Hence we write (accounted by multiplicity) $\sigma(B) = \bigcup_{p\in\mathbb{Z}} \Lambda^p$. Make the family of GDD be as follows:

$$E_p(t) = \{[\mu_1^p](t), [\mu_1^p, \mu_2^p](t), \ldots, [\mu_1^p, \mu_2^p, \ldots, \mu^p_{\tilde{m}^p_{N^p}}](t)\}. \tag{13}$$

Now we are in a position to prove the main result of this section.

THEOREM 2.5. *If there exists a $T>0$ such that the family of GDD $\{E_p(t)\}_1^\infty$ defined by* (13) *forms a Riesz basis in the closed subspace spanned by itself in $L^2(0,T)$, then*

(a)

$$S_\infty(B) = \left\{ x \in \mathbf{H} | x = \sum_{p\in\mathbb{Z}} \sum_{j=1}^{N^p} \mathbb{P}_{\nu_j^p} x \right\} = \mathbf{H}, \tag{14}$$

where $\mathbb{P}_{\nu_j^p}$ denotes the eigen-projection of B corresponding to eigenvalue ν_j^p;

(b) *there are constants $M_1, M_2 > 0$ such that*

$$M_1 \sum_{p\in\mathbb{Z}} \left\| \sum_{j=1}^{N^p} \mathbb{P}_{\nu_j^p} x \right\|^2 \le \|x\|^2 \le M_2 \sum_{p\in\mathbb{Z}} \left\| \sum_{j=1}^{N^p} \mathbb{P}_{\nu_j^p} x \right\|^2 \qquad \forall\, x \in \mathbf{H}; \tag{15}$$

(c) *the spectrum-determined growth condition holds:*

$$\omega(B) = \inf\{\omega | \textit{there exists } M>1 \textit{ such that } \|e^{Bt}\| \le Me^{\omega t}\} = S(B) = \sup_{\nu\in\sigma_p(B)} \mathrm{Re}\nu.$$

Proof. We first prove (15). Since B generates a C_0-group, there are $M, \omega > 0$ such that

$$\|e^{Bt}\| \le Me^{\omega|t|} \qquad \forall\, t \in \mathbb{R}.$$

Take $x_0 \in S_\infty(B),\ x_0 = \sum_{p\in\mathbb{Z}} \sum_{j=1}^{N^p} \mathbb{P}_{\nu_j^p} x_0$. Then

$$e^{Bt}x_0 = \sum_{p\in\mathbb{Z}} \sum_{j=1}^{N^p} e^{\nu_j^p t} \sum_{i=1}^{m_j^p} \frac{(B-\nu_j^p)^{i-1}}{(i-1)!} t^{i-1} \mathbb{P}_{\nu_j^p} x_0 = \sum_{p\in\mathbb{Z}} \sum_{j=1}^{N^p} e^{\nu_j^p t} \sum_{i=1}^{m_j^p} a_{ij}^p t^{i-1}, \tag{16}$$

where we set $a_{ij}^p = \frac{(B-\nu_j^p)^{i-1}}{(i-1)!} \mathbb{P}_{\nu_j^p} x_0$. By Proposition 2.2, we can write

$$e^{Bt}x_0 = \sum_{p\in\mathbb{Z}} \sum_{j=1}^{N^p} \sum_{i=1}^{m_j^p} G^p_{i+\tilde{m}^p_{j-1}}(x_0)[\mu_1^p, \mu_2^p, \ldots, \mu^p_{i+\tilde{m}^p_{j-1}}](t). \tag{17}$$

Further, by assumption and Lemma 2.4, there are constants $C_1, C_2 > 0$ such that

$$C_1 \sum_{p\in\mathbb{Z}} \sum_{j=1}^{N^p} \sum_{i=1}^{m_j^p} \|G^p_{i+\tilde{m}^p_{j-1}}(x_0)\|^2 \le \int_0^T \|e^{Bt}x_0\|^2 dt \le C_2 \sum_{p\in\mathbb{Z}} \sum_{j=1}^{N^p} \sum_{i=1}^{m_j^p} \|G^p_{i+\tilde{m}^p_{j-1}}(x_0)\|^2. \tag{18}$$

In particular,

$$C_1\sum_{p\in\mathbb{Z}}\|G_1^p(x_0)\|^2=C_1\sum_{p\in\mathbb{Z}}\left\|\sum_{j=1}^{N^p}\mathbb{P}_{\nu_j^p}x_0\right\|^2\le\int_0^T\|e^{Bt}x_0\|^2dt\le\frac{M^2}{2\omega}(e^{2\omega T}-1)\|x_0\|^2. \tag{19}$$

Next, since $Sp(B)=\mathbf{H}, Sp(B)\subset\overline{S_\infty(B)}=\mathbf{H}$, we see that (19) holds for all $x_0\in\mathbf{H}$. Setting $M_1=C_1\frac{2\omega}{M^2}(e^{2\omega T}-1)^{-1}$, we obtain the first inequality of (15). Finally, from (18),

$$C_1\sum_{j=1}^{N^p}\sum_{i=1}^{m_j^p}\|G^p_{i+\tilde m^p_{j-1}}(x_0)\|^2\le\int_0^T\|e^{Bt}x_0\|^2dt\le\frac{M^2}{2\omega}(e^{2\omega T}-1)\|x_0\|^2\text{ as }x_0\in S_\infty(B).$$

In particular, by letting $x_0=\sum_{j=1}^{N^p}\mathbb{P}_{\nu_j^p}x\in S_\infty(B), p\ge1, x\in H$, we obtain

$$\begin{aligned}C_1\sum_{j=1}^{N^p}\sum_{i=1}^{m_j^p}\|G^p_{i+\tilde m^p_{j-1}}(x)\|^2&=C_1\sum_{j=1}^{N^p}\sum_{i=1}^{m_j^p}\|G^p_{i+\tilde m^p_{j-1}}(x_0)\|^2\\&\le\int_0^T\|e^{Bt}x_0\|^2dt\le\frac{M^2}{2\omega}(e^{2\omega T}-1)\left\|\sum_{j=1}^{N^p}\mathbb{P}_{\nu_j^p}x\right\|^2\qquad\forall\,x\in\mathbf{H}.\end{aligned} \tag{20}$$

Hence, from (18) and (20), we see that for any $x\in\mathbf{H}$

$$\begin{aligned}\|x\|^2&=\frac1T\int_0^T\|e^{-Bt}e^{Bt}x\|^2dt\le\frac{M^2e^{2\omega T}}{T}\int_0^T\|e^{Bt}x\|^2dt\\&\le\frac{M^2e^{2\omega T}}{T}\frac{M^2}{2\omega}(e^{2\omega T}-1)C_1^{-1}C_2\sum_{p\in\mathbb{Z}}\left\|\sum_{j=1}^{N^p}\mathbb{P}_{\nu_j^p}x\right\|^2.\end{aligned} \tag{21}$$

Setting $M_2=C_2M^4e^{2\omega T}(2\omega TC_1)^{-1}(e^{2\omega T}-1)$ yields the second inequality of (15).

To show (14), we need to show that $S_\infty(B)$ is a closed subspace of $\mathbf{H}$. By virtue of Theorem 3.5 of [13], it suffices to show that there exists $M_0>0$ such that

$$\left\|\sum_{p=\tilde M}^{\tilde N}\sum_{j=1}^{N^p}\mathbb{P}_{\nu_j^p}x\right\|^2\le M_0\|x\|^2\qquad\forall x\in\mathbf{H}\ \text{ and integers }\tilde M,\tilde N.$$

Let $x=\sum_{p=\tilde M}^{\tilde N}\sum_{j=1}^{N^p}P_{\nu_j^p}z, z\in\mathbf{H}$. Then $\sum_{j=1}^{N^p}\mathbb{P}_{\nu_j^p}x=\sum_{j=1}^{N^p}\mathbb{P}_{\nu_j^p}z$. From (15)

$$\frac{M_1}{M_2}\|x\|^2\le M_1\sum_{p=\tilde M}^{\tilde N}\left\|\sum_{j=1}^{N^p}P_{\nu_j^p}x\right\|^2=M_1\sum_{p=\tilde M}^{\tilde N}\left\|\sum_{j=1}^{N^p}P_{\nu_j^p}z\right\|^2\le\|z\|^2\qquad\forall\,z\in\mathbf{H},$$

and hence

$$\left\| \sum_{p=\tilde{M}}^{\tilde{N}} \sum_{j=1}^{N^p} P_{\nu_j^p} z \right\|^2 \le \frac{M_2}{M_1} \|z\|^2 \qquad \forall z \in \mathbf{H},\ \tilde{N} > 1.$$

Equation (14) is thus proved. Finally, since $\mathrm{Re}\mu_1^p \ge \mathrm{Re}\mu_2^p \ge \cdots \ge \mathrm{Re}\mu_{i+\tilde{m}_{j-1}^p}^p$, it follows from (3) that

$$|[\mu_1^p, \mu_2^p, \ldots, \mu_{i+\tilde{m}_{j-1}^p}^p](t)| \le t^N e^{S(B)t} \qquad \forall\, t \ge 1. \tag{22}$$

From (15), (16), (17), and (22), we get

$$\begin{aligned}
\|e^{Bt}x_0\|^2 &\le M_2 \sum_{p\in\mathbb{Z}} \left\| \sum_{j=1}^{N^p} \mathbb{P}_{\nu_j^p} e^{Bt} x_0 \right\|^2 = M_2 \sum_{p\in\mathbb{Z}} \left\| \sum_{j=1}^{N^p} e^{\nu_j^p t} \sum_{i=1}^{m_j^p} \frac{(B-\nu_j^p)^{i-1}}{(i-1)!} t^{i-1} \mathbb{P}_{\nu_j^p} x_0 \right\|^2 \\
&= M_2 \sum_{p\in\mathbb{Z}} \left\| \sum_{j=1}^{N^p} \sum_{i=1}^{m_j^p} G_{i+\tilde{m}_{j-1}^p}^p(x_0) [\mu_1^p, \mu_2^p, \ldots, \mu_{i+\tilde{m}_{j-1}^p}^p](t) \right\|^2 \\
&\le M_2 \sum_{p\in\mathbb{Z}} \sum_{j=1}^{N^p} \sum_{i=1}^{m_j^p} \|G_{i+\tilde{m}_{j-1}^p}^p(x_0)\|^2 N t^{2N} e^{2S(B)t} \qquad \forall\, x_0 \in \mathbf{H}.
\end{aligned}$$

This, together with (20) and (15), gives

$$\|e^{Bt}x_0\|^2 \le C t^{2N} e^{2S(B)t} \sum_{p\in\mathbb{Z}} \left\| \sum_{j=1}^{N^p} \mathbb{P}_{\nu_j^p} x_0 \right\|^2 \le C M_1^{-1} t^{2N} e^{2S(B)t} \|x_0\|^2,$$

where C is a constant. The proof is complete. □

3. Application to an N-connected string equation. In this section, we are concerned with the following system of $N+1$ serially connected strings under joint feedback controls:

$$\begin{cases}
y_{tt}(x,t) - c_i^2 y_{xx}(x,t) = 0, \quad i-1 < x < i, i = 1, 2, \ldots, N+1, \\
y(0,t) = y(N+1,t) = 0, \\
y(i^-,t) = y(i^+,t), \\
c_i^2 y_x(i^-,t) - c_{i+1}^2 y_x(i^+,t) = k_i y_t(i,t), \quad i = 1, \ldots, N, \\
y(x,0) = y_0(x), \quad y_t(x,0) = y_1(x),
\end{cases} \tag{23}$$

where $t > 0, k_i \in \mathbb{R}, c_i > 0, i = 1, 2, \ldots, N$.

Let the underlying state Hilbert space $\mathcal{H} = H_0^1(0, N+1) \times L^2(0, N+1)$. Define an inner product in $\mathcal{H}$ as

$$\langle (f_1,g_1),(f_2,g_2)\rangle_H = \sum_{i=1}^{N+1}\int_{i-1}^{i}[c_i^2 f_1'(x)\overline{f_2'(x)}+g_1(x)\overline{g_2(x)}]dx \qquad \forall\,(f_1,g_1),(f_2,g_2)\in\mathcal{H}. \tag{24}$$

Introduce operator $\mathcal{A}$ in $\mathcal{H}$:

$$\mathcal{A}(f,g)=(g(x),c_i^2 f''(x)),\quad i-1<x<i,\ i=1,2,\dots,N+1,\quad \forall\,(f,g)\in D(\mathcal{A}), \tag{25}$$

where

$$\begin{aligned} D(\mathcal{A})=\{(f,g)\in H\mid f,g\in H_0^1(0,N+1), f|_{[j-1,j]}\in H^2(j-1,j), 1\le j\le N+1,\\ c_i^2 f'(i^-)-c_{i+1}^2 f'(i^+)=k_i g(i), 1\le i\le N\} \end{aligned} \tag{26}$$

with $f|_{[a,b]}$ denoting the restriction of f on $[a,b]$. Set $Y(t)=(y(\cdot,t),y_t(\cdot,t)), Y_0=(y_0(\cdot),y_1(\cdot))$. Then system (23) can be written as an evolutionary equation in $\mathcal{H}$:

$$\begin{cases} \dfrac{dY(t)}{dt}=\mathcal{A}Y(t), \quad t>0,\\ Y(0)=Y_0. \end{cases} \tag{27}$$

Throughout the paper, we always make the following assumption on the string system (23):

$$|k_i|\neq c_i+c_{i+1},\quad i=1,2,\dots,N. \tag{28}$$

Note that this condition is very minor. It holds for almost all constants c_i, k_i.

THEOREM 3.1. *Under condition* (28)*, the operator* $\mathcal{A}$ *defined by* (25) *and* (26) *generates a* C_0*-group on* $\mathcal{H}$.

Proof. For any $(f_1,g_1),(f_2,g_2)\in\mathcal{H}$, define a new inner product of the following in $\mathcal{H}$:

$$\begin{aligned} \langle (f_1,g_1),(f_2,g_2)\rangle_* = \sum_{i=1}^{N+1}\int_{i-1}^{i}\Bigg[A_i(x)\cdot\frac{c_i f_1'(x)+g_1(x)}{2}\cdot\frac{c_i\overline{f_2'(x)}+\overline{g_2(x)}}{2}\\ +B_i(x)\cdot\frac{c_i f_1'(x)-g_1(x)}{2}\cdot\frac{c_i\overline{f_2'(x)}-\overline{g_2(x)}}{2}\Bigg]dx, \end{aligned} \tag{29}$$

where $A_i(x), B_i(x)$, defined on $[i-1,i], i=1,2,\dots,N+1$, are positive differentiable functions to be determined later. It is obvious that the norm induced by the new inner product is equivalent to that induced by (24). We claim that, under this new inner product, $\mathcal{A}$ is a densely defined m-dissipative operator and hence generates a C_0-group in $\mathcal{H}$ (see [17]). Indeed, for any $(f,g)\in D(\mathcal{A}), (f,g)\neq 0$,

$$\begin{aligned}
\operatorname{Re}\langle \mathcal{A}(f,g),(f,g)\rangle_* &= \sum_{i=1}^{N+1} \operatorname{Re}\int_{i-1}^{i} \left[A_i(x)\cdot \frac{c_i g'(x)+c_i^2 f''(x)}{2}\cdot \frac{c_i\overline{f'(x)}+\overline{g(x)}}{2} \right. \\
&\qquad \left. + B_i(x)\cdot \frac{c_i g'(x)-c_i^2 f''(x)}{2}\cdot \frac{c_i\overline{f'(x)}-\overline{g(x)}}{2} \right] dx \\
&= \sum_{i=1}^{N+1} \operatorname{Re}\int_{i-1}^{i} \left\{ c_i A_i(x)\cdot \left[\frac{c_i f'(x)+g(x)}{2}\right]'\cdot \left[\frac{c_i\overline{f'(x)}+\overline{g(x)}}{2}\right] \right. \\
&\qquad \left. - c_i B_i(x)\cdot \left[\frac{c_i f'(x)-g(x)}{2}\right]'\cdot \left[\frac{c_i\overline{f'(x)}-\overline{g(x)}}{2}\right] \right\} dx \\
&= \sum_{i=1}^{N+1} \left[\frac{c_i A_i(x)}{2}\left|\frac{c_i f'(x)+g(x)}{2}\right|^2\Bigg|_{i-1}^{i} - \frac{c_i}{2}\int_{i-1}^{i} A_i'(x)\left|\frac{c_i f'(x)+g(x)}{2}\right|^2 dx \right. \\
&\qquad \left. - \frac{c_i B_i(x)}{2}\left|\frac{c_i f'(x)-g(x)}{2}\right|^2\Bigg|_{i-1}^{i} + \frac{c_i}{2}\int_{i-1}^{i} B_i'(x)\left|\frac{c_i f'(x)-g(x)}{2}\right|^2 dx \right] \\
&= I_1 + I_2,
\end{aligned} \tag{30}$$

where

$$\begin{aligned}
I_2 &= \sum_{i=1}^{N+1} \frac{c_i}{2}\int_{i-1}^{i} \left[B_i'(x)\left|\frac{c_i f'(x)-g(x)}{2}\right|^2 - A_i'(x)\left|\frac{c_i f'(x)+g(x)}{2}\right|^2 \right] dx \\
&\le M\,\|(f,g)\|_*^2
\end{aligned} \tag{31}$$

with M being a positive constant, and

$$\begin{aligned}
I_1 &= -\frac{1}{2}\sum_{i=1}^{N+1} c_i \left[B_i(x)\left|\frac{c_i f'(x)-g(x)}{2}\right|^2\Bigg|_{i-1}^{i} - A_i(x)\left|\frac{c_i f'(x)+g(x)}{2}\right|^2\Bigg|_{i-1}^{i} \right] \\
&= -\frac{1}{2}\left\{ c_1\left[A_1(0)\left|\frac{c_1 f'(0)+g(0)}{2}\right|^2 - B_1(0)\left|\frac{c_1 f'(0)-g(0)}{2}\right|^2 \right] \right. \\
&\qquad + c_{N+1}\left[B_{N+1}(N+1)\left|\frac{c_{N+1} f'(N+1)-g(N+1)}{2}\right|^2 \right. \\
&\qquad\qquad \left. - A_{N+1}(N+1)\left|\frac{c_{N+1} f'(N+1)+g(N+1)}{2}\right|^2 \right] \\
&\qquad + \sum_{i=1}^{N}\left[c_i B_i(i)\left|\frac{c_i f'(i^-)-g(i^-)}{2}\right|^2 + c_{i+1}A_{i+1}(i)\left|\frac{c_{i+1} f'(i^+)+g(i^+)}{2}\right|^2 \right. \\
&\qquad \left. \left. - c_i A_i(i)\left|\frac{c_i f'(i^-)+g(i^-)}{2}\right|^2 - c_{i+1}B_{i+1}(i)\left|\frac{c_{i+1} f'(i^+)-g(i^+)}{2}\right|^2 \right] \right\} \\
&= -\frac{1}{2}[c_1 I_1' + c_{N+1} I_2' + I_3'].
\end{aligned} \tag{32}$$

From the first boundary condition of (23), we obtain

$$\begin{cases} \dfrac{c_1 f'(0) - g(0)}{2} = \dfrac{c_1 f'(0) + g(0)}{2}, \\ \dfrac{c_{N+1} f'(N+1) - g(N+1)}{2} = \dfrac{c_{N+1} f'(N+1) + g(N+1)}{2}. \end{cases} \tag{33}$$

Substituting (33) into (32) yields

$$\begin{aligned} I_1' &= A_1(0) \left| \frac{c_1 f'(0) + g(0)}{2} \right|^2 - B_1(0) \left| \frac{c_1 f'(0) - g(0)}{2} \right|^2 \\ &= [A_1(0) - B_1(0)] \left| \frac{c_1 f'(0) + g(0)}{2} \right|^2, \\ I_2' &= B_{N+1}(N+1) \left| \frac{c_{N+1} f'(N+1) - g(N+1)}{2} \right|^2 \\ &\quad - A_{N+1}(N+1) \left| \frac{c_{N+1} f'(N+1) + g(N+1)}{2} \right|^2 \\ &= [B_{N+1}(N+1) - A_{N+1}(N+1)] \left| \frac{c_{N+1} f'(N+1) - g(N+1)}{2} \right|^2. \end{aligned} \tag{34}$$

Now we choose $A_1(x), B_1(x), A_{N+1}(x), B_{N+1}(x)$ so that $A_1(0) \geq B_1(0), B_{N+1}(N+1) \geq A_{N+1}(N+1)$. This implies $I_1' \geq 0, I_2' \geq 0$. Similarly, from the third boundary condition of (23) we obtain

$$\begin{aligned} &\frac{c_i f'(i^-) - g(i^-)}{2} - \frac{c_i f'(i^-) + g(i^-)}{2} = \frac{c_{i+1} f'(i^+) - g(i^+)}{2} - \frac{c_{i+1} f'(i^+) + g(i^+)}{2}, \\ &c_i \left[\frac{c_i f'(i^-) - g(i^-)}{2} + \frac{c_i f'(i^-) + g(i^-)}{2} \right] - c_{i+1} \left[\frac{c_{i+1} f'(i^+) - g(i^+)}{2} + \frac{c_{i+1} f'(i^+) + g(i^+)}{2} \right] \\ &\quad = -k_i \left[\frac{c_i f'(i^-) - g(i^-)}{2} - \frac{c_i f'(i^-) + g(i^-)}{2} \right]; \end{aligned} \tag{35}$$

that is,

$$\begin{pmatrix} 1 & 1 \\ c_i + k_i & -c_{i+1} \end{pmatrix} \begin{pmatrix} \frac{c_i f'(i^-) - g(i^-)}{2} \\ \frac{c_{i+1} f'(i^+) + g(i^+)}{2} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -c_i + k_i & c_{i+1} \end{pmatrix} \begin{pmatrix} \frac{c_i f'(i^-) + g(i^-)}{2} \\ \frac{c_{i+1} f'(i^+) - g(i^+)}{2} \end{pmatrix}. \tag{36}$$

Under condition (28),

$$\begin{aligned} \begin{pmatrix} \frac{c_i f'(i^-) + g(i^-)}{2} \\ \frac{c_{i+1} f'(i^+) - g(i^+)}{2} \end{pmatrix} &= \begin{pmatrix} 1 & 1 \\ -c_i + k_i & c_{i+1} \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 \\ c_i + k_i & -c_{i+1} \end{pmatrix} \begin{pmatrix} \frac{c_i f'(i^-) - g(i^-)}{2} \\ \frac{c_{i+1} f'(i^+) + g(i^+)}{2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{c_{i+1} - c_i - k_i}{-k_i + c_i + c_{i+1}} & \frac{2c_{i+1}}{-k_i + c_i + c_{i+1}} \\ \frac{2c_i}{-k_i + c_i + c_{i+1}} & \frac{c_i - c_{i+1} - k_i}{-k_i + c_i + c_{i+1}} \end{pmatrix} \begin{pmatrix} \frac{c_i f'(i^-) - g(i^-)}{2} \\ \frac{c_{i+1} f'(i^+) + g(i^+)}{2} \end{pmatrix}. \end{aligned} \tag{37}$$

Thus,

$$\begin{aligned}
&\left|\frac{c_i f'(i^-)+g(i^-)}{2}\right|^2 \\
&\quad=\left|\frac{c_{i+1}-c_i-k_i}{-k_i+c_i+c_{i+1}}\cdot\frac{c_i f'(i^-)-g(i^-)}{2}+\frac{2c_{i+1}}{-k_i+c_i+c_{i+1}}\cdot\frac{c_{i+1}f'(i^+)+g(i^+)}{2}\right|^2 \\
&\quad\le 2\left[\left(\frac{c_{i+1}-c_i-k_i}{-k_i+c_i+c_{i+1}}\right)^2\left|\frac{c_i f'(i^-)-g(i^-)}{2}\right|^2\right. \\
&\qquad\left.+\left(\frac{2c_{i+1}}{-k_i+c_i+c_{i+1}}\right)^2\left|\frac{c_{i+1}f'(i^+)+g(i^+)}{2}\right|^2\right].
\end{aligned} \tag{38}$$

Similarly,

$$\begin{aligned}
\left|\frac{c_{i+1}f'(i^+)-g(i^+)}{2}\right|^2 &\le 2\left[\left(\frac{2c_i}{-k_i+c_i+c_{i+1}}\right)^2\left|\frac{c_i f'(i^-)-g(i^-)}{2}\right|^2\right. \\
&\quad\left.+\left(\frac{c_i-c_{i+1}-k_i}{-k_i+c_i+c_{i+1}}\right)^2\left|\frac{c_{i+1}f'(i^+)+g(i^+)}{2}\right|^2\right].
\end{aligned} \tag{39}$$

Next, substituting (38) into (32) produces

$$\begin{aligned}
I_3' &= \sum_{i=1}^{N}\left[c_iB_i(i)\left|\frac{c_i f'(i^-)-g(i^-)}{2}\right|^2+c_{i+1}A_{i+1}(i)\left|\frac{c_{i+1}f'(i^+)+g(i^+)}{2}\right|^2\right. \\
&\qquad\left.-c_{i+1}B_{i+1}(i)\left|\frac{c_{i+1}f'(i^+)-g(i^+)}{2}\right|^2-c_iA_i(i)\left|\frac{c_i f'(i^-)+g(i^-)}{2}\right|^2\right] \\
&\le \sum_{i=1}^{N}\left[c_iB_i(i)-2c_iA_i(i)\left(\frac{c_{i+1}-c_i-k_i}{-k_i+c_i+c_{i+1}}\right)^2\right. \\
&\qquad\left.-2c_{i+1}B_{i+1}(i)\left(\frac{2c_i}{-k_i+c_i+c_{i+1}}\right)^2\right]\left|\frac{c_i f'(i^-)-g(i^-)}{2}\right|^2 \\
&\quad+\left[c_{i+1}A_{i+1}(i)-2c_iA_i(i)\left(\frac{2c_{i+1}}{-k_i+c_i+c_{i+1}}\right)^2\right. \\
&\qquad\left.-2c_{i+1}B_{i+1}(i)\left(\frac{c_i-c_{i+1}-k_i}{-k_i+c_i+c_{i+1}}\right)^2\right]\left|\frac{c_{i+1}f'(i^+)+g(i^+)}{2}\right|^2.
\end{aligned} \tag{40}$$

Choosing $A_i(i), A_{i+1}(i), B_i(i), B_{i+1}(i), i=1,\dots,N$, properly so that $I_3'\ge 0$, we get

$$\operatorname{Re}\langle\mathcal{A}(f,g),(f,g)\rangle_* \le M\,\|(f,g)\|_*^2 \qquad \forall\,(f,g)\in D(\mathcal{A}). \tag{41}$$

That is, $\mathcal{A}-\lambda$ is dissipative for any $\lambda\ge M$. Referring to Lemma 3.3 below, we can take real $\lambda>M, \lambda\in\rho(\mathcal{A})$ so that $\mathcal{A}-\lambda$ satisfies all conditions of the Lumer–Phillips theorem (see, e.g., [17, p. 14]). Therefore, $\mathcal{A}-\lambda$ generates a C_0-semigroup on $\mathcal{H}$ and so does $\mathcal{A}$. Meanwhile, it is seen also that the above analysis is still valid after exchanging k_i with $-k_i$ under the condition (28).

We accomplish the proof by showing that $-\mathcal{A}$ also generates a C_0-semigroup on $\mathcal{H}$. To this end, consider the following equation:

$$\begin{cases} z_{tt}(x,t) - c_i^2 z_{xx}(x,t) = 0, \quad i-1 < x < i, i = 1, 2, \ldots, N+1, \\ z(0,t) = z(N+1,t) = 0, \\ z(i^-,t) = z(i^+,t), \\ c_i^2 z_x(i^-,t) - c_{i+1}^2 z_x(i^+,t) = -k_i z_t(i,t), \quad i = 1, \ldots, N, \\ z(x,0) = z_0(x), \quad z_t(x,0) = z_1(x). \end{cases} \tag{42}$$

System (42) can be written as an evolutionary equation in $\mathcal{H}$:

$$\begin{cases} \dfrac{dZ(t)}{dt} = \mathcal{B}Z(t), \quad t > 0, \\ Z(0) = Z_0 = (z_0, z_1), \end{cases} \tag{43}$$

where $Z(t) = (z(\cdot,t), z_t(\cdot,t))$ and $\mathcal{B}$ is given by

$$\begin{cases} \mathcal{B}(f,g) = (g(x), c_i^2 f''(x)), \quad i-1 < x < i, i = 1, 2, \ldots, N+1, \qquad \forall\, (f,g) \in D(\mathcal{A}), \\ D(\mathcal{B}) = \{(f,g) \in H | f, g \in H_0^1(0, N+1), f|_{[j-1,j]} \in H^2(j-1, j), 1 \le j \le N+1, \\ \qquad\qquad c_i^2 f'(i^-) - c_{i+1}^2 f'(i^+) = -k_i g(i), 1 \le i \le N\}. \end{cases} \tag{44}$$

Since the assumption (28) is also valid for system (42), it follows from the result justified that $\mathcal{B}$ generates a C_0-semigroup on $\mathcal{H}$. Hence for any $Z(0) \in D(\mathcal{B})$ there exists a unique classical solution $Z(t)$ to (43), since such a $Z(t)$ is well defined; that is to say, both $z(\cdot,t)$ and $z_t(\cdot,t)$ make sense for any $t \ge 0$. Let

$$W(t) = (w_1(\cdot,t), w_2(\cdot,t)) = (z(\cdot,t), -z_t(\cdot,t)).$$

Then it is seen that

$$\dot{W}(t) = -\mathcal{A}W(t),\ W(0) = (z_0, -z_1). \tag{45}$$

On the other hand, it is seen that $(z_0, z_1) \in D(\mathcal{B})$ if and only if $(z_0, -z_1) \in D(\mathcal{A}) = D(-\mathcal{A})$. We thus have proved that for any $W(0) \in D(-\mathcal{A})$ there exists a unique classical solution to (45). Since $\rho(-\mathcal{A}) \neq \emptyset$ from Lemma 3.3 below, it follows from Theorem 1.3 of [17, p. 102] that $-\mathcal{A}$ generates a C_0-semigroup on $\mathcal{H}$. Therefore, $\mathcal{A}$ generates a C_0-group on $\mathcal{H}$. □

Remark 3.2. It should be pointed out that the proof of Theorem 3.1 is similar to the approach used in [16]. However, since (23) and the equations studied in [16] are not always equivalent (see, e.g., [14]), we cannot transform (23) into the form of the latter. That is why we start directly from (23).

LEMMA 3.3. *Under condition* (28), $\mathcal{A}$ *is a discrete operator, and hence the spectrum* $\sigma(\mathcal{A})$ *of* $\mathcal{A}$ *consists of isolated eigenvalues only, and each eigenvalue is geometrically simple.*

Proof. Let $\lambda \in \mathbb{C}$. For any given $(f,g) \in \mathcal{H}$, find a pair $(u,v) \in D(\mathcal{A})$ so that

$$(\lambda - \mathcal{A})(u,v) = (f,g). \tag{46}$$

We then have that $v = \lambda u - f$ and u satisfies the following equation:

$$\begin{cases} \lambda^2 u - c_i^2 u'' = \lambda f + g \qquad \forall\, x \in (i-1,i),\ i = 1,2,\dots,N+1, \\ u(0) = u(N+1) = 0, \\ u(i^-) = u(i^+), \\ c_i^2 u'(i^-) - c_{i+1}^2 u'(i^+) = \lambda k_i u(i) - k_i f(i), \quad i = 1,\dots,N. \end{cases} \tag{47}$$

Solving (47) for $u(x)$ produces

$$u(x) = \begin{cases} a_1 \left(e^{\frac{\lambda}{c_1}x} - e^{-\frac{\lambda}{c_1}x} \right) - \displaystyle\int_0^x \frac{e^{\frac{\lambda}{c_1}(x-s)} - e^{-\frac{\lambda}{c_1}(x-s)}}{2\lambda c_1} (\lambda f + g)(s)ds, \quad x \in (0,1), \\ a_{i1} e^{-\frac{\lambda}{c_i}x} + a_{i2} e^{\frac{\lambda}{c_i}x} - \displaystyle\int_{i-1}^x \frac{e^{\frac{\lambda}{c_i}(x-s)} - e^{-\frac{\lambda}{c_i}(x-s)}}{2\lambda c_i} (\lambda f + g)(s)ds, \\ \qquad\qquad x \in (i-1,i),\ i = 2,\dots,N, \\ a_{N+1} \left[e^{\frac{\lambda}{c_{N+1}}(x-(N+1))} - e^{-\frac{\lambda}{c_{N+1}}(x-(N+1))} \right] \\ - \displaystyle\int_N^x \frac{e^{\frac{\lambda}{c_{N+1}}(x-s)} - e^{-\frac{\lambda}{c_{N+1}}(x-s)}}{2\lambda c_{N+1}} (\lambda f + g)(s)ds, \quad x \in (N,N+1), \end{cases} \tag{48}$$

where $a_1, a_{N+1}, a_{i1}, a_{i2}, i = 2,3,\dots,N$, are constants depending on λ. Substituting $u(x)$ into the boundary condition of (47), we see that $a_1, a_{N+1}, a_{i1}, a_{i2}, i = 2,3,\dots,N$, satisfy the following algebraic system of equations:

$$\begin{cases} a_1(e^{\frac{\lambda}{c_1}} - e^{-\frac{\lambda}{c_1}}) - a_{21}e^{-\frac{\lambda}{c_2}} - a_{22}e^{\frac{\lambda}{c_2}} = \dfrac{1}{\lambda}(f_{11} + g_{11}), \\ a_1 c_1 (e^{\frac{\lambda}{c_1}} + e^{-\frac{\lambda}{c_1}}) + a_{21}e^{-\frac{\lambda}{c_2}}(-k_1 + c_2) + a_{22}e^{\frac{\lambda}{c_2}}(-k_1 - c_2) = \dfrac{c_1}{\lambda}(f_{12} + g_{12}) - k_1 f(1), \\ a_{i1}e^{-\frac{i\lambda}{c_i}} + a_{i2}e^{\frac{i\lambda}{c_i}} - a_{i+1,1}e^{-\frac{i\lambda}{c_{i+1}}} - a_{i+1,2}e^{\frac{i\lambda}{c_{i+1}}} = \dfrac{1}{\lambda}(f_{i1} + g_{i1}), \\ -a_{i1}c_i e^{-\frac{i\lambda}{c_i}} + a_{i2}c_i e^{\frac{i\lambda}{c_i}} + a_{i+1,1}e^{-\frac{i\lambda}{c_{i+1}}}(-k_i + c_{i+1}) + a_{i+1,2}e^{\frac{i\lambda}{c_{i+1}}}(-k_i - c_{i+1}) \\ \qquad = \dfrac{c_i}{\lambda}(f_{i2} + g_{i2}) - k_i f(i), \quad i = 2,\dots,N-1, \\ a_{N1}e^{-\frac{N\lambda}{c_N}} + a_{N2}e^{\frac{N\lambda}{c_N}} - a_{N+1}(e^{-\frac{\lambda}{c_{N+1}}} - e^{\frac{\lambda}{c_{N+1}}}) = \dfrac{1}{\lambda}(f_{N1} + g_{N1}), \\ -a_{N1}c_N e^{-\frac{N\lambda}{c_N}} + a_{N2}c_N e^{\frac{N\lambda}{c_N}} + a_{N+1}[e^{-\frac{\lambda}{c_{N+1}}}(-k_N - c_{N+1}) - e^{\frac{\lambda}{c_{N+1}}}(-k_N + c_{N+1})] \\ \qquad = \dfrac{c_N}{\lambda}(f_{N2} + g_{N2}) - k_N f(N), \end{cases} \tag{49}$$

where

$$
\begin{cases}
f_{i1} = \displaystyle\int_{i-1}^{i} \frac{e^{\frac{\lambda}{c_i}(i-s)} - e^{-\frac{\lambda}{c_i}(i-s)}}{2c_i}\lambda f(s)ds \\
\quad = \displaystyle\int_{0}^{1} \frac{e^{\frac{\lambda}{c_i}(1-s)} - e^{-\frac{\lambda}{c_i}(1-s)}}{2c_i}\lambda f(i-1+s)ds, \\
f_{i2} = \displaystyle\int_{i-1}^{i} \frac{e^{\frac{\lambda}{c_i}(i-s)} + e^{-\frac{\lambda}{c_i}(i-s)}}{2c_i}\lambda f(s)ds, \\
g_{i1} = \displaystyle\int_{i-1}^{i} \frac{e^{\frac{\lambda}{c_i}(i-s)} - e^{-\frac{\lambda}{c_i}(i-s)}}{2c_i} g(s)ds, \\
g_{i2} = \displaystyle\int_{i-1}^{i} \frac{e^{\frac{\lambda}{c_i}(i-s)} + e^{-\frac{\lambda}{c_i}(i-s)}}{2c_i} g(s)ds \\
\quad = \displaystyle\int_{0}^{1} \frac{e^{\frac{\lambda}{c_i}(1-s)} + e^{-\frac{\lambda}{c_i}(1-s)}}{2c_i} g(i-1+s)ds, \\
\qquad\qquad\qquad\qquad i = 1, 2, \ldots, N.
\end{cases} \tag{50}
$$

Consider the determinant of the coefficients matrix $\Delta(\lambda) = [\Delta_1(\lambda), \Delta_2(\lambda)]$, where

$$
\begin{cases}
\Delta_1(\lambda) = \begin{pmatrix}
e^{\frac{\lambda}{c_1}} - e^{-\frac{\lambda}{c_1}} & -e^{-\frac{\lambda}{c_2}} & -e^{\frac{\lambda}{c_2}} \\
c_1(e^{\frac{\lambda}{c_1}} + e^{-\frac{\lambda}{c_1}}) & (-k_1 + c_2)e^{-\frac{\lambda}{c_2}} & (-k_1 - c_2)e^{\frac{\lambda}{c_2}} \\
0 & e^{-\frac{2\lambda}{c_2}} & e^{\frac{2\lambda}{c_2}} \\
0 & -c_2 e^{-\frac{2\lambda}{c_2}} & c_2 e^{\frac{2\lambda}{c_2}} \\
\vdots & \vdots & \vdots \\
0 & 0 & 0
\end{pmatrix}, \\
\Delta_2(\lambda) = \begin{pmatrix}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
-e^{-\frac{2\lambda}{c_3}} & -e^{\frac{2\lambda}{c_3}} & \cdots & 0 \\
(-k_2 + c_3)e^{-\frac{2\lambda}{c_3}} & (-k_2 - c_3)e^{\frac{2\lambda}{c_3}} & \cdots & 0 \\
\vdots & \vdots & \vdots & 0 \\
0 & 0 & \cdots & \theta
\end{pmatrix},
\end{cases} \tag{51}
$$

with $\theta = -e^{-\frac{\lambda}{c_{N+1}}}(k_N + c_{N+1}) + e^{\frac{\lambda}{c_{N+1}}}(k_N - c_{N+1})$. Taking $e^{\frac{\lambda}{c_1}}$ out of the first column, $e^{-\frac{\lambda}{c_2}}$ from the second column, $e^{2\frac{\lambda}{c_2}}$ from the third column of $\Delta(\lambda)$, and so on, we obtain

$$
\begin{aligned}
& e^{-\lambda\sum_{i=1}^{N+1}\frac{1}{c_i}} \det(\Delta(\lambda)) \\
& \quad = \det \begin{pmatrix}
1 & -1 & 0 & 0 & \cdots & 0 & 0 \\
c_1 & -k_1 + c_2 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & -1 & \cdots & 0 & 0 \\
0 & 0 & c_2 & -k_2 + c_3 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 1 & 1 \\
0 & 0 & 0 & 0 & \cdots & c_N & k_N - c_{N+1}
\end{pmatrix} + o(1) \\
& \quad = -\prod_{j=1}^{N}(-k_j + c_j + c_{j+1}) + o(1) \qquad \text{as } \mathrm{Re}\lambda \longrightarrow +\infty.
\end{aligned} \tag{52}
$$

Hence, if condition (28) holds, $\det(\Delta(\lambda))$ is an entire function not identical to zero. Additionally, when $\det(\Delta(\lambda)) \neq 0$, the system of equations (49) admits a solution $a_1, a_{N+1}, a_{i1}, a_{i2}, i = 2, \ldots, N$,

$$
\begin{cases}
a_1 = \dfrac{1}{\det(\Delta(\lambda))}\left[\displaystyle\sum_{i=1}^{N} \Delta_{2i-1,1}(\lambda)\frac{f_{i1}+g_{i1}}{\lambda} - \Delta_{2i,1}(\lambda)\left[\frac{c_i(f_{i2}+g_{i2})}{\lambda} - k_i f(i)\right]\right], \\
a_{N+1} \\
\quad = \dfrac{1}{\det(\Delta(\lambda))}\left[\displaystyle\sum_{i=1}^{N} -\Delta_{2i-1,2N}(\lambda)\frac{f_{i1}+g_{i1}}{\lambda} + \Delta_{2i,2N}(\lambda)\left[\frac{c_i(f_{i2}+g_{i2})}{\lambda} - k_i f(i)\right]\right], \\
a_{l1} \\
\quad = \dfrac{1}{\det(\Delta(\lambda))}\left[\displaystyle\sum_{i=1}^{N} -\Delta_{2i-1,2(l-1)}(\lambda)\frac{f_{i1}+g_{i1}}{\lambda} + \Delta_{2i,2(l-1)}(\lambda)\left[\frac{c_i(f_{i2}+g_{i2})}{\lambda} - k_i f(i)\right]\right], \\
a_{l2} \\
\quad = \dfrac{1}{\det(\Delta(\lambda))}\left[\displaystyle\sum_{i=1}^{N} \Delta_{2i-1,2l-1}(\lambda)\frac{f_{i1}+g_{i1}}{\lambda} - \Delta_{2i,2l-1}(\lambda)\left[\frac{c_i(f_{i2}+g_{i2})}{\lambda} - k_i f(i)\right]\right], \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad l = 2, \ldots, N,
\end{cases}
\tag{53}
$$

where $\Delta_{ij}(\lambda), 1 \leq i, j \leq 2N$, is the algebraic cofactor of ith row and jth column of the matrix $\Delta(\lambda)$.

From the discussion above and the Sobolev embedding theorem, we see that under the condition (28), for any $\lambda \in \rho(\mathcal{A})$, $(\lambda - \mathcal{A})^{-1}$ is compact and $(u, v) = (\lambda - \mathcal{A})^{-1}$ (f, g) is given by (48) with the coefficients $a_1, a_{N+1}, a_{i1}, a_{i2}, i = 2, 3, \ldots, N$, in (53). Therefore, the spectrum $\sigma(\mathcal{A})$ of $\mathcal{A}$ consists of isolated eigenvalues only, and $\lambda \in \sigma(\mathcal{A})$ if and only if $\det(\Delta(\lambda)) = 0$.

Moreover, it is seen from (48) and (53) that for any $\lambda \in \rho(\mathcal{A})$, $(\lambda - \mathcal{A})^{-1}$ is of the form

$$
(\lambda - \mathcal{A})^{-1}(f, g) = \frac{G(\lambda, (f, g))}{\det(\Delta(\lambda))} \qquad \forall\ (f, g) \in \mathcal{H}, \tag{54}
$$

where $G(\lambda, (f, g))$ is an $\mathcal{H}$-valued entire function with order at most 1. The order of $\det(\Delta(\lambda))$ is just 1 (see [25]).

Finally, simple computations show that the rank of the matrix $\Delta(\lambda)$ is less than its order by one. Hence, each eigenvalue of $\mathcal{A}$ is geometrically simple. □

THEOREM 3.4. *Under condition* (28)*, the root subspace of $\mathcal{A}$ is complete in H: $Sp(\mathcal{A}) = \mathcal{H}$.*

Proof. From Theorem 3.1, $\mathcal{A}$ generates a C_0-group and so does $\mathcal{A}^*$. Hence $R(\lambda, \mathcal{A}^*)$ is uniformly bounded along the negative real axis as $\lambda \to -\infty$. Notice that $(\overline{\lambda} - \mathcal{A}^*)^{-1}(f, g) = G^*(\lambda, (f, g))/\overline{\det(\Delta(\lambda))}$ $\forall\ (f, g) \in \mathcal{H}$, and the eigenvalues of $\mathcal{A}$ are symmetric about the real axis. Taking $\rho_2 = \rho = 1, n = 2, \gamma_1 = \{\lambda | \arg \lambda = \pi\}$, we see that all conditions of Theorem 4 of [24] are satisfied. The result follows. □

Finally, we go to the main result of this section: the Riesz basis generation of the root subspace of operator $\mathcal{A}$. Before doing this, we introduce the concept of a sine-type function.

Recall that an entire function f is called exponential type (see, e.g., [25]) if there are constants $C, D > 0$ such that

$$|f(z)| \leq Ce^{D|z|}$$

for all complex numbers z. The function of exponential type f is called *sine-type* if the following two conditions are satisfied [1, Definition II.1.27]:

(a) The zeros of f lie in a strip $\{z \in \mathbb{C} || \text{Im} z| \leq h\}$ for some $h > 0$.

(b) There are $c_1, c_2 > 0$, $y_0 \in \mathbb{R}$ such that for all $x \in \mathbb{R}$, $c_1 \leq |f(x + iy_0)| \leq c_2$.

The class of sine-type functions was first introduced in [11] to deal with problems of interpolation by entire functions and in studying Riesz basis in L^2 space. The distribution of the zeros of a sine-type function is given by the next important proposition (Proposition II.1.28 in [1]; see also [12]).

PROPOSITION 3.5. *Let f be a sine-type function. Its set of zeros (a multiple zero is repeated in a number of times equal to its multiplicity) is a finite unification of separable sets. Consequently, the multiplicities of zeros of sine-type functions are uniformly upper bounded.*

Now let

$$F(z) = \det(\Delta(z)). \tag{55}$$

Then the zeros of F consist of eigenvalues of $\mathcal{A}$. (However, the multiplicity as zero of F may be different from algebraic multiplicity as eigenvalue of $\mathcal{A}$.) Since $\mathcal{A}$ generates a C_0-group, all zeros of F lie in a strip parallel to the imaginary axis. From (51) and (52), it is seen that $F(iz)$ is a function of exponential type, and $|F(iz)|$ has positive lower and upper bound along a straight line parallel to the real axis in the lower half complex plane. Therefore, $F(iz)$ is a sine-type function. It then follows from Proposition 3.5 that the zeros of $F(iz)$ can be decomposed into a finite union of separable sets (accounted according to their multiplicity as zeros of F). Furthermore, since each $\lambda_n \in \sigma(\mathcal{A})$ is geometrically simple, the algebraic multiplicity of λ_n is equal to the order of $R(\lambda, \mathcal{A})$ at λ_n (see [15]), which is less than or equal to the multiplicity of λ_n as the zero of F by the decomposition (54). Therefore, from Proposition 3.5, we know that for all $\lambda_n \in \sigma(\mathcal{A})$ the algebraic multiplicity m_n of each is uniformly upper bounded:

$$\sup_{\lambda_n \in \sigma(\mathcal{A})} m_n < \infty. \tag{56}$$

Thus the eigenvalues of $\mathcal{A}$ can be decomposed into a finite union of separable sets (accounted according to their algebraic multiplicity as eigenvalues of $\mathcal{A}$):

$$\text{eigenvalues of } \mathcal{A} = \Lambda = \bigcup_{n=1}^{N} \Lambda_n, \qquad \inf_{i \neq j, \lambda_i, \lambda_j \in \Lambda_n} |\lambda_i - \lambda_j| > 0 \quad \forall\, 1 \leq n \leq N. \tag{57}$$

Let $\delta = \min_{1 \leq n \leq N} \inf_{i \neq j, \lambda_i, \lambda_j \in \Lambda_n} |\lambda_i - \lambda_j| > 0$. Then for any $r < r_0 = \delta/(2N)$, by the discussion in section 2, there exist $\Lambda^p = \{\lambda_{j,p}\}_{j=1}^{M^p}, M^p \leq N, p \in \mathbb{Z}$, the pth connected component of intersection of Λ with $\bigcup_{n \in \mathbb{Z}} D_{\lambda_n}(r)$, where $D_{\lambda_n}(r)$ is the circle centered at λ_n with radius r such that

$$\sigma(A) = \bigcup_{p \in \mathbb{Z}} \Lambda^p. \tag{58}$$

We may assume without loss of generality that $\{\lambda_n\}$ are arranged so that $\mathrm{Im}\lambda_n$ are nondecreasing and $\mathrm{Re}\lambda_{1,p} \geq \mathrm{Re}\lambda_{2,p} \geq \cdots \geq \mathrm{Re}\lambda_{M^p,p}$. Form a family of GDD as follows:

$$E^p(\Lambda, r) = \{[\lambda_{1,p}](t), [\lambda_{1,p}, \lambda_{2,p}](t), \ldots, [\lambda_{1,p}, \lambda_{2,p}, \ldots, \lambda_{M^p,p}](t)\}, \quad p \in \mathbb{Z}.$$

It is seen from (4) that $D^+(\Lambda) < \infty$. From Proposition 2.3, for any $T > 2\pi D^+(\Lambda)$, the family of GDD $\{E^p(\Lambda, r)\}_{p\in\mathbb{Z}}$ forms a Riesz basis in the closed subspace spanned by itself in $L^2(0,T)$. Specifically, suppose that each $\Lambda^p = \{\lambda_j^p\}_1^{N^p}$ has N^p different elements and that each appears m_{pj} times, $\sum_{j=1}^{N^p} m_{pj} = M^p$. Since $M^p \leq N$, all conditions of Theorem 2.5 are satisfied. We thus have proved the following result on the Riesz basis generation of the root subspace of $\mathcal{A}$.

THEOREM 3.6. *Under condition* (28)*, the operator* $\mathcal{A}$ *defined by* (25) *and* (26) *has the following properties:*

(a) *There exists* $\epsilon > 0$ *such that* $\sigma(\mathcal{A}) = \bigcup_{p\in\mathbb{Z}}\{\lambda_i^p\}_{i=1}^{N^p}$ *(the multiplicity is not accounted), where* $\sup_p N^p < \infty$, $|\lambda_i^p - \lambda_j^q| \geq \epsilon\ \forall p, q \in \mathbb{Z},\ p \neq q, 1 \leq i \leq N^p$, $1 \leq j \leq N^q$.

(b) *The algebraic multiplicity* m_{pi} *of* $\lambda_i^p \in \sigma(A)$ *is uniformly upper bounded:*

$$\sup_{p\in\mathbb{Z}, 1\leq i\leq N^p} m_{pi} < \infty.$$

(c)

$$Y = \sum_{p\in\mathbb{Z}} \sum_{i=1}^{N^p} \mathbb{P}_{\lambda_i^p} Y \qquad \forall\, Y \in \mathcal{H}, \tag{59}$$

where $\mathbb{P}_{\lambda_i^p}$ *is the eigen-projection of* $\mathcal{A}$ *corresponding to* λ_i^p.

(d) *There are constants* $M_1, M_2 > 0$ *such that*

$$M_1 \sum_{p\in\mathbb{Z}} \left\| \sum_{i=1}^{N^p} \mathbb{P}_{\lambda_i^p} Y \right\|^2 \leq \|Y\|^2 \leq M_2 \sum_{p\in\mathbb{Z}} \left\| \sum_{i=1}^{N^p} \mathbb{P}_{\lambda_i^p} Y \right\|^2 \qquad \forall\, Y \in \mathcal{H}. \tag{60}$$

(e) *The spectrum-determined growth condition holds:*

$$\omega(\mathcal{A}) = S(\mathcal{A}).$$

Remark 3.7. Assertion (e) in Theorem 3.6 was proved in [14], [16] by completely different arguments, while in this article it becomes an immediate consequence of Riesz basis generation. It should be mentioned that if condition (28) is not satisfied, there may be cases like $\sigma(\mathcal{A}) = \emptyset$ when, e.g., $N = 1, c_1 = c_2 = 2, k_1 = 4$. In this case, we certainly could not talk about a Riesz basis. Finally, we point out that the method presented in this paper can be applied to system (23) with different boundary conditions as well.

The property expressed by (59) and (60) is nothing other than the *Riesz basis with parentheses* [20]. Further, if we know that the eigenvalues of $\mathcal{A}$ are separable, the Riesz basis with parentheses reduces the usual Riesz basis. Unfortunately, it is not clear whether or not separability holds.

REFERENCES

[1] S. A. Avdonin and S. A. Ivanov, *Families of Exponentials: The Method of Moments in Controllability Problems for Distributed Parameter Systems*, Cambridge University Press, Cambridge, UK, 1995.

[2] S. A. Avdonin and S. A. Ivanov, *Riesz bases of exponentials and divided differences*, St. Petersburg Math. J., 13 (2002), pp. 339–351.

[3] S. A. Avdonin and W. Moran, *Ingham-type inequalities and Riesz bases of divided differences*, Int. J. Appl. Math. Comput. Sci., 11 (2001), pp. 803–820.

[4] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts in Appl. Math. 21, Springer-Verlag, New York, 1995.

[5] B.-Z. Guo, *Riesz basis property and exponential stability of controlled Euler–Bernoulli beam equations with variable coefficients*, SIAM J. Control Optim., 40 (2002), pp. 1905–1923.

[6] B. Z. Guo and K. Y. Chan, *Riesz basis generation, eigenvalues distribution, and exponential stability for a Euler–Bernoulli beam with joint feedback control*, Rev. Mat. Complut., 14 (2001), pp. 205–229.

[7] B. Z. Guo and Y. H. Luo, *Riesz basis property of a second order hyperbolic system with collocated scalar input/output*, IEEE Trans. Automat. Control, 47 (2002), pp. 693–698.

[8] B. Z. Guo and R. Yu, *The Riesz basis property of discrete operators and application to a Euler–Bernoulli beam equation with boundary linear feedback control*, IMA J. Math. Control Inform., 18 (2001), pp. 241–251.

[9] S. V. Hruschev, N. K. Nikolski, and B. S. Pavlov, *Unconditional bases of exponentials and reproducing kernels*, in Complex Analysis and Spectral Theory, Lecture Notes in Math. 864, Springer-Verlag, New York, 1981, pp. 214–335.

[10] B. Jacob and H. Zwart, *Exact controllability of C_0-groups with one-dimensional input operators*, in Advances in Mathematical Systems Theory, F. Colonius, ed., Birkhäuser Boston, Cambridge, MA, 2000, pp. 221–242.

[11] B. Ya. Levin, *On bases of exponential functions in L_2*, Zapiski Math. Otd. Phys. Math. Facul. Khark. Univ., 27 (1961), pp. 39–48 (in Russian).

[12] B. YA. Levin and I. V. Ostrovskii, *Small perturbations of the set of roots of sine-type functions*, Izv. Akad. Nauk SSSR Ser. Mat., 43 (1979), pp. 87–110 (in Russian).

[13] J. Locker, *Spectral Theory of Non–Self-Adjoint Two-Point Differential Operators*, American Mathematical Society, Providence, RI, 2000.

[14] Z. H. Luo, B. Z. Guo, and O. Mörgul, *Stability and Stabilization of Infinite Dimensional Systems with Applications*, Springer-Verlag, London, 1999.

[15] R. Nagel, Ed., *One Parameter Semigroups of Positive Operators*, Lecture Notes in Math. 1184, Springer-Verlag, New York, 1986.

[16] A. F. Neves, H. D. S. Ribeiro, and O. Lopes, *On the spectrum of evolution operators generated by hyperbolic systems*, J. Funct. Anal., 67 (1986), pp. 320–344.

[17] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[18] P. Rideau, *Contrôle d'un Assemblage de Poutres Flexibles par des Capteurs-Actionneurs Ponctuels: Étude du Spectre du Système*, Thèse, Ecole Nationale Supérieure des Mines de Paris, Sophia-Antipolis, France, 1985.

[19] D. L. Russell, *Nonharmonic Fourier series in control theory of distributed systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–559.

[20] A. A. Shklikov, *Boundary problems for ordinary differential equations with parameter in the boundary conditions*, J. Soviet Math., 33 (1986), pp. 1311–1342.

[21] M. A. Shubov, *Basis property of eigenfunctions of nonselfadjoint operator pencils generated by the equation of nonhomogeneous damped string*, Integral Equations Operations Theory, 25 (1996), pp. 289–328.

[22] M. A. Shubov, *Spectral operators generated by damped hyperbolic equations*, Integral Equations Operations Theory, 28 (1997), pp. 358–372.

[23] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

[24] G.-Q. Xu and B.-Z. Guo, *Riesz basis property of evolution equations in Hilbert spaces and application to a coupled string equation*, SIAM J. Control Optim., 42 (2003), pp. 966–984.

[25] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, London, 1980.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1253–1271

FINITE TIME STABILITY AND ROBUST CONTROL SYNTHESIS OF UNCERTAIN SWITCHED SYSTEMS*

Y. ORLOV†

Abstract. Stability analysis is developed for uncertain nonlinear switched systems. While being asymptotically stable and homogeneous of degree $q < 0$, these systems are shown to approach the equilibrium point in finite time. Restricted to second order systems, this feature is additionally demonstrated to persist regardless of inhomogeneous perturbations. Based on this fundamental property, switched control algorithms are then developed to globally stabilize uncertain minimum phase systems of uniform m-vector relative degree $(2, \dots, 2)^T$. The controllers constructed do not rely on the generation of sliding motions while providing robustness features similar to those possessed by their sliding mode counterparts. The proposed synthesis procedure is illustrated via application to a friction servo-motor.

Key words. robust stability, nonautonomous switched system, homogeneity, finite time stabilization

AMS subject classifications. 49K24, 93D09, 93B12

DOI. 10.1137/S0363012903425593

1. Introduction.

Finite time stability of asymptotically stable homogeneous systems has been well recognized for only continuous vector fields [16]. Extending this result to switched systems presents a formidable problem and requires proceeding differently because a smooth homogeneous Lyapunov function, whose existence was proven in [27] for continuous asymptotically stable homogeneous vector fields, can no longer be brought into play.

In the present work the finite time stability property is established for homogeneous asymptotically stable switched systems of homogeneity degree $q < 0$. Exemplified with a second order system, the finite time stability is additionally demonstrated to remain in force regardless of inhomogeneous perturbations. This fundamental property forms a basis for subsequent robust synthesis of globally stabilizing controllers of uncertain nonlinear minimum phase systems of uniform m-vector relative degree $(2, \dots, 2)^T$ (see [9] for the relative degree concept).

The strategy of the discontinuous controllers constructed is to drive the system to the zero dynamics manifold in finite time and maintain it there in spite of the parameter uncertainties and external disturbances, both with a priori known norm bounds. Desired robustness properties are thus provided and asymptotic stability of the closed-loop system is guaranteed.

The sliding mode control technique has long been recognized as a powerful control method to counteract nonvanishing external disturbances and unmodelled dynamics [32, 33]. In contrast to this standard technique, the present synthesis does not rely on the generation of sliding modes, while exhibiting an infinite number of switches on a finite time interval. This phenomenon, called the Fuller phenomenon, is now fully understood from optimal control theory [13, 34] where it has been extensively exploited for higher (greater than 1) relative degree systems to attain optimal

*Received by the editors April 2, 2003; accepted for publication (in revised form) April 26, 2004; published electronically January 5, 2005.
http://www.siam.org/journals/sicon/43-4/42559.html
†CICESE Research Center, Electronics and Telecommunications Department, P.O. Box 434944 San Diego, CA 92143-4944 (yorlov@cicese.mx).

performance. A controller which exhibits the Fuller phenomenon is referred to as a chattering controller [34] that has become standard in the literature.

Typically, optimal chattering controllers appear to minimize a degenerate quadratic criterion, viewed over an infinite horizon under instant control constraints and dependent upon the output of the system. Due to this, the input-output stability property is only guaranteed while admitting instability of the overall system. In order to ensure asymptotic stability of the closed-loop system, suboptimal (twisting, supertwisting, and others) chattering control algorithms, also referred to as second order sliding mode control algorithms, have subsequently been developed for SISO (Single Input-Single Output) systems (see the very recent survey [12] and references therein). While retaining useful robustness features against matching disturbances, these algorithms have greatly reduced undesired high-frequency oscillations of the closed-loop system compared to those caused by their standard sliding mode counterparts (see [3, 4] for further details).

Until recently, attempts to extend the aforementioned algorithms to MIMO (Multi Input-Multi Output) systems were hampered by the lack of suitable analysis tools for switched systems and, consequently, the lack of consistent control synthesis methodology. In the present work chattering control synthesis is constructively developed for uncertain nonlinear MIMO systems, regardless of whether the optimal performance is achieved. To facilitate the exposition, the model chosen for treatment is confined to that of uniform m-vector relative degree $(2, \dots, 2)^T$. A possible extension of chattering control synthesis to MIMO systems of higher relative degree does not seem trivial and is beyond the scope of the paper.

Attractive features of the chattering controllers proposed are illustrated by application to finite time stabilization of a servo-motor. It should be noted that allowing relatively strong Coulomb friction in this application precludes the use of continuous regulators. Indeed, the closed-loop system in that case would have a nontrivial set of equilibrium points and it would therefore be driven to a wrong endpoint. As opposed to continuous controllers, the chattering controllers are demonstrated to be capable of providing the desired system performance in spite of significant uncertainties in the system description as is typically the case in control of electromechanical systems with complex backlash/friction phenomena.

The paper is organized as follows. Basic definitions are given in section 2. Finite time stability of nonautonomous switched systems is proven in section 3. Perturbation analysis of second order homogeneous switched systems is developed in section 4. Chattering control synthesis and its interpretation in an electromechanical application are proposed in section 5. Finally, section 6 presents some conclusions.

2. Basic definitions. The model of a nonautonomous switched system in question is given by

$$\dot{x} = \varphi(x, t), \tag{2.1}$$

where $x = (x_1, \dots, x_n)^T$ is the state vector, $t \in \mathbf{R}$ is the time variable, and the function $\varphi = (\varphi_1, \dots, \varphi_n)^T$ is piece-wise continuous. Recall that the function $\varphi : \mathbf{R}^{n+1} \to \mathbf{R}^n$ is piece-wise continuous iff $\mathbf{R}^{n+1}$ is partitioned into a finite number of domains $G_j \subset \mathbf{R}^{n+1}$, $j = 1, \dots, N$, with disjoint interiors and boundaries ∂G_j of measure zero such that φ is continuous within each of these domains and for all $j = 1, \dots, N$ it has a finite limit $\varphi^j(x, t)$ as the argument $(x^*, t^*) \in G_j$ approaches a boundary point $(x, t) \in \partial G_j$.

Throughout, the precise meaning of the differential equation (2.1) with a piecewise continuous right-hand side is defined in the sense of Filippov [10].

DEFINITION 2.1. *Given the differential equation* (2.1) *let us introduce for each point* $(x,t) \in \mathbf{R}^n \times \mathbf{R}$ *the smallest convex closed set* $\Phi(x,t)$ *which contains all the limit points of* $\varphi(x^*,t)$ *as* $x^* \to x$, $t = const$, *and* $(x^*,t) \in \mathbf{R}^{n+1} \setminus (\cup_{j=1}^N \partial G_j)$. *An absolutely continuous function* $x(\cdot)$, *defined on an interval* I, *is said to be a solution of* (2.1) *if it satisfies the differential inclusion*

$$\dot{x} \in \Phi(x,t) \tag{2.2}$$

almost everywhere on I.

By Theorem 8 of [10, p. 85], system (2.1) has a solution for arbitrary initial conditions $x(t_0) = x^0 \in \mathbf{R}^n$. This solution is locally defined on some time interval $[t_0, t_1)$; however, generally speaking it is nonunique.

It is worth noticing that the above model (2.1) admits a sliding motion on the boundary set $\mathcal{N} = \bigcup_{j=1}^N \partial G_j$ with an infinite number of switches on a finite time interval. Thus, a larger class of switched systems is captured in comparison to that of [8, 22, 23] where infinitely fast switching is explicitly ruled out.

Along with the differential equation (2.1), we deal with its perturbed version

$$\dot{x} = \varphi(x,t) + \psi(x,t), \tag{2.3}$$

where ψ is a piece-wise continuous function whose components $\psi_i, \dots, \psi_n$ are locally uniformly bounded within a ball B_δ, centered at the origin with radius δ; i.e.,

$$|\psi_i(x,t)| \le M_i, \quad i = 1, \dots, n, \tag{2.4}$$

for almost all $(x,t) \in B_\delta \times \mathbf{R}$ and some constants $M_i \ge 0$, fixed a priori. The above equation is further viewed as a differential equation with rectangular uncertainties. A solution concept for such an uncertain differential equation is introduced as follows.

DEFINITION 2.2. *An absolutely continuous function* $x(\cdot)$, *defined on an interval* I, *is said to be a solution of the uncertain differential equation* (2.3) *with the rectangular uncertainty constraints* (2.4) *iff it is a solution of* (2.3) *on the interval* I *in the sense of Definition* 2.1 *for some piece-wise continuous function* ψ *subject to* (2.4).

As a matter of fact, the differential equation (2.1) with no uncertain term ψ is particularly represented in the form of (2.3), (2.4) with $M_i = 0$ for all $i = 1, \dots, n$.

It should be pointed out that an uncertain system (2.3) with uncertainty constraints (2.4) can be represented as a differential inclusion of the form

$$\dot{x} \in \Phi(x,t) + \Psi, \tag{2.5}$$

where $\Phi(x,t)$ is the same as it appears in Definition 2.1, Ψ is the Cartesian product of the intervals $\Psi_i = [-M_i, M_i]$, $i = 1, \dots, n$, and the set

$$\Phi(x,t) + \Psi = \{\phi + \psi : \phi \in \Phi(x,t),\ \psi \in \Psi\}.$$

If $\varphi(x,t) = \varphi(x)$ is time-independent, the uncertain system (2.3), (2.4) is governed by the autonomous differential inclusion (2.5), in spite of the presence of the uncertain time-varying term $\psi(x,t)$.

Stability of a switched system (2.1) with possibly nonuniquely defined trajectories and stability of an uncertain system (2.3), (2.4) to be uniform in the uncertainty are introduced by means of the corresponding differential inclusion. In turn, stability of

a differential inclusion (2.2) is defined as follows. Suppose $x = 0$ is an equilibrium point of the differential inclusion (2.2) and $x(\cdot, t_0, x^0)$ denotes a solution $x(\cdot)$ of (2.2) under the initial conditions $x(t_0) = x^0$.

DEFINITION 2.3. *The equilibrium point $x = 0$ of the differential inclusion* (2.2) *is stable* (*uniformly stable*) *iff for each $t_0 \in \mathbf{R}$, $\varepsilon > 0$, there is $\delta = \delta(\varepsilon, t_0) > 0$, dependent on ε and possibly dependent on t_0* (*respectively, independent of t_0*) *such that each solution $x(\cdot, t_0, x^0)$ of* (2.2) *with the initial data $x^0 \in B_\delta$ exists on the semi-infinite time interval $[t_0, \infty)$ and satisfies the inequality*

$$\|x(t, t_0, x^0)\| < \varepsilon \quad \textit{for all } t \in [t_0, \infty).$$

DEFINITION 2.4. *The equilibrium point $x = 0$ of the differential inclusion* (2.2) *is said to be* (*uniformly*) *asymptotically stable if it is* (*uniformly*) *stable and the convergence*

$$\lim_{t \to \infty} \|x(t, t_0, x^0)\| = 0 \tag{2.6}$$

holds for all solutions of (2.2) *initialized within some B_δ* (*uniformly in the initial data t_0 and x^0, and all the solutions $x(\cdot, t_0, x^0)$*). *If the above convergence remains in force for all solutions of* (2.2) *regardless of the choice of the initial data* (*and, respectively, for each $\delta > 0$ the convergence is uniform in t_0 and $x^0 \in B_\delta$, and all $x(\cdot, t_0, x^0)$*), *the equilibrium point is said to be globally* (*uniformly*) *asymptotically stable.*

DEFINITION 2.5. *The equilibrium point $x = 0$ of the differential inclusion* (2.2) *is said to be globally* (*uniformly*) *finite time stable if, in addition to the global* (*uniform*) *asymptotical stability, the limiting relation*

$$x(t, t_0, x^0) = 0 \tag{2.7}$$

holds for each solution $x(\cdot, t_0, x^0)$ and all $t \geq t_0 + T(t_0, x^0)$, where the settling time function

$$T(t_0, x^0) = \sup_{x(\cdot, t_0, x^0)} \inf\{T \geq 0 : x(t, t_0, x^0) = 0 \textit{ for all } t \geq t_0 + T\} \tag{2.8}$$

is such that

$$T(t_0, x^0) < \infty \quad \textit{for all } t_0 \in \mathbf{R} \textit{ and } x^0 \in \mathbf{R}^n$$

(*respectively,* $\mathcal{T}(B_\delta) = \sup_{t_0 \in \mathbf{R},\, x^0 \in B_\delta} T(t_0, x^0) < \infty$ *for each* $\delta > 0$).

In application to the uncertain system (2.3), (2.4) the above definitions are specified as follows. In order to emphasize that these definitions require such a system to be uniformly stable not only in the initial data but also in the uncertainty, the corresponding system will be referred to as equiuniformly stable. Suppose that $x = 0$ is an equilibrium point of the uncertain system (2.3), (2.4) (i.e., $x = 0$ is a solution of (2.3) for some function ψ_0, admissible in the sense of (2.4)) and let $x_\psi(\cdot, t_0, x^0)$ denote a solution $x(\cdot)$ of (2.2) for some admissible function ψ under the initial conditions $x(t_0) = x^0$.

DEFINITION 2.6. *The equilibrium point $x = 0$ of the uncertain system* (2.3), (2.4) *is equiuniformly stable iff for each $t_0 \in \mathbf{R}$, $\varepsilon > 0$, there is $\delta = \delta(\varepsilon) > 0$, dependent on ε and independent of t_0 and ψ, such that each solution $x_\psi(\cdot, t_0, x^0)$ of* (2.3), (2.4) *with the initial data $x^0 \in B_\delta$ exists on the semi-infinite time interval $[t_0, \infty)$ and satisfies the inequality*

$$\|x_\psi(t, t_0, x^0)\| < \varepsilon \quad \textit{for all } t \in [t_0, \infty).$$

DEFINITION 2.7. *The equilibrium point $x = 0$ of the uncertain system* (2.3), (2.4) *is said to be equiuniformly asymptotically stable if it is equiuniformly stable and the convergence*

$$\lim_{t\to\infty} \|x_\psi(t, t_0, x^0)\| = 0 \tag{2.9}$$

holds for all solutions of (2.3), (2.4) *initialized within some B_δ, uniformly in the initial data t_0 and x^0, and all the solutions $x_\psi(\cdot, t_0, x^0)$. If this convergence remains in force for each $\delta > 0$, the equilibrium point is said to be globally equiuniformly asymptotically stable.*

DEFINITION 2.8. *The equilibrium point $x = 0$ of the uncertain system* (2.3), (2.4) *is said to be globally equiuniformly finite time stable if, in addition to the global equiuniform asymptotical stability, the limiting relation*

$$x_\psi(t, t_0, x^0) = 0 \tag{2.10}$$

holds for each solution $x_\psi(\cdot, t_0, x^0)$ and all $t \geq t_0 + T(t_0, x^0)$, where the settling time function

$$T(t_0, x^0) = \sup_{x_\psi(\cdot, t_0, x^0)} \inf\{T \geq 0 : x_\psi(t, t_0, x^0) = 0 \quad \textit{for all } t \geq t_0 + T\} \tag{2.11}$$

is such that

$$\mathcal{T}(B_\delta) = \sup_{t_0\in\mathbf{R},\, x^0\in B_\delta} T(t_0, x^0) < \infty \quad \textit{for each } \delta > 0.$$

In the present paper, we focus our analysis on the global equiuniform finite time stability of switched systems. The concept of homogeneity, studied earlier in [15, 35] and [27] for continuously differentiable and, respectively, continuous vector fields, plays a central role in our analysis. This concept is now generalized for differential inclusions and, particularly, for nonautonomous switched systems.

DEFINITION 2.9. *The differential inclusion* (2.2) (*the differential equation* (2.1) *or the uncertain system* (2.3), (2.4)) *is called locally homogeneous of degree $q \in \mathbf{R}$ with respect to dilation $(r_1, \dots, r_n)$, where $r_i > 0$, $i = 1, \dots, n$, if there exist a constant $c_0 > 0$, called a lower estimate of the homogeneity parameter, and a ball $B_\delta \subset \mathbf{R}^n$, called a homogeneity ball, such that any solution $x(\cdot)$ of* (2.2) (*respectively, that of* (2.1) *or* (2.3), (2.4)), *evolving within the ball B_δ, generates a parameterized set of solutions $x^c(\cdot)$ with components*

$$x_i^c(t) = c^{r_i} x_i(c^q t) \tag{2.12}$$

and parameter $c \geq c_0$.

DEFINITION 2.10. *A piece-wise continuous function $\varphi : \mathbf{R}^{n+1} \to \mathbf{R}^n$ is called locally homogeneous of degree $q \in \mathbf{R}$ with respect to dilation $(r_1, \dots, r_n)$, where $r_i > 0$, $i = 1, \dots, n$, if there exist a constant $c_0 > 0$ and a ball $B_\delta \subset \mathbf{R}^n$ such that*

$$\varphi_i(c^{r_1} x_1, \dots, c^{r_n} x_n, c^{-q} t) = c^{q+r_i} \varphi_i(x_1, \dots, x_n, t) \tag{2.13}$$

for all $c \geq c_0$ and almost all $(x, t) \in B_\delta \times \mathbf{R}$.

The global homogeneity concept for the differential inclusion (2.2) and that for the piece-wise continuous function φ are formally introduced by setting $\delta = \infty$ in the above definitions.

It is worth noting that Definitions 2.9 and 2.10 are consistent in the sense that homogeneity of the function φ ensures homogeneity of the corresponding differential equation (2.1).

LEMMA 2.11. *Let a piece-wise continuous function φ be locally homogeneous of degree $q \in \mathbf{R}$ with respect to dilation $(r_1, \dots, r_n)$. Then the corresponding differential equation* (2.1) *is locally homogeneous of the same degree $q \in \mathbf{R}$ with respect to the same dilation $(r_1, \dots, r_n)$.*

Proof. Let $x(\cdot)$ be a solution of (2.1), evolving within B_δ. Then it is straightforward to verify that due to (2.13), the function $x^c(\cdot)$ with components (2.12) is also a solution of (2.1) for all $c \geq c_0$. Thus, the differential equation (2.1), whose right-hand side is locally homogeneous in the sense of Definition 2.10, is also locally homogeneous in the sense of Definition 2.9. Lemma 2.11 is proved. □

To this end, we present conditions for the uncertain system (2.3), (2.4) to be locally homogeneous.

LEMMA 2.12. *Let the following conditions be satisfied:*

1. *a piece-wise continuous function φ is locally homogeneous of degree $q \in \mathbf{R}$ with respect to dilation $(r_1, \dots, r_n)$;*
2. *components ψ_i, $i = 1, \dots, n$, of a piece-wise continuous function ψ are locally uniformly bounded by constants $M_i \geq 0$;*
3. *$M_i = 0$ whenever $q + r_i > 0$.*

Then the uncertain differential equation (2.3) *with the uncertainty constraints* (2.4) *is locally homogeneous of degree $q \in \mathbf{R}$ with respect to dilation $(r_1, \dots, r_n)$.*

Proof. Let $x(\cdot) = (x_1(\cdot), \dots, x_n(\cdot))^T$ be a solution of (2.3) under some piece-wise continuous function ψ, satisfying (2.4), and let $x(\cdot)$ evolve within a ball B_δ where the homogeneity condition (2.13) holds almost everywhere for all $c \geq c_0$. Then it is straightforward to verify that for arbitrary $c \geq \max(1, c_0)$ the function $x^c(\cdot)$ with components $x_i^c(t) = c^{r_i} x_i(c^q t)$, $i = 1, \dots, n$, is a solution of (2.3) with the piece-wise continuous function $\psi = \psi^c$ whose components are as follows:

$$\psi_i^c(x, t) = c^{q+r_i} \psi_i(c^{-r_1} x_1, \dots, c^{-r_n} x_n, c^q t).$$

Since by condition 3 of the lemma one has $c^{q+r_i} \leq 1$ whenever $M_i > 0$, the function ψ^c is also admissible in the sense of (2.4).

Thus, any solution of the uncertain differential equation (2.3), evolving within a homogeneity ball B_δ, generates a parameterized set of solutions $x^c(t)$ with the parameter c large enough. Hence, (2.3), (2.4) is locally homogeneous of degree $q \in \mathbf{R}$ with respect to dilation $(r_1, \dots, r_n)$. Lemma 2.12 is proved. □

3. Finite time stability of homogeneous systems. Finite time stability of continuous autonomous systems has recently been studied in [7]. When continuous, a globally homogeneous time-invariant vector field $\varphi(x)$ of degree $q < 0$ with respect to dilation $(r_1, \dots, r_n)$ is known to be globally finite time stable whenever it is globally asymptotically stable. The proof of this fact, given in [16], is based on the result from [27] that an autonomous continuous homogeneous system, if asymptotically stable, possesses a homogeneous Lyapunov function. However, the existence of a homogeneous Lyapunov function is no longer guaranteed for systems governed by differential inclusions (even the converse of Lyapunov's second theorem has not been successfully extended to this case). Therefore, the global finite time stability of these systems is established by going through a different route, which is closely related to rescaling of the time and state variables.

Going through this route allows one to additionally obtain an upper estimate

$$T(t_0, x^0) \le \tau(x^0, E_R) + \frac{1}{1-2^q}(\delta R^{-1})^q s(\delta) \tag{3.1}$$

of the settling-time function (2.8) via the reaching-time functions

$$\tau(x^0, E_R) = \sup_{x(\cdot, t_0, x^0)} \inf\{T \ge 0 : x(t, t_0, x^0) \in E_R \text{ for all } t_0 \in \mathbf{R},\ t \ge t_0 + T\} \tag{3.2}$$

and

$$s(\delta) = \sup_{x^0 \in E_\delta} \tau(x^0, E_{\frac{1}{2}\delta}), \tag{3.3}$$

where E_R denotes an ellipsoid of the form

$$E_R = \left\{ x \in \mathbf{R}^n : \sqrt{\sum_{i=1}^{n} \left(\frac{x_i}{R^{r_i}}\right)^2} \le 1 \right\}, \tag{3.4}$$

E_R is located within a homogeneity ball, $\delta \ge c_0 R$, and $c_0 > 0$ is a lower estimate of the homogeneity parameter.

THEOREM 3.1. *Let the differential inclusion* (2.2) *be locally homogeneous of degree* $q < 0$ *with respect to dilation* $(r_1, \dots, r_n)$ *and let the equilibrium point* $x = 0$ *of* (2.2) *be globally uniformly asymptotically stable. Then the differential inclusion* (2.2) *is globally uniformly finite time stable, and an upper estimate of the settling-time function* (2.8) *is given by* (3.1).

Proof. Due to the global uniform asymptotic stability of (2.2), all the trajectories of the differential inclusion, initialized within a compact set, are uniformly steered toward an arbitrarily small ellipsoid (3.4). Then the condition

$$x(t) \in E_R \quad \text{for } t \ge t_0 + \tau(x^0, E_R) \tag{3.5}$$

holds for an arbitrary solution $x(\cdot)$ of (2.2) initialized with $x(t_0) = x^0$, where the ellipsoid E_R has been assumed to be small enough to be located within a homogeneity ball.

Furthermore, given an a priori fixed $\delta \ge c_0 R$, where $c_0 > 0$ is a lower estimate of the homogeneity parameter, there exists $s(\delta) > 0$ such that for each initial time moment $\tilde{t}_0$ and all the solutions $\tilde{x}(\cdot)$ with $\tilde{x}(\tilde{t}_0) \in E_\delta$ one has $\tilde{x}(t) \in E_{\frac{1}{2}\delta}$ for $t \ge \tilde{t}_0 + s(\delta)$. Since the function $x^c(\cdot)$, whose components (2.12) are specified with $c = \delta R^{-1} \ge c_0$, is a solution of (2.2) by homogeneity, and, in addition, $x^c(\tilde{t}_0) \in E_\delta$ at $\tilde{t}_0 = c^{-q}(t_0 + \tau(x^0, E_R))$, it follows that

$$x^c(t) \in E_{\frac{1}{2}\delta} \quad \text{for } t \ge c^{-q}(t_0 + \tau(x^0, E_R)) + s(\delta). \tag{3.6}$$

The latter relation, rewritten in terms of $x(t)$ by means of (2.12) subject to $c = \delta R^{-1}$, is represented as follows:

$$x(t) \in E_{\frac{1}{2}R} \quad \text{for } t \ge t_1 = t_0 + \tau(x^0, E_R) + (\delta R^{-1})^q s(\delta). \tag{3.7}$$

Now, by applying the same derivation to a solution of (2.2) with $x(t_1) \in E_{\frac{1}{2}R}$, one obtains

$$x(t) \in E_{\frac{1}{4}R} \quad \text{for } t \ge t_2 = t_1 + 2^q(\delta R^{-1})^q s(\delta). \tag{3.8}$$

In general, the following relations are derived:

$$x(t) \in E_{2^{-(i+1)}R} \quad \text{for } t \geq t_{i+1} = t_i + 2^{qi}(\delta R^{-1})^q s(\delta),\ i = 1, 2, \dots, \tag{3.9}$$

by iterating on i. Since $\lambda = 2^q < 1$ by virtue of $q < 0$, the convergence of the time instants t_k, $k = 1, 2, \dots$, to a finite limit takes place:

$$\begin{aligned} \lim_{k\to\infty} t_k &= t_0 + \tau(x^0, E_R) + \lim_{k\to\infty} \sum_{i=0}^{k-1} \lambda^i (\delta R^{-1})^q s(\delta) \\ &= t_0 + \tau(x^0, E_R) + \lim_{k\to\infty} \frac{1-\lambda^k}{1-\lambda} (\delta R^{-1})^q s(\delta) \\ &= t_0 + \tau(x^0, E_R) + \frac{1}{1-\lambda} (\delta R^{-1})^q s(\delta) < \infty. \end{aligned} \tag{3.10}$$

Hence, relations (3.9) result in

$$x(t) \in \bigcap_{i=1}^{\infty} E_{2^{-i}R} = \{0\} \quad \text{for } t \geq t_0 + \tau(x^0, E_R) + \frac{1}{1-2^q} (\delta R^{-1})^q s(\delta), \tag{3.11}$$

thereby establishing both the required finite time convergence property for the locally homogeneous differential inclusion (2.2) and the upper estimate (3.1) of the settling-time function (2.8). Theorem 3.1 is thus proved. □

Remark 1. For a globally homogeneous differential inclusion (2.2) one can choose $\delta_0 = c_0 R_0$ and R_0 sufficiently large to guarantee that $x(t, t_0, x^0) \in E_{R_0}$ for all $t \geq t_0$. Then $\tau(x^0, E_{R_0}) = 0$, and the upper estimate (3.1) of the settling-time function (2.8) is simplified to

$$T(t_0, x^0) \leq \frac{c_0^q}{1-2^q} s(\delta_0) \tag{3.12}$$

with $\delta_0 = \delta_0(x^0)$ dependent on x_0.

An important corollary of Theorem 3.1 is obtained if the differential inclusion (2.2) is generated by an uncertain differential equation (2.3) with piece-wise continuous functions φ and ψ, which are locally homogeneous and locally uniformly bounded, respectively.

THEOREM 3.2. *Let the following conditions be satisfied:*

1. *the right-hand side of an uncertain differential equation* (2.3) *consists of a locally homogeneous piece-wise continuous function φ of degree $q < 0$ with respect to dilation $(r_1, \dots, r_n)$ and a piece-wise continuous function ψ whose components ψ_i, $i = 1, \dots, n$, are locally uniformly bounded by constants $M_i \geq 0$ within a homogeneity ball;*
2. *$M_i = 0$ whenever $q + r_i > 0$;*
3. *the uncertain system* (2.3), (2.4) *is globally equiuniformly asymptotically stable around the origin.*

Then the uncertain system (2.3), (2.4) *is globally equiuniformly finite time stable and its settling-time function* (2.11) *is estimated as* (3.1).

Proof. By Lemma 2.12, conditions 1 and 2 of the theorem guarantee that the uncertain system (2.3), (2.4) is locally homogeneous of degree $q < 0$ with respect to dilation $(r_1, \dots, r_n)$. Thus, coupled to condition 3 of the theorem, these conditions ensure that Theorem 3.1 is applicable to the uncertain system (2.3), (2.4). By applying Theorem 3.1 to this system, the proof of Theorem 3.2 is completed. □

Apparently, Remark 1 remains valid for an uncertain system (2.3), (2.4), the right-hand side of which consists of a globally homogeneous piece-wise continuous function φ of degree $q < 0$ and a globally uniformly bounded piece-wise continuous function ψ.

4. Stability analysis of uncertain second order systems. The aim of this section is to illustrate, by means of a simple example, that the finite time stability of an autonomous switched system remains in force regardless of some nonlinear time-varying perturbations.

A second order switched system of the form

$$\dot{x} = y, \quad \dot{y} = -a \operatorname{sign} x - b \operatorname{sign} y \tag{4.1}$$

is presently under study. Due to the discontinuous functions $\operatorname{sign} x$ and $\operatorname{sign} y$ that appear in the right-hand side of (4.1), solutions of this system are defined in the Filippov sense. By Definition 2.1, these solutions do not depend on how the discontinuous functions are specified on the switching lines, which is why there is no need to specify the function sign at 0.

System (4.1) turns out to be globally uniformly finite time stable if the inequalities

$$a > b > 0 \tag{4.2}$$

hold for the parameters of the system.

THEOREM 4.1. *Let the parameters of the switched system* (4.1) *be such that condition* (4.2) *is satisfied. Then system* (4.1) *is globally uniformly finite time stable around the origin.*

Proof. First, let us note that no motion appears on the axes $x = 0$ and $y = 0$ except the origin $x = y = 0$, which proves to be the only equilibrium point of the switched system (4.1). Indeed, if $x(t) = 0$ on a trajectory of (4.1), then it follows from the first equation of (4.1) that $y(t) = 0$ along the trajectory. In turn, if $y(t) = 0$ on a trajectory of (4.1), then due to the parameter subordination (4.2), the second equation of (4.1) fails to hold for $x \neq 0$.

Next, let us prove that system (4.1) is globally uniformly asymptotically stable. For this purpose, let us introduce the function $V(x, y) = a|x| + \frac{1}{2}y^2$, which is Lipschitz continuous, radially unbounded, and positive definite. The time derivative $\dot{V}(x(t), y(t))$ of the composite function $V(x(t), y(t))$, computed along the trajectories of system (4.1), is as follows:

$$\dot{V}(x(t), y(t)) = -b|y(t)| \tag{4.3}$$

everywhere but on the vertical axis y where $x = 0$ and the function $V(x, y)$ is not differentiable. Since no sliding motion appears on the vertical axis except the origin $x = y = 0$, where $\dot{V}(x(t), y(t)) = 0$, relation (4.3) remains in force for almost all t.

As mentioned before, the trajectories of (4.1) cross the switching lines $x = 0$ and $y = 0$ everywhere except the origin $x = y = 0$ so that all the system trajectories are uniquely continuable on the right. Hence, the extended version [1, 28] of Krasovskii–LaSalle's invariance principle [17, 18, 19] is applicable to the switched system (4.1). Since the equilibrium point $x = y = 0$ is the only trajectory of (4.1) on the invariance manifold $y = 0$ where $\dot{V}(x(t), y(t)) = 0$, this system is globally uniformly asymptotically stable by the aforementioned extension of the invariance principle.

Moreover, it is straightforwardly verified that the right-hand side of system (4.1) is globally homogeneous of degree $q = -1$ with respect to dilation $(2, 1)$. By applying Theorem 3.1 it follows that the autonomous switched system (4.1) is globally uniformly finite time stable around the origin. Theorem 4.1 is proved. □

Remark 2. It is of interest to note that under the parameter subordination

$$b \geq a > 0, \tag{4.4}$$

opposite to (4.2), system (4.1) is not even asymptotically stable. Indeed, the velocity vectors in the case of (4.4) are normal to and directed toward each other while approaching the horizontal axis x from different sides. By Definition 2.1, it follows that the trajectories of (4.1) cannot leave this axis. Thus, the horizontal axis consists of equilibrium points of system (4.1), which is why (4.1) is not asymptotically stable.

Along with system (4.1), we shall also study its nonlinear time-varying perturbation of the form

$$\dot{x} = y, \quad \dot{y} = -a \operatorname{sign} x - b \operatorname{sign} y - hx - py + \omega(x, y, t), \tag{4.5}$$

where h and p are parameters of the linear gain, and $\omega(x, y, t)$ is a piece-wise continuous nonlinear perturbation, uniformly bounded

$$|\omega(x, y, t)| \leq M \tag{4.6}$$

for all continuity points (x, y, t) and some $M > 0$. If the bound M is small enough, namely,

$$0 < M < b < a - M, \tag{4.7}$$

and the linear gain is nonpositive, i.e.,

$$h \geq 0, \qquad p \geq 0, \tag{4.8}$$

the uncertain system (4.5), (4.6) proves to be globally equiuniformly asymptotically stable. The global equiuniform finite time stability of this system can then be demonstrated by invoking Theorem 3.2, which turns out to be applicable to the uncertain system (4.5), (4.6) because its nominal model (4.1) has a globally homogeneous right-hand side of degree $q = -1$ with respect to dilation $r = (2, 1)$, thus satisfying condition 2 of Theorem 3.2. Summarizing, the following result is in force.

THEOREM 4.2. *Let conditions* (4.7), (4.8) *be satisfied. Then the uncertain switched system* (4.5), (4.6) *is globally equiuniformly finite time stable around the origin.*

The qualitative behavior of system (4.5) is depicted in Figure 1. Due to the parameter subordination (4.7), the velocity vectors of (4.5) point toward the same region in the switching lines

$$\begin{aligned} S_1 &= \{(x, y) \in \mathbf{R}^2 : x > 0,\ y = 0\}, \\ S_2 &= \{(x, y) \in \mathbf{R}^2 : x = 0,\ y < 0\}, \\ S_3 &= \{(x, y) \in \mathbf{R}^2 : x < 0,\ y = 0\}, \\ S_4 &= \{(x, y) \in \mathbf{R}^2 : x = 0,\ y > 0\}, \end{aligned} \tag{4.9}$$

regardless of uncertainty (4.6) affecting the system. Hence, the uncertain system (4.5)–(4.8) and, particularly, its unperturbed version (4.1) rotate around the origin

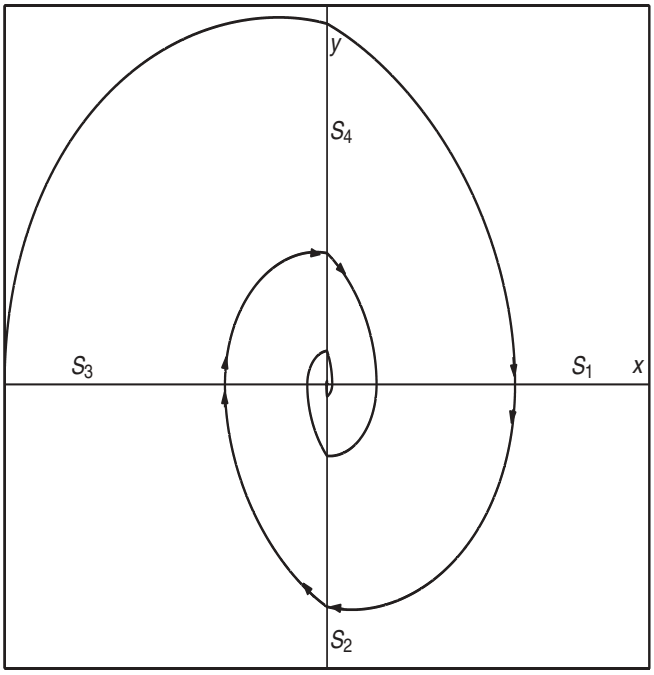


FIG. 1. *Phase portrait of the second order switched system* (4.5).

$x = y = 0$, while approaching the origin in a finite time. Thus, both the nominal system (4.1) and its uncertain counterpart (4.5)–(4.8) exhibit chattering modes with an infinite number of switches on a finite time interval. These systems do not generate sliding motions anywhere except the origin. If a trajectory starts there at any given finite time, there appears the so-called sliding mode of the second order (see [4, 11, 12, 20, 21] for advanced results on second order sliding modes).

4.1. Proof of Theorem 4.2. We break the proof into several simple steps.

1. *Equiuniform stability.* To demonstrate that the uncertain system (4.5)–(4.8) is equiuniformly stable we introduce the Lyapunov function

$$\tilde{V}(x,y) = a|x| + \frac{1}{2}(y^2 + hx^2). \tag{4.10}$$

Similar to (4.3), the time derivative of $\tilde{V}(x,y)$ along the trajectories $(x_\omega(\cdot), y_\omega(\cdot))$ of the uncertain system is negative semidefinite:

$$\begin{aligned} \dot{\tilde{V}}(x_\omega(t), y_\omega(t)) &= -b|y_\omega(t)| - py_\omega^2(t) + y_\omega(t)\omega(x_\omega(t), y_\omega(t), t) \\ &\le -(b - M)|y_\omega(t)| \end{aligned} \tag{4.11}$$

for almost all t. It follows that the uncertain system (4.5)–(4.8) is equiuniformly stable. Indeed, due to (4.11) the positive definite function (4.10) does not increase along the solutions of (4.5)–(4.8). Thus, initialized in an arbitrarily small vicinity

$$D_\epsilon = \{(x,y) \in \mathbf{R}^2 : \tilde{V}(x,y) \le \epsilon\} \tag{4.12}$$

of the origin, the uncertain system (4.5)–(4.8) cannot leave this vicinity, regardless of which admissible uncertainty ω affects the system.

2. *Semiglobal Lyapunov functions.* Our next goal is to construct a parameterized family of local Lyapunov functions $V_R(x,y)$, $R>0$, such that each $V_R(x,y)$ is well-posed on the corresponding compact set

$$D_R = \{(x,y) \in \mathbf{R}^2 : \tilde{V}(x,y) \le R\}. \tag{4.13}$$

In other words, $V_R(x,y)$ is to be positive definite on D_R whereas its time derivative, computed along the trajectories of the uncertain system (4.5)–(4.8) with initial conditions within D_R, is to be equiuniformly negative definite in the sense that

$$\dot{V}_R(x,y) \le -W_R(x,y) \tag{4.14}$$

for all $(x,y) \in D_R$ and some $W_R(x,y)$, positive definite on D_R.

Parameterized Lyapunov functions $V_R(x,y)$, $R>0$, with the properties above are constructed by combining the aforegiven Lyapunov function (4.10), whose time derivative along the system motion is only semidefinite, with the indefinite function $U(x,y)=xy$:

$$V_R(x,y) = \tilde{V}(x,y) + \kappa_R U(x,y) = a|x| + \frac{1}{2}(y^2+hx^2) + \kappa_R xy, \tag{4.15}$$

where the weight parameter $\kappa_R > 0$ is chosen small enough, namely,

$$\kappa_R < \min\left\{1, \frac{2a^2}{R}, \frac{a(b-M)}{a\sqrt{2R}+pR}\right\}. \tag{4.16}$$

Apparently, the function $V_R(x,y)$, thus constructed, is positive definite on compacta (4.13) because (4.13) implies that $|x| \le \frac{R}{a}$, $|y| \le \sqrt{2R}$, and therefore

$$\begin{aligned} a|x| + \frac{1}{2}(y^2+hx^2) + \kappa_R xy &\ge a|x| + \frac{1}{2}(y^2+hx^2) - \frac{1}{2}\kappa_R x^2 - \frac{1}{2}\kappa_R y^2 \\ &\ge \left(a - \frac{R}{2a}\kappa_R\right)|x| + \frac{1}{2}(1-\kappa_R)y^2 + \frac{1}{2}hx^2 > 0 \end{aligned} \tag{4.17}$$

for all $(x,y) \in D_R \setminus \{(0,0)\}$ and $\kappa_R > 0$, satisfying (4.16).

Moreover, the time derivative of (4.15), computed along the trajectories of the uncertain system (4.5)–(4.8) with the initial conditions within D_R, is equiuniformly negative definite. Indeed, differentiating the auxiliary function $U(x,y)=xy$ along the trajectories of the uncertain system (4.5)–(4.8) yields

$$\begin{aligned} \dot{U}(x_\omega(t), y_\omega(t)) &= y_\omega^2(t) - a|x_\omega(t)| - bx_\omega(t)\operatorname{sign} y_\omega(t) - hx_\omega^2(t) \\ &\quad - px_\omega(t)y_\omega(t) + x_\omega(t)\omega(x_\omega(t), y_\omega(t), t) \\ &\le y_\omega^2(t) - |x_\omega(t)|[a + b\operatorname{sign} x_\omega(t)\operatorname{sign} y_\omega(t) \\ &\qquad - \omega(x_\omega(t), y_\omega(t), t)\operatorname{sign} x_\omega(t)] - px_\omega(t)y_\omega(t) \\ &\le y_\omega^2(t) - (a-b-M)|x_\omega(t)| - px_\omega(t)y_\omega(t). \end{aligned} \tag{4.18}$$

Then, by employing (4.11) and (4.18) one has

$$\begin{aligned} \dot{V}_R(x_\omega(t), y_\omega(t)) &\le -(b-M)|y_\omega(t)| \\ &\quad + \kappa_R y_\omega^2(t) - \kappa_R(a-b-M)|x_\omega(t)| - \kappa_R p x_\omega(t)y_\omega(t). \end{aligned} \tag{4.19}$$

Since due to (4.11) all possible solutions of (4.5)–(4.8), initialized at $t_0 \in \mathbf{R}$ within the compact set (4.13), are a priori estimated by

$$\sup_{t\in[t_0,\infty)} \tilde{V}(x_\omega(t), y_\omega(t)) \le R, \tag{4.20}$$

it follows that

$$\begin{aligned}\dot{V}_R(x_\omega(t), y_\omega(t)) \le &-\left[b - M - \kappa_R\left(\sqrt{2R} + \frac{pR}{a}\right)\right]|y_\omega(t)| \\ &-\kappa_R(a - b - M)|x_\omega(t)| \le -c_R[|y_\omega(t)| + |x_\omega(t)|],\end{aligned} \tag{4.21}$$

where $c_R = \min\{b - M - \kappa_R(\sqrt{2R} + \frac{pR}{a}),\ \kappa_R(a - b - M)\} > 0$ by virtue of (4.16).

To this end, (4.21) results in

$$\dot{V}_R(x_\omega(t), y_\omega(t)) \le -K_R V_R(x_\omega(t), y_\omega(t)), \tag{4.22}$$

where

$$K_R = \frac{2ac_R}{\max\{2a^2 + hR,\ a\sqrt{2R} + 2\kappa_R R\}} > 0$$

and the upper estimate

$$V_R(x, y) \le \frac{2a^2 + hR}{2a}|x| + \frac{a\sqrt{2R} + 2\kappa_R R}{2a}|y|$$

of the Lyapunov function (4.15) on compacta (4.13) has been used. Thus, the desired equiuniform negative definiteness (4.14) is obtained with $W_R(x, y) = K_R V_R(x, y)$.

3. *Global equiuniform asymptotic stability.* Since the differential inequality (4.22) holds on the solutions of the uncertain system (4.5)–(4.8), initialized within the compact set (4.13), the function $V_R(x_\omega(t), y_\omega(t))$ exponentially decays

$$V_R(x_\omega(t), y_\omega(t)) \le V_R(x_\omega(t_0), y_\omega(t_0))e^{-K_R(t-t_0)} \tag{4.23}$$

on these solutions with the decay rate K_R, independent of the uncertainty ω. While being viewed on compacta (4.13), the functions $V_R(x, y)$ and $\tilde{V}(x, y)$ are equivalent in the sense that

$$L_R\tilde{V}(x, y) \le V_R(x, y) \le M_R\tilde{V}(x, y) \tag{4.24}$$

for all $(x, y) \in D_R$ and positive constants L_R, M_R, satisfying

$$L_R < \min\left\{\frac{2a^2 - R\kappa_R}{2a^2}, 1 - \kappa_R\right\}, \qquad M_R > \max\left\{\frac{2a^2 + R\kappa_R}{2a^2}, 1 + \kappa_R\right\}. \tag{4.25}$$

The above relations (4.23) and (4.24), coupled together, ensure that the function $\tilde{V}(x, y)$ exponentially decays

$$\begin{aligned}\tilde{V}(x_\omega(t), y_\omega(t)) &\le L_R^{-1}M_R\tilde{V}(x_\omega(t_0), y_\omega(t_0))e^{-K_R(t-t_0)} \\ &\le L_R^{-1}M_R R e^{-K_R(t-t_0)}\end{aligned} \tag{4.26}$$

on the solutions of (4.5)–(4.8), equiuniformly in the uncertainty ω and the initial data, located within an arbitrarily large set (4.13). Clearly, this proves that the uncertain system (4.5)–(4.8) is globally equiuniformly asymptotically stable.

4. *Global equiuniform finite time stability.* Due to (4.6) the piece-wise continuous uncertainty $\omega(x,y,t) - hx - py$ is locally uniformly bounded, whereas the right-hand side of the nominal model (4.1) is piece-wise continuous and globally homogeneous of degree $q = -1$ with respect to dilation $r = (2,1)$. Hence, the condition $q + r_2 \le 0$, required by Theorem 3.2, is satisfied, and Theorem 3.2 is applicable to the globally equiuniformly asymptotically stable uncertain system (4.5)–(4.8). By applying Theorem 3.2, the uncertain system (4.5)–(4.8) is thus globally equiuniformly finite time stable. The proof of Theorem 4.2 is completed.

5. Chattering control synthesis. The present section investigates the capability of a nonlinear system of the form

$$\begin{aligned} \dot z &= g(z,\xi,\dot\xi,t),\\ \ddot\xi &= f(z,\xi,\dot\xi,t) + u \end{aligned} \tag{5.1}$$

to be globally asymptotically stabilizable in spite of significant model uncertainties with a priori known norm bounds. Hereafter, $t \in \mathbf{R}$ is the time variable, $z \in \mathbf{R}^n$, $\xi, \dot\xi \in \mathbf{R}^m$ are the state components, $u \in \mathbf{R}^m$ is the input, $\xi \in \mathbf{R}^m$ is the output, and the nonlinear functions g and f have appropriate dimensions and involve the system uncertainties whose influence on the control process should be rejected.

The above system often arises in practice, e.g., to describe controlled electromechanical plants. It can be obtained via a nonlinear change of state coordinates and a feedback transformation from a general affine control system

$$\begin{aligned} &\dot x = a(x,t) + b(x,t)u, \quad x \in \mathbf{R}^{n+2m},\ u \in \mathbf{R}^m,\\ &y = h(x,t), \quad y \in \mathbf{R}^m, \end{aligned} \tag{5.2}$$

of uniform m-vector relative degree $(2,\dots,2)^T$ with the involutive distribution

$$B = \operatorname{span}\{b_1,\dots,b_m\},$$

the span of the columns of $b(x)$ (see [9] for details).

The following assumptions on system (5.1) are made throughout.

1. The functions $g(z,\xi,\dot\xi,t)$ and $f(z,\xi,\dot\xi,t)$ are piece-wise continuous in all the arguments, and, in addition, the function $g(z,\xi,\dot\xi,t)$ is continuous in $(\xi,\dot\xi)$ locally around $(\xi,\dot\xi) = 0$ for almost all z and t.

2. The function $g(z,\xi,\dot\xi,t)$ satisfies the linear growth condition

$$\|g(z,\xi,\dot\xi,t)\| \le k(\xi,\dot\xi,t)(1 + \|z\|) \tag{5.3}$$

in z everywhere in its domain (i.e., (5.3) can only be violated on the set where g undergoes discontinuities) with some continuous function $k(\xi,\dot\xi,t)$.

3. The system

$$\dot z = g(z,0,0,t) \tag{5.4}$$

has 0 as a globally uniformly asymptotically stable equilibrium.

Solutions of the state and zero dynamics differential equations (5.1) and (5.4), both with piece-wise continuous right-hand sides, are defined in the sense of Filippov according to Definition 2.1. Under assumption 1 on (5.1) the existence of a solution (possibly nonunique) of either equation with an arbitrary initial condition is

guaranteed by Theorem 8 of [10, p. 85]. Other assumptions are made for technical reasons. Assumption 2 is introduced to avoid the destabilizing effect of the peaking phenomenon (a detailed treatment of the peaking phenomenon in the continuous setting can be found in [29]). Assumption 3 means that (5.1) is a globally uniformly minimum phase system. The role of this notion is well known from the theory of smooth fields [9] and it is now under study for switched nonautonomous systems.

System (5.1) is operating under uncertainty conditions that imply imperfect knowledge of the nonlinearities f and g. The nonlinear gain g cannot destabilize the closed-loop system because of the minimum phase hypothesis, which is why no more information is required for this gain. The destabilizing term

$$f(z,\xi,\dot\xi,t) = f^{nom}(z,\xi,\dot\xi,t) + f^b(z,\xi,\dot\xi,t) \tag{5.5}$$

typically consists of a nominal part f^{nom} to be handled through nonlinear damping and an uncertain bounded gain f^b to be rejected. It is assumed that the nominal part f^{nom} is known a priori, whereas the components f^b_j, $j = 1,\dots,m$, of the bounded gain f^b are upper estimated:

$$|f^b_j(z,\xi,\dot\xi,t)| \le F_j < \infty \quad \text{for almost all } (z,\xi,\dot\xi,t) \in \mathbf{R}^{n+2m+1} \tag{5.6}$$

by constants F_j, also known a priori. Apart from this, both functions f^{nom} and f^b are assumed to be piece-wise continuous.

The following switched control law

$$u(z,\xi,\dot\xi,t) = -f^{nom}(z,\xi,\dot\xi,t) - \mu\dot\xi - \nu\xi - \beta\operatorname{sign}\xi - \gamma\operatorname{sign}\dot\xi, \tag{5.7}$$

with the parameter gains

$$\mu = \operatorname{diag}(\mu_j), \qquad \nu = \operatorname{diag}(\nu_j), \qquad \beta = \operatorname{diag}(\beta_j), \qquad \gamma = \operatorname{diag}(\gamma_j)$$

subject to

$$\mu_j \ge 0, \quad \nu_j \ge 0, \quad \beta_j - F_j > \gamma_j > F_j, \quad j = 1,\dots,m, \tag{5.8}$$

is proposed to stabilize the uncertain system (5.1), (5.5), (5.6) whose state $(z,\xi,\dot\xi)$ is available for measurements. Hereafter, the notation diag is used to denote a diagonal matrix of an appropriate dimension; $\operatorname{sign}\xi$ with a vector $\xi = (\xi_1,\dots,\xi_m)^T$ stands for the column vector $(\operatorname{sign}\xi_1,\dots,\operatorname{sign}\xi_m)^T$.

In what follows, the switched control law (5.7), (5.8) is proven to exhibit chattering modes, while driving the uncertain system (5.1) to the zero dynamics manifold $\xi = \dot\xi = 0$ in finite time. Desired stability properties are thus imposed on the closed-loop system.

THEOREM 5.1. *Let assumptions* 1–3 *be satisfied and let the uncertain system* (5.1), (5.5), (5.6) *be driven by the state feedback* (5.7) *such that condition* (5.8) *holds. Then the closed-loop system* (5.1), (5.5)–(5.8) *is globally equiuniformly asymptotically stable.*

Proof. The closed-loop system (5.1) driven by (5.7) is represented as follows:

$$\dot z = g(z,\xi,\zeta,t), \tag{5.9}$$

$$\dot\xi_j = \zeta_j, \qquad \dot\zeta_j = f^b_j(z,\xi,\zeta,t) - \nu_j\xi_j - \mu_j\zeta_j - \beta_j\operatorname{sign}\xi_j - \gamma_j\operatorname{sign}\zeta_j, \quad j = 1,\dots,m. \tag{5.10}$$

Due to assumption 1, Theorem 8 of [10, p. 85] is applicable to system (5.9), (5.10), and by applying this theorem, the system has a local solution for all initial data and uncertainties (5.6). Let us demonstrate that each solution of (5.9), (5.10) is globally continuable on the right.

Similarly to (4.11), the time derivative of the function $V_j(\xi_j, \zeta_j) = \beta_j|\xi_j| + \frac{1}{2}(\zeta_j^2 + \nu_j\xi_j^2)$, $j = 1, \dots, m$, computed along the trajectories of the corresponding subsystem (5.10), is negative semidefinite:

$$\dot{V}_j(\xi_j(t), \zeta_j(t)) \le -(\gamma_j - F_j)|\zeta_j(t)|. \tag{5.11}$$

As in the proof of Theorem 4.1, it follows that each solution of subsystem (5.10) subject to (5.6) is uniformly bounded in t.

In turn, for given continuous, uniformly bounded functions $\xi(t)$, $\zeta(t)$, all possible solutions of subsystem (5.9) remain bounded on any finite time interval due to the linear growth assumption (5.3) (assumption 2) on the right-hand side of (5.9). Indeed, an arbitrary solution $z(\cdot)$ of (5.9) is a priori estimated as

$$\|z(t)\| \le \|z(t_0)\| + \int_{t_0}^{t} k(\xi(t), \zeta(t), t)(1 + \|z(t)\|)dt, \tag{5.12}$$

and by applying Bellman–Gronwall's lemma to the integral inequality (5.12), $z(\cdot)$ is bounded on any finite time interval (t_0, t_1).

Thus, all possible solutions of the overall uncertain system (5.6), (5.9), (5.10) remain bounded on any finite time interval, and by property B of Theorem 9 of [10, p. 86], these solutions are globally continuable on the right.

Next let us observe that due to the uniform boundedness (5.6) of the uncertain terms $f_j^b(z, \xi, \zeta, t)$, $j = 1, \dots, m$, and by virtue of the parameter subordination (5.8), each subsystem (5.10) satisfies all the conditions of Theorem 4.2, and by applying this theorem, (5.10) is globally equiuniformly finite time stable. It follows that starting from a finite time moment, the overall uncertain system (5.6), (5.9), (5.10) evolves in the chattering mode on the zero dynamics manifold $\xi = \zeta = 0$, where its behavior is governed by the zero dynamics equation (5.4).

Now, to complete the proof, it remains to note that by assumption 3, the zero dynamics (5.4) is globally uniformly asymptotically stable. Coupled to the global equiuniform finite time stability of (5.10), this ensures that the closed-loop system (5.9), (5.10) is globally equiuniformly asymptotically stable, too. The proof of Theorem 5.1 is thus completed. □

5.1. Application to friction electromechanical systems. A simple physical interpretation of the chattering controller (5.7) is revealed by its application to the servo-motor governed by

$$j\ddot{\xi} = u - F(\dot{\xi}) + \omega(t), \quad \xi, u \in \mathbf{R}^1. \tag{5.13}$$

In the above equation, ξ denotes displacement, u is the control input, j is the inertia, ω is the external disturbance involving the transmitted torque (as a matter of fact, this function can be state-dependent; however, for simplicity it is viewed as an integrable function of time, computed along the system trajectories), and $F(\dot{\xi})$ is the friction force governed by the classical model

$$F(\dot{\xi}) = \alpha_v\dot{\xi} + \alpha_c \operatorname{sign}\dot{\xi} \tag{5.14}$$

with the viscous friction coefficient $\alpha_v > 0$ and the Coulomb friction level $\alpha_c > 0$. Solutions of the state equation (5.13) are defined in the sense of Filippov, and by virtue of Definition 2.1, there is no need to specify the above friction model on the discontinuity manifold $\dot{\xi} = 0$.

Since the phenomenon of friction is not yet completely understood and it is hard to model, the uncertain term $\omega(t)$ has been incorporated into the motor equation (5.13) to account for destabilizing model discrepancies such as the Stribeck effect and backlash, whose upper bounds are computed experimentally (see, e.g., [2]). Thus, an upper bound $F_0 > 0$ for the magnitude of the uncertain term is normally known a priori:

$$|\omega(t)| \le F_0 \quad \text{for all } t. \tag{5.15}$$

Motivated by application requirements, the control objective is to stabilize the servo-motor in finite time in spite of the uniformly bounded external disturbances whose influence on the control process should be rejected.

It is worth noting that continuous control algorithms do not admit even asymptotic stabilization of (5.13), (5.14) due to the presence of the discontinuous term $\alpha_c \operatorname{sign} \dot{\xi}$, corresponding to the Coulomb friction in (5.14). Indeed, the servo-motor (5.13), (5.14) enforced by an arbitrary continuous controller u has a nontrivial set of equilibrium points around the position $\xi = 0$ (see, e.g., [25]), thereby yielding an inappropriate solution to the stabilization problem. Thus, continuous controllers, particularly those especially developed in [6, 16] to stabilize frictionless oscillators in finite time, become unable to asymptotically stabilize the servo-motor (5.13), (5.14) with a nonzero Coulomb level α_c.

Hence, in order to attain asymptotic stability of the servo-motor (5.13), (5.14) with a nontrivial Coulomb friction level, discontinuous controllers have to be brought into play. Since standard sliding mode controllers are capable only of providing ultimate boundedness of the second order dynamic system in question [24], the chattering control algorithm (5.7), specified for the servo-motor (5.13), (5.14) by

$$u = -\mu\dot{\xi} - \nu\xi - \beta \operatorname{sign} \xi - \gamma \operatorname{sign} \dot{\xi}, \tag{5.16}$$

is chosen to stabilize the servo-motor in finite time. Due to the presence of the Coulomb friction, the subordination rule (5.8) for the parameters of the closed-loop system (5.13), (5.14), (5.16) is modified as follows:

$$\mu, \nu \ge 0, \qquad \beta - F_0 > \alpha_c + \gamma > F_0, \tag{5.17}$$

where μ, ν, β, γ are the parameter gains, α_c is the Coulomb level of the friction force (5.14), and F_0 is the upper bound (5.15) for the external disturbances.

In the particular case of $\mu = \nu = 0$ and $\beta > \gamma > 0$, the above controller (5.16) is nothing more than the so-called twisting controller from [20]. Adding a linear part with nonzero μ and ν into this twisting algorithm allows one to reduce the convergence time of the controller without enlarging the amplitude of switching used in the controller. Furthermore, in contrast to the twisting algorithm the pure switched position feedback

$$u = -\beta \operatorname{sign} \xi \tag{5.18}$$

is admitted by the proposed control law (5.16) when $\mu = \nu = \gamma = 0$. The switched position controller (5.18) still stabilizes the servo-motor (5.13), (5.14) in finite time

while rejecting the external disturbance (5.15) only when the upper bound F_0 is smaller than the Coulomb friction level α_c.

The performance of the chattering controller (5.16) and that of its position feedback version (5.18) were studied in [26] by simulation. In this study, a servo-motor was required to move from an initial static position to the origin $\xi = 0$, $\dot{\xi} = 0$, while the external disturbance was composed by the backlash model from [30] and the Stribeck friction force from Tustin's model [31].

As predicted by the theory, both closed-loop systems were driven to the origin in finite time in spite of the presence of backlash and Stribeck effect whereas better performance (settling time) was yielded by the state feedback controller. Thus, to decrease the settling time one should utilize a chattering state feedback controller (5.16) and increase the parameter gains of the controller. This may, however, excite unmodelled high-frequency dynamics of the system, thereby limiting the achievable performance.

6. Conclusions. Finite time stability of uncertain switched systems has been studied. These systems were shown to be globally equiuniformly finite time stable whenever they are globally equiuniformly asymptotically stable and homogeneous of degree $q < 0$. Restricted to second order systems, the finite time stability feature was additionally demonstrated to remain in force regardless of some inhomogeneous perturbations.

Based on these tools, discontinuous control algorithms have been developed to globally asymptotically stabilize uncertain minimum phase systems of uniform m-vector relative degree $(2, \dots, 2)^T$. The stabilizing strategy was to drive the system to the zero dynamics manifold in a finite time and maintain it there in spite of the parameter uncertainties and external disturbances. Desired robustness properties were thus provided and asymptotic stability of the closed-loop system was guaranteed.

The controllers constructed do not rely on the generation of sliding motions on the switching manifolds but on their intersections, thus exhibiting the so-called second order sliding modes [11] with an infinite number of switches on a finite time interval. In analogy to optimal control theory [34], such controllers have been referred to as chattering controllers regardless of whether the optimal performance is achieved.

Attractive features of chattering controllers have been illustrated by application to a servo-motor. The controllers were demonstrated to be capable of providing the desired system performance in spite of significant uncertainties in the system description, as is typically the case in control of electromechanical systems with complex hard-to-model nonlinear phenomena.

Acknowledgment. The author wishes to thank the anonymous reviewers for their valuable comments on earlier versions of this manuscript.

REFERENCES

[1] J. Alvarez, I. Orlov, and L. Acho, *An invariance principle for discontinuous dynamic systems with applications to a Coulomb friction oscillator*, Trans. ASME J. Dynam. Systems Measurement Control, 122 (2000), pp. 687–690.

[2] B. Armstrong-Hélouvry, P. Dupont, and C. Canudas de Wit, *A survey of models, analysis tools and compensation methods for the control of machines with friction*, Automatica J. IFAC, 30 (1994), pp. 1083–1138.

[3] G. Bartolini, *Chattering phenomena in discontinuous control systems*, Internat. J. Systems Sci., 20 (1989), pp. 2471–2481.

[4] G. Bartolini, A. Ferrara, and E. Usai, *Chattering avoidance by second-order sliding mode control*, IEEE Trans. Automat. Control, 43 (1998), pp. 241–246.

[5] H. Berghuis and H. Nijmeijer, *Global regulation of robots using only position measurements*, Systems Control Lett., 21 (1993), pp. 289–293.
[6] S. P. Bhat and D. S. Bernstein, *Continuous finite-time stabilization of the translational and rotational double integrators*, IEEE Trans. Automat. Control, 43 (1998), pp. 678–682.
[7] S. P. Bhat and D. S. Bernstein, *Finite-time stability of continuous autonomous systems*, SIAM J. Control Optim., 38 (2000), pp. 751–766.
[8] M. S. Branicky, *Multiple Lyapunov functions and other analysis tools for switched and hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 475–482.
[9] C. I. Byrnes and A. Isidori, *Asymptotic stabilization of minimum phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1122–1137.
[10] A. F. Filippov, *Differential Equations with Discontinuous Right-Hand Sides*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.
[11] L. Fridman and A. Levant, *Higher order sliding modes as a natural phenomenon in control theory*, in Robust Control Via Variable Structure and Lyapunov Techniques, Lecture Notes in Control and Inform. Sci. 217, F. Garafalo and L. Glielmo, eds., Springer-Verlag, London, 1996, pp. 107–133.
[12] L. Fridman and A. Levant, *Higher order sliding modes*, in Sliding Mode Control in Engineering, W. Perruquetti and J.-P. Barbout, eds., Marcel Dekker, New York, 2002, pp. 53–102.
[13] A. T. Fuller, *Relay control systems optimized for various performance criteria*, in Proceedings of the First World Congress IFAC, Vol. 1, Moscow, 1960, Butterworth, London, 1961, pp. 510–519.
[14] V. T. Haimo, *Finite time controllers*, SIAM J. Control Optim., 24 (1986), pp. 760–770.
[15] W. Hahn, *Stability of Motion*, Springer-Verlag, Berlin, 1967.
[16] Y. Hong, J. Huang, and Y. Xu, *On an output feedback finite-time stabilization problem*, IEEE Trans. Automat. Control, 46 (2001), pp. 305–309.
[17] N. N. Krasovskii, *Certain Problems in the Theory of Stability of Motion*, Gosudarstv. Izdat. Fiz.-Mat. Lit., Moscow, 1959 (in Russian).
[18] N. N. Krasovskii, *Stability of Motion. Applications of Lyapunov's Second Method to Differential Systems and Equations with Delay*, Stanford University Press, Stanford, CA, 1963 (in English).
[19] J. P. LaSalle, *Some extensions of Lyapunov's second method*, IRE Trans. Circuit Theory, CT-7 (1960), pp. 520–527.
[20] A. Levant, *Sliding order and sliding accuracy in sliding mode control*, Internat. J. Control, 58 (1993), pp. 1247–1263.
[21] A. Levant, *Variable measurement step in* 2*-sliding control*, Kybernetika (Prague), 36 (2000), pp. 77–93.
[22] D. Liberzon, *Switching in Systems and Control*, Birkhäuser Boston, Boston, 2003.
[23] D. Liberzon and A. S. Morse, *Basic problems in stability and design of switched systems*, IEEE Control Systems, 19 (1999), pp. 59–70.
[24] X. Y. Lu and S. K. Spurgeon, *Robust sliding mode control of uncertain nonlinear systems*, Systems Control Lett., 32 (1997), pp. 75–90.
[25] Y. Orlov, *Extended invariance principle for nonautonomous switched systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 1448–1452.
[26] Y. Orlov, L. Aguilar, and J. C. Cadiou, *Switched chattering control vs. backlash/friction phenomena in electrical servo-motors*, Internat. J. Control, 76 (2003), pp. 959–967.
[27] L. Rosier, *Homogeneous Lyapunov function for homogeneous continuous vector field*, Systems Control Lett., 19 (1992), pp. 467–473.
[28] D. Shevitz and B. Paden, *Lyapunov stability theory of nonsmooth systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1910–1914.
[29] H. J. Sussman and P. V. Kokotovic, *The peaking phenomenon and the global stabilization of nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 424–440.
[30] G. Tao and P. V. Kokotovic, *Adaptive Control of Systems with Actuator and Sensor Nonlinearities*, Wiley, New York, 1996.
[31] A. Tustin, *The effects of backlash and speed-dependent friction on the stability of closed-cycle control system*, J. Institution of Electrical Engineers, 94 (1947), pp. 143–151.
[32] V. I. Utkin, *Sliding Modes in Control Optimization*, Springer-Verlag, Berlin, 1992.
[33] V. I. Utkin, J. Guldner, and J. Shi, *Sliding Modes in Electromechanical Systems*, Taylor and Francis, London, 1999.
[34] M. Zelikin and V. Borisov, *Theory of Chattering Control. With Applications to Astronautics, Robotics, Economics, and Engineering*, Birkhäuser Boston, Boston, 1994.
[35] V. I. Zubov, *Methods of A. M. Lyapunov and Their Applications*, P. Noordhoff, Groningen, The Netherlands, 1964.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1272–1297

QUENCHED LARGE DEVIATIONS FOR ONE DIMENSIONAL NONLINEAR FILTERING*

ÉTIENNE PARDOUX† AND OFER ZEITOUNI‡

Abstract. Consider the standard, one dimensional, nonlinear filtering problem for diffusion processes observed in small additive white noise: $dX_t = b(X_t)dt + dB_t$, $dY_t^\varepsilon = \gamma(X_t)dt + \varepsilon dV_t$, where $B.,V.$ are standard independent Brownian motions. Denote by $q_1^\varepsilon(\cdot)$ the density of the law of Ξ_1 conditioned on $\sigma(Y_t^\varepsilon : 0 \le t \le 1)$. We provide "quenched" large deviation estimates for the random family of measures $q_1^\varepsilon(x)dx$: there exists a continuous, explicit mapping $\bar{\mathcal{J}} : \mathbb{R}^2 \to \mathbb{R}$ such that for almost all $B.,V.$, $\bar{\mathcal{J}}(\cdot, X_1)$ is a good rate function, and for any measurable $G \subset \mathbb{R}$,

$$-\inf_{x\in G^o} \bar{\mathcal{J}}(x, X_1) \le \liminf_{\varepsilon\to 0} \varepsilon \log \int_G q_1^\varepsilon(x)dx \le \limsup_{\varepsilon\to 0} \varepsilon \log \int_G q_1^\varepsilon(x)dx \le -\inf_{x\in \bar{G}} \bar{\mathcal{J}}(x, X_1).$$

Key words. nonlinear filtering, large deviations

AMS subject classifications. 93E11, 60F10

DOI. 10.1137/S0363012903365032

1. Introduction and statement of results.

Consider the following one dimensional filtering problem, where the signal process $X.$ and the observation process $Y.^\varepsilon$, parametrized by a "small noise intensity" ε, are

$$\begin{aligned} dX_t &= b(X_t)dt + dB_t, \quad X_0 \sim p_0(\cdot), \\ dY_t^\varepsilon &= h(X_t)dt + \varepsilon dV_t. \end{aligned} \tag{1.1}$$

Here, $B.,V.$ are independent standard one dimensional Brownian motions, and the functions b, h, p_0 satisfy the following assumptions:[1]

(A-1) b, h, b', h' are Lipschitz functions,
(A-2) $h'(\cdot) \ge h_0 > 0$,
(A-3) $|\log p_0(x) - \log p_0(y)| \le c(1 + |x| + |y|)|x - y|$, $x, y \in \mathbb{R}$, and p_0 is uniformly bounded.

For technical reasons, we need to impose the following additional restriction:

(A-4) $h'b, h'h, h'', hb$ are Lipschitz functions and $\lim_{|x|\to\infty} h''(x) = 0$.

(A-4) implies that, outside large compacts, the observation function h is essentially linear. Let $\Omega_1 = \Omega_2 = C([0,1];\mathbb{R})$, $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F}_i$ be the Borel σ-algebra on Ω_i,

*Received by the editors June 1, 2003; accepted for publication (in revised form) June 3, 2004; published electronically January 5, 2005.
http://www.siam.org/journals/sicon/43-4/36503.html

†LATP, Université de Provence and CNRS, CMI, 39 rue Joliot Curie, 13453 Marseille Cedex 13, France (pardoux@cmi.univ-mrs.fr); member of the Institut Universitaire de France.

‡Departments of Electrical Engineering and of Mathematics, Technion, Haifa 32000, Israel, and Department of Mathematics, Vincent Hall, University of Minnesota, Minneapolis, MN 55455 (zeitouni@math.umn.edu). Part of this work was done while this author was visiting the LATP, University of Provence and CNRS, Marseille. This author was also supported by NSF grant DMS-0302230.

[1]Due to the one dimensional nature of our model, no generality is lost in assuming the diffusion coefficient of the signal process to be one. Indeed, if the signal process satisfies $d\Xi_t = \beta(\Xi_t)dt + \sigma(\Xi_t)dB_t$, with σ uniformly bounded away from zero, then the transformation $X_t = \bar{G}(\Xi_t)$, with $\bar{G}(x) = \int_0^x (1/\sigma)(u)du$, allows one to rewrite the problem in the form (1.1).

$i = 1, 2$, and $\mathcal{F}$ be the Borel σ-algebra on Ω; let P_1, P_2 denote the Wiener measure on Ω_1, Ω_2, and $P = P_1 \otimes P_2$. We define $B_t(\omega) = \omega_1(t)$, $V_t(\omega) = \omega_2(t)$, $0 \le t \le 1$. The pair (B, V) is then distributed according to P. The solution (X, Y^ε) of the SDE (1.1) is then an $\mathcal{F}$-measurable, $C([0,1]; \mathbb{R}^2)$-valued, random variable.

Let $\mu_t^\varepsilon(\cdot)$ denote the conditional law of X_t conditioned on $\mathcal{Y}_t^\varepsilon = \sigma\{Y_s^\varepsilon,\, 0 \le s \le t\}$, which we consider as an $\mathcal{F}$-measurable map from Ω to $M_1(\mathbb{R})$, the space of probability measures on $\mathbb{R}$. Note that μ_t^ε is in fact measurable with respect to the ε-dependent σ-algebra $\mathcal{Y}_t^\varepsilon \subset \mathcal{F}$.

It is known that μ_t^ε is absolutely continuous, with $\mu_t^\varepsilon(dx) = q_t^\varepsilon(x)dx$, and that as $\varepsilon \to 0$, the conditional law $\mu_1^\varepsilon(dx) = q_1^\varepsilon(x)dx$ of X_1, given $\mathcal{Y}_1^\varepsilon$, converges to the Dirac measure δ_{X_1}. (All these facts can be found, e.g., in [7].) In particular, X_1 is measurable with respect to the limiting σ-algebra $\mathcal{Y}_1^0$, since h is one-to-one. It is known from the results of Picard [7] that the conditional law μ_1^ε has a variance of order ε and can be well approximated by a Gaussian law, which is given by an extended Kalman filter.

Our goal in this paper is to establish a large deviations result in the following sense. Let G be a measurable subset of $\mathbb{R}$. By the above remarks, we know that on the event $\{X_1 \notin \overline{G}\}$, $\mu_1^\varepsilon(G) \to 0$, P-almost surely. It turns out that it goes to zero at exponential speed, i.e., roughly like $\exp[-c_1(G)/\varepsilon]$. What is the value of $c_1(G) = -\lim \varepsilon \log \mu_1^\varepsilon(G)$ (if this limit exists), the "rate function" that tells us at which speed the quantity $P(X_1 \in G \,|\, \mathcal{Y}_1^\varepsilon)$ goes to zero, whenever $X_1 \notin \overline{G}$? Clearly $c_1(G)$ must depend on X_1 (at least intuitively through its distance to $\overline{G}$), and we shall see that this is indeed the case. There is no surprise in the fact that $c_1(\cdot)$ is random, since it tells us at which exponential speed the random measures μ_1^ε converge to the random measure $\mu_1^0 = \delta_{X_1}$. Our results show that c_1 does not depend on anything else, in the sense that, conditionally on $\sigma(X_1)$, it is P-almost surely constant.

We call our result "quenched" (borrowing that terminology from the theory of random media), meaning that the randomness of the observation process is frozen. One could also discuss a "semiquenched" large deviations statement by computing the P_1-almost sure limit (if it exists) of

$$\varepsilon \log \int \int_G q_1^\varepsilon(x + X_1) dx dP_2,$$

while an "annealed" large deviations result would describe the asymptotic behavior of

$$\varepsilon \log E \int_G q_1^\varepsilon(x + X_1) dx.$$

Finally, one could also consider large deviations questions at the level of the conditional measure itself, for example questions concerning the rate of decay of probabilities of the form $P(q_1^\varepsilon(x)dx \in A)$, with A a measurable subset of the space of probability measures on $\mathbb{R}$. We hope to study all these elsewhere.

Let us now state our result. Define

$$\bar{\mathcal{J}}(x, X_1) = \int_{X_1}^x (h(y) - h(X_1)) dy.$$

Our main result is the following theorem. For standard definitions concerning the large deviation principle (LDP), see [3]. For a set $G \subset \mathbb{R}$, we denote by G^o its interior and by $\bar{G}$ its closure.

THEOREM 1.1. *Assume* (A-1)–(A-4). *Then the family of (random) probability measures $q_1^\varepsilon(x)dx$ satisfies a quenched LDP (on the space $\mathbb{R}$ equipped with the standard Euclidean norm) with continuous good rate function $\bar{\mathcal{J}}(\cdot, X_1)$. That is, for any measurable set $G \subset \mathbb{R}$,*

$$
\begin{aligned}
-\inf_{x\in G^o} \bar{\mathcal{J}}(x, X_1) &\le \liminf_{\varepsilon\to 0} \varepsilon \log \int_G q_1^\varepsilon(x)dx \le \limsup_{\varepsilon\to 0} \varepsilon \log \int_G q_1^\varepsilon(x)dx \\
&\le -\inf_{x\in \bar{G}} \bar{\mathcal{J}}(x, X_1), \quad P-a.s. \qquad (1.2)
\end{aligned}
$$

In fact, we have the estimate, valid for any fixed compact set $K_0 \subset \mathbb{R}$,

$$
\lim_{\varepsilon\to 0} \sup_{x\in K_0} |\varepsilon \log q_1^\varepsilon(x) + \bar{\mathcal{J}}(x, X_1)| = 0, \quad P-a.s. \qquad (1.3)
$$

(It will be obvious from the proof that the fixed time 1 can be replaced by any fixed time $t \in (0, \infty)$; that is, the statement of Theorem 1.1 remains true with q_t^ε and X_t replacing q_1^ε and X_1.)

Remarks.

1. In the particular case $h(x) = x$, Theorem 1.1 can be deduced from the results of [10].

2. The reader could wonder why the statement (1.2) is equivalent to the large deviations principle on $\mathbb{R}$ for P-almost ω, since in (1.2), the null set on which the statement does not hold true may depend on G. Note, however, that once the inequalities in (1.2) hold true for each interval $G = (a, b)$ on a set of full measure $\Omega_{a,b}$, we can set

$$
\Omega' = \cap_{a,b\in\mathbb{Q}} \Omega_{a,b}
$$

and conclude that $P(\Omega') = 1$ while (1.2) holds true for all $\omega \in \Omega'$ and all open intervals G with rational endpoints. Since the latter are a base for the topology on $\mathbb{R}$, one concludes (see, e.g., [3, Theorem 4.1.11]) that the full LDP holds for each $\omega \in \Omega'$.

We conclude this introduction with some comments about previous work and possible applications and extensions of our result. Our motivation for the study of the large deviations of the optimal filter is their utility in a variety of applications such as tracking (see [9]) or the study of the filter memory length (see [1]). In the one dimensional linear observation case studied in [10], precise pointwise estimates can be derived by comparison with the linear filtering problem, whose (Gaussian) solution is known explicitly. In contrast, here, the main tool used in the proof of Theorem 1.1 is the representation, due to Picard [7], of the density q_1^ε in terms of an auxiliary suboptimal filter, and the availability of good estimates on the performance of this suboptimal filter. These results are not available in the general multidimensional case. When they are, e.g., in the setup discussed in [8], we believe our analysis can be carried through. Hence, while our result is presently limited to one dimension, we expect that its multidimensional extension to the case where the dimensions of the state and observation coincide, and the observation function is one-to-one, could be deduced from the results of [8]. Extension to the case where the dimension of the observation is smaller than the dimension of the state (which is the most relevant case for applications) would require completely new additional ideas, since the result would be of a completely different nature (the limiting measure is no longer necessarily a Dirac measure, and even when it is, the convergence to the Dirac measure is at different speeds for different coordinates).

We finally note that Hijab [4] has derived a (path) quenched large deviations for the conditional density for systems in which both the signal and the observation noises are small. This is related, by a time change, to looking at short times (of order εT) of the filtering equations

$$dX_t^\varepsilon = \frac{1}{\varepsilon}\bar{b}(X_t^\varepsilon)dt + dB_t, \quad X_0^\varepsilon = x,$$
$$dY_t^\varepsilon = h(X_t^\varepsilon)dt + \varepsilon dV_t.$$

(Hijab's results are not stated in this way, but are equivalent to the description given here. Note that his setup is more general than ours in that it applies to the multi-dimensional setup and allows for general regular diffusion coefficients.) Hijab's results are not directly comparable with the LDP we derive here because of the different time interval on which they apply, and also because of the different type of conditioning. (His statement looks at the conditional density as a continuous functional of the observation trajectory, and considers the LDP when this trajectory is frozen. It is thus not directly applicable as a quenched statement.)

We refer the reader to [5] for a general introduction to stochastic calculus and stochastic differential equations, and to [6] for an exposition of nonlinear filtering theory.

Convention. Throughout the paper, when relevant, we make explicit on what parameters constants depend, even if the actual value of the constant may change from line to line. When nothing explicit is mentioned, i.e., a generic constant C is used, it is understood that it may depend on the trajectories $\{X_\cdot\}$, $\{V_\cdot\}$, but not on ε. For $\infty > t > 0$, we use the notation $||f||_t = \sup_{s\le t}|f(s)|$, with $||f|| := ||f||_{1/\varepsilon}$. Finally, we use θ^t to denote the shift operator, e.g., $\theta^t\tilde{m}(\cdot) = \tilde{m}(t+\cdot)$.

2. Picard's formulation and a path integral. The filtering problem we are going to analyze is (1.1), and the assumptions (A-1)–(A-4) will be assumed to hold throughout the paper. We also note that since nothing is changed (in terms of the filtering problem) by adding a constant to the observation function h, we may and will assume throughout the paper that $h(0) = 0$.

It is known from the results of Picard [7] that the conditional law $q_1^\varepsilon(x)dx$ has a small variance, and that there exist finite dimensional filters that provide good approximations of the unknown state. We shall now recall the formula derived by Picard [7] for $q_1^\varepsilon(x)$, which was used there to study approximate filters. It will be an essential tool for our large deviation results.

Define the approximate filter

$$dM_t^\varepsilon = b(M_t^\varepsilon)dt + \frac{1}{\varepsilon}(dY_t^\varepsilon - h(M_t^\varepsilon)dt),$$

with $M_0^\varepsilon = 0$, and let $\bar{m}_s = M_{1-s}^\varepsilon$ and $\tilde{m}_s = \bar{m}_{\varepsilon s}$, $s \in [0, 1/\varepsilon]$.

One of the main contributions of Picard in [7, Proposition 4.2] was to express the conditional density $q_1^\varepsilon(x)$ in terms of the law of an auxiliary process $\{\bar{X}_{1-t}^x, 0 \le t \le 1\}$, which fluctuates backward in time, starting at time 1 from the position x, around the trajectory of the approximate filter $M_\cdot^\varepsilon$. Performing a time change and a Girsanov transformation, Picard's result can be rewritten as follows.[2] Define the process

$$d\tilde{Z}_s^{\varepsilon,x} = \left[-h(\tilde{Z}_s^{\varepsilon,x}) + \tilde{m}_s h'(\tilde{Z}_s^{\varepsilon,x}) - \varepsilon b(\tilde{Z}_s^{\varepsilon,x})\right] ds + \sqrt{\varepsilon}d\tilde{W}_s, \quad \tilde{Z}_0^{\varepsilon,x} = x,$$

[2]For completeness, and since the computations involved are somewhat lengthy, we present the derivation in an appendix at the end of the paper.

with $\widetilde{W}$ a standard Brownian motion, independent of $B_\cdot, V_\cdot$. Throughout, we let $\mathbb{E}$ and $\mathbb{P}$ denote expectations and probabilities with respect to the law of the Brownian motion $\widetilde{W}_\cdot$. Then a version of the conditional density of X_1, given $\mathcal{Y}_1^\varepsilon$, is given by

$$q_1^\varepsilon(x) = \frac{\rho_1^\varepsilon(x)}{\int_{\mathbb{R}} \rho_1^\varepsilon(x)dx}, \tag{2.1}$$

where

$$\rho_1^\varepsilon(x) := e^{-F(x,\tilde{m}_0)/\varepsilon}\mathbb{E}\left[\exp\left(I_\varepsilon(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x},0) + \int_0^{1/\varepsilon} g_1(\tilde{Z}_s^{\varepsilon,x},\tilde{m}_s)ds + \frac{1}{\varepsilon}\int_0^{1/\varepsilon} g_2(\tilde{Z}_s^{\varepsilon,x},\tilde{m}_s)ds\right)\right], \tag{2.2}$$

and

$$\begin{aligned}
F(z,m) &= \int_0^z (h(y)-h(m))dy - mh(z) + h(m)z,\\
I_\varepsilon(z,m) &= \log p_0(z) + \frac{1}{\varepsilon}F(z,m),\\
g_1(z,m) &= -mh'(z)b(z) + \frac{mh''(z)}{2} + h(z)b(z) - \frac{h'(z)}{2} - \varepsilon b'(z) - h(z)b(m),\\
g_2(z,m) &= h(z)h(m) - \frac{h^2(m)}{2} - mh(z)h'(z) + \frac{m^2h'(z)^2}{2}.
\end{aligned}$$

Note that assumptions (A-1)–(A-4) ensure that, for each given m, $g_1(\cdot,m)$, $g_2(\cdot,m)$ are Lipschitz functions with Lipschitz constant uniformly bounded for m in compacts.

It is important to note that, above and throughout the paper, expressions of the form $\mathbb{E}(\cdot)$ may still be random, due to their possible dependence on $B_\cdot, V_\cdot$. Thus, any equality between such expressions is to be understood in an almost sure sense. We will not explicitly mention this in what follows.

Equipped with (2.2), one is tempted to apply standard tools of large deviations theory, viz. the large deviations principle for $\tilde{Z}_\cdot^{\varepsilon,x}$ and Varadhan's lemma, to the analysis of the exponential rate of decay of the $\mathbb{P}$ expectation in (2.2). This temptation is quenched when one realizes that, in fact, the rate of growth of ρ_1^ε is exponential in $1/\varepsilon^2$, and it is only after normalization that one can hope to obtain the relevant $1/\varepsilon$ asymptotics. This fact, unfortunately, makes the analysis slightly more subtle. In the next section, we present several lemmas, whose proof is deferred to section 4, and show how to deduce Theorem 1.1 from these lemmas. Before closing this section, however, we state the following easy a priori estimates. Recall that, according to our convention, $||X||_1 = \sup_{s\le 1}|X_s|$.

LEMMA 2.1. *$||X||_1 < \infty$, P-almost surely,*

$$|||\tilde{m}||| := \limsup_{\varepsilon\to 0}\sup_{t\in[0,1/\varepsilon]}|\tilde{m}_t| < \infty, \quad P-a.s.,$$

and for $T_\varepsilon = \log(1/\varepsilon)$, $|||\tilde{m}_X||| := \sup_{s\in[0,T_\varepsilon]}|\tilde{m}_s - X_1|$,

$$\limsup_{\varepsilon\to 0}|||\tilde{m}_X||| = 0, \quad P-a.s. \tag{2.3}$$

Further, there exists a constant $C_{V,X}$ depending only on $\{X_\cdot, V_\cdot\}$ such that

$$\sup_{s\in[0,T_\varepsilon]}|\tilde{m}_s - X_1| \le C_{V,X}/\sqrt{T_\varepsilon}, \quad P-a.s.$$

Proof of Lemma 2.1. The statement that $||X||_1 < \infty$ is part of the statement concerning existence of solutions to the SDE (1.1). Next, we prove that

$$\limsup_{\varepsilon\to 0} \sup_{t\le 1} |M_t^\varepsilon| < \infty. \tag{2.4}$$

Indeed, fix constants $C = C(||X||_1)$ and ε_0 such that $h(y) - h(x) + \sup_{\varepsilon\le\varepsilon_0} \varepsilon b(x) < 0$ for all $x \ge C$ and $|y| \le ||X||_1$ (this is always possible because b, h are Lipschitz and $h' > h_0$). Define the stopping times $\tau_0 = 0,\ \theta_0 = 0$ and

$$\tau_i = \inf\{t > \theta_{i-1} : M_t^\varepsilon = C\}, \qquad \theta_i = \inf\{t > \tau_i : M_t^\varepsilon = C + 1\}.$$

By definition, $M_t^\varepsilon \le C + 1$ for $t \in [\tau_i, \theta_i]$, while for $t \in [\theta_i, \tau_{i+1}]$ it holds that for all $\varepsilon < \varepsilon_0$,

$$M_t^\varepsilon = M_{\theta_i}^\varepsilon + \int_{\theta_i}^t \left[b(M_s^\varepsilon) + \frac{1}{\varepsilon}(h(X_s) - h(M_s^\varepsilon)) \right] ds + V_t - V_{\theta_i} \le C + 1 + 2||V||_1.$$

We conclude that $\sup_{t\le 1} M_t^\varepsilon \le C+1+2||V||_1 < \infty$ for all $\varepsilon < \varepsilon_0$. A similar argument shows that $\inf_{t\le 1} M_t^\varepsilon \ge -(C + 1 + 2\|V\|_1)$.

To see the stated convergence of $\tilde{m}_s$ to X_1, recall that X_t and V_t are almost surely Hölder(η) continuous for all $\eta < 1/2$. Fix $t_0 = 1 - 2\varepsilon T_\varepsilon$, $t_1 = 1 - \varepsilon T_\varepsilon$, $\delta_\varepsilon = 1/\sqrt{T_\varepsilon}$, and write $Y_t = M_t^\varepsilon - X_1$. With these notations,

$$Y_t = Y_{t_0} + \int_{t_0}^t \left[b(M_s^\varepsilon) + \frac{h(X_s) - h(X_1)}{\varepsilon} \right] ds + \frac{1}{\varepsilon} \int_{t_0}^t (h(X_1) - h(M_s^\varepsilon)) ds + (V_t - V_{t_0}).$$

By the first part of the lemma, it holds that $|Y_{t_0}| \le C$. We first show that for some $\tau \in (t_0, t_1)$ it holds that $|Y_\tau| \le \delta_\varepsilon$. Indeed, assume without loss of generality that $Y_{t_0} > \delta_\varepsilon$. Then, by the Hölder property of $X.$ and $V.$, it holds that

$$\sup_{t\in(t_0,t_1)} |V_t - V_{t_0}| \le C(\varepsilon T_\varepsilon)^\eta, \qquad \sup_{t\in(t_0,t_1)} |X_t - X_{t_0}| \le C(\varepsilon T_\varepsilon)^\eta.$$

Hence, if a τ as defined above does not exist, then necessarily, using the Lipschitz continuity of h,

$$-C \le C_1 \varepsilon T_\varepsilon \left(1 + \frac{(\varepsilon T_\varepsilon)^\eta}{\varepsilon} \right) - h_0 \delta_\varepsilon T_\varepsilon + C_1 (\varepsilon T_\varepsilon)^\eta,$$

which is clearly impossible unless $\varepsilon \ge \varepsilon_0$ for some $\varepsilon_0 > 0$. Now, for $\tau < t \le 1$ we claim that it is impossible that $Y_t > 2\delta_\varepsilon$. Indeed, let $\theta' = \inf\{\tau < t \le 1 : Y_t = 2\delta_\varepsilon\}$. Repeating the argument above, we now obtain that if such a θ' exists, it must hold that for some $\theta < 2\varepsilon T_\varepsilon$,

$$\delta_\varepsilon \le C_1 \theta + C_1 \frac{\theta^{\eta+1}}{\varepsilon} + C_1 \theta^\eta - \frac{h_0 \delta_\varepsilon \theta}{\varepsilon},$$

which again is impossible, unless $\varepsilon \ge \varepsilon_0'$, for some $\varepsilon_0' > 0$. The case of $Y_t < -2\delta_\varepsilon$ for some $t > t_0$ being handled similarly, the conclusion follows. □

3. Auxiliary lemmas and proof of Theorem 1.1. Let us set $J_\varepsilon(x) := \rho_1^\varepsilon(x) e^{F(x,\tilde{m}_0)/\varepsilon}$ and

$$\bar{L}_\varepsilon(x,t) = \exp\left(\int_0^t \left(g_1(\tilde{Z}_s^{\varepsilon,x}, \tilde{m}_s) + \frac{1}{\varepsilon} g_2(\tilde{Z}_s^{\varepsilon,x}, \tilde{m}_s)\right) ds\right) \tag{3.1}$$

and

$$L_\varepsilon(x,t) = \exp(I_\varepsilon(\tilde{Z}_t^{\varepsilon,x}, 0)) \bar{L}_\varepsilon(x,t). \tag{3.2}$$

Although both $\bar{L}_\varepsilon(x,t)$ and $L_\varepsilon(x,t)$ depend on the path $\tilde{m}.$, we omit this dependence when no confusion occurs, while $L_\varepsilon(x,t,m.)$ will denote the quantity $L_\varepsilon(x,t)$ with $\tilde{m}.$ replaced by $m.$, and similarly for $\bar{L}_\varepsilon$.

The following are the auxiliary lemmas alluded to above. The proof of the first, Lemma 3.1, is standard, combining large deviations estimates for solutions of SDEs (see, e.g., [2, Theorem 2.13, p. 91]) with Varadhan's lemma (see, e.g., [3, Theorem 4.3.1, p. 137]), and is omitted.

LEMMA 3.1 (finite horizon LDP). *Fix $T < \infty$ and a compact $K \subset\subset \mathbb{R}$. Define*

$$I_T(x,z) := \sup_{\phi \in H^1 : \phi_0 = x, \phi_T = z} \int_0^T g_2(\phi_s, X_1) ds - \frac{1}{2} \int_0^T \left[\dot{\phi}_s + h(\phi_s) - X_1 h'(\phi_s)\right]^2 ds.$$

Then, uniformly in $x, z \in K$, P-almost surely,

$$\limsup_{\delta \to 0} \limsup_{\varepsilon \to 0} \left| \varepsilon \log \mathbb{E}\left[\bar{L}_\varepsilon(x,T) \mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x} - z| < \delta\}}\right] - I_T(x,z)\right| = 0.$$

It is worth noting the following simpler representation of $I_T(x,z)$:

$$\begin{aligned} I_T(x,z) = \sup_{\phi \in H^1 : \phi_0 = x, \phi_T = z} \Bigg[& X_1(h(z) - h(x)) - h(X_1)(z - x) \\ & - \frac{1}{2} \int_0^T \left[\dot{\phi}_s - (h(X_1) - h(\phi_s))\right]^2 ds \Bigg]. \end{aligned} \tag{3.3}$$

From this representation, the following is immediate:

$$I_T(X_1, X_1) = 0, \tag{3.4}$$

and, with $V_T(x) := I_T(x, X_1)$, it holds that

$$V_T(x) \to_{T \to \infty} -X_1 h(x) + h(X_1) x. \tag{3.5}$$

This, and standard large deviations considerations, give the next result.

COROLLARY 3.2. *Fix a compact set $K \subset\subset \mathbb{R}$. Then uniformly in $x, z \in K$, P-almost surely,*

$$\begin{aligned} & \limsup_{T \to \infty} \limsup_{\delta \to 0} \limsup_{\varepsilon \to 0} \Bigg| \varepsilon \log \mathbb{E}\left[\bar{L}_\varepsilon(x,T) \mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x} - z| < \delta/2\}} \mathbf{1}_{\{|\tilde{Z}_{T/2}^{\varepsilon,x} - X_1| < \delta/2\}}\right] \\ & \qquad - h(X_1)x + h(x)X_1 - I_{T/2}(X_1, z) \Bigg| \\ & = \limsup_{T \to \infty} \limsup_{\delta \to 0} \limsup_{\varepsilon \to 0} \left| \varepsilon \log \mathbb{E}\left[\bar{L}_\varepsilon(x,T) \mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x} - z| < \delta/2\}} \mathbf{1}_{\{|\tilde{Z}_{T/2}^{\varepsilon,x} - X_1| < \delta/2\}}\right] - I_T(x,z)\right| \\ & = \limsup_{T \to \infty} \limsup_{\delta \to 0} \limsup_{\varepsilon \to 0} \left| \varepsilon \log \mathbb{E}\left[\bar{L}_\varepsilon(x,T) \mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x} - z| < \delta/2\}}\right] - I_T(x,z)\right| = 0. \end{aligned}$$

The key to the proof of Theorem 1.1 is a localization procedure that allows one to restrict attention to compact (in time and space) subsets. A first coarse step in that direction is provided by the next two lemmas.

LEMMA 3.3 (coarse localization 1). *For each $\eta > 0$ there exist constants $M_1 = M_1(|||\tilde{m}|||, \eta, |X_1|)$ and $\varepsilon_{00} = \varepsilon_{00}(|||\tilde{m}|||, \eta, |X_1|)$ such that, for all $\varepsilon < \varepsilon_{00}$,*

$$\int \rho_1^\varepsilon(x)\mathbf{1}_{\{|x|>M_1/\sqrt{\varepsilon}\}}dx \le e^{-\eta/\varepsilon}\inf_{|x|<1}\rho_1^\varepsilon(x) \le e^{-\eta/\varepsilon}\int \rho_1^\varepsilon(x)\mathbf{1}_{\{|x|\le M_1/\sqrt{\varepsilon}\}}dx, \quad P-a.s. \tag{3.6}$$

LEMMA 3.4 (coarse localization 2). *For each $\eta > 0$ and M_1, ε_{00} as in Lemma 3.3, there exist constants $M_i = M_i(|||\tilde{m}|||, \eta, |X_1|)$, $i = 2, 3$, with $M_3 \le M_2$ and $\varepsilon_0 = \varepsilon_0(|||\tilde{m}|||, \eta, |X_1|) < \varepsilon_{00}$, such that for all $\varepsilon < \varepsilon_0$, uniformly in $|x| \le M_1/\sqrt{\varepsilon}$,*

$$J_\varepsilon(x) \le 2\mathbb{E}\left[L_\varepsilon\left(x, \frac{1}{\varepsilon}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||\le M_3/\varepsilon\}}\right], \tag{3.7}$$

and uniformly in $|z| \le M_3/\varepsilon$, $T < 1/\varepsilon$,

$$\mathbb{E}\left[L_\varepsilon\left(z, \frac{1}{\varepsilon} - T, \theta^T\tilde{m}\right)\right] \le 2\mathbb{E}\left[L_\varepsilon\left(z, \frac{1}{\varepsilon} - T, \theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,z}||_{1/\varepsilon-T}\le M_2/\varepsilon\}}\right]. \tag{3.8}$$

The following comparison lemma is also needed.

LEMMA 3.5. *There exists a function $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$, depending on $|||\tilde{m}|||, |X_1|, \eta$ only, with $g(\delta) \to_{\delta\to 0} 0$, and an $\varepsilon_1 = \varepsilon_1(|||\tilde{m}|||, X_1, \eta) < \varepsilon_0$ such that for all $\varepsilon < \varepsilon_1$, $t \in [1/2\varepsilon, 1/\varepsilon]$, and $|x|, |y| \le M_3/\varepsilon$, $|x - y| < \delta$,*

$$\varepsilon \log\left(\frac{\mathbb{E}(L_\varepsilon(x, t, \theta^{1/\varepsilon-t}\tilde{m})\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||_t\le M_2/\varepsilon\}})}{\mathbb{E}(L_\varepsilon(y, t, \theta^{1/\varepsilon-t}\tilde{m})\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,y}||_t\le M_2/\varepsilon\}})}\right) \le g(\delta), \tag{3.9}$$

and there exists a constant $C_1(|||\tilde{m}|||, X_1, \eta)$ such that, for all $\varepsilon < \varepsilon_1$,

$$\sup_{t\in[1/2\varepsilon,1/\varepsilon]} \varepsilon\left|\log\left(\frac{\mathbb{E}\left[L_\varepsilon(x, t, \theta^{1/\varepsilon-t}\tilde{m})\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||_t\le M_2/\varepsilon\}}\right]}{\mathbb{E}\left[L_\varepsilon(X_1, t, \theta^{1/\varepsilon-t}\tilde{m})\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,X_1}||_t\le M_2/\varepsilon\}}\right]}\right)\right| \le C_1(1 + |x|). \tag{3.10}$$

The last step needed in order to carry out the localization procedure is the following.

LEMMA 3.6 (localization). *Fix a sequence T_ε as in Lemma 2.1. Then there exist constants $C_i = C_i(|||\tilde{m}|||, M_1, M_2, M_3, X_1) > 0$, $i \ge 2$, and $\varepsilon_2 = \varepsilon_2(|||\tilde{m}|||, M_1, M_2, M_3, X_1) < \varepsilon_1$ such that, for all $\varepsilon < \varepsilon_2$, $|x| \le M_1/\sqrt{\varepsilon}$, $|z| \le M_3/\varepsilon$, $\delta < 1$, and $1 \le T \le T_\varepsilon$,*

$$\begin{aligned}&\mathbb{E}\left[\bar{L}_\varepsilon(x, T)\mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x}-z|<\delta\}}\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||_T\le M_3/\varepsilon\}}\right]\\ &\quad\le \exp\left(\frac{C_2}{\varepsilon} - \frac{C_3(|z| - |x|)_+^2}{\varepsilon} + \frac{C_4(|x| + |z|)}{\varepsilon}\right),\end{aligned} \tag{3.11}$$

and, uniformly for $|z - X_1| < 1, |x - X_1| < 1$,

$$\mathbb{E}\left[\bar{L}_\varepsilon(x, T)\mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x}-z|<\delta\}}\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||_T\le M_3/\varepsilon\}}\right] \ge \exp\left(-\frac{C_2}{\varepsilon}\right). \tag{3.12}$$

We may now proceed to the proof of Theorem 1.1, as a consequence of the above lemmas. Fix an $\eta > 0$ as in Lemma 3.3, and for $\delta > 0$, $T > 0$ to be chosen below, with $T < T_\varepsilon$, T_ε as in Lemma 2.1, define

$$\begin{aligned}\tilde{J}_\varepsilon(x) &= \mathbb{E}\left(L_\varepsilon\left(x, \frac{1}{\varepsilon}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||_T \le M_3/\varepsilon, ||\tilde{Z}^{\varepsilon,x}|| \le M_2/\varepsilon\}}\right)\\ &= \sum_{i=-M_3/\varepsilon\delta}^{M_3/\varepsilon\delta} \mathbb{E}\left(L_\varepsilon\left(x, \frac{1}{\varepsilon}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}|| \le M_2/\varepsilon, ||\tilde{Z}^{\varepsilon,x}||_T \le M_3/\varepsilon, |\tilde{Z}_T^{\varepsilon,x} - i\delta| \le \delta/2\}}\right)\\ &=: \sum_{i=-M_3/\varepsilon\delta}^{M_3/\varepsilon\delta} \tilde{J}_{\varepsilon,T}(x, i\delta). \qquad (3.13)\end{aligned}$$

Set $\mathcal{Z}_T^{\varepsilon,x} = \sigma(\tilde{Z}_t^{\varepsilon,x}, t \le T)$. Using the Markov property, and the fact that $M_3 < M_2$, one may write, for $|z| < M_3/\varepsilon$,

$$\begin{aligned}\tilde{J}_{\varepsilon,T}(x,z) = \mathbb{E}\Bigg[&\bar{L}_\varepsilon(x,T)\mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x} - z| \le \delta/2\}}\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||_T \le M_3/\varepsilon\}}\\ &\cdot \mathbb{E}\left(L_\varepsilon\left(\tilde{Z}_T^{\varepsilon,x}, \frac{1}{\varepsilon} - T, \theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}|| \le M_2/\varepsilon\}} \mid \mathcal{Z}_T^{\varepsilon,x}\right)\Bigg]. \qquad (3.14)\end{aligned}$$

Applying (3.9) and the Markov property, it follows that on the event $\{|\tilde{Z}_T^{\varepsilon,x} - z| \le \delta/2\} \cap \{||\tilde{Z}^{\varepsilon,x}||_T \le M_3/\varepsilon\}$, one has for $\varepsilon < \varepsilon_1$, and $|x| \le M_1/\sqrt{\varepsilon}$, $|z| \le M_3/\varepsilon$,

$$\begin{aligned}&\mathbb{E}\left(L_\varepsilon\left(\tilde{Z}_T^{\varepsilon,x}, \frac{1}{\varepsilon} - T, \theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}|| \le M_2/\varepsilon\}} \mid \mathcal{Z}_T^{\varepsilon,x}\right)\\ &= \mathbb{E}\left(L_\varepsilon\left(\tilde{Z}_T^{\varepsilon,x}, \frac{1}{\varepsilon} - T, \theta^T\tilde{m}\right)\mathbf{1}_{\{\sup_{T \le t \le 1/\varepsilon} |\tilde{Z}_t^{\varepsilon,x}| \le M_2/\varepsilon\}} \mid \mathcal{Z}_T^{\varepsilon,x}\right)\\ &\le e^{g(\delta)/\varepsilon}\mathbb{E}\left(L_\varepsilon\left(z, \frac{1}{\varepsilon} - T, \theta^T\tilde{m}\right)\mathbf{1}_{\{\sup_{0 \le 1/\varepsilon - T} |\tilde{Z}_t^{\varepsilon,z}| \le M_2/\varepsilon\}}\right)\\ &= e^{g(\delta)/\varepsilon}\mathbb{E}\left(L_\varepsilon\left(z, \frac{1}{\varepsilon} - T, \theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,z}||_{1/\varepsilon - T} \le M_2/\varepsilon\}}\right).\end{aligned}$$

Substituting in (3.14), one concludes that for all $\varepsilon < \varepsilon_1$ and $|x| \le M_1/\sqrt{\varepsilon}$, $|z| \le M_3/\varepsilon$,

$$\begin{aligned}\tilde{J}_{\varepsilon,T}(x,z)e^{-g(\delta)/\varepsilon} &\le \mathbb{E}\left[\bar{L}_\varepsilon(x,T)\mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x} - z| \le \delta/2\}}\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||_T \le M_3/\varepsilon\}}\right]\\ &\quad \cdot \mathbb{E}\left[L_\varepsilon\left(z, \frac{1}{\varepsilon} - T, \theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,z}||_{1/\varepsilon - T} \le M_2/\varepsilon\}}\right] \qquad (3.15)\\ &:= \hat{J}_{\varepsilon,T}(x,z) \le \tilde{J}_{\varepsilon,T}(x,z)e^{g(\delta)/\varepsilon}.\end{aligned}$$

Next, using (3.10) in the first inequality and Lemma 3.6 in the second, it follows that for all $\varepsilon < \varepsilon_2$; $T \in (1, T_\varepsilon)$, T_ε as in Lemma 2.1; and some constants C_i independent of T, ε,

$$\begin{aligned}\hat{J}_{\varepsilon,T}(x,z) &\le \mathbb{E}\left[\bar{L}_\varepsilon(x,T)\mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x}-z|\le\delta/2\}}\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||_T\le M_3/\varepsilon\}}\right]\\ &\quad\cdot\mathbb{E}\left[L_\varepsilon\left(X_1,\frac{1}{\varepsilon}-T,\theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,X_1}||_{1/\varepsilon-T}\le M_2/\varepsilon\}}\right]e^{C_1(|z|+1)/\varepsilon}\\ &\le \exp\left(\frac{C_2}{\varepsilon}-\frac{C_3(|z|-|x|)_+^2}{\varepsilon}+\frac{C_5(|x|+|z|)}{\varepsilon}\right)\\ &\quad\cdot\mathbb{E}\left[L_\varepsilon\left(X_1,\frac{1}{\varepsilon}-T,\theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,X_1}||_{1/\varepsilon-T}\le M_2/\varepsilon\}}\right].\end{aligned}\tag{3.16}$$

Similarly, for all $\varepsilon<\varepsilon_2$ and $|x-X_1|\le 1$, $|z-X_1|\le 1$,

$$\hat{J}_{\varepsilon,T}(x,z)\ge\exp\left(-\frac{C_2}{\varepsilon}\right)\mathbb{E}\left[L_\varepsilon\left(X_1,\frac{1}{\varepsilon}-T,\theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,X_1}||_{1/\varepsilon-T}\le M_2/\varepsilon\}}\right].\tag{3.17}$$

We next note that, due to the quadratic growth of $F(x,X_1)$ as $|x|\to\infty$, there exists a compact set $\mathcal{K}_1$, depending on $|||m|||, X_1, \eta, C_i$ only, such that

$$\begin{aligned}&\sup_{(x,z)\in(\mathcal{K}_1\times\mathcal{K}_1)^c}\frac{C_2}{\varepsilon}-\frac{C_3(|z|-|x|)_+^2}{\varepsilon}+\frac{C_5(|x|+|z|)}{\varepsilon}-\frac{F(x,X_1)}{\varepsilon}\\ &\qquad\le-\frac{F(X_1,X_1)}{\varepsilon}-\frac{C_2}{\varepsilon}.\end{aligned}\tag{3.18}$$

Thus, using (3.16) in the first inequality, (3.18) in the second, and (3.17) in the third,

$$\begin{aligned}&\sup_{\substack{|x|\le M_1/\sqrt{\varepsilon},|z|\le M_3/\varepsilon,\\(x,z)\in(\mathcal{K}_1\times\mathcal{K}_1)^c}}\hat{J}_{\varepsilon,T}(x,z)e^{-F(x,X_1)/\varepsilon}\\ &\le\mathbb{E}\left[L_\varepsilon\left(X_1,\frac{1}{\varepsilon}-T,\theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,X_1}||_{1/\varepsilon-T}\le M_2/\varepsilon\}}\right]\\ &\quad\cdot\sup_{\substack{|x|\le M_1/\sqrt{\varepsilon},|z|\le M_3/\varepsilon,\\(x,z)\in(\mathcal{K}_1\times\mathcal{K}_1)^c}}\exp\left(\frac{C_2}{\varepsilon}-\frac{C_3(|z|-|x|)_+^2}{\varepsilon}+\frac{C_5(|x|+|z|)}{\varepsilon}-\frac{F(x,X_1)}{\varepsilon}\right)\\ &\le\mathbb{E}\left[L_\varepsilon\left(X_1,\frac{1}{\varepsilon}-T,\theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,X_1}||_{1/\varepsilon-T}\le M_2/\varepsilon\}}\right]\exp\left(-\frac{C_2}{\varepsilon}-\frac{F(X_1,X_1)}{\varepsilon}\right)\\ &\le\hat{J}_{\varepsilon,T}(X_1,X_1)e^{-F(X_1,X_1)/\varepsilon}.\end{aligned}\tag{3.19}$$

It follows by substituting (3.19) into (3.15) that for all ε small enough and any $T\in(0,T_\varepsilon)$,

$$\sup_{|x|\le M_1/\sqrt{\varepsilon},|z|\le M_3/\varepsilon}\tilde{J}_{\varepsilon,T}(x,z)e^{-F(x,X_1)/\varepsilon}\le e^{2g(\delta)/\varepsilon}\sup_{x\in\mathcal{K}_1,z\in\mathcal{K}_1}\tilde{J}_{\varepsilon,T}(x,z)e^{-F(x,X_1)/\varepsilon}.\tag{3.20}$$

We may, by enlarging $\mathcal{K}_1$ if necessary, assume also that $[-1,1]\subset\mathcal{K}_1$. With η and $\mathcal{K}_1$ as above, next choose T large enough, δ small enough (with $g(\delta)<\eta/8$), and $\varepsilon_3(\delta,T,\eta,|||\tilde{m}|||,|||\tilde{m}_X|||,X_1)<\varepsilon_2$ such that, for all $\varepsilon<\varepsilon_3$, the following hold:

- The errors in the expression in Corollary 3.2 and in (3.5) are each bounded above by $\eta/8$, uniformly in $x,z\in\mathcal{K}_1$.
- $|F(x,\tilde{m}_0)-F(x,X_1)|\le\eta/8$, uniformly in $x\in\mathcal{K}_1$ (which is possible by Lemma 2.1 and the uniform continuity of $F(x,\cdot)$ for x in compacts).

- $\varepsilon \log 2 \le \eta/8$.
- $\varepsilon \log(2M_3/\varepsilon\delta) \le \eta/8$.

Hence, for $x \in \mathcal{K}_1$ and all $\varepsilon < \varepsilon_3$,

$$
\begin{aligned}
\varepsilon \log \rho_1^\varepsilon(x) &= -F(x,\tilde m_0) + \varepsilon \log \mathbb{E}\left(L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\right) \quad \texttt{by (2.2)} \\
&\le -F(x,\tilde m_0) + \varepsilon \log \mathbb{E}\left(L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{||\tilde Z^{\varepsilon,x}||\le M_3/\varepsilon\}}\right) + \varepsilon\log 2 \quad \texttt{by (3.7)} \\
&\le -F(x,X_1) + \varepsilon \log \mathbb{E}\left(L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{||\tilde Z^{\varepsilon,x}||\le M_3/\varepsilon\}}\right) + \frac{\eta}{4} \quad \texttt{by } \varepsilon < \varepsilon_3 \\
&\le -F(x,X_1) + \varepsilon \log \tilde J_\varepsilon(x) + \frac{\eta}{4} \quad \texttt{by (3.13)} \\
&\le -F(x,X_1) + \varepsilon \log \sup_{z\in\mathcal{K}_1} \tilde J_{\varepsilon,T}(x,z) + \frac{\eta}{2} \quad \texttt{by (3.13) and (3.20)} \\
&\le -F(x,X_1) + \sup_{z\in\mathcal{K}_1}\Bigg[\varepsilon \log \mathbb{E}(\bar L_\varepsilon(x,T)\mathbf{1}_{\{|\tilde Z_T^{\varepsilon,x}-z|\le\delta/2\}}) \\
&\quad + \varepsilon \log \mathbb{E}\left(L_\varepsilon\left(z,\frac{1}{\varepsilon}-T,\theta^T\tilde m\right)\mathbf{1}_{\{||\tilde Z^{\varepsilon,z}||_{1/\varepsilon-T}\le M_2/\varepsilon\}}\right)\Bigg] + \frac{5\eta}{8} \quad \texttt{by (3.15)} \\
&\le -F(x,X_1) + \sup_{z\in\mathcal{K}_1}\Bigg[h(X_1)x - h(x)X_1 + I_{T/2}(X_1,z) \\
&\qquad\qquad + \varepsilon \log \mathbb{E}\left(L_\varepsilon\left(z,\frac{1}{\varepsilon}-T,\theta^T\tilde m\right)\mathbf{1}_{\{||\tilde Z^{\varepsilon,z}||_{1/\varepsilon-T}\le M_2/\varepsilon\}}\right)\Bigg] \\
&\quad + \frac{7\eta}{8} \quad \texttt{by Corollary 3.2} \\
&\le h(X_1)x - h(x)X_1 - F(x,X_1) + \eta \\
&\quad + \sup_{z\in\mathcal{K}_1}\left[I_{T/2}(X_1,z) + \varepsilon \log \mathbb{E}\left(L_\varepsilon\left(z,\frac{1}{\varepsilon}-T,\theta^T\tilde m\right)\right)\right] \\
&=: -\bar{\mathcal{J}}(x,X_1) + \eta + \mathcal{C}_\varepsilon, \qquad (3.21)
\end{aligned}
$$

where $\mathcal{C}_\varepsilon$ depends only on ε, and not on x, and is defined by the last equality. Similarly, for all $x \in \mathcal{K}_1$ and all $\varepsilon < \varepsilon_3$,

$$
\begin{aligned}
\varepsilon \log \rho_1^\varepsilon(x) &= -F(x,\tilde m_0) + \varepsilon \log \mathbb{E}\left(L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\right) \quad \texttt{by (2.2)} \\
&\ge -F(x,\tilde m_0) + \varepsilon \log \mathbb{E}\left(L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{||\tilde Z^{\varepsilon,x}||_T\le M_3/\varepsilon, ||\tilde Z^{\varepsilon,x}||\le M_2/\varepsilon\}}\right) \\
&\ge -F(x,X_1) + \varepsilon \log \mathbb{E}\left(L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{||\tilde Z^{\varepsilon,x}||_T\le M_3/\varepsilon, ||\tilde Z^{\varepsilon,x}||\le M_2/\varepsilon\}}\right) \\
&\quad - \frac{\eta}{4} \quad \texttt{by } \varepsilon < \varepsilon_3 \\
&= -F(x,X_1) + \varepsilon \log \tilde J_\varepsilon(x) - \frac{\eta}{4} \quad \texttt{by definition} \\
&\ge -F(x,X_1) + \varepsilon \log \sup_{z\in\mathcal{K}_1} \tilde J_{\varepsilon,T}(x,z) - \frac{\eta}{4} \quad \texttt{by definition}
\end{aligned}
$$

$$
\begin{aligned}
&\geq -F(x,X_1)+\sup_{z\in\mathcal{K}_1}\left[\varepsilon\log\mathbb{E}(\bar{L}_\varepsilon(x,T)\mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x}-z|\leq\delta/2\}})\right.\\
&\quad\left.+\varepsilon\log\mathbb{E}\left(L_\varepsilon\left(z,\frac{1}{\varepsilon}-T,\theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,z}||_{1/\varepsilon-T}\leq M_2/\varepsilon\}}\right)\right]-\frac{5\eta}{8}\quad\texttt{by (3.15)}\\
&\geq -F(x,X_1)+\sup_{z\in\mathcal{K}_1}\left[h(X_1)x-h(x)X_1+I_{T/2}(X_1,z)\right.\\
&\qquad\left.+\varepsilon\log\mathbb{E}\left(L_\varepsilon\left(z,\frac{1}{\varepsilon}-T,\theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,z}||_{1/\varepsilon-T}\leq M_2/\varepsilon\}}\right)\right]\\
&\quad-\frac{7\eta}{8}\quad\texttt{by Corollary(3.2)}\\
&\geq h(X_1)x-h(x)X_1-F(x,X_1)-\eta\\
&\quad+\sup_{z\in\mathcal{K}_1}\left[I_{T/2}(X_1,z)+\varepsilon\log\mathbb{E}\left(L_\varepsilon\left(z,\frac{1}{\varepsilon}-T,\theta^T\tilde{m}\right)\right)\right]\\
&=-\bar{\mathcal{J}}(x,X_1)-\eta+\mathcal{C}_\varepsilon, \qquad (3.22)
\end{aligned}
$$

where $\mathcal{C}_\varepsilon$ is the same as in (3.21). Since $\bar{\mathcal{J}}(\cdot,X_1)$ is continuous and $\bar{\mathcal{J}}(X_1,X_1)=0$, it follows from (3.22) that

$$
\liminf_{\varepsilon\to0}\varepsilon\log\int_{\mathbb{R}}\rho_1^\varepsilon(x)dx-\mathcal{C}_\varepsilon\geq-2\eta. \qquad (3.23)
$$

On the other hand, for $\varepsilon<\varepsilon_3$,

$$
\begin{aligned}
\varepsilon\log\int_{\mathbb{R}}\rho_1^\varepsilon(x)dx&\leq\varepsilon\log(1+e^{-\eta/\varepsilon})+\varepsilon\log\int_{|x|\leq M_1/\sqrt{\varepsilon}}\rho_1^\varepsilon(x)dx\quad\texttt{by Lemma 3.3}\\
&\leq\varepsilon\log(1+e^{-\eta/\varepsilon})+\varepsilon\log2+\varepsilon\log\left(\frac{2M_3}{\varepsilon\delta}\right)\\
&\qquad+\sup_{|x|\leq M_1/\sqrt{\varepsilon},|z|\leq M_3/\varepsilon}\varepsilon\log\left(\tilde{J}_{\varepsilon,T}(x,z)e^{-F(x,X_1)/\varepsilon}\right)\\
&\qquad\qquad\texttt{by Lemma 3.4 and (3.13)}\\
&\leq\frac{5\eta}{8}+\sup_{x,z\in\mathcal{K}_1}\varepsilon\log\left(\tilde{J}_{\varepsilon,T}(x,z)e^{-F(x,X_1)/\varepsilon}\right)\quad\texttt{by (3.20)}\\
&\leq\frac{5\eta}{8}+\varepsilon\log\left(\sup_{x\in\mathcal{K}_1}\rho_1^\varepsilon(x)\right)\\
&\leq2\eta+\mathcal{C}_\varepsilon-\inf_x\bar{\mathcal{J}}(x,X_1)=2\eta+\mathcal{C}_\varepsilon \qquad (3.24)\\
&\qquad\qquad\texttt{by (3.21) and }\bar{\mathcal{J}}(x,X_1)\geq0.
\end{aligned}
$$

Consider now an open ball $B(x_0,r)\subset\mathbb{R}$. Then, using (3.24) in the first inequality and (3.22) in the last,

$$
\begin{aligned}
\liminf_{\varepsilon\to0}\varepsilon\log\int_{B(x_0,r)}q_1^\varepsilon(x)dx&=\liminf_{\varepsilon\to0}\left[\varepsilon\log\int_{B(x_0,r)}\rho_1^\varepsilon(x)dx-\varepsilon\log\int_{\mathbb{R}}\rho_1^\varepsilon(x)dx\right]\\
&\geq\liminf_{\varepsilon\to0}\left[\varepsilon\log\int_{B(x_0,r)}\rho_1^\varepsilon(x)dx-\mathcal{C}_\varepsilon-2\eta\right]\\
&\geq-\bar{\mathcal{J}}(x_0,X_1)-3\eta.
\end{aligned}
$$

Since η is arbitrary, one deduces that

$$\liminf_{\varepsilon\to 0} \varepsilon \log \int_{B(x_0,r)} q_1^\varepsilon(x)dx \geq -\bar{\mathcal{J}}(x_0, X_1). \tag{3.25}$$

To see the complementary upper bound for the ball $B(x_0, r)$, enlarge $\mathcal{K}_1$ if necessary so that $B(x_0, r) \subset \mathcal{K}_1$ (decreasing ε_3 above as a byproduct). Then, using (3.23) in the first inequality and (3.21) in the last,

$$\begin{aligned}
\limsup_{\varepsilon\to 0} \varepsilon \log \int_{B(x_0,r)} q_1^\varepsilon(x)dx &= \limsup_{\varepsilon\to 0} \left[\varepsilon \log \int_{B(x_0,r)} \rho_1^\varepsilon(x)dx - \varepsilon \log \int_{\mathbb{R}} \rho_1^\varepsilon(x)dx \right] \\
&\leq \limsup_{\varepsilon\to 0} \left[\varepsilon \log \int_{B(x_0,r)} \rho_1^\varepsilon(x)dx - \mathcal{C}_\varepsilon + 2\eta \right] \\
&\leq - \sup_{x\in B(x_0,r)} \bar{\mathcal{J}}(x, X_1) + 3\eta + \limsup_{\varepsilon\to 0} \varepsilon \log(2r).
\end{aligned}$$

Since η is arbitrary, the above, (3.25), and the continuity of $\bar{\mathcal{J}}(\cdot, X_1)$ imply that

$$\lim_{r\to 0} \limsup_{\varepsilon\to 0} \varepsilon \log \int_{B(x_0,r)} q_1^\varepsilon(x)dx = \lim_{r\to 0} \liminf_{\varepsilon\to 0} \varepsilon \log \int_{B(x_0,r)} q_1^\varepsilon(x)dx = \bar{\mathcal{J}}(x_0, X_1).$$

Next, [3, Theorem 4.1.11], the above, Remark 2 following Theorem 1.1, and the continuity of $\bar{\mathcal{J}}(\cdot, X_1)$ imply that the weak LDP holds for the sequence of (random) measures $\mu_1^\varepsilon(dx) = q_1^\varepsilon(x)dx$ on $\mathbb{R}$. To prove the full large deviations principle, it remains, by [3, Lemma 1.2.8], to prove the exponential tightness of the sequence μ_1^ε. That is, for each given L we must find a constant C_L such that

$$\limsup_{\varepsilon\to 0} \varepsilon \log \int_{[-L,L]^c} q_1^\varepsilon(x)ds < -L. \tag{3.26}$$

Since the proof of (3.26) uses some estimates from the proof of Lemma 3.3, to avoid repetitions we postpone it to the end of section 4.

Finally, we note that (1.3) is an immediate consequence of the estimates (3.21), (3.22), (3.24), and (3.23). □

4. Proofs of auxiliary lemmas. Throughout this section, C denotes a positive constant that depends on $|||\tilde{m}|||, |||\tilde{m}_X|||, C_{V,X}, X$ only, and whose value may change from line to line.

Proof of Lemma 3.3. The right-hand inequality is a trivial consequence of the left-hand one. To prove the latter, we first need an upper bound for the left-hand side of (3.6). A subsequent, easily derived lower bound on the middle term will conclude the proof. Define the function

$$H(x) = \int_0^x h(y)dy. \tag{4.1}$$

We note that

$$I_\varepsilon(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x}, 0) - \frac{F(x, \tilde{m}_0)}{\varepsilon} = \frac{1}{\varepsilon} \left(H(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x}) - H(x) \right) + \log p_0(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x}) + \frac{\tilde{m}_0 h(x)}{\varepsilon}.$$

We first rewrite the $\tilde{Z}_t^{\varepsilon,x}$ equation as

$$\tilde{Z}_t^{\varepsilon,x} = x + \int_0^t [-h(\tilde{Z}_s^{\varepsilon,x}) + g(s, \tilde{Z}_s^{\varepsilon,x})]ds + \sqrt{\varepsilon}\tilde{W}_t,$$

and next deduce from Itô's formula that

$$\begin{aligned} &H(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x}) - H(x) \\ &\quad = \int_0^{1/\varepsilon} \left[-h^2(\tilde{Z}_s^{\varepsilon,x}) + (hg)(s, \tilde{Z}_s^{\varepsilon,x}) + \frac{\varepsilon}{2}h'(\tilde{Z}_s^{\varepsilon,x}) \right] ds + \sqrt{\varepsilon} \int_0^{1/\varepsilon} h(\tilde{Z}_s^{\varepsilon,x}) d\tilde{W}_s. \end{aligned}$$

It now follows from (2.2) and the (uniform in m in compacts) linear growth of $g_1(z,m)$ and $g_2(z,m)$ in z that for some C (depending on $|||\tilde{m}|||$ and X only) and all $\varepsilon \le 1$, $\delta > 0$,

$$\begin{aligned} &\rho_1^\varepsilon(x) \le \exp\left[\frac{C}{\varepsilon^2} + \frac{\tilde{m}_0 h(x)}{\varepsilon}\right] \left(\mathbb{E}\left[p_0(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x}) \right]^{\frac{1+\delta}{\delta}} \right)^{\frac{\delta}{1+\delta}} \\ &\times \left(\mathbb{E} \exp\left[\frac{1+\delta}{\sqrt{\varepsilon}} \int_0^{1/\varepsilon} h(\tilde{Z}_s^{\varepsilon,x}) d\tilde{W}_s - \frac{1+\delta}{\varepsilon} \int_0^{1/\varepsilon} h^2(\tilde{Z}_s^{\varepsilon,x}) ds + \frac{C}{\varepsilon} \int_0^{1/\varepsilon} |\tilde{Z}_s^{\varepsilon,x}| ds \right] \right)^{\frac{1}{1+\delta}}. \end{aligned}$$

Now, provided $\delta < 1$, we have $1 + \delta > \frac{(1+\delta)^2}{2}$, and thus there exists a $p > 1$ and a $p' > 0$ such that

$$1 + \delta = \frac{p(1+\delta)^2}{2} + p'.$$

Thus, with $q = p/(p-1)$,

$$\begin{aligned} &\left(\mathbb{E} \exp\left[\frac{1+\delta}{\sqrt{\varepsilon}} \int_0^{1/\varepsilon} h(\tilde{Z}_s^{\varepsilon,x}) d\tilde{W}_s - \frac{(1+\delta)^2 p}{2\varepsilon} \int_0^{1/\varepsilon} h^2(\tilde{Z}_s^{\varepsilon,x}) ds \right.\right. \\ &\qquad \left.\left. - \frac{p'}{\varepsilon} \int_0^{1/\varepsilon} h^2(\tilde{Z}_s^{\varepsilon,x}) ds + \frac{C}{\varepsilon} \int_0^{1/\varepsilon} |\tilde{Z}_s^{\varepsilon,x}| ds \right] \right)^{\frac{1}{1+\delta}} \\ &\le \left(\mathbb{E} \exp\left[\frac{p(1+\delta)}{\sqrt{\varepsilon}} \int_0^{1/\varepsilon} h(\tilde{Z}_s^{\varepsilon,x}) d\tilde{W}_s - \frac{(1+\delta)^2 p^2}{2\varepsilon} \int_0^{1/\varepsilon} h^2(\tilde{Z}_s^{\varepsilon,x}) ds \right] \right)^{\frac{1}{p(1+\delta)}} \\ &\quad \times \left(\mathbb{E} \exp\left[-\frac{p'q}{2\varepsilon} \int_0^{1/\varepsilon} h^2(\tilde{Z}_s^{\varepsilon,x}) ds + \frac{Cq}{\varepsilon} \int_0^{1/\varepsilon} |\tilde{Z}_s^{\varepsilon,x}| ds \right] \right)^{\frac{1}{q(1+\delta)}} \\ &= \left(\mathbb{E} \exp\left[-\frac{p'q}{2\varepsilon} \int_0^{1/\varepsilon} h^2(\tilde{Z}_s^{\varepsilon,x}) ds + \frac{Cq}{\varepsilon} \int_0^{1/\varepsilon} |\tilde{Z}_s^{\varepsilon,x}| ds \right] \right)^{\frac{1}{q(1+\delta)}}. \end{aligned}$$

Since $h(z)^2 \ge h_0^2 z^2$ (recall that $h(0) = 0$!), there exist $C(\delta) > 0$, $C_1(\delta)$ such that $p'qh(z)^2/2 - Cq|z| \ge C(\delta)z^2 - C_1(\delta)$, and hence, with $C_2(\delta) = C + C_1(\delta)\delta/p(1+\delta)$

(all constants here being positive and depending on $|||\tilde{m}|||, X$ only!),

$$\begin{aligned}\rho_1^\varepsilon(x) &\le \exp\left[\frac{C_2(\delta)}{\varepsilon^2}+\frac{\tilde{m}_0 h(x)}{\varepsilon}\right]\left(\mathbb{E}\left[p_0(\tilde{Z}^{\varepsilon,x}_{1/\varepsilon})\right]^{\frac{1+\delta}{\delta}}\right)^{\frac{\delta}{1+\delta}} \\ &\times\left(\mathbb{E}\exp\left[-\frac{C(\delta)}{\varepsilon}\int_0^{1/\varepsilon}|\tilde{Z}^{\varepsilon,x}_s|^2 ds\right]\right)^{\frac{\delta}{q(1+\delta)}} \\ &\le \exp\left[\frac{C_3(\delta)}{\varepsilon^2}\right]\left(\mathbb{E}\exp\left[-\frac{C(\delta)}{\varepsilon}\int_0^{1/\varepsilon}|\tilde{Z}^{\varepsilon,x}_s|^2 ds\right]\right)^{\frac{\delta}{q(1+\delta)}}.\end{aligned} \tag{4.2}$$

It thus remains to estimate the last factor in the above right-hand side. Define $\tau = \inf\{t>0 : |\tilde{Z}^{\varepsilon,x}_s| < x/2\}$ and fix $\eta>0$. We claim that for some $\eta>0$ small enough, it holds that for some $C_\eta>0$, x_0 and all $|x|\ge x_0$,

$$\mathbb{P}(\tau<\eta)\le \exp\left[-\frac{C_\eta x^2}{\varepsilon}\right]. \tag{4.3}$$

Assume (4.3), which will be proved below, and note that on the event $\{\tau\ge\eta\}$ we have that $\inf_{s\in(0,\eta]}|\tilde{Z}^{\varepsilon,x}_s| > x/2$. We deduce from (4.2)

$$\rho_1^\varepsilon(x)\le \exp\left[\frac{C_3(\delta)}{\varepsilon^2}\right]\times\left(\exp\left[-\frac{C_\eta x^2}{\varepsilon}\right]+\exp\left[-\frac{C(\delta)x^2\eta\delta}{4q(1+\delta)\varepsilon}\right]\right), \tag{4.4}$$

from which one easily concludes the bound

$$\rho_1^\varepsilon(x)\le \exp\left[\frac{C_4(\delta)}{\varepsilon^2}-\frac{Cx^2}{\varepsilon}\right] \tag{4.5}$$

for some constants $C_4(\delta)$ and C depending on $\delta, |||\tilde{m}|||, X$ only.

On the other hand, define the event

$$A_C=\left\{\sup_{t\in(0,1/\varepsilon)}\sqrt{\varepsilon}|\tilde{W}_t|\le C\right\}.$$

Then there exists a constant $C_3>0$ depending on C such that $\mathbb{P}(A_C)\ge C_3$. Note that on the event A_C, because $h'(\cdot)>0$ and h, b are Lipschitz, Gronwall's inequality implies that $\sup_{|x|\le 1, s\le 1/\varepsilon}|\tilde{Z}^{\varepsilon,x}_s|\le C'$ for some constant C' depending on $C, \tilde{m}, X$ only. Thus, on the event A_C,

$$\left|I_\varepsilon(\tilde{Z}^{\varepsilon,x}_{1/\varepsilon},0)+\int_0^{1/\varepsilon}g_1(\tilde{Z}^{\varepsilon,x}_s,\tilde{m}_s)ds+\frac{1}{\varepsilon}\int_0^{1/\varepsilon}g_2(\tilde{Z}^{\varepsilon,x}_s,\tilde{m}_s)ds\right|\le\frac{C_4}{\varepsilon^2},$$

where C_4 depends only on $\tilde{m}, X$ and the constants in Assumptions (A-1)–(A-4). Hence (cf. (2.2)), there exists a constant C_2 (again, depending on the same quantities only) such that, uniformly in $|x|<1$,

$$\rho_1^\varepsilon(x)\ge\exp\left[-\frac{C_2}{\varepsilon^2}\right]. \tag{4.6}$$

Equations (4.6) and (4.5) complete the proof of the lemma, once we prove (4.3).

Toward this end, assume without loss of generality that $x>0$, and set $\hat{h} = 2\sup_{y>0} h'(y)$. Using the Itô formula, one has

$$\begin{aligned}&\tilde{Z}_t^{\varepsilon,x} e^{\hat{h}t}\\ &= x + \int_0^t \left(\hat{h}\tilde{Z}_s^{\varepsilon,x} - h(\tilde{Z}_s^{\varepsilon,x}) + \tilde{m}_s h'(\tilde{Z}_s^{\varepsilon,x}) - \varepsilon b(\tilde{Z}_s^{\varepsilon,x})\right) e^{\hat{h}s} ds + \sqrt{\varepsilon}\int_0^t e^{\hat{h}s} d\tilde{W}_s.\end{aligned} \tag{4.7}$$

Hence, denoting $C_3 = |||\tilde{m}|||\sup_x h'(x)$, it follows that the event $\{\tau<\eta\}$ is contained in the event

$$\left\{\sup_{t\in(0,\eta)}\left|\sqrt{\varepsilon}\int_0^t e^{\hat{h}s} d\tilde{W}_s\right| \geq x - C_3\frac{e^{\hat{h}\eta}-1}{\hat{h}} - \frac{xe^{\hat{h}\eta}}{2}\right\} \subset \left\{\sup_{t\in(0,\eta)}\left|\sqrt{\varepsilon}\int_0^t e^{\hat{h}s} d\tilde{W}_s\right| \geq \frac{x}{4}\right\} =: B$$

if one chooses η small enough and x large enough. We have that

$$\mathbb{P}(B) \leq 4\exp\left(-\frac{Cx^2}{\varepsilon}\right)$$

for some constant C, which completes the proof of (4.3). □

Proof of Lemma 3.4. We prove only (3.7), the proof of (3.8) being similar. All we need to show is that for all $\varepsilon\leq\varepsilon_0$, $|x|\leq M_1/\sqrt{\varepsilon}$, and some M_2,

$$\mathbb{E}\left[L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|>M_2/\varepsilon\}}\right] \leq \mathbb{E}\left[L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|\leq M_2/\varepsilon\}}\right]. \tag{4.8}$$

We first bound the left-hand side of (4.8) for $\varepsilon\leq 1$. Recall the function H introduced in (4.1), and apply Itô's formula to develop $H(\tilde{Z}_t^{\varepsilon,x})$ between $t=0$ and $t=1/\varepsilon$, obtaining

$$\begin{aligned}\log L_\varepsilon\left(x,\frac{1}{\varepsilon}\right) - \frac{H(x)}{\varepsilon} &= -\frac{1}{2\varepsilon}\int_0^{1/\varepsilon}|h(\tilde{Z}_t^{\varepsilon,x}) - h(\tilde{m}_t)|^2 dt - \frac{1}{2\varepsilon}\int_0^{1/\varepsilon}|h(\tilde{Z}_t^{\varepsilon,x})|^2 dt\\ &\quad + \frac{1}{\sqrt{\varepsilon}}\int_0^{1/\varepsilon} h(\tilde{Z}_t^{\varepsilon,x}) d\tilde{W}_t + \int_0^{1/\varepsilon} g_{3,\varepsilon}(\tilde{Z}_t^{\varepsilon,x},\tilde{m}_t) dt + \log p_0(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x}),\end{aligned}$$

where

$$g_{3,\varepsilon}(z,m) = g_1(z,m) - b(z)h(z) + \frac{1}{2}h'(z) + \frac{1}{2\varepsilon}m^2(h'(z))^2.$$

Note that $\log p_0(\cdot)$ is bounded above, and

$$|g_{3,\varepsilon}(z,\tilde{m}_t)| \leq C\left(\frac{1}{\varepsilon} + |z|\right).$$

Now since, for any $p>1$,

$$\mathbb{E}\left[\exp\left(-\frac{p^2}{2\varepsilon}\int_0^{1/\varepsilon}|h(\tilde{Z}_t^{\varepsilon,x})|^2 dt + \frac{p}{\sqrt{\varepsilon}}\int_0^{1/\varepsilon} h(\tilde{Z}_t^{\varepsilon,x}) d\tilde{W}_t\right)\right] = 1,$$

it follows from Hölder's inequality that for any $q > p > 1$ satisfying $1/p + 1/q = 1$,

$$
\begin{aligned}
& e^{-H(x)/\varepsilon}\ \mathbb{E}\left[L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|>M_2/\varepsilon\}}\right] \\
& \le \left(\mathbb{E}\left[\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|>M_2/\varepsilon\}}\exp\left(C\int_0^{1/\varepsilon}\left(\frac{1}{\varepsilon}+|\tilde{Z}_t^{\varepsilon,x}|^2\right)dt\right)\right.\right. \\
& \left.\left.\times\exp\left(-\frac{q}{2\varepsilon}\int_0^{1/\varepsilon}|h(\tilde{Z}_t^{\varepsilon,x})-h(\tilde{m}_t)|^2dt+\frac{p}{2\varepsilon}\int_0^{1/\varepsilon}|h(\tilde{Z}_t^{\varepsilon,x})|^2dt\right)\right]\right)^{1/q},
\end{aligned}
\tag{4.9}
$$

where $C > 0$. However, note that, due to $h' \ge h_0$, there exists a constant C depending on $|||\tilde{m}|||$ such that

$$
\sup_{z\in\mathbb{R},|m|\le|||\tilde{m}|||}|z|^2-\frac{q}{2}|h(z)-h(m)|^2+\frac{p}{2}|h(z)|^2\le C.
$$

Substituting this into (4.9), one deduces that

$$
e^{-H(x)/\varepsilon}\mathbb{E}\left[L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|>M_2/\varepsilon\}}\right]\le\left(\mathbb{E}\left[\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|>M_2/\varepsilon\}}\exp\left(\frac{C}{\varepsilon^2}\right)\right]\right)^{1/q}.
\tag{4.10}
$$

(Recall that the value of C may change from line to line!)

We prove below that, provided M_2 is large enough, there exists a $c > 0$ such that

$$
\mathbb{E}\left[\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|>M_2/\varepsilon\}}\right]\le\exp\left(-\frac{c}{\varepsilon^3}\right).
\tag{4.11}
$$

Combined with (4.10), this implies that, uniformly in $|x| \le M_1/\sqrt{\varepsilon}$,

$$
\mathbb{E}\left[L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|>M_2/\varepsilon\}}\right]\le\exp\left(-\frac{c}{\varepsilon^3}\right).
\tag{4.12}
$$

To see (4.11), let $H = \sup|h'|$, define $\theta_0 = 0$, and let

$$
\tau_i=\inf\left\{t>\theta_{i-1}:|\tilde{Z}_t^{\varepsilon,x}|>\frac{M_2}{2\varepsilon}\right\},\quad \theta_i=\inf\left\{t>\tau_i:|\tilde{Z}_t^{\varepsilon,x}|<\frac{M_2}{4\varepsilon}\right\}.
$$

Setting $f(z,m) = -h(z) + mh'(z) - \varepsilon b(z)$, we have that, for $|z| \in [M_2/4\varepsilon, M_2/\varepsilon]$, $t \le 1/\varepsilon$, and ε small enough, it holds that $h_0M_2/8\varepsilon \le |f(z,\tilde{m}_t)| \le 2HM_2/\varepsilon$ and $\mathrm{sign} f(z,\tilde{m}_t) = -\mathrm{sign}(z)$. Then, choosing $\eta = (16H)^{-1}$ for each i, it holds that

$$
\begin{aligned}
\mathbb{P}\left(\theta_i-\tau_i<\eta,\ \sup_{t\in[\tau_i,\theta_i]}|\tilde{Z}_t^{\varepsilon,x}|<\frac{M_2}{\varepsilon}\right)&\le\mathbb{P}\left(\sqrt{\varepsilon}\sup_{0\le t\le\eta}|W_t|\ge\frac{M_2}{4\varepsilon}-\frac{2H\eta M_2}{\varepsilon}\right)\\
&\le\mathbb{P}\left(\sqrt{\varepsilon}\sup_{0\le t\le\eta}|W_t|\ge\frac{M_2}{8\varepsilon}\right)\\
&\le\exp\left(-\frac{cM_2^2}{\varepsilon^3\eta}\right).
\end{aligned}
\tag{4.13}
$$

Similarly

$$
\begin{aligned}
\mathbb{P}\left(\theta_i-\tau_i\ge\eta,\ |\tilde{Z}_{\tau_i+\eta}^{\varepsilon,x}|\ge\frac{M_2}{2\varepsilon}\right)&\le\mathbb{P}\left(\sqrt{\varepsilon}W_\eta\ge\frac{h_0M_2\eta}{8\varepsilon}\right)\\
&\le\exp\left(-\frac{cM_2^2\eta}{\varepsilon^3}\right)
\end{aligned}
\tag{4.14}
$$

and

$$\mathbb{P}\left(\sup_{t\in[\tau_i,(\tau_i+\eta)\wedge\theta_i]}|\tilde{Z}_t^{\varepsilon,x}|>\frac{M_2}{\varepsilon}\right)\le\mathbb{P}\left(\sqrt{\varepsilon}\sup_{0\le t\le\eta}|W_t|\ge\frac{M_2}{2\varepsilon}\right)\le\exp\left(-\frac{cM_2^2}{\varepsilon^3\eta}\right). \tag{4.15}$$

Hence, using (4.13), (4.14), and (4.15),

$$\mathbb{E}\left[\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|>M_2/\varepsilon\}}\right]\le\frac{1}{\varepsilon\eta}\left(\exp\left(-\frac{cM_2^2}{\varepsilon^3\eta}\right)+\exp\left(-\frac{cM_2^2}{\varepsilon^3}\right)+\exp\left(-\frac{cM_2^2\eta}{\varepsilon^3\eta}\right)\right),$$

completing the proof of (4.11).

We now turn to the lower bound of the right-hand side of (4.8). Let, with $M_1'=M_1+1$,

$$\varepsilon_0=1\wedge\left(\frac{M_2}{M_1'}\right)^2.$$

For $\varepsilon\le\varepsilon_0$,

$$\left\{\|\tilde{Z}^{\varepsilon,x}\|\le\frac{M_1'}{\sqrt{\varepsilon}}\right\}\subset\left\{\|\tilde{Z}^{\varepsilon,x}\|\le\frac{M_2}{\varepsilon}\right\},$$

so that for some $c'>0$

$$\mathbb{E}\left[L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|\le M_2/\varepsilon\}}\right]\ge\mathbb{E}\left[L_\varepsilon\left(x,\frac{1}{\varepsilon}\right)\mathbf{1}_{\{\|\tilde{Z}^{\varepsilon,x}\|\le M_1'/\sqrt{\varepsilon}\}}\right]\ge\exp\left(-\frac{c'}{\varepsilon^{5/2}}\right)\mathbb{P}\left(\|\tilde{Z}^{\varepsilon,x}\|\le\frac{M_1'}{\sqrt{\varepsilon}}\right). \tag{4.16}$$

Finally (4.8) follows from (4.12), (4.16), and the estimate

$$\mathbb{P}\left(\|\tilde{Z}^{\varepsilon,x}\|\le\frac{M_1'}{\sqrt{\varepsilon}}\right)\ge\mathbb{P}\left(\sqrt{\varepsilon}\|\tilde{W}\|\le C\right)\ge c''>0. \qquad \square$$

Proof of Lemma 3.5. Note first that because of (A-4), there exists a constant $\kappa=\kappa(|||\tilde{m}|||)$ such that for all $z\notin[-\kappa,\kappa]$, all $\varepsilon<1/\kappa$, all $|m|\le|||\tilde{m}|||$, and all z',

$$\Delta(z,z',m)=-h(z)+h(z')+m[h'(z)-h'(z')]-\varepsilon[b(z)-b(z')]$$

satisfies $\operatorname{sign}(\Delta(z,z',m))=\operatorname{sign}(z'-z)$, while $|\Delta(z,z',m)|\ge h_0|z-z'|/2$.

Assume, without loss of generality, that $x<y$. Fix, for δ given, a smooth, even, nonnegative function $c(z)$ such that $c(|z|)$ is nonincreasing, $c(z)=\sqrt{\delta}$ for $|z|\le\kappa$, and $c(z)=0$ for $|z|>2\kappa$, with $||c'||\le10\sqrt{\delta}$. Define next the diffusions

$$\begin{aligned}d\xi_s^1&=[-h(\xi_s^1)+\tilde{m}_sh'(\xi_s^1)-\varepsilon b(\xi_s^1)+c(\xi_s^1)\mathbf{1}_{\{\tau>s\}}]ds+\sqrt{\varepsilon}dB_s,\quad\xi_0^1=x,\\ d\xi_s^2&=[-h(\xi_s^2)+\tilde{m}_sh'(\xi_s^2)-\varepsilon b(\xi_s^2)]dt+\sqrt{\varepsilon}dB_s,\quad\xi_0^2=y,\end{aligned}$$

where B is a Brownian motion independent of the process $\tilde{m}$, and $\tau=\min\{t:\xi_t^1=\xi_t^2\}\wedge1/\varepsilon$. Note that ξ^2 coincides in distribution with $\tilde{Z}^{\varepsilon,y}$, whereas the law of ξ^1

is absolutely continuous with respect to the law of $\tilde{Z}^{\varepsilon,x}$ with the Radon–Nikodym derivative given by

(4.17)
$$\Lambda = \exp\left(\frac{1}{\varepsilon}\int_0^\tau c(\xi_s^1)d\xi_s^1 - \frac{1}{2\varepsilon}\int_0^\tau c^2(\xi_s^1)ds - \frac{1}{\varepsilon}\int_0^\tau c(\xi_s^1)g(s,\xi_s^1)ds\right)$$
$$= \exp\left(\frac{1}{\varepsilon}[\bar{c}(\xi_\tau^1) - \bar{c}(\xi_0^1)] - \frac{1}{2\varepsilon}\int_0^\tau c^2(\xi_s^1)ds - \frac{1}{\varepsilon}\int_0^\tau c(\xi_s^1)g(s,\xi_s^1)ds - \frac{1}{2}\int_0^\tau c'(\xi_s^1)ds\right),$$

where $g(s,z) = -h(z) + \tilde{m}_s h'(z) - \varepsilon b(z)$ and $\bar{c}(z) = \int_0^z c(y)dy$.

Next, note that with $\zeta_s = \xi_s^1 - \xi_s^2$, and using that $x < y$, it holds that $\zeta_s \le 0$ for all s, while by definition, $|\zeta_0| \le \delta$. Hence, by the definition of $c(\cdot)$ and of κ, it holds that for all $\delta < \delta_1(\kappa, |||m|||)$,

$$\frac{d\zeta_s}{ds} \ge -\frac{h_0\zeta_s}{2} + \frac{c(\xi_s^1)\mathbf{1}_{s<\tau}}{2},$$

from which one concludes that $\zeta_s \ge -\delta e^{-hs/2}$. In particular, this implies that for all such δ,

$$\int_0^\tau c(\xi_s^1)\mathbf{1}_{\{\tau>s\}}ds = \int_0^\tau c(\xi_s^1)ds \le C\delta$$

for some constant $C = C(\kappa, |||\tilde{m}|||)$. Since $c(z) = 0$ for $|z| > 2\kappa$, and since $|g(s,z)|$ is bounded uniformly in $s \le 1/\varepsilon$ and $|z| \le 2\kappa$ (by a bound that depends only on $|||\tilde{m}|||$), the last inequality implies that

$$\left|\int_0^\tau c(\xi_s^1)g(s,\xi_s^1)ds\right| \le C\delta,$$

again for some constant C depending on $\kappa, |||\tilde{m}|||$ only. Finally, note that

$$\int_0^\tau c^2(\xi_s^1)ds \le \sqrt{\delta}\int_0^\tau c(\xi_s^1)ds \le C\delta^{3/2},$$

and that $|\bar{c}(z)| \le 2\kappa\sqrt{\delta}$. Substituting back into (4.17) and recalling that $\kappa = \kappa(|||\tilde{m}|||)$, one concludes the existence of a constant $C_2 = C_2(|||\tilde{m}|||)$ such that for all $\delta < \delta_1$,

$$e^{-C_2\sqrt{\delta}/\varepsilon} \le \Lambda \le e^{C_2\sqrt{\delta}/\varepsilon}. \tag{4.18}$$

Therefore, with $\mathbb{E}_B$ denoting expectation with respect to $B_\cdot$, and using the bound on Λ in the second inequality, and the Lipschitz property of g_1, g_2 together with the exponential decay of ζ_s in the third, and omitting the dependence on $\theta^{1/\varepsilon - t}\tilde{m}$

everywhere, it holds, for all $t > 1/2\varepsilon$, that

(4.19)
$$\begin{aligned}
&\mathbb{E}L_\varepsilon(x,t) \le 2\mathbb{E}L_\varepsilon(x,t)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||<M_2/\varepsilon\}}\\
&= 2\mathbb{E}_B\left(\mathbf{1}_{\{||\xi^1||<M_2/\varepsilon\}}\Lambda^{-1}\exp\left(I_\varepsilon(\xi_t^1,0)+\int_0^t\left(g_1(\xi_s^1,\tilde{m}_s)+\frac{1}{\varepsilon}g_2(\xi_s^1,\tilde{m}_s)\right)ds\right)\right)\\
&\le 2\mathbb{E}_B\Bigg(\mathbf{1}_{\{||\xi^2||<(M_2+1)/\varepsilon\}}\ \exp\Bigg(\frac{C_2\sqrt{\delta}}{\varepsilon}+I_\varepsilon(\xi_t^2+\zeta_t,0)\\
&\qquad\qquad +\int_0^t\left(g_1(\xi_s^2+\zeta_s,\tilde{m}_s)+\frac{1}{\varepsilon}g_2(\xi_s^2+\zeta_s,\tilde{m}_s)\right)ds\Bigg)\Bigg)\\
&\le 2\mathbb{E}_B\left(\exp\left(\frac{C_3\sqrt{\delta}}{\varepsilon}+I_\varepsilon(\xi_t^2,0)+\int_0^t\left(g_1(\xi_s^2,\tilde{m}_s)+\frac{1}{\varepsilon}g_2(\xi_s^2,\tilde{m}_s)\right)ds\right)\right)\\
&= 2\mathbb{E}\left(\exp\left(\frac{C_3\sqrt{\delta}}{\varepsilon}+I_\varepsilon(\tilde{Z}_t^{\varepsilon,y},0)+\int_0^t\left(g_1(\tilde{Z}_s^{\varepsilon,y},\tilde{m}_s)+\frac{1}{\varepsilon}g_2(\tilde{Z}_s^{\varepsilon,y},\tilde{m}_s)\right)ds\right)\right)\\
&= 2\exp\left(\frac{C_3\sqrt{\delta}}{\varepsilon}\right)\mathbb{E}L_\varepsilon(y,t)\\
&\le 4\exp\left(\frac{C_3\sqrt{\delta}}{\varepsilon}\right)\mathbb{E}\left(\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,y}||<M_2/\varepsilon\}}L_\varepsilon(y,t)\right),
\end{aligned}$$

yielding (3.9) for $x < y$ and $\delta < \delta_1$, with $g(\delta) = C_3\sqrt{\delta}$. Further, the same computation gives

$$4\mathbb{E}\left(L_\varepsilon(x,t)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||<M_2/\varepsilon\}}\right)\ge\exp\left(\frac{-C_3\sqrt{\delta}}{\varepsilon}\right)\mathbb{E}\left(L_\varepsilon(y,t)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,y}||<M_2/\varepsilon\}}\right),$$

yielding, by exchanging the roles of x and y, (3.9) for $x > y$ and $\delta < \delta_1$ with the same $g(\delta)$. Finally, for $\delta > \delta_1$, iterate this procedure to obtain (3.9) with $g(\delta) = C_3\sqrt{\delta\wedge\delta_1}\lceil\delta/\delta_1\rceil$. Substituting $y = X_1$ into the latter version of (3.9) then gives (3.10). □

Proof of Lemma 3.6. Throughout the proof, we fix once and for all the sequence T_ε. All constants C_i used in the proof may depend on the choice of the sequence but not explicitly on ε.

We begin with the proof of (3.11). Using Girsanov's theorem, one finds that with $\bar{Z}_t^{\varepsilon,x} = x + \sqrt{\varepsilon}\tilde{W}_t$,

$$\begin{aligned}
&\mathbb{E}\left[\bar{L}_\varepsilon(x,T)\mathbf{1}_{\{|\tilde{Z}_T^{\varepsilon,x}-z|<\delta\}}\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,x}||_T\le M_3/\varepsilon\}}\right]\\
&= \mathbb{E}\Bigg[\mathbf{1}_{\{|\bar{Z}_T^{\varepsilon,x}-z|<\delta\}}\mathbf{1}_{\{||\bar{Z}^{\varepsilon,x}||_T\le M_3/\varepsilon\}}\\
&\quad\cdot\exp\left(\frac{1}{\varepsilon}\int_0^T\left[-h(\bar{Z}_s^{\varepsilon,x})+\tilde{m}_s h'(\bar{Z}_s^{\varepsilon,x})-\varepsilon b(\bar{Z}_s^{\varepsilon,x})\right]d\bar{Z}_s^{\varepsilon,x}\right.\\
&\qquad\left.-\frac{1}{\varepsilon}\int_0^T\left(\frac{[h(\bar{Z}_s^{\varepsilon,x})-h(\tilde{m}_s)]^2}{2}+\frac{b^2(\bar{Z}_s^{\varepsilon,x})\varepsilon^2}{2}+\varepsilon b(\bar{Z}_s^{\varepsilon,x})h(\bar{Z}_s^{\varepsilon,x})\right.\right.\\
&\qquad\left.\left.-\varepsilon h'(\bar{Z}_s^{\varepsilon,x})b(\bar{Z}_s^{\varepsilon,x})\tilde{m}_s-\varepsilon g_1(\bar{Z}_s^{\varepsilon,x},\tilde{m}_s)\right)ds\right)\Bigg].
\end{aligned}\tag{4.20}$$

We consider the different terms in (4.20) separately. Note first that one may, exactly as in the course of the proof of Lemma 3.5, move from starting point x to starting point X_1 in the right-hand side of (4.20), with the effect of picking up a term bounded by $\exp(C|x|/\varepsilon)$ and widening the allowed region where $\bar{Z}_T^{\varepsilon,x}$ need to be; namely, for all $T_\varepsilon \geq T > 1$, the right-hand side of (4.20) is bounded by

$$\exp\left(\frac{C_1 + C_2|x|}{\varepsilon}\right) \mathbb{E}\Bigg[\mathbf{1}_{\{|\bar{Z}_T^{\varepsilon,X_1} - z| < \delta + |x| + |X_1|\}} \mathbf{1}_{\{||\bar{Z}^{\varepsilon,X_1}||_T \leq |x| + (M_3+1)/\varepsilon\}} \cdot \exp\left(\frac{1}{\varepsilon}\int_0^T \Big[-h(\bar{Z}_t^{\varepsilon,X_1}) + \tilde{m}_s h'(\bar{Z}_s^{\varepsilon,X_1}) - \varepsilon b(\bar{Z}_s^{\varepsilon,X_1})\Big] d\bar{Z}_s^{\varepsilon,X_1}\right)\Bigg]. \tag{4.21}$$

An integration by parts gives that

$$-\int_0^T h(\bar{Z}_t^{\varepsilon,X_1}) d\bar{Z}_t^{\varepsilon,X_1} = -\bar{\mathcal{J}}(\bar{Z}_T^{\varepsilon,X_1}, X_1) - h(X_1)(\bar{Z}_T^{\varepsilon,X_1} - X_1) + \frac{\varepsilon}{2}\int_0^T h'(\bar{Z}_t^{\varepsilon,X_1}) dt,$$

and hence, on the event $\{|\bar{Z}_T^{\varepsilon,X_1} - z| < \delta + |x| + |X_1|\}$, it holds that

$$-\int_0^T h(\bar{Z}_t^{\varepsilon,X_1}) d\bar{Z}_t^{\varepsilon,X_1} \leq -C(|z| - |x| - |X_1| - \delta)_+^2 + C. \tag{4.22}$$

Similarly, with $B(z) = \int_{X_1}^z b(x)dx$,

$$\int_0^T b(\bar{Z}_s^{\varepsilon,X_1}) d\bar{Z}_s^{\varepsilon,X_1} = B(\bar{Z}_T^{\varepsilon,X_1}) - \frac{\varepsilon}{2}\int_0^T b'(\bar{Z}_s^{\varepsilon,X_1}) ds \leq C(|z|^2 + |x|^2 + 1). \tag{4.23}$$

Finally, rewrite

$$\int_0^T \tilde{m}_s h'(\bar{Z}_s^{\varepsilon,X_1}) d\bar{Z}_s^{\varepsilon,X_1} = X_1 \int_0^T h'(\bar{Z}_s^{\varepsilon,X_1}) d\bar{Z}_s^{\varepsilon,X_1} + \int_0^T (\tilde{m}_s - X_1) h'(\bar{Z}_s^{\varepsilon,X_1}) d\bar{Z}_s^{\varepsilon,X_1}.$$

The first stochastic integral in the above expression is handled exactly as in (4.23), and substituting into (4.21) one concludes that the right-hand side of (4.20) is bounded by

$$\begin{aligned}
&\exp\left(\frac{C + C(|x| + |z|) - C(|z| - |x|)_+^2}{\varepsilon}\right) \mathbb{E}\left[\exp\left(\frac{1}{\varepsilon}\int_0^T (\tilde{m}_s - X_1) h'(\bar{Z}_s^{\varepsilon,X_1}) d\bar{Z}_s^{\varepsilon,X_1}\right)\right] \\
&\quad \leq \exp\left(\frac{C + C(|x| + |z|) - C(|z| - |x|)_+^2}{\varepsilon} + \frac{1}{2\varepsilon}\int_0^{T_\varepsilon} C|\tilde{m}_s - X_1|^2 ds\right) \\
&\quad \leq \exp\left(\frac{C + C(|x| + |z|) - C(|z| - |x|)_+^2}{\varepsilon}\right),
\end{aligned}$$

where in the last inequality we have used the last part of Lemma 2.1. This completes the proof of (3.11).

The proof of (3.12) proceeds along similar lines. The starting point is the change of measure leading to (4.20). Define the function

$$\Psi_t = \begin{cases} x + 2(X_1 - x)t, & t \leq \frac{1}{2}, \\ X_1, & T - \frac{1}{2} > t \geq \frac{1}{2}, \\ z + 2(z - X_1)(t - T), & T \geq t \geq T - \frac{1}{2}. \end{cases}$$

Let D denote the event

$$D := \left\{ \sup_{t \le T} |\bar{Z}_t^{\varepsilon,x} - \Psi_t| < \sqrt{\varepsilon} \right\}.$$

We will prove below that, for $|x - X_1| \le 1$ and $T < T_\varepsilon$, there exists a constant C independent of T and ε such that

$$\mathbb{P}(D) \ge e^{-C/\varepsilon}. \tag{4.24}$$

We can clearly bound the right-hand side of (4.20) from below by

$$\begin{aligned}
\mathbb{E}\Bigg[& \mathbf{1}_{\{|\bar{Z}_T^{\varepsilon,x} - z| < \delta\}} \mathbf{1}_{\{||\bar{Z}^{\varepsilon,x}||_T \le M_3/\varepsilon\}} \mathbf{1}_D \\
& \cdot \exp\left(\frac{1}{\varepsilon} \int_0^T \left[-h(\bar{Z}_t^{\varepsilon,x}) + \tilde{m}_s h'(\bar{Z}_s^{\varepsilon,x}) - \varepsilon b(\bar{Z}_s^{\varepsilon,x}) \right] d\bar{Z}_s^{\varepsilon,x} \right. \\
& \qquad - \frac{1}{\varepsilon} \int_0^T \left(\frac{[h(\bar{Z}_t^{\varepsilon,x}) - h(\tilde{m}_t)]^2}{2} + \frac{b^2(\bar{Z}_t^{\varepsilon,x})\varepsilon^2}{2} + \varepsilon b(\bar{Z}_s^{\varepsilon,x}) h(\bar{Z}_s^{\varepsilon,x}) \right. \\
& \qquad\qquad \left. \left. - \varepsilon h'(\bar{Z}_s^{\varepsilon,x}) b(\bar{Z}_s^{\varepsilon,x}) \tilde{m}_s - \varepsilon g_1(\bar{Z}_s^{\varepsilon,x}, \tilde{m}_s) \right) ds \right) \Bigg].
\end{aligned}$$

We now assume that (4.24) and $|z - X_1| \le 1$ hold. Then using the same integration by parts as in the proof of the upper bound, one concludes that the right-hand side of (4.20) is bounded from below by

$$\mathbb{E}\left[\mathbf{1}_D \exp\left(\frac{-C}{\varepsilon} + \frac{1}{\varepsilon} \int_0^T (\tilde{m}_s - X_1) h'(\bar{Z}_s^{\varepsilon,x}) d\bar{Z}_s^{\varepsilon,x} \right) \right]. \tag{4.25}$$

However, since

$$\mathrm{Var}\left(\int_0^T (\tilde{m}_s - X_1) h'(\bar{Z}_s^{\varepsilon,x}) d\bar{Z}_s^{\varepsilon,x} \right) \le C\varepsilon,$$

one gets, using Chebyshev's inequality, that

$$\mathbb{P}\left[\int_0^T (\tilde{m}_s - X_1) h'(\bar{Z}_s^{\varepsilon,x}) d\bar{Z}_s^{\varepsilon,x} < -c \right] \le \exp\left(-\frac{C_2 c^2}{\varepsilon} \right).$$

Hence,

$$\mathbb{P}\left[\int_0^T (\tilde{m}_s - X_1) h'(\bar{Z}_s^{\varepsilon,x}) d\bar{Z}_s^{\varepsilon,x} < -c | D \right] \le \frac{\exp\left(-\frac{C_2 c^2}{\varepsilon} \right)}{\mathbb{P}(D)} \le \frac{1}{2}$$

if c is chosen large, where in the last inequality we used (4.24). In particular, it follows that

$$\mathbb{E}\left[\exp\left(\frac{1}{\varepsilon} \int_0^T (\tilde{m}_s - X_1) h'(\bar{Z}_s^{\varepsilon,x}) d\bar{Z}_s^{\varepsilon,x} \right) \,\middle|\, D \right] \ge \exp\left(-\frac{C}{\varepsilon} \right)$$

for some $C > 0$. Substituting back into (4.25), the required lower bound follows.

It thus remains only to prove (4.24). This, however, is immediate from a martingale argument: first, perform the change of measure making $S_t := \bar{Z}_t^{\varepsilon,x} - \Psi_t$ into a Brownian motion of variance ε. Then, for $1 \le T \le T_\varepsilon$,

$$\mathbb{P}(D) = \mathbb{E}\left(\mathbf{1}_{\{\sup_{t\le T}|S_t|\le\sqrt{\varepsilon}\}} \exp\left(-\frac{1}{\varepsilon}\int_0^T \dot{\Psi}_t dS_t - \frac{1}{2\varepsilon}\int_0^T \dot{\Psi}_t{}^2 dt\right)\right).$$

Integrating the stochastic integral by parts and using that $\dot{\Psi}(t) = 0$ for $t \in (1/2, T - 1/2)$, (4.24) follows, which completes the proof of the lemma. □

Proof of (3.26). We let $\eta > 0$ as before. Note first that, by (4.5) and (4.6), there is a constant M depending on $|||\tilde{m}|||$ only such that

$$\limsup_{\varepsilon\to 0} \varepsilon \log \int_{[-M/\sqrt{\varepsilon}, M/\sqrt{\varepsilon}]^c} q_1^\varepsilon(x)dx = -\infty. \tag{4.26}$$

We may and will in what follows assume that $M = M_1$, where M_1 is defined in Lemma 3.3, and we use M_3 and M_2 as in Lemma 3.4.

Next, set ε_4 such that $\varepsilon_4 \log 2 < \eta/8$ and $\varepsilon \log(2M_3/\varepsilon\delta) \le \eta/8$ for $\varepsilon < \varepsilon_4$. Repeating the arguments in (3.21), without using the compact set $\mathcal{K}_1$, one has for $\varepsilon < \varepsilon_4$ and $|x| \le M_1/\sqrt{\varepsilon}$,

$$\begin{aligned}
\varepsilon \log \rho_1^\varepsilon(x) &\le -F(x,\tilde{m}_0) + \varepsilon \log \tilde{J}_\varepsilon(x) + \frac{\eta}{4} \quad \texttt{as in (3.21)} \\
&\le -F(x,\tilde{m}_0) + \varepsilon \log \sup_{|z|\le M_3/\varepsilon} \hat{J}_{\varepsilon,T}(x,z) + \frac{\eta}{2} \quad \texttt{by (3.13) and (3.15)} \\
&\le -F(x,\tilde{m}_0) + \frac{\eta}{2} + C_2 - C_3(|z|-|x|)_+^2 + C_5(|x|+|z|) \\
&\quad + \varepsilon \log \mathbb{E}\left[L_\varepsilon\left(X_1, \frac{1}{\varepsilon} - T, \theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,X_1}||_{1/\varepsilon - T} \le M_2/\varepsilon\}}\right].
\end{aligned} \tag{4.27}$$

A similar argument shows that for $|x - X_1| < 1$ and some constant C_6 depending only on $X, |||\tilde{m}|||$,

$$\begin{aligned}
\varepsilon \log \rho_1^\varepsilon(x) \ge &-F(X_1,X_1) - C_6 \\
&+ \varepsilon \log \mathbb{E}\left[L_\varepsilon\left(X_1, \frac{1}{\varepsilon} - T, \theta^T\tilde{m}\right)\mathbf{1}_{\{||\tilde{Z}^{\varepsilon,X_1}||_{1/\varepsilon - T} \le M_2/\varepsilon\}}\right].
\end{aligned} \tag{4.28}$$

Fixing now an L, and using as in (3.18) the uniform quadratic growth of $F(x,m)$ as $|x| \to \infty$ and $|m| < |||\tilde{m}|||$, one finds a compact set $\mathcal{K}^L$ such that

$$\begin{aligned}
&\sup_{|m|<|||\tilde{m}|||} \sup_{x\in(\mathcal{K}^L)^c, z\in\mathbb{R}} \frac{C_2}{\varepsilon} - \frac{C_3(|z|-|x|)_+^2}{\varepsilon} + \frac{C_5(|x|+|z|)}{\varepsilon} - F(x,m) \\
&\qquad \le -F(X_1,X_1) - \frac{C_6 + L}{\varepsilon},
\end{aligned} \tag{4.29}$$

and hence, from (4.27) and (4.28), for $x \in (\mathcal{K}^L)^c \cap [-M_1/\sqrt{\varepsilon}, M_1/\sqrt{\varepsilon}]$,

$$\varepsilon \log \rho_1^\varepsilon(x) \le \inf_{|y - X_1|\le 1} \varepsilon \log \rho_1^\varepsilon(y) - L. \tag{4.30}$$

Hence,

(4.31)

$$\begin{aligned}
\limsup_{\varepsilon\to 0}\varepsilon\log\int_{(\mathcal{K}^L)^c} q_1^\varepsilon(x)dx &= \limsup_{\varepsilon\to 0}\varepsilon\log\int_{(\mathcal{K}^L)^c\cap[-M_1/\sqrt{\varepsilon},M_1/\sqrt{\varepsilon}]} q_1^\varepsilon(x)dx \quad \texttt{by (4.26)}\\
&\le \limsup_{\varepsilon\to 0}\left[\varepsilon\log\int_{(\mathcal{K}^L)^c\cap[-M_1/\sqrt{\varepsilon},M_1/\sqrt{\varepsilon}]}\rho_1^\varepsilon(x)dx - \varepsilon\log\int_{[X_1-1,X_1+1]}\rho_1^\varepsilon(x)dx\right]\\
&\le \limsup_{\varepsilon\to 0}\left[\varepsilon\log\left(\frac{2M_1}{\sqrt{\varepsilon}}\right) + \inf_{|y-X_1|\le 1}\varepsilon\log\rho_1^\varepsilon(y) - L - \inf_{|y-X_1|\le 1}\varepsilon\log\rho_1^\varepsilon(y)\varepsilon\log 2\right]\\
&\qquad\qquad \texttt{by (4.30)}\\
&\le -L\,.
\end{aligned}$$

This completes the proof. □

Appendix. Derivation of (2.1). We first recall Picard's theorem [7, Proposition 4.2]: under the assumptions of the current paper and with the same notation, a version of the conditional unnormalized density is given by

$$\tilde{q}(1,x) = \exp\left\{\frac{1}{2\varepsilon^2}\int_0^1 h^2(\bar{m}_s)ds - \frac{1}{\varepsilon}F(x,\tilde{m}_0)\right\}\tilde{\mathbb{E}}'\left[\exp\rho_1^{y,x}\right], \tag{A.1}$$

where

$$\begin{aligned}
\rho_1^{x,y} &= \log p_0(\bar{X}_1^x) + \frac{1}{\varepsilon}F(\bar{X}_1^x,0) - \frac{1}{\varepsilon}\int_0^1 h(\bar{m}_s)d\bar{X}_s^x - \frac{1}{\varepsilon}\int_0^1 h(\bar{X}_s^x)b(\bar{m}_s)ds\\
&\quad + \frac{1}{\varepsilon}\int_0^1 \bar{m}_s h'(\bar{X}_s^x)d\bar{X}_s^x + \frac{1}{2\varepsilon}\int_0^1 \bar{m}_s h''(\bar{X}_s^x)ds\\
&\quad + \frac{1}{\varepsilon}\int_0^1\left[b(\bar{X}_s^x)(h(\bar{X}_s^x)-h(\bar{m}_s)) - \frac{1}{2}h'(\bar{X}_s^x) - \varepsilon b'(\bar{X}_s^x)\right]ds,\\
d\bar{X}_s^x &= -\frac{1}{\varepsilon}(h(\bar{X}_s^x)-h(\bar{m}_s))ds - b(\bar{X}_s^x)ds + dW_s, \quad \bar{X}_0^x = x,
\end{aligned}$$

$W_\cdot$ is a Brownian motion, and $\tilde{\mathbb{E}}'$ denotes expectation with respect to this Brownian motion. Performing a time change $t\mapsto\varepsilon t$ and setting $\tilde{W}_t = \frac{1}{\sqrt{\varepsilon}}W_{\varepsilon t}$, we have that $\tilde{W}_t$ is again a standard Brownian motion and, with $\bar{X}_t^{\varepsilon,x} = \bar{X}_{\varepsilon t}^x$,

$$\begin{aligned}
\rho_1^{x,y} &= \log p_0(\bar{X}_{1/\varepsilon}^{\varepsilon,x}) + \frac{1}{\varepsilon}F(\bar{X}_{1/\varepsilon}^{\varepsilon,x},0) - \frac{1}{\varepsilon}\int_0^{1/\varepsilon} h(\tilde{m}_s)d\bar{X}_s^{\varepsilon,x} - \int_0^{1/\varepsilon} h(\bar{X}_s^{\varepsilon,x})b(\tilde{m}_s)ds\\
&\quad + \frac{1}{\varepsilon}\int_0^{1/\varepsilon}\tilde{m}_s h'(\bar{X}_s^{\varepsilon,x})d\bar{X}_s^{\varepsilon,x} + \frac{1}{2}\int_0^{1/\varepsilon}\tilde{m}_s h''(\bar{X}_s^{\varepsilon,x})ds\\
&\quad + \int_0^{1/\varepsilon}\left[b(\bar{X}_s^{\varepsilon,x})(h(\bar{X}_s^{\varepsilon,x})-h(\tilde{m}_s)) - \frac{1}{2}h'(\bar{X}_s^{\varepsilon,x}) - \varepsilon b'(\bar{X}_s^{\varepsilon,x})\right]ds,\\
d\bar{X}_s^{\varepsilon,x} &= -(h(\bar{X}_s^{\varepsilon,x})-h(\tilde{m}_s))ds - \varepsilon b(\bar{X}_s^{\varepsilon,x})ds + \sqrt{\varepsilon}d\tilde{W}_s\,, \quad \bar{X}_0^{\varepsilon,x} = x,
\end{aligned}$$

and

$$\tilde{q}(1,x) = \exp\left\{\frac{1}{2\varepsilon}\int_0^{1/\varepsilon} h^2(\tilde{m}_s))ds - \frac{1}{\varepsilon}F(x,\tilde{m}_0)\right\}\tilde{\mathbb{E}}\left[\exp\rho_1^{y,x}\right], \tag{A.2}$$

where the expectation now is with respect to the Brownian motion $\tilde{W}_t$.

Observe next that, by Girsanov's theorem, the law of the process $\bar{X}_t^{\varepsilon,x}$ is absolutely continuous with respect to that of the process $\tilde{Z}_t^{\varepsilon,x}$, with the Radon–Nikodym derivative given by

$$
\begin{aligned}
e^{\Lambda} = \exp\Bigg[& \frac{1}{\varepsilon}\int_0^{1/\varepsilon} [h(\tilde{m}_s) - \tilde{m}_s h'(\tilde{Z}_s^{\varepsilon,x})] d\tilde{Z}_s^{\varepsilon,x} \\
& - \frac{1}{2\varepsilon}\int_0^{1/\varepsilon} [h(\tilde{Z}_s^{\varepsilon,x}) - h(\tilde{m}_s) + \varepsilon b(\tilde{Z}_s^{\varepsilon,x})]^2 ds \\
& + \frac{1}{2\varepsilon}\int_0^{1/\varepsilon} [h(\tilde{Z}_s^{\varepsilon,x}) - \tilde{m}_s h'(\tilde{Z}_s^{\varepsilon,x}) + \varepsilon b(\tilde{Z}_s^{\varepsilon,x})]^2 ds \Bigg].
\end{aligned} \tag{A.3}
$$

Hence, with $\mathbb{E}$ denoting expectations with respect to the Brownian motion $\tilde{W}_t$ appearing in the definition of $\tilde{Z}_t^{\varepsilon,x}$, (A.2) transforms to

$$
\tilde{q}(1,x) = \exp\left\{ \frac{1}{2\varepsilon}\int_0^{1/\varepsilon} h^2(\tilde{m}_s))ds - \frac{1}{\varepsilon}F(x,\tilde{m}_0) \right\} \mathbb{E}\exp[\Lambda_1(x)],
$$

where

$$
\begin{aligned}
\Lambda_1(x) = {} & \Lambda + \log p_0(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x}) + \frac{1}{\varepsilon}F(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x},0) - \frac{1}{\varepsilon}\int_0^{1/\varepsilon} h(\tilde{m}_s) d\tilde{Z}_s^{\varepsilon,x} - \int_0^{1/\varepsilon} h(\tilde{Z}^{\varepsilon,x}) b(\tilde{m}_s) ds \\
& + \frac{1}{\varepsilon}\int_0^{1/\varepsilon} \tilde{m}_s h'(\tilde{Z}_s^{\varepsilon,x}) d\tilde{Z}_s^{\varepsilon,x} + \frac{1}{2}\int_0^{1/\varepsilon} \tilde{m}_s h''(\tilde{Z}_s^{\varepsilon,x}) ds \\
& + \int_0^{1/\varepsilon} \left[b(\tilde{Z}_s^{\varepsilon,x})(h(\tilde{Z}_s^{\varepsilon,x}) - h(\tilde{m}_s)) - \frac{1}{2}h'(\tilde{Z}_s^{\varepsilon,x}) - \varepsilon b'(\tilde{Z}_s^{\varepsilon,x}) \right] ds \\
= {} & \log p_0(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x}) + \frac{1}{\varepsilon}F(\tilde{Z}_{1/\varepsilon}^{\varepsilon,x},0) + \int_0^{1/\varepsilon} g_1(\tilde{Z}_s^{1/\varepsilon},\tilde{m}_s) ds + \frac{1}{\varepsilon}\int_0^{1/\varepsilon} g_2(\tilde{Z}_s^{1/\varepsilon},\tilde{m}_s) ds.
\end{aligned}
$$

Since $\int_0^{1/\varepsilon} h^2(\tilde{m}_s)ds$ does not depend on x, taking

$$
\rho_1^{\varepsilon}(x) = \tilde{q}(1,x)\exp\left\{ -\frac{1}{2\varepsilon}\int_0^{1/\varepsilon} h^2(\tilde{m}_s))ds \right\}
$$

gives a version of the unnormalized conditional density that coincides with (2.2). □

Acknowledgment. We thank Ki-Jung Lee for a careful reading of a preliminary version of this paper. We also thank an anonymous referee for a detailed reading of the paper and many useful and important comments.

REFERENCES

[1] R. Atar, *Exponential stability for nonlinear filtering of diffusion processes in a noncompact domain*, Ann. Probab., 26 (1998), pp. 1552–1574.

[2] R. Azencott, *Grande deviations et applications*, VIII, *Summer school of probability (St. Flour)*, Lecture Notes in Math. 774, Springer, New York, 1980.

[3] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., Springer, New York, 1998.

[4] O. Hijab, *Asymptotic Bayesian estimation of a first order equation with small diffusion*, Ann. Probab., 12 (1984), pp. 890–902.

[5] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North–Holland, Amsterdam, 1981.

[6] E. Pardoux, *Filtrage non linéaire et équations aux dérivées partielles stochastiques associées*, in Ecole d'Été de Probabilité de St. Flour XIX, Lecture Notes in Math. 1464, Springer, New York, 1991, pp. 67–163.

[7] J. Picard, *Nonlinear filtering of one-dimensional diffusions in the case of a high signal-to-noise ratio*, SIAM J. Appl. Math., 46 (1986), pp. 1098–1125.

[8] J. Picard, *Nonlinear filtering and smoothing with high signal-to-noise ratio*, in Stochastic Processes in Physics and Engineering (Bielefeld, 1986), Math. Appl., 42, Reidel, Dordrecht, Boston, 1988, pp. 237–251.

[9] O. Zeitouni and M. Zakai, *On the optimal tracking problem*, SIAM J. Control. Optim, 30 (1992), pp. 426–439; *Erratum*, SIAM J. Control. Optim, 32 (1994), p. 1194.

[10] O. Zeitouni, *Approximate and limit results for nonlinear filters with small observation noise: The linear sensor and constant diffusion coefficient case*, IEEE Trans. Automat. Control, 33 (1988), pp. 595–599.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1298–1312

WELL-POSEDNESS OF NONCONVEX INTEGRAL FUNCTIONALS*

SILVIA VILLA†

Abstract. We find a sufficient condition which does not involve global convexity, guaranteeing well-posedness in a strong sense of the minimization of a multiple integral on the Sobolev space $W^{1,1}(\Omega;\mathbb{R}^m)$ with boundary datum equal to zero. In particular we apply our result to some nonconvex problems recently studied: functionals depending only on the gradient and radially symmetric functionals.

Key words. calculus of variations, well-posedness, nonconvex integrals, extreme points

AMS subject classifications. 49K40, 49A50

DOI. 10.1137/S0363012903437289

1. Introduction.

We consider well-posedness properties of the problem of minimizing the integral

$$J(u) = \int_\Omega L(x, u(x), Du(x))\, dx \tag{1.1}$$

in some class of vector-valued Sobolev functions with boundary datum equal to zero.

There are at least two different concepts of well-posedness. We recall that a minimization problem is called Tikhonov well-posed if there exists a unique minimizer to which every minimizing sequence converges. On the other hand, a problem is well-posed by perturbations if the solution depends continuously on the problem's data: the idea is to embed the given minimization problem J defined by (1.1) in a suitable family of problems of the same kind depending on a parameter which varies in a convenient topological space $\mathcal{A}$. In this manner, the functional that we are minimizing corresponds to a fixed parameter $a \in \mathcal{A}$, and we can denote it by J_a. For every sequence $a_n \in \mathcal{A}$ such that $a_n \to a$ we consider the corresponding functional J_{a_n}, assuming that $\inf J_a$ and $\inf J_{a_n}$ are finite. We say that a sequence u_n is asymptotically minimizing for J_{a_n} if $J_{a_n}(u_n) - \inf J_{a_n} \to 0$. Now, more precisely, well-posedness by perturbations of J_a means that J_a has a unique minimizer u_0, that any asymptotically minimizing sequence converges to u_0, and that $\inf J_{a_n} \to \inf J_a$ for any $a_n \to a$. We observe in particular that if a minimization problem is well-posed by perturbations, it is also Tikhonov well-posed. See section 4 for rigorous definitions, [12] for a survey, and [24]. In our context we deal with perturbations of the integrand (see Corollary 4.12) and, in section 5, with perturbations of the boundary data, as was done in [8] and [23].

In the recent work [15] Ioffe and Zaslavski proved a variational principle, thanks to which it is possible to show that well-posedness by perturbations with respect to a suitable topology on the space of the integrands is a generic property. Moreover, in [22] a stronger result is proved: the set of ill-posed problems of the calculus of variations is not only of the first category, but it is also σ-porous.

*Received by the editors November 6, 2003; accepted for publication (in revised form) May 12, 2004; published electronically January 5, 2005.
http://www.siam.org/journals/sicon/43-4/43728.html
†Department of Mathematics "Ulisse Dini," University of Florence, Viale Morgagni 67a, 50134 Florence, Italy (villa@math.unifi.it).

Surprisingly, the classical hypotheses of convexity and superlinearity at infinity are not needed generically to have well-posedness, and therefore existence and uniqueness of the minimizer, for an integral functional.

In this context it is useful to find a characterization of well-posed problems, or at least some sufficient conditions which imply well-posedness without involving global convexity. We note that uniqueness of the solution does not imply well-posedness of classical problems of the calculus of variations: examples of ill-posed problems with a unique solution can be found in [8].

There are not many well-posedness results in the classical calculus of variations; however, results on this subject are proved in [4], [8], [12], [23], and [24] with respect to different types of perturbations.

In [4] a Tikhonov well-posed one-dimensional integral functional is considered. In particular it is proved that under suitable hypotheses this is enough to get the strong convergence in some Sobolev space of the asymptotically minimizing sequences corresponding to perturbations of the integrand with respect to the variational convergence not involving the derivative.

In [23] sufficient and necessary conditions are obtained for well-posedness by perturbations of the boundary data for integral functionals depending only on the gradient. The integrand is assumed to be continuous in the one-dimensional case and, moreover, to have polynomial growth in the multidimensional one.

We find a sufficient condition of local character for Tikhonov well-posedness of an integral functional defined on the Sobolev space $W_0^{1,1}(\Omega;\mathbb{R}^m)$ by extending some results of [18] and [21]. More precisely we extend the definition of strict convexity at a point given by Sychev for continuous functions to the case of lower semicontinuous functions. Then we use this local definition to generalize the results of Visintin in [21] without global strict convexity assumptions, and to get new well-posedness results. We also remark that results analogous to the ones contained in [21] were obtained by Olech in [17].

The strategy of the proof is completely different from that in [18]; in fact, we investigate the geometric properties of the point at which the function is strictly convex, and this fact allows us to drop out the hypothesis of continuity of the integrand. This is remarkable because in this way it is possible to consider integrands also assuming the value $+\infty$ and therefore to treat problems with constraints.

Well-posedness by perturbations of the same problem follows from [19], where the equivalence between Tikhonov well-posedness and well-posedness by perturbations is proved under suitable hypotheses.

This paper is organized as follows. After introducing in section 2 definitions and preliminaries, in section 3 we state some properties of convex and strictly convex functions at a point. In fact we require these hypotheses on the integrand in order to get well-posedness of the integral functional.

In particular, Lemma 3.4 is useful because it relates different hypotheses made in the literature in order to obtain strong convergence of a sequence which is only weakly convergent.

In section 4 we obtain the main result, Corollary 4.11, which establishes in particular that if the integrand is strictly convex at the point $Du(x)$ for almost every $x \in \Omega$ (where u is the minimizer), then the minimizing sequences are in fact strongly convergent to u in $W_0^{1,1}(\Omega;\mathbb{R}^m)$. This is a consequence of Theorem 4.2, which partially extends Theorem 2 of [18] to lower semicontinuous integrands. In other words we get that every coercive integral functional with a unique minimizer u is Tikhonov

well-posed and well-posed by perturbations if the integrand is strictly convex at the point $Du(x)$ for almost every $x \in \Omega$. Moreover, as a consequence of the representation formula for the relaxed functional given, for instance, in [16], we give a sufficient condition for well-posedness which depends only on the problem's data, namely, on the integrand L and on its convex regularization with respect to the last variable, denoted by L^{**}. More precisely in Theorem 4.13 we prove that an integral functional is Tikhonov well-posed if it has a unique minimizer and it is strictly convex at every point belonging to the set where the integrands L and L^{**} coincide.

Finally, in section 5, we apply our results to study well-posedness of two classes of integral functionals. The first is the case of integrands depending only on the gradient, and with linear boundary datum: for this class we extend a previous result of [8] and we prove stability of the solution by perturbations of the boundary data. The second is the case of radially symmetric functionals, treated, for instance, in [7] and [10]. For this class we prove that the hypotheses which ensure existence and uniqueness of the solution guarantee also well-posedness in the sense of Tikhonov and therefore also well-posedness by perturbations with respect to the bounded Hausdorff topology (which we review in section 4).

2. Definitions and preliminaries. As we said in the introduction, very few results, not only regarding well-posedness but also existence and uniqueness of the minimizer of an integral functional, are known without the classical hypotheses of convexity and superlinearity of the integrand. As is well known, these hypotheses guarantee lower semicontinuity of the functional (1.1) and compactness of its sublevel sets with respect to the weak topology of the relevant Sobolev space.

Since we aim to avoid the global convexity requirement and to replace it with a local property, we start with the following definition, which is equivalent to the one given in [18] in the case of continuous functions.

DEFINITION 2.1. *Let U be a closed convex subset of $\mathbb{R}^n$ and $u_0 \in U$. We say that a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is* convex at u_0 *with respect to the set U if*

$$\sum_{i=1}^{m} c_i f(v_i) \geq f(u_0) \tag{2.1}$$

for every $m > 0$, and for every $v_i \neq u_0$ and $c_i > 0$ such that $v_i \in U$, $\sum_{i=1}^m c_i = 1$, and $\sum_{i=1}^m c_i v_i = u_0$.

If inequality (2.1) *is always strict, we say that the function f is* strictly convex at u_0 *with respect to U.*

It is very easy to find examples of convex and strictly convex functions at a point which are not globally convex: e.g., the function $x \mapsto \sqrt{|x|}$ is convex at the point 0 with respect to $\mathbb{R}$.

In the case that $U = \mathbb{R}^n$ we do not specify the set with respect to which f is convex.

Remark 2.2. In the definition of strict convexity at a point we require

(a) $\sum_{i=1}^m c_i f(v_i) > f(u_0)$ for every $m > 0$, and for every $v_i \neq u_0$ and $c_i > 0$ such that $\sum_{i=1}^m c_i = 1$ and $\sum_{i=1}^m c_i v_i = u_0$. We observe that it would be equivalent to require that

(b) $\sum_{i=1}^m c_i f(v_i) > f(u_0)$ for every $m > 0$, for every v_i, $i = 1, \ldots, m$, such that there exists at least some $v_j \neq u_0$, and for every $c_i > 0$ such that $\sum_{i=1}^m c_i = 1$ and $\sum_{i=1}^m c_i v_i = u_0$.

The existence theory of minimizers for nonconvex integrands is based on the idea of relaxation (see, e.g., [16]). In this context we recall the definition of the convex regularization, or the second conjugate, of a given function f.

DEFINITION 2.3. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. The* convex regularization of f *(or* second conjugate *of f) is the function f^{**} defined by*

$$f^{**}(x) = \sup\{g(x) \mid g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\},\ g \text{ l.s.c. and convex},\ g \le f\}.$$

If the function f^{**} is proper, the following equality holds (see Proposition 3.2 of [13]):

$$\overline{\text{co}}(\text{epi} f) = \text{epi} f^{**}. \tag{2.2}$$

Here $\overline{\text{co}}$ denotes the closed convex hull. See [13] for further properties of f^{**}.

3. Properties of convex functions at a point. In this section we collect some general results about convex and strictly convex functions at a point that we will need later in the case of integrands arising in the calculus of variations.

We begin with a lemma which establishes a formula for computing the convex regularization of the restriction to an arbitrary closed ball of a given function defined on $\mathbb{R}^n$, useful in characterizing the points where f and f^{**} coincide, according to the following corollary. In the following we denote by f_r the function defined on $\mathbb{R}^n$ by setting $f_r = f$ on the closed ball $\overline{B}(u_0, r)$ with center u_0 and radius r and $f_r = +\infty$ otherwise.

LEMMA 3.1. *Let $f : \mathbb{R}^n \to [0, +\infty]$ be lower semicontinuous. Then the following formula holds:*

$$(f_r)^{**}(u) = \min\left\{ \sum_{i=1}^{m} c_i f(v_i) \;\middle|\; m \in \mathbb{N}; c_i > 0; v_i \in \overline{B}(u_0, r); \sum_{i=1}^{m} c_i = 1; \sum_{i=1}^{m} c_i v_i = u \right\} \tag{3.1}$$

for every $u_0 \in \mathbb{R}^n$ and for every $r > 0$.

Proof. Fix $u_0 \in \mathbb{R}^n$ and $r > 0$. Since f is lower semicontinuous, then f_r is lower semicontinuous too. In addition f_r is a superlinear function. From [13, Lemma 3.3, p. 280], it follows that formula (3.1) holds. □

A direct consequence of the preceding lemma is the following corollary.

COROLLARY 3.2. *Let $f : \mathbb{R}^n \to [0, +\infty]$ be lower semicontinuous. Then f is convex at $u_0 \in \mathbb{R}^n$ if and only if $f(u_0) = f^{**}(u_0)$.*

Proof. Suppose that f is convex at $u_0 \in \mathbb{R}^n$. Fix $r > 0$ and let $m \in \mathbb{N}$, $c_i > 0$, and $v_i \in \overline{B}(u_0, r)$ such that $\sum_{i=1}^m c_i = 1$ and $\sum_{i=1}^m c_i v_i = u_0$. Then by the definition of convexity at a point we get $f(u_0) \le \sum_{i=1}^m c_i f(v_i)$. Passing to the infimum on the convex combinations of points lying in $\overline{B}(u_0, r)$ and using Lemma 3.1, we obtain the inequality $f(u_0) \le (f_r)^{**}(u_0)$. Passing to the limit for $r \to +\infty$ (see, for instance, [13, Lemma 3.1, p. 329]), we get $f(u_0) \le f^{**}(u_0)$. Since $f^{**} \le f$, we conclude that $f(u_0) = f^{**}(u_0)$.

The other implication follows immediately from Lemma 3.1. □

We now recall the definition of an extreme point, since we will need this concept in what follows.

DEFINITION 3.3. *Let C be a convex subset of $\mathbb{R}^n$ and let $z \in C$. We say that z is an* extreme point of C *if $\lambda x + (1-\lambda)y = z$ for $\lambda \in (0,1)$ and $x, y \in C$ imply $x = y = z$.*

It's easy to verify that

$$z \text{ is an extreme point of } C \iff C \setminus \{z\} \text{ is convex.} \tag{3.2}$$

The next lemma shows the relation between strict convexity of a function at a point u_0 and extremality of the corresponding point of $(u_0, f(u_0))$ with respect to the epigraph of the second conjugate of f.

One implication is always true: if the point $(u_0, f^{**}(u_0))$ is extremal for $\mathrm{epi}f^{**}$, then the function f is strictly convex at the point u_0.

The converse does not hold in general, although it is true that the point $(u_0, f(u_0))$ is extremal for the epigraph of the convex regularization of the function f restricted to any closed ball of center u_0 and arbitrary radius.

LEMMA 3.4. *Let $f : \mathbb{R}^n \to [0, +\infty]$ be lower semicontinuous and proper. Then the following hold:*

1. *If $(u_0, f^{**}(u_0))$ is an extreme point of $\mathrm{epi}f^{**}$, then f is strictly convex at u_0.*
2. *If f is strictly convex at u_0, then $(u_0, f(u_0))$ is an extreme point of $\mathrm{epi}(f_r)^{**}$ for all $r > 0$.*

Proof. Suppose that $(u_0, f^{**}(u_0))$ is an extreme point of $\mathrm{epi}f^{**}$. By Remark 5.3 of [10] we get $f(u_0) = f^{**}(u_0)$.

Let $v_i \in \mathbb{R}^n$, $c_i > 0$ for $i = 1, \dots, m$ such that $\sum_{i=1}^m c_i v_i = u_0$. Since $f^{**} \le f$ we have that $(v_i, f(v_i)) \in \mathrm{epi}f^{**}$, and by the convexity of $\mathrm{epi}f^{**}$ it follows that

$$\sum_{i=1}^{m} c_i f(v_i) \ge f(u_0).$$

If the equality occurs, we get $v_i = u_0$ for every i, because $(u_0, f(u_0))$ is an extreme point. This implies that f is strictly convex at u_0.

Conversely, suppose that f is strictly convex at u_0. By definition we immediately obtain that also f_r is strictly convex at u_0 for every $r > 0$. According to (3.2) it's enough to prove that the set $\mathrm{epi}(f_r)^{**} \setminus (u_0, f(u_0))$ is a convex set. We will show that

(i) $\mathrm{co}(\mathrm{epi}f_r) \setminus (u_0, f(u_0))$ is a convex set;
(ii) $\mathrm{co}(\mathrm{epi}f_r)$ is closed.

Then the thesis will follow from the equality $\mathrm{epi}(f_r)^{**} = \overline{\mathrm{co}}(\mathrm{epi}f_r)$.

Let us first prove assertion (i). Fix $(v, s), (w, t) \in \mathrm{co}(\mathrm{epi}f_r) \setminus (u_0, f(u_0))$ and let $\lambda \in (0, 1)$. Since (v, s) and (w, t) are in the convex envelope of a subset of $\mathbb{R}^{n+1}$, by Carathéodory's theorem (see [11, p. 42]) they can be written as a convex combination of $(n + 2)$ points of $\mathrm{epi}f_r$:

$$(v, s) = \sum_{i=1}^{n+2} c_i(v_i, s_i); \quad (w, t) = \sum_{i=1}^{n+2} d_i(w_i, t_i). \tag{3.3}$$

Now consider the point $\lambda(v, s) + (1 - \lambda)(w, t)$. Certainly this point is in $\mathrm{co}(\mathrm{epi}f_r)$ by the definition of a convex envelope. We have to prove that it is not $(u_0, f(u_0))$. For this purpose assume that $\lambda v + (1 - \lambda)w = u_0$; we are going to show that $\lambda s + (1 - \lambda)t \ne f(u_0)$. From (3.3) we get

$$\sum_{i=1}^{n+2} \lambda c_i v_i + (1 - \lambda) d_i w_i = u_0; \tag{3.4}$$

i.e., u_0 is a convex combination of the points v_i and w_i, which are not all equal to u_0.

Recalling the strict convexity at the point u_0 and Remark 2.2, by (3.4) it follows that

$$\begin{aligned} f(u_0) &< \sum_{i=1}^{n+2} \lambda c_i f(v_i) + (1-\lambda) d_i f(w_i) \\ &\le \sum_{i=1}^{n+2} \lambda c_i s_i + (1-\lambda) d_i t_i \\ &= \lambda s + (1-\lambda) t. \end{aligned}$$

We obtain therefore that $\lambda(v,s) + (1-\lambda)(w,t)$ is in $\operatorname{co}(\operatorname{epi} f_r) \setminus \{(u_0, f(u_0))\}$ and assertion (i) is proved.

Now, it is enough to show that assertion (ii) holds.

Since f_r is lower semicontinuous and superlinear, (ii) is a consequence of [13, Lemma 3.3, p. 280] and the proof is complete. □

Remark 3.5. As we observed before, the converse of the first statement of the previous lemma is not true in general. Consider, for instance, the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} |x| & \text{if } |x| \le 1, \\ 1 & \text{if } |x| > 1. \end{cases}$$

Clearly f is strictly convex at 0, but $(0,0)$ is not extremal for $\operatorname{epi} f^{**} = \mathbb{R} \times [0, +\infty]$.

4. Well-posedness results. In this section we obtain a sufficient condition that guarantees well-posedness of the minimization problem of an integral functional. First of all we recall the definition of a normal integral.

DEFINITION 4.1. *Let $g : \Omega \times \mathbb{R}^m \times \mathbb{R}^l \to [-\infty, +\infty]$, with Ω an open subset of $\mathbb{R}^n$. We say that g is a* normal integrand *if*

(i) *for almost every $x \in \Omega$, $g(x, \cdot, \cdot)$ is lower semicontinuous;*

(ii) *there exists a Borel function $\hat{g} : \Omega \times \mathbb{R}^m \times \mathbb{R}^l \to [-\infty, +\infty]$ such that $\hat{g}(x, \cdot, \cdot) = g(x, \cdot, \cdot)$ for almost every $x \in \Omega$.*

More precisely we consider the problem cited in the introduction of minimizing the functional

$$J(u) = \int_\Omega L(x, Du(x))\, dx \tag{4.1}$$

in the class of functions $u \in W_0^{1,1}(\Omega; \mathbb{R}^m)$.

We always assume that Ω is an open and bounded subset of $\mathbb{R}^n$ and that $L : \Omega \times \mathbb{R}^{nm} \to [0, +\infty]$ is a normal integrand possibly taking the value $+\infty$. When the integrand is not summable, we define $J(u) = +\infty$.

In order to study the semicontinuity of (4.1) at a point and the convergence of minimizing sequences, we also deal with the more general functional

$$G(u, \xi) = \int_\Omega g(x, u(x), \xi(x))\, dx, \tag{4.2}$$

where $g : \Omega \times \mathbb{R}^m \times \mathbb{R}^l \to [0, +\infty]$ is a normal integrand, $u \in L^1(\Omega; \mathbb{R}^m)$, and $\xi \in L^1(\Omega; \mathbb{R}^l)$. In the rest of the paper we will denote by $\rightharpoonup$ the weak convergence in the spaces $L^1(\Omega; \mathbb{R}^m)$ and $W^{1,1}(\Omega; \mathbb{R}^m)$.

Now we are able to state a partial generalization of Theorem 2 of [18] to the case of lower semicontinuous integrands.

THEOREM 4.2. *Let $L_1 : \Omega \times \mathbb{R}^l \to [0, +\infty]$ be a normal integrand and let J_1 be defined by $J_1(\xi) := \int_\Omega L_1(x, \xi(x))\,dx$. Consider $\xi_h, \xi_0 \in L^1(\Omega; \mathbb{R}^l)$ and suppose that the function $v \mapsto L_1(x, v)$ is strictly convex at the point $\xi_0(x)$ for almost every $x \in \Omega$.*

If $\xi_h \rightharpoonup \xi_0$ in $L^1(\Omega; \mathbb{R}^l)$ and $J_1(\xi_h) \to J_1(\xi_0)$, then

$$\|\xi_h - \xi_0\| \to 0 \ \text{ in } L^1(\Omega; \mathbb{R}^l).$$

We will apply Theorem 4.2 in order to obtain well-posedness results in the two senses mentioned in the introduction, namely, Tikhonov well-posedness and well-posedness by perturbations. We also remark that the integrand can assume the value $+\infty$: we emphasize that in this way Theorem 4.2 can be applied also to constrained problems. The strategy of the proof is different from the one adopted by Sychev in [18] for the case of Carathéodory integrands. In fact Sychev's proof is based on the compactness (which fails in our hypotheses) of the union of the subdifferentials with respect to the last variable evaluated at some special points.

First we prove a semicontinuity result, using the equality furnished by Corollary 3.2, and then, extending some results of [21], we obtain Theorem 4.2.

Results of strong convergence implied by weak were obtained also in [21] and in [17]. We remark that Theorem 3 of [21] is a special case of Theorem 4.2 under the more restrictive hypothesis of global strict convexity. The proof of Theorem 3 of [21] relies on Theorem 1 of the same paper, which states that a sequence weakly convergent to an extreme point of the convex hull of the whole sequence is in fact strongly convergent. The proof of Theorem 3 is thus reduced to show that from strict convexity, the extremality of the corresponding point for the epigraph of the second conjugate follows. If we do not require global strict convexity, but only strict convexity at a point, as we saw in Remark 3.5, this point is in general not extremal for epif^{**}. Moreover, it is not easy to check the extremality of a point for epif^{**}, since it is already difficult to calculate f^{**}. But reviewing the proof of Theorem 1 of [21] it turns out that for some special sequences the thesis is still valid under less restrictive hypotheses (see Lemma 4.5).

Before proving Theorem 4.2, we need some preliminary lemmas.

LEMMA 4.3. *Let $g : \Omega \times \mathbb{R}^m \times \mathbb{R}^l \to [0, +\infty]$ be a normal integrand and let G be defined as in (4.2). Consider $u_h, u_0 \in L^1(\Omega; \mathbb{R}^m)$, $\xi_h, \xi_0 \in L^1(\Omega; \mathbb{R}^l)$, and suppose that the function $v \mapsto g(x, u_0(x), v)$ is convex at the point $\xi_0(x)$ for almost every $x \in \Omega$.*

If $\|u_h - u_0\| \to 0$ in $L^1(\Omega; \mathbb{R}^m)$ and $\xi_h \rightharpoonup \xi_0$ in $L^1(\Omega; \mathbb{R}^l)$, then

$$\liminf G(u_h, \xi_h) \geq G(u_0, \xi_0).$$

Proof. First of all, by the convexity at the point $\xi_0(x)$ and by Corollary 3.2, we obtain

$$g^{**}(x, u_0(x), \xi_0(x)) = g(x, u_0(x), \xi_0(x))$$

for almost every $x \in \Omega$, where g^{**} is the convex regularization of g with respect to the last variable. We also note that by standard facts (see, for instance, [13, Proposition 1.3, p. 238]) the function $(x, u, v) \mapsto g^{**}(x, u, v)$ is a normal integrand, and therefore the functional

$$\hat{G}(u, \xi) := \int_\Omega g^{**}(x, u(x), \xi(x))\,dx$$

is well defined and takes values in $[0, +\infty]$.

Moreover $\hat{G}$ verifies

$$\hat{G}(u, \xi) \leq G(u, \xi)$$

for every $u \in L^1(\Omega; \mathbb{R}^m)$ and $\xi \in L^1(\Omega; \mathbb{R}^l)$. Finally we observe that $\hat{G}$ is sequentially lower semicontinuous with respect to the weak topology at every point because g^{**} is convex with respect to the last variable, so

$$\liminf_{n \to +\infty} G(u_h, \xi_h) \geq \liminf \hat{G}(u_h, \xi_h) \geq \hat{G}(u_0, \xi_0) = G(u_0, \xi_0),$$

and this concludes the proof. □

If g and L_1 are normal integrands as before, and $u \in L^1(\Omega; \mathbb{R}^m)$, $\xi \in L^1(\Omega; \mathbb{R}^l)$, we write

$$g \circ (u, \xi) \quad \text{and} \quad L_1 \circ \xi$$

for the maps $x \mapsto g(x, u(x), \xi(x))$ and $x \mapsto L_1(x, \xi(x))$, respectively.

LEMMA 4.4. *Let $g : \Omega \times \mathbb{R}^m \times \mathbb{R}^l \to [0, +\infty]$ be a normal integrand and let $u_h, u_0 \in L^1(\Omega; \mathbb{R}^m)$, $\xi_h, \xi_0 \in L^1(\Omega; \mathbb{R}^l)$. Suppose that the function $v \mapsto g(x, u_0(x), v)$ is convex at the point $\xi_0(x)$ for almost every $x \in \Omega$.*

If $\|u_h - u_0\|_{L^1} \to 0$, $\xi_h \rightharpoonup \xi_0$, and $G(u_h, \xi_h) \to G(u_0, \xi_0)$, then $g \circ (u_h, \xi_h) \rightharpoonup g \circ (u_0, \xi_0)$ in $L^1(\Omega; \mathbb{R})$.

Proof. Using Lemma 4.3 the proof goes exactly as that of Lemma 3 of [21], since the only property that is used there is the semicontinuity of the functional G at the point (u_0, ξ_0) and not the semicontinuity at every point. □

LEMMA 4.5. *Assume that $\xi_h \rightharpoonup \xi$ and $L_1 \circ \xi_h \rightharpoonup L_1 \circ \xi$ in $L^1(\Omega; \mathbb{R}^l)$, where L_1 is defined as above. Set $K(x) = \mathrm{epi} L_1(x, \cdot)$ and $C(x, r) = \overline{B}(\xi(x), r) \times \mathbb{R}$ for almost every $x \in \Omega$. Suppose that $(\xi(x), L_1(x, \xi(x)))$ is an extremal point of $\overline{\mathrm{co}}(K(x) \cap C(x, r))$ for almost every $x \in \Omega$ and for all $0 < r < R$ for some fixed $R > 0$.*

Then $\xi_h \to \xi$ strongly in $L^1(\Omega; \mathbb{R}^l)$.

Proof. Without loss of generality we can suppose that $\xi(x) = 0$ and $L_1(x, \xi(x)) = 0$ for almost every $x \in \Omega$; in fact it is always possible to reduce the problem to this case simply by substituting $\mathrm{epi} L_1^{**}(x, \cdot)$ with the set $\mathrm{epi} L_1^{**}(x, \cdot) - (\xi(x), L_1(x, \xi(x)))$.

Mimicking the proof of Theorem 1 of [21], fix any ϵ such that $0 < \epsilon < R$.

Set

$$w_h^\epsilon(x) = \begin{cases} \xi_h(x) & \text{if } |\xi_h(x)| \leq \epsilon, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad t_h^\epsilon(x) = \begin{cases} L(x, \xi_h(x)) & \text{if } |\xi_h(x)| \leq \epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

From the definition, we have $(w_h^\epsilon(x), t_h^\epsilon(x)) \in K(x)$ for almost every $x \in \Omega$.

Define also $v_h^\epsilon = \xi_h - w_h^\epsilon$ and $s_h^\epsilon = L_1 \circ \xi_h - t_h^\epsilon$; it follows that also $(v_h^\epsilon(x), s_h^\epsilon(x))$ is in $K(x)$ for almost every $x \in \Omega$.

Since w_h^ϵ is bounded in the L^∞-norm, there exists a subsequence $w_{h'}^\epsilon$ and w^ϵ such that $w_{h'}^\epsilon \rightharpoonup w^\epsilon$ in $L^1(\Omega)$; hence $v_{h'}^\epsilon \rightharpoonup -w^\epsilon := v^\epsilon$ in $L^1(\Omega)$. Moreover, by hypothesis, we also have

$$t_h^\epsilon \rightharpoonup 0 \quad \text{and} \quad s_h^\epsilon \rightharpoonup 0.$$

As in Theorem 1 of [21] we get $\frac{(v^\epsilon, 0)}{2} + \frac{(w^\epsilon, 0)}{2} = (0, 0)$ almost everywhere in Ω and $(v^\epsilon(x), 0), (w^\epsilon(x), 0) \in \overline{\mathrm{co}}[K(x) \cap C(0, r)]$ for almost every $x \in \Omega$. Since $(0, 0)$ is

an extremal point of $\overline{\text{co}}[K(x) \cap C(0,r)]$, it follows that $v^\epsilon = w^\epsilon = 0$ and the whole sequence v_h^ϵ converges.

From now on, exactly as in [21] we obtain that $\|\xi_h\| \to 0$. □

Now we are able to prove the theorem stated at the beginning of this section.

Proof of Theorem 4.2. From Lemma 3.4, the point $(\xi_0(x), L_1(x, \xi_0(x)))$ is extremal for $\text{epi}(L_1)_r^{**}(x,\cdot)$, where $(L_1)_r(x,\cdot)$ is the function $L_1(x,\cdot)$ restricted to the closed ball of center $\xi_0(x)$ and radius r for every $r > 0$.

Consider the sequence $(\xi_h, L_1 \circ \xi_h)$. By Lemma 4.4 we have

$$(\xi_h, L_1 \circ \xi_h) \rightharpoonup (\xi_0, L_1 \circ \xi_0)$$

in $L^1(\Omega;\mathbb{R}^m) \times L^1(\Omega;\mathbb{R})$. Then the conclusion follows from Lemma 4.5. □

Theorem 4.2 can be applied to the functional (4.1).

COROLLARY 4.6. *Let $L : \Omega \times \mathbb{R}^{nm} \to [0, +\infty]$ be a normal integrand and let J be defined as in* (4.1).

Let $u_h, u \in W_0^{1,1}(\Omega;\mathbb{R}^m)$ be such that $Du_h \rightharpoonup Du$ in $L^1(\Omega;\mathbb{R}^{nm})$.

Suppose that the function $v \mapsto L(x,v)$ is strictly convex at the point $Du(x)$ for almost every $x \in \Omega$ and further that $J(u_h) \to J(u)$. Then

$$\|Du_h - Du\| \to 0$$

in $L^1(\Omega;\mathbb{R}^{mn})$.

Now we apply Theorem 4.2 to the special case of minimizing sequences in order to obtain a sufficient condition for Tikhonov well-posedness of a nonconvex integral functional. Finally, using the results of [19] we will also get a corollary about well-posedness by perturbations. More precisely, we deal with perturbations of the integrand, and in this framework it is reasonable to use the so-called variational convergences. In particular our result deals with the bounded Hausdorff convergence. See [1], [2], [3] for the properties of this convergence. We start by giving the definitions that we need.

Let X be a normed space.

DEFINITION 4.7. *We say that the problem of minimizing $J : X \to (-\infty, +\infty]$ is Tikhonov well-posed if it satisfies the following conditions:*

1. *There exists a unique global minimum point x_0 for J.*
2. *If x_h is any minimizing sequence, i.e., a sequence in X such that $J(x_h) \to J(x_0)$, then $\|x_h - x_0\| \to 0$.*

Let $J_h : X \to (-\infty, +\infty]$ be a sequence of functions. We say that J_h is equicoercive if there exists $\phi : [0, +\infty) \to \mathbb{R}$ such that

$$\lim_{y \to +\infty} \phi(y) = +\infty \quad \text{and} \quad J_h(x) \geq \phi(|x|)$$

for every $x \in X$.

Now we give the definition of well-posedness by perturbations for an abstract global minimization problem (see [24]).

We consider metric spaces X and $\mathcal{A}$, a fixed point $a_0 \in \mathcal{A}$, and $B(a_0, R)$, a given ball around a_0 with positive radius R in $\mathcal{A}$. We are given the proper extended real-valued functions

$$J : X \to (-\infty, +\infty], \quad F : B(a_0, R) \times X \to (-\infty, +\infty]$$

such that

$$J(x) = F(a_0, x), \ x \in X.$$

The corresponding value function is given by

$$V(a) = \inf\{F(a,x) \,|\, x \in X\},\ a \in \mathcal{A}.$$

DEFINITION 4.8. *The problem of minimizing J on X is called* well-posed by perturbations *(with respect to the embedding defined by F) if and only if*

1. *$V(a) > -\infty$ for all $a \in B(a_0, R)$;*
2. *there exists a unique global minimizer x_0 for J;*
3. *for every sequence $a_h \to a_0$ in $\mathcal{A}$ such that*

$$F(a_h, x_h) - V(a_h) \to 0 \text{ as } n \to +\infty \tag{4.3}$$

we have $x_h \to x_0$ in X.

Sequences as in (4.3) will be called asymptotically minimizing corresponding to the sequence a_h.

Now for $A, C \subset X$, the excess of A on C is given by

$$e(A,C) := \sup_{x \in A} d(x,C),$$

with the convention that $e(A,C) = 0$ if $A = \emptyset$ and $e(A,C) = +\infty$ if $C = \emptyset$. Then the ρ-Hausdorff distance between A and C is defined by

$$\mathrm{haus}_\rho(A,C) := \max\{e(A \cap B(0,\rho), C), e(C \cap B(0,\rho), A)\}.$$

DEFINITION 4.9. *For $\rho \geq 0$, the ρ-(Hausdorff) distance between $f, g : X \to \mathbb{R} \cup \{+\infty\}$ is*

$$\mathrm{haus}_\rho(f,g) := \mathrm{haus}_\rho(\mathrm{epi} f, \mathrm{epi} g),$$

where the unit ball of $X \times \mathbb{R}$ is the set $B(0,1) := \{(x,\alpha) : \|x\| \leq 1, |\alpha| \leq 1\}$.

DEFINITION 4.10. *Let $f, f_h : X \to [-\infty, +\infty]$ be lower semicontinuous functions. We say that f_h converges to f with respect to the bounded Hausdorff convergence (or in the sense of Attouch–Wets), and we write $f = \tau_{aw} - \lim f_h$ if and only if*

$$\exists \rho_0 > 0 \text{ such that } \mathrm{haus}_\rho(f, f_h) \to 0 \text{ as } h \to +\infty \ \ \forall \rho > \rho_0.$$

Examples of sequences of functionals of the calculus of variations converging with respect to the bounded Hausdorff convergence can be found in [9]. Now we are able to prove the main results of the paper, which are in fact corollaries of Theorem 4.2.

COROLLARY 4.11. *Let L and J be defined as in Corollary* 4.6. *Suppose that J is coercive having a unique minimum point $u_0 \in W_0^{1,1}(\Omega; \mathbb{R}^m)$.*

Moreover, assume that $v \mapsto L(x,v)$ is strictly convex at the point $Du_0(x)$ for almost every $x \in \Omega$. Then J is Tikhonov well-posed in $W_0^{1,1}(\Omega; \mathbb{R}^m)$.

Equivalently, if $u_h \in W_0^{1,1}(\Omega; \mathbb{R}^m)$ is any minimizing sequence, then

$$\|u_h - u_0\| \to 0 \ \text{ in } W_0^{1,1}(\Omega; \mathbb{R}^m).$$

Proof. Let $u_h \in W_0^{1,1}(\Omega; \mathbb{R}^m)$ be a minimizing sequence for J. Up to subsequences, from the coercivity hypothesis it follows that u_h is weakly convergent in $W_0^{1,1}(\Omega; \mathbb{R}^m)$ to a function v. Since the functional J is lower semicontinuous, v is a minimizer of J, and therefore we get $v = u_0$.

Now, applying Theorem 4.2, we obtain that the weakly convergent sequence Du_h is actually strongly convergent in $L^1(\Omega;\mathbb{R}^{nm})$ to Du_0, and therefore

$$\|u_h - u_0\| \to 0 \text{ in } W_0^{1,1}(\Omega;\mathbb{R}^m). \qquad \square$$

See, for instance, [11] for the hypotheses which ensure coercivity of an integral functional.

The last corollary is about well-posedness by perturbations, and it follows directly from Theorem 3.5 of [19].

COROLLARY 4.12. *Let $L, L_h : \Omega \times \mathbb{R}^{nm} \to [0,+\infty]$ be normal integrands and let J be defined as in* (4.1). *Suppose that J is coercive having a unique minimum point $u_0 \in W_0^{1,1}(\Omega;\mathbb{R}^m)$.*

Moreover, assume that $v \mapsto L(x,v)$ is strictly convex at the point $Du_0(x)$ for almost every $x \in \Omega$. Consider the sequence J_h defined by $J_h(u) := \int_\Omega L_h(x, Du(x))\,dx$ on $W_0^{1,1}(\Omega,\ \mathbb{R}^m)$. Assume that

(i) *$J_h \to J$ with respect to the bounded Hausdorff convergence on $W_0^{1,1}(\Omega;\mathbb{R}^m)$;*

(ii) *J_h is equicoercive.*

Then

$$\|u_h - u_0\| \to 0 \text{ in } W_0^{1,1}(\Omega;\mathbb{R}^m)$$

for every asymptotically minimizing sequence u_h.

In some cases it is possible to put some natural topology on the space of integrands and to relate the convergence of the sequence of integrands with the bounded Hausdorff convergence of the sequence of the associated integral functionals. Results on this subject can be found in [9] for quadratic integrals of elliptic type and in [19] and [20] in a more general setting.

We observe that the condition imposed on the integrand in Corollary 4.11 is very precise but not easy to check: if we want to know that the integral functional is well-posed, we first have to find out the minimizer u_0 and then prove that the integrand is strictly convex at the point $Du_0(x)$. But under the conditions that allow us to represent J^{**} as an integral functional, it is possible to state a result which is less precise but depends only on the integrand L and on L^{**}, the convex regularization of L with respect to the second variable.

We shall deal with scalar functions; therefore we can use the fact that $J^{**}(u)$ is given by $\int_\Omega L^{**}(x, Du(x))\,dx$. Hence, from now on, we consider only the case $m = 1$ and the functional J defined on the space $W_0^{1,1}(\Omega;\mathbb{R})$ by (4.1).

Let us introduce the following set of points:

$$\mathcal{M}(x) := \{z \in \mathbb{R}^n \mid L(x,z) = L^{**}(x,z)\}.$$

THEOREM 4.13. *Let L and J be as in Corollary* 4.6, *with $m = 1$. Assume that $L(x,\cdot)$ has superlinear growth and that it is strictly convex at z for every $z \in \mathcal{M}(x)$ for almost every $x \in \Omega$. Suppose moreover that J has a unique minimizer on the space $W_0^{1,1}(\Omega;\mathbb{R})$.*

Then J is Tikhonov well-posed on $W_0^{1,1}(\Omega;\mathbb{R})$.

Proof. Let u be the unique minimizer for J. By Remark (iv) following Theorem 5.2 of [11], we obtain

$$L(x, Du(x)) = L^{**}(x, Du(x))$$

for almost every $x \in \Omega$, i.e., $Du(x) \in \mathcal{M}(x)$ for almost every x, and therefore $L(x,\cdot)$ is strictly convex at $Du(x)$. The thesis then follows from Corollary 4.11. $\square$

5. Two classes of well-posed functionals.

5.1. Functionals depending only on the gradient. In the papers [5], [6], [8], [14] the following problem of minimizing a functional of the gradient under linear boundary conditions is studied:

$$(P_a) \quad \text{Minimize} \int_\Omega L(Du(x))dx \quad \text{subject to } u \in u_a + W_0^{1,1}(\Omega).$$

Here $\Omega \subset \mathbb{R}^n$ is open, bounded, and with piecewise C^1 boundary, and $u_a(x) := \langle a, x\rangle$, where $a \in \mathbb{R}^n$ and $\langle\cdot,\cdot\rangle$ denotes the standard scalar product in $\mathbb{R}^n$. The function $L : \mathbb{R}^n \to \mathbb{R}$ is supposed to be lower semicontinuous and bounded from below. Moreover the function L satisfies the growth condition

$$(G) \qquad L(y) \geq \Phi(|y|) \text{ for any } y \in \mathbb{R}^n,$$

where $\Phi : [0, +\infty) \to \mathbb{R}$ is such that $\lim_{t\to+\infty} \frac{\Phi(t)}{t} = +\infty$.

In particular the paper [8] studies the continuous dependence of the solutions on the boundary data. For this purpose we need to introduce the problem

$$(P_a^{**}) \quad \text{Minimize} \int_\Omega L^{**}(Du(x))\,dx \quad \text{subject to } u \in u_a + W_0^{1,1}(\Omega).$$

As is well known (see Chapter X of [13]), it turns out that

$$\inf P_a = L^{**}(a)|\Omega| \tag{5.1}$$

and that the function u_a is a minimizer. The following theorem, which is a particular case of Theorem 4.1 of [8], holds.

THEOREM 5.1. *Let L be lower semicontinuous satisfying the growth condition (G). Suppose that $L(a) = L^{**}(a)$ and that $(a, L^{**}(a))$ belongs to the relative interior of a proper face F_1 of an n-dimensional face F of* epiL^{**}, *and let $\{a_k\}_{k\in\mathbb{N}}$ be any sequence such that $(a_k, L^{**}(a_k))$ belongs to the relative interior of F for any $k \in \mathbb{N}$ and $\lim_{k\to+\infty} a_k = a$.*

(i) *If any sequence u_k of solutions of P_{a_k} converges strongly to u_a in $W_0^{1,1}(\Omega)$, then $\dim(F_1) = 0$ and so $(a, L^{**}(a))$ is an extreme point of* epiL^{**}.

(ii) *If $(a, L^{**}(a))$ is an extreme point of* epiL^{**}, *then any sequence u_k of solutions of $P_{a_k}^{**}$ converges strongly to u_a in $W_0^{1,1}(\Omega)$.*

The previous theorem can be improved using the results of section 4. We establish the strong convergence also of asymptotically minimizing sequences without imposing constraints on the position of the sequence a_k with respect to epiL^{**}, and, according to the definition given in the introduction, this means that problem (P_a) is well-posed by perturbations of the boundary data.

THEOREM 5.2. *Let $L : \mathbb{R}^n \to \mathbb{R}$ be lower semicontinuous, bounded from below, and satisfying the growth condition (G). Suppose that $(a, L^{**}(a))$ is an extreme point of* epiL^{**}.

Then any asymptotically minimizing sequence of P_{a_k} converges strongly to u_a in $W^{1,1}(\Omega)$.

Proof. Since $(a, L^{**}(a))$ is an extreme point of epiL^{**} and L is lower semicontinuous, it follows that $L(a) = L^{**}(a)$ by Lemma 3.2. Then by [6] problem (P_a) has the only solution u_a. Let a_k be a sequence in $\mathbb{R}^n$ such that $a_k \to a$. We observe that problem (P_{a_k}) is equivalent to the following one:

$$(\hat{P}_{a_k}) \quad \text{Minimize} \quad G_k(u) := \int_\Omega L(Du(x) + a_k)dx \quad \text{subject to } u \in W_0^{1,1}(\Omega).$$

In fact $\int_\Omega L(Du(x))dx = \int_\Omega L(Dv(x)+a)dx$, where $u \in u_a + W_0^{1,1}(\Omega)$ and $v = u - u_a$, $v \in W_0^{1,1}(\Omega)$. In this manner we do not move the set on which we consider the minimization problems, and we can apply the previous results.

Now, if v_k is an asymptotically minimizing sequence for G_k and $u_k := v_k + u_{a_k}$, then there exists a sequence $\epsilon_k \to 0$ such that

$$G_k(v_k) \le \inf G_k + \epsilon_k = L^{**}(a_k)|\Omega| + \epsilon_k, \tag{5.2}$$

where the equality is a well-known fact in the calculus of variations as previously noted in (5.1). We observe that L^{**} is convex and finite, since L is finite, and therefore it is continuous at the point a. Hence the last quantity in (5.2) is bounded and this implies, by condition (G) and the Dunford–Pettis theorem, that the sequence Dv_k is weakly compact in $L^1(\Omega)$.

Let Dv_{k_h} be a weakly convergent subsequence and let U be its weak limit.

We aim to prove that there exists $v \in W_0^{1,1}(\Omega)$ such that $U = Dv$.

Since the sequence Dv_{k_h} is weakly convergent, there exists $M > 0$ such that

$$\|Dv_{k_h}\|_{L^1} \le M.$$

If the dimension of the space is $n = 1$, we are done. If $n > 1$, by the Sobolev embedding theorem (see, for instance, [11, p. 25]) we obtain

$$\|v_{k_h}\|_{L^{\frac{n}{n-1}}} \le C$$

for a suitable constant $C > 0$. Since $\frac{n}{n-1} > 1$ there exists a further subsequence of v_{k_h}, which we still denote by v_{k_h}, weakly convergent to a function v in $L^1(\Omega)$. We shall prove that $Dv = U$. To this purpose consider any function $\psi \in C_c^\infty(\Omega)$. By the definition of a weak derivative (denoted by ∂_i), the equality

$$\int_\Omega v_{k_h}(x)\partial_i\psi(x)\,dx = -\int_\Omega \partial_i v_{k_h}(x)\psi(x)\,dx$$

holds for every $i \in \{1,\dots,n\}$. Now, the right-hand side converges to $\int_\Omega v(x)\partial_i\psi(x)\,dx$, whereas the left-hand side converges to $-\int_\Omega U_i(x)\psi(x)\,dx$, and therefore

$$\int_\Omega v(x)\partial_i\psi(x)\,dx = -\int_\Omega U_i(x)\psi(x)\,dx \tag{5.3}$$

by the uniqueness of the limit. Since (5.3) holds for every $i \in \{1,\dots,n\}$ and for every $\psi \in C_c^\infty(\Omega)$, we get that $U = Dv$, and hence the function v is in $W^{1,1}(\Omega)$. In order to obtain that $v \in W_0^{1,1}(\Omega)$ it is enough to observe that $v_{k_h} \in W_0^{1,1}(\Omega)$, $v_{k_h} \rightharpoonup v$ in $W^{1,1}(\Omega)$ and $W_0^{1,1}(\Omega)$ is weakly closed.

Now define the functional

$$G^{**}(v) := \int_\Omega L^{**}(Dv + a)\,dx.$$

As a direct consequence of the Jensen inequality we know that

$$\inf_{v \in W_0^{1,1}(\Omega)} \int_\Omega L^{**}(Dv + a)\,dx = L^{**}(a)|\Omega|. \tag{5.4}$$

Since the functional G^{**} is lower semicontinuous on $W_0^{1,1}(\Omega)$ with respect to the weak convergence, by (5.4) we obtain

$$\begin{aligned}\inf P_a &= L^{**}(a)|\Omega| \\ &\le \int_\Omega L^{**}(Dv + a)dx \\ &\le \liminf \int_\Omega L(Dv_{k_h}(x) + a_{k_h}) \\ &\le \liminf L^{**}(a_{k_h})|\Omega| + \epsilon_{k_h} \\ &= L^{**}(a)|\Omega|.\end{aligned}$$

Then the function $v = 0$ because u_a is the unique minimizer of P_a.

Now using the fact that $(a, L^{**}(a))$ is an extreme point of epiL^{**}, by Lemma 3.4 and Theorem 4.2 of the preceding section, the subsequence Dv_{k_h} is actually strongly convergent to 0.

Since from any subsequence of Dv_k we can extract a strongly convergent subsequence with limit 0, then the whole sequence is strongly convergent to 0.

This means that the sequence u_k is strongly convergent to the function u_a, and this completes the proof. □

5.2. Radially symmetric functionals. In this section we consider the well-posedness of the variational problem

$$\text{(5.5)}\quad \text{Minimize } \int_{B(0,R)} [L(|Du(x)|) + h(u(x))]\,dx \quad \text{subject to } u \in W_0^{1,1}(B(0,R)),$$

where $B(0,R)$ is the ball of $\mathbb{R}^n$ of radius R centered at the origin; the function $L : [0,+\infty[\to [0,+\infty]$ is lower semicontinuous and $h : \mathbb{R} \to [0,+\infty[$ is a convex function.

Problems of this kind arise in various fields as nonlinear elasticity, fluid dynamics, and optimal design. We refer to [7] and [10] for results about existence and uniqueness of minimizers, and we remark that well-posedness was not considered there.

In order to prove well-posedness of (5.5) we need an elementary result about convergence of sequences.

LEMMA 5.3. *Let $\{a_h\}$, $\{b_h\}$ be two sequences and let $a, b \in [0,+\infty)$ such that*

(i) *$a_h \ge 0$, $b_h \ge 0$ for every n;*

(ii) *there exist $a, b \ge 0$ such that $\lim(a_h + b_h) = a + b$;*

(iii) *$\liminf a_h \ge a$; $\liminf b_h \ge b$.*

Then $\lim a_h = a$ and $\lim b_h = b$.

The proof of Lemma 5.3 is trivial, so we omit it.

Applying the results of [7] and Corollary 4.11 we obtain the following theorem.

THEOREM 5.4. *Assume that*

(h1) *$L : [0,+\infty[\to [0,+\infty]$ is lower semicontinuous and has superlinear growth;*

(h2) *$h : \mathbb{R} \to [0,+\infty[$ is convex and monotonic;*

and that either h or L^{**} is strictly monotonic. Then problem (5.5) *is well-posed in the sense of Tikhonov on the space $W_0^{1,1}(B(0,R))$.*

Proof. One of the steps of the proof of Theorem 5 of [7] consists in showing that if u is a solution of the minimum problem, then $(Du(x), L^{**}(Du(x)))$ is an extremal point of epiL^{**} for almost every $x \in B(0,R)$. Thus by Lemma 3.4 we deduce that L is strictly convex at the point $Du(x)$ for almost every $x \in \Omega$.

Now let u_h be a minimizing sequence; we want to apply Lemma 5.3 with $a_h = \int_\Omega L(|Du_h(x)|)\,dx$ and $b_h = \int_\Omega h(u_h(x))\,dx$. All the hypotheses are satisfied; in fact (i) is obvious; (ii) follows directly from the definition of minimizing sequence, with

$a = \int_\Omega L(|Du(x)|)\,dx$ and $b = \int_\Omega h(u(x))\,dx$; and (iii) comes from Lemma 4.3. We therefore obtain

$$\int_\Omega L(|Du_h(x)|)\,dx \to \int_\Omega L(|Du(x)|)\,dx.$$

Then the thesis follows from Corollary 4.11 of the preceding section. □

We remark that, applying Corollary 4.12, the problem defined by (5.5) is also well-posed by perturbations with respect to the bounded Hausdorff convergence.

REFERENCES

[1] H. Attouch and R. J.-B. Wets, *Quantitative stability of variational systems.* I. *The epigraphical distance*, Trans. Amer. Math. Soc., 3 (1991), pp. 695–729.

[2] H. Attouch and R. J.-B. Wets, *Quantitative stability of variational systems.* II. *A framework for nonlinear conditioning*, SIAM J. Optim., 3 (1993), pp. 359–381.

[3] G. Beer, *Topologies on Closed and Closed Convex Sets*, Kluwer, Dordrecht, The Netherlands, 1993.

[4] S. Bertirotti, *Wellposedness in the calculus of variations*, J. Convex Anal., 7 (2000), pp. 299–318.

[5] A. Cellina, *On minima of a functional of the gradient: Necessary conditions*, Nonlinear Anal., 20 (1993), pp. 337–341.

[6] A. Cellina, *On minima of a functional of the gradient: Sufficient conditions*, Nonlinear Anal., 20 (1993), pp. 343–347.

[7] A. Cellina and S. Perrotta, *On minima of radially symmetric functionals of the gradient*, Nonlinear Anal., 23 (1994), pp. 239–249.

[8] A. Cellina and S. Zagatti, *A version of Olech's lemma in a problem of the calculus of variations*, SIAM J. Control Optim., 32 (1994), pp. 1114–1127.

[9] Z. Chbani, *Caractérisation de la convergence d'integrales définies à partir d'opérateurs elliptiques par la convergence des coéfficients*, Seminaire d'Analyse Convexe, Montpellier, 1991, Exposé n. 12.

[10] G. Crasta, *Existence, uniqueness, and qualitative properties of minima to radially symmetric non-coercive non-convex variational problems*, Math. Z., 235 (2000), pp. 569–589.

[11] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Appl. Math. Sci. 78, Springer-Verlag, Berlin, 1988.

[12] A. L. Dontchev and T. Zolezzi, *Well-posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, Berlin, 1993.

[13] I. Ekeland and R.Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[14] G. Friesecke, *A necessary and sufficient condition for non-attainment and formation of microstructure almost everywhere in scalar variational problems*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 437–471.

[15] A. D. Ioffe and A. J. Zaslavski, *Variational principles and well-posedness in optimization and calculus of variations*, SIAM J. Control. Optim., 38 (2000), pp. 566–581.

[16] P. Marcellini, *Nonconvex integrals of the calculus of variations*, in Methods in Nonconvex Analysis (Varenna, 1989), Lecture Notes in Math., Springer-Verlag, Berlin, pp. 16–57.

[17] C. Olech, *The Lyapunov Theorem: Its Extensions and Applications*, Lecture Notes in Math. 1446, Springer-Verlag, Berlin, 1990.

[18] M. A. Sychev, *Necessary and sufficient conditions in semicontinuity and convergence theorems with a functional*, Mat. Sb., 186 (1995), pp. 847–877.

[19] S. Villa, *AW-convergence and well-posedness of non convex functions*, J. Convex Anal., 10 (2003), pp. 351–364.

[20] S. Villa, *Wellposedness under perturbations of integral functionals*, ESAIM Control Optim. Calc. Var., submitted.

[21] A. Visintin, *Strong convergence results related to strict convexity*, Comm. Partial Differential Equations, 9 (1984), pp. 439–466.

[22] A. J. Zaslavski, *Well posedness and porosity in the calculus of variations without convexity assumptions*, Nonlinear Anal., 53 (2003), pp. 1–22.

[23] T. Zolezzi, *Wellposed problems of the calculus of variations for nonconvex integrals*, J. Convex Anal., 2 (1995), pp. 375–383.

[24] T. Zolezzi, *Well posedness criteria in optimization with application to the calculus of variations*, Nonlinear Anal., 25 (1995), pp. 437–453.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1313–1327

NULL CONTROLLABILITY AND THE ALGEBRAIC RICCATI EQUATION IN BANACH SPACES*

J. M. A. M. VAN NEERVEN†

Abstract. By a recent result of Priola and Zabczyk, a null controllable linear system

$$y'(t) = Ay(t) + Bu(t)$$

in a Hilbert space E is null controllable with vanishing energy if and only if it is null controllable and the only positive self-adjoint solution of the associated algebraic Riccati equation

$$XA + A^*X - XBB^*X = 0$$

is the trivial solution $X = 0$. In this paper we extend this result to Banach spaces with an elementary proof which uses only reproducing kernel Hilbert space techniques. We also show that null controllability with vanishing energy implies null controllability.

Key words. null controllability with vanishing energy, algebraic Riccati equation, reproducing kernel Hilbert space

AMS subject classifications. Primary, 93B05; Secondary, 47D06, 93B03, 93C25

DOI. 10.1137/S0363012903437058

Let A be the generator of a C_0-semigroup on a real Banach space E and let B be a bounded linear operator from a real Hilbert space H into E. The pair (A, B) is said to be *null controllable with vanishing energy* if for all $x \in E$ and all $\varepsilon > 0$ there exists a time $t > 0$ and a function $u \in L^2(0, t; H)$ satisfying $\|u\|_{L^2(0,t;H)} < \varepsilon$ such that the mild solution $y^{u,x}$ of the linear control problem

$$\begin{aligned} y'(s) &= Ay(s) + Bu(s) \qquad (s \in [0, t]), \\ y(0) &= x \end{aligned} \tag{0.1}$$

satisfies $y^{u,x}(t) = 0$. The pair (A, B) is said to be *null controllable in finite time* if there exists a fixed time $t_0 > 0$ such that for all $x \in E$ there exists a function $u \in L^2(0, t_0; H)$ such that the mild solution of the problem (0.1) satisfies $y^{u,x}(t_0) = 0$.

For Hilbert spaces E, Priola and Zabczyk recently proved that a pair (A, B), which is null controllable in finite time, is null controllable with vanishing energy if and only if the only positive self-adjoint solution to the algebraic Riccati equation

$$XA + A^*X - XBB^*X = 0$$

is the trivial solution $X = 0$ [10]. One of the main ingredients of the proof is the fact that a certain differential Riccati equation is solved in terms of a minimal energy functional. In this paper we extend the Priola–Zabczyk result to Banach spaces with a different proof which is based on reproducing kernel Hilbert space techniques, and we

*Received by the editors November 3, 2003; accepted for publication (in revised form) June 4, 2004; published electronically January 5, 2005. This work was supported by the Research Training Network HPRN-CT-2002-00281 and a "VIDI subsidie" in the "Vernieuwingsimpuls" program of the Netherlands Organization for Scientific Research (NWO).

http://www.siam.org/journals/sicon/43-4/43705.html

†Department of Applied Mathematical Analysis, Technical University of Delft, P.O. Box 5031, 2600 GA Delft, The Netherlands (J.vanNeerven@math.tudelft.nl). This work was started while the author was visiting the Institute of Mathematics of the Polish Academy of Sciences in Warsaw. The paper was written while the author stayed at the School of Mathematics at the University of New South Wales.

show that null controllability with vanishing energy in fact implies null controllability in finite time. Our approach relies upon the identification of the space H_t of points that are reachable in time t as the reproducing kernel Hilbert space associated with the operator $Q_t \in \mathcal{L}(E^*, E)$ defined by

$$Q_t x^* := \int_0^t S(s)BB^*S^*(s)x^*\,ds \qquad (x^* \in E^*).$$

The square norm $\|h\|_{H_t}^2$ can be interpreted as the minimal energy needed to reach the state $h \in H_t$ in time t starting from the origin. The basic problem is then to understand how this minimal energy varies with h and t. Our main result in this direction is Theorem 2.5, which describes the instantaneous rate of change of the minimal energy along curves in H_t as time progresses. It is used to obtain an explicit positive symmetric solution $X(t)$ for a differential Riccati equation. As in [10], the weak operator limit $X = \lim_{t\to\infty} X(t)$ then turns out to be the maximal positive symmetric solution of the algebraic Riccati equation, and null controllability with vanishing energy is equivalent to the condition that $X = 0$.

For more information about null controllability and Riccati equations as well as applications to various control systems we refer to [1, 2, 3, 4, 7, 8, 12, 13].

1. Reachable states and reproducing kernels. The mild solution of the problem (0.1) will be denoted by $y^{u,x}$. Thus,

$$y^{u,x}(s) := S(s)x + \int_0^s S(s-r)Bu(r)\,dr \qquad (s \in [0,t]).$$

An element $h \in E$ is *reachable in time* t if there exists a control $u \in L^2(0,t;H)$ such that $y^{u,0}(t) = h$. The collection H_t of all elements that are reachable in time t is a linear subspace of E which is a Hilbert space with norm

$$\|h\|_{H_t}^2 := \inf\big\{\|u\|_{L^2(0,t;H)}^2 :\ u \in L^2(0,t;H),\ y^{u,0}(t) = h\big\}.$$

Thus, $\|h\|_{H_t}^2$ is the minimal energy needed to steer the system from 0 to h in time t. Notice that H_t equals the range of the operator $L_t \in \mathcal{L}(L^2(0,t;H), E)$ defined by

$$L_t f := \int_0^t S(t-s)Bf(s)\,ds.$$

It is easy to check that $L_t^* x^* = B^*S^*(t-\cdot)x^*$ for all $x^* \in E^*$. Consequently, $L_t \circ L_t^* = Q_t$, where $Q_t \in \mathcal{L}(E^*, E)$ is defined by

$$Q_t x^* := \int_0^t S(s)BB^*S^*(s)x^*\,ds. \tag{1.1}$$

It follows from this that H_t can be identified with the reproducing kernel Hilbert space of Q_t. Denoting the inclusion operator $H_t \hookrightarrow E$ by i_t, we have the operator identity

$$i_t \circ i_t^* = Q_t. \tag{1.2}$$

Moreover, by general results on reproducing kernel Hilbert spaces, the range of i_t^* is dense in H_t.

We insert a simple result on controls with minimal energy. It will not be needed in what follows and is included for reasons of completeness only. We write Λ_t for the L_t when we regard it as an operator from $L^2(0,t;H)$ onto H_t.

PROPOSITION 1.1 (control with minimal energy). *For all $h \in H_t$ we have $\Lambda_t\Lambda_t^* h = h$ and $\|\Lambda_t^* h\|^2_{L^2(0,t;H)} = \|h\|^2_{H_t}$.*

Upon identifying $h \in H_t$ with $i_t h \in E$, we have $L_t(\Lambda_t^* h) = h$. Thus, the lemma states that the control $\Lambda_t^* h$ steers 0 to h in time t with minimal energy.

Proof. For all $x^* \in E^*$ we have $\Lambda_t^* i_t^* x^* = L_t^* x^* = B^* S^*(t-\cdot)x^*$. Hence,

$$i_t\Lambda_t\Lambda_t^* i_t^* x^* = L_t\Lambda_t^* i_t^* x^* = \int_0^t S(t-s)BB^*S^*(t-s)x^*\,ds = Q_t x^* = i_t i_t^* x^*.$$

Since i_t is injective and the range of i_t^* is dense in H_t, this implies that $\Lambda_t\Lambda_t^* h = h$ for all $h \in H_t$. This proves the first assertion. The second follows from

$$\|\Lambda_t^* i_t^* x^*\|^2_{L^2(0,t;H)} = \|L_t^* x^*\|^2_{L^2(0,t;H)} = \langle L_t L_t^* x^*, x^*\rangle = \langle Q_t x^*, x^*\rangle = \|i_t^* x^*\|^2_{H_t}$$

and another density argument. □

It will be helpful to recall some elementary facts about the spaces H_t; for the proofs we refer to [9, 13]. The inequality $\langle Q_t x^*, x^*\rangle \leqslant \langle Q_{t+s}x^*, x^*\rangle$, valid for all $x^* \in E^*$, $t > 0$ and $s \geqslant 0$, implies that $H_t \subseteq H_{t+s}$ (as subsets of E) with a contractive inclusion mapping

$$i_{t,t+s} : H_t \hookrightarrow H_{t+s}, \quad i_{t,t+s}h = h \quad (h \in H_t).$$

Moreover, $S(s)$ restricts to a contraction from H_t into H_{t+s}. We will denote this restriction by $S_{t,t+s}(s)$. Thus,

$$S_{t,t+s}(s) : H_t \to H_{t+s}, \quad S_{t,t+s}(s)h = S(s)h \quad (h \in H_t).$$

2. Null controllability. The pair (A,B) is said to be *null controllable in finite time* if there exists a time $t_0 > 0$ such that for any $x \in E$ there exists a control $u \in L^2(0,t_0;H)$ such that $y^{u,x}(t_0) = 0$. If we want to stress the role of t_0, we say that (A,B) is *null controllable in time t_0*.

From the trivial identity $y^{u,x}(t_0) = S(t_0)x + y^{u,0}(t_0)$ we see that (A,B) is null controllable in time t_0 if and only if

$$S(t_0)x \in H_{t_0} \quad \text{for all } x \in E.$$

As an operator from E into H_{t_0}, we shall denote $S(t_0)$ by $\Sigma(t_0)$. Thus,

$$S(t_0) = i_{t_0} \circ \Sigma(t_0). \tag{2.1}$$

If (A,B) is null controllable in time t_0, then (A,B) is null controllable in time t for all $t \geqslant t_0$. Indeed, from $S(t_0)x \in H_{t_0}$ and the fact that $S(t-t_0)$ maps H_{t_0} into H_t we see that $S(t)x \in H_t$ for all $x \in E$. As subsets of E, the spaces of reachable points agree:

$$H_t = H_{t_0} \quad \text{with equivalent norms.}$$

The inclusion $H_{t_0} \hookrightarrow H_t$ always holds. To prove the converse inclusion $H_t \hookrightarrow H_{t_0}$, we first note that (1.1) implies the operator identity

$$Q_t = Q_{t_0} + S(t_0)Q_{t-t_0}S^*(t_0) \qquad (t \geqslant t_0). \tag{2.2}$$

Using this identity, for all $t \geqslant t_0$ and $x^* \in E^*$ we have

$$\begin{aligned}\langle Q_t x^*, x^*\rangle &= \langle Q_{t_0} x^*, x^*\rangle + \langle Q_{t-t_0} S^*(t_0)x^*, S^*(t_0)x^*\rangle \\ &= \langle Q_{t_0} x^*, x^*\rangle + \langle Q_{t-t_0}\Sigma^*(t_0) i_{t_0}^* x^*, \Sigma^*(t_0) i_{t_0}^* x^*\rangle \\ &\leqslant \langle Q_{t_0} x^*, x^*\rangle + \|Q_{t-t_0}\| \cdot \|\Sigma(t_0)\|^2 \cdot \|i_{t_0}^* x^*\|_{H_{t_0}}^2 \\ &= \big(1 + \|Q_{t-t_0}\| \cdot \|\Sigma(t_0)\|^2\big) \cdot \langle Q_{t_0} x^*, x^*\rangle.\end{aligned}$$

The inclusion $H_t \hookrightarrow H_{t_0}$ now follows from [9, Proposition 1.1]. In general, H_{t_0} and H_t will be different as Hilbert spaces, and for this reason we will distinguish between these spaces carefully.

For the rest of this section we fix $t_0 > 0$ and assume that the pair (A, B) is null controllable in time t_0.

Since (A, B) is null controllable in any time $t \geqslant t_0$, for $t \geqslant t_0$ we define $\Sigma(t)$ as $S(t)$, regarded as an operator from E into H_t. Notice that $\|\Sigma(t)x\|_{H_t}^2$ is the minimal energy to steer from x to 0 in time t. The function $t \mapsto \|\Sigma(t)x\|_{H_t}^2$ is nonincreasing on $[t_0, \infty)$: this follows from

$$\|\Sigma(t+s)x\|_{H_{t+s}}^2 = \|S_{t,t+s}(s)\Sigma(t)x\|_{H_{t+s}}^2 \leqslant \|\Sigma(t)x\|_{H_t}^2. \tag{2.3}$$

By a similar argument, for each $t \geqslant t_0$ the function $s \mapsto \|i_{t,t+s}\Sigma(t)x\|_{H_{t+s}}^2$ is nonincreasing on $[0, \infty)$. The main result of this section, Theorem 2.5, will show that this function is in fact differentiable at $s = 0$, and its derivative will be computed explicitly.

To prepare for the proof we need a series of lemmas. The first uses the identity

$$i_{t+s}^* = i_{t,t+s} i_t^* + i_{t,t+s}\Sigma(t) Q_s \Sigma^*(t) i_t^*, \tag{2.4}$$

which follows from (2.2) by using (1.2), (2.1), the trivial identity $i_t = i_{t+s} \circ i_{t,t+s}$, and the injectivity of i_{t+s}.

LEMMA 2.1. *For all $h \in H_{t_0}$ the function $t \mapsto i_{t_0,t}^* i_{t_0,t} h$ is continuous on the interval $[t_0, \infty)$.*

Proof. Fix $t' \geqslant t \geqslant t_0$ arbitrary. Since $\|i_{t_0,t}\| \leqslant 1$, for all $h \in H_{t_0}$ we have

$$\|i_{t_0,t'}^* i_{t_0,t'} h - i_{t_0,t}^* i_{t_0,t} h\|_{H_{t_0}} = \|i_{t_0,t}^*(i_{t,t'}^* i_{t,t'} - I) i_{t_0,t} h\|_{H_{t_0}} \leqslant \|(i_{t,t'}^* i_{t,t'} - I) i_{t_0,t} h\|_{H_t}.$$

Hence it suffices to prove that $\lim_{t'-t\downarrow 0} \|i_{t,t'}^* i_{t,t'} g - g\|_{H_t} = 0$ for all $g \in H_t$. We first take $g = i_t^* x^*$ with $x^* \in E^*$. Then by (2.4),

$$i_{t,t'}^* i_{t,t'} g = i_{t,t'}^*\big(i_{t'}^* x^* - i_{t,t'}\Sigma(t) Q_{t'-t}\Sigma^*(t) i_t^* x^*\big) = g - i_{t,t'}^* i_{t,t'}\Sigma(t) Q_{t'-t}\Sigma^*(t) g.$$

Since the range of i_t^* is dense in H_t, a limiting argument shows that this identity holds for all $g \in H_t$. Using (2.3), for all $g \in H_t$ we have

$$\begin{aligned}\|i_{t,t'}^* i_{t,t'} g - g\|_{H_t} &= \|i_{t,t'}^* i_{t,t'}\Sigma(t) Q_{t'-t}\Sigma^*(t) g\|_{H_t} \\ &\leqslant \|\Sigma(t)\|^2 \, \|Q_{t'-t}\| \, \|g\|_{H_t} \leqslant \|\Sigma(t_0)\|^2 \, \|Q_{t'-t}\| \, \|g\|_{H_t}.\end{aligned}$$

Since $\lim_{t'-t\downarrow 0} \|Q_{t'-t}\| = 0$, this proves that $\lim_{t'\downarrow t} \|i_{t,t'}^* i_{t,t'} g - g\|_{H_t} = 0$. ∎

The adjoint T^* of a C_0-semigroup T on a Banach space X may fail to be strongly continuous on X^*. To overcome this problem, one defines

$$X^\odot := \Big\{x^* \in X^* : \lim_{t\downarrow 0} \|T^*(t)x^* - x^*\| = 0\Big\}.$$

This is a norm closed, weak*-dense, S^*-invariant subspace of X^*, and the restricted semigroup $T^\odot = T^*\big|_{X^\odot}$ is strongly continuous on $X^\odot$. If X is reflexive, then $X^\odot$ is norm closed and weakly dense in X^*, and therefore we have $X^\odot = X^*$.

LEMMA 2.2. *For all $t \geqslant t_0$ the space H_t is S-invariant and the restricted semigroup $S_t := S|_{H_t}$ is strongly continuous on H_t.*

Proof. Invariance follows from the fact that $S(s)$ maps H_t into H_{t+s} and the fact that both H_t and H_{t+s} equal H_{t_0} as subsets of E.

Let $\delta > 0$ be arbitrary and fixed. For all $x^* \in E^*$ and $s \in [0, \delta]$ we have

$$\begin{aligned}
\|S_t^*(s) i_t^* x^*\|_{H_t}^2 &= \|i_t^* S^*(s) x^*\|_{H_t}^2 \\
&= \langle Q_{t+s} x^*, x^* \rangle - \langle Q_s x^*, x^* \rangle \\
&\leqslant \langle Q_{t+s} x^*, x^* \rangle \\
&= \langle Q_t x^*, x^* \rangle + \int_0^s \langle BB^* S^*(t+r) x^*, S^*(t+r) x^* \rangle \, dr \\
&= \|i_t^* x^*\|_{H_t}^2 + \int_0^s \langle BB^* S^*(r) \Sigma^*(t) i_t^* x^*, S^*(r) \Sigma^*(t) i_t^* x^* \rangle \, dr \\
&\leqslant \left(1 + \delta \cdot \|BB^*\| \cdot \|\Sigma(t)\|^2 \cdot \sup_{r \in [0,\delta]} \|S(r)\|^2 \right) \cdot \|i_t^* x^*\|_{H_t}^2 .
\end{aligned}$$

Hence,

$$\limsup_{s \downarrow 0} \|S_t(s)\| \leqslant \left(1 + \delta \cdot \|BB^*\| \cdot \|\Sigma(t)\|^2 \cdot \sup_{r \in [0,\delta]} \|S(r)\|^2 \right).$$

On the other hand, for all $h \in H_t$ and $x^* \in E^*$ we have

$$\lim_{s \downarrow 0} [S_t(s) h - h, i_t^* x^*]_{H_t} = \lim_{s \downarrow 0} \langle S(t) i_t h - i_t h, x^* \rangle = 0.$$

It follows that S_t is weakly continuous. By a general result from semigroup theory, this implies that S_t is strongly continuous. □

We note two immediate consequences of this lemma.

LEMMA 2.3. *For all $x \in E$ the function $t \mapsto \Sigma^*(t)\Sigma(t)x$ is continuous on the interval $[t_0, \infty)$.*

Proof. By the observations preceding Lemma 2.2, the adjoint semigroup S_t^* is strongly continuous. The lemma now follows from the identity

$$\Sigma^*(t)\Sigma(t)x = \Sigma^*(t_0) S_{t_0}^*(t - t_0) i_{t_0,t}^* i_{t_0,t} S_{t_0}(t - t_0) \Sigma(t_0) x$$

and Lemmas 2.1 and 2.2. □

LEMMA 2.4. *For all $h \in H_t$ we have $\Sigma^*(t)h \in E^\odot$.*

Proof. This follows from

$$\lim_{s \downarrow 0} \|S^*(s)\Sigma^*(t)h - \Sigma^*(t)h\| = \lim_{s \downarrow 0} \big\| \Sigma^*(t) \big(S_t^*(s)h - h \big) \big\| = 0,$$

where we used again the strong continuity of S_t^*. □

We are now ready for the main result of this section, which describes the instantaneous rate of the change of the minimal energy along curves in the space of reachable states as time progresses.

THEOREM 2.5 (rate of change of minimal energy). *Let the pair (A,B) be null controllable in time t_0. Fix $t \geqslant t_0$ and let $f : [0,\infty) \to H_t$ be differentiable at 0. The function $\phi : [0,\infty) \to [0,\infty)$ defined by*

$$\phi(s) := \|i_{t,t+s} f(s)\|^2_{H_{t+s}}$$

is differentiable at 0, with derivative

$$\phi'(0) = 2[f'(0), f(0)]_{H_t} - \|B^*\Sigma^*(t)f(0)\|^2_H.$$

Notice that the first term on the right-hand side accounts for the speed and direction of leaving $f(0)$, while the second term describes the energy savings resulting from the extra time available.

Proof. Upon writing $f(s) = f(0) + sf'(0) + g(s)$ with $\lim_{s\downarrow 0} g(s)/s = 0$ we have

$$\begin{aligned}\lim_{s\downarrow 0} \frac{1}{s}\Big[\|f(s)\|^2_{H_t} - \|f(0)\|^2_{H_t}\Big] \\ = \lim_{s\downarrow 0} \frac{1}{s}\Big[2[sf'(0)+g(s), f(0)]_{H_t} + \|sf'(0)+g(s)\|^2_{H_t}\Big] = 2[f'(0), f(0)]_{H_t}.\end{aligned}$$

Consequently, it remains to prove that

$$\lim_{s\downarrow 0} \frac{1}{s}\Big[\|i_{t,t+s}f(s)\|^2_{H_{t+s}} - \|f(s)\|^2_{H_t}\Big] = -\|B^*\Sigma^*(t)f(0)\|^2_H.$$

Let $x^* \in E^*$ be fixed. Noting that

$$\|i^*_{t+s}x^*\|^2_{H_{t+s}} - \|i^*_t x^*\|^2_{H_t} = \langle Q_{t+s}x^*, x^*\rangle - \langle Q_t x^*, x^*\rangle = \langle Q_s\Sigma^*(t)i^*_t x^*, \Sigma^*(t)i^*_t x^*\rangle,$$

from identity (2.4) we have

$$\begin{aligned}&\|i_{t,t+s}i^*_t x^*\|^2_{H_{t+s}} - \|i^*_t x^*\|^2_{H_t} \\ &\quad = \|i^*_{t+s}x^*\|^2_{H_{t+s}} - \|i^*_t x^*\|^2_{H_t} \\ &\qquad - 2[i^*_{t+s}x^*, i_{t,t+s}\Sigma(t)Q_s\Sigma^*(t)i^*_t x^*]_{H_{t+s}} + \|i_{t,t+s}\Sigma(t)Q_s\Sigma^*(t)i^*_t x^*\|^2_{H_{t+s}} \\ &\quad = \langle Q_s\Sigma^*(t)i^*_t x^*, \Sigma^*(t)i^*_t x^*\rangle \\ &\qquad - 2[i^*_t x^*, \Sigma(t)Q_s\Sigma^*(t)i^*_t x^*]_{H_t} + \|i_{t,t+s}\Sigma(t)Q_s\Sigma^*(t)i^*_t x^*\|^2_{H_{t+s}}.\end{aligned}$$

By approximation, for all $s \geqslant 0$ we obtain

$$\begin{aligned}&\|i_{t,t+s}f(s)\|^2_{H_{t+s}} - \|f(s)\|^2_{H_t} \\ &\quad = \langle Q_s\Sigma^*(t)f(s), \Sigma^*(t)f(s)\rangle \\ &\qquad - 2[f(s), \Sigma(t)Q_s\Sigma^*(t)f(s)]_{H_t} + \|i_{t,t+s}\Sigma(t)Q_s\Sigma^*(t)f(s)\|^2_{H_{t+s}}.\end{aligned}$$

Next, for any $y^\odot \in E^\odot$ we have, by strong continuity,

$$\lim_{s\downarrow 0} \frac{1}{s} Q_s y^\odot = \lim_{s\downarrow 0} \frac{1}{s}\int_0^s S(r)BB^*S^*(r)y^\odot\, dr = BB^*y^\odot.$$

Hence, using the continuity of f at 0, the fact that $\limsup_{s\downarrow 0} \frac{1}{s}\|Q_s\| < \infty$, and the fact that $\Sigma^*(t)f(0) \in E^{\odot}$ by Lemma 2.4, we obtain

$$\begin{aligned}
&\limsup_{s\downarrow 0} \left\| \frac{1}{s} Q_s \Sigma^*(t) f(s) - BB^* \Sigma^*(t) f(0) \right\| \\
&\quad \leqslant \limsup_{s\downarrow 0} \left\| \frac{1}{s} Q_s \Sigma^*(t) f(s) - \frac{1}{s} Q_s \Sigma^*(t) f(0) \right\| \\
&\qquad\qquad + \limsup_{s\downarrow 0} \left\| \frac{1}{s} Q_s \Sigma^*(t) f(0) - BB^* \Sigma^*(t) f(0) \right\| \\
&\quad \leqslant \|\Sigma^*(t)\| \cdot \limsup_{s\downarrow 0} \left(\frac{1}{s} \|Q_s\| \right) \cdot \limsup_{s\downarrow 0} \|f(s) - f(0)\| \\
&\qquad\qquad + \limsup_{s\downarrow 0} \left\| \frac{1}{s} Q_s \Sigma^*(t) f(0) - BB^* \Sigma^*(t) f(0) \right\| = 0.
\end{aligned}$$

It follows that

$$\lim_{s\downarrow 0} \frac{1}{s} Q_s \Sigma^*(t) f(s) = BB^* \Sigma^*(t) f(0).$$

As a consequence,

$$\begin{aligned}
&\lim_{s\downarrow 0} \frac{1}{s} \Big[\langle Q_s \Sigma^*(t) f(s), \Sigma^*(t) f(s) \rangle - 2[f(s), \Sigma(t) Q_s \Sigma^*(t) f(s)]_{H_t} \\
&\qquad\qquad + \|i_{t,t+s} \Sigma(t) Q_s \Sigma^*(t) f(s)\|^2_{H_{t+s}} \Big] \\
&= \lim_{s\downarrow 0} \left\langle \frac{1}{s} Q_s \Sigma^*(t) f(s), \Sigma^*(t) f(s) \right\rangle - 2 \lim_{s\downarrow 0} \left[f(s), \Sigma(t) \left(\frac{1}{s} Q_s \Sigma^*(t) f(s) \right) \right]_{H_t} \\
&\qquad\qquad + \lim_{s\downarrow 0} s \left\| i_{t,t+s} \Sigma(t) \left(\frac{1}{s} Q_s \Sigma^*(t) f(s) \right) \right\|^2_{H_{t+s}} \\
&= \langle BB^* \Sigma^*(t) f(0), \Sigma^*(t) f(0) \rangle - 2[f(0), \Sigma(t) BB^* \Sigma^*(t) f(0)]_{H_t} + 0 \\
&= -\|B^* \Sigma^*(t) f(0)\|^2_H;
\end{aligned}$$

in the next to last step we used that $\|i_{t,t+s}\| \leqslant 1$. □

For the convenience of those readers familiar with the Hilbert space formalism as used, e.g., in [10], we add a reformulation of Theorem 2.5 for Hilbert spaces E. In this setting we identify E and its dual in the usual way and identify Q_t with a positive self-adjoint operator on E. As is well known, the reproducing kernel Hilbert space of Q_t is then given by

$$i_t(H_t) = \operatorname{Im} Q_t^{1/2}. \tag{2.5}$$

In what follows we identify $i_t(H_t)$ and H_t and abuse notation by regarding both $Q_t^{1/2}$ and Q_t as operators from E to H_t whenever this is convenient. Denoting the closure of H_t in E by E_t, it follows from (2.5) and a standard argument that $Q_t^{1/2}$ is unitary as an operator from E_t to H_t.

By (2.5), the pair (A, B) is null controllable in time t_0 if and only if $\operatorname{Im} S(t_0) \subseteq \operatorname{Im} Q_{t_0}^{1/2}$. Since the restriction of $Q_{t_0}^{1/2}$ to E_{t_0} is injective, the inverse $Q_{t_0}^{-1/2}$ is well-defined on the linear subspace H_{t_0} of E. Then by null controllability, the operator

$\Gamma(t_0) := Q_{t_0}^{-1/2} S(t_0)$ is well-defined as a bounded operator from E to E_{t_0}. For all $h = Q_{t_0}^{1/2} y \in H_{t_0}$ we have

$$[\Gamma(t_0)x, h]_E = [S(t_0)x, y]_E = [x, S^*(t_0)y]_E = [x, S^*(t_0)Q_{t_0}^{-1/2}h]_E.$$

Since H_{t_0} is dense in E_{t_0} we see that $\Gamma^*(t_0) := (\Gamma(t_0))^*$ is the unique extension of $S^*(t_0)Q_{t_0}^{-1/2}$ to a bounded operator from E_{t_0} to E.

COROLLARY 2.6. *Let the pair (A, B) be null controllable in time t_0. Fix $t \geqslant t_0$ and let $g : [0, \infty) \to E_t$ be differentiable at 0. The function $\phi : [0, \infty) \to [0, \infty)$ defined by*

$$\phi(s) := \|Q_t^{1/2} g(s)\|^2_{H_{t+s}} \tag{2.6}$$

is differentiable at 0, with derivative

$$\phi'(0) = 2[g'(0), g(0)]_E - [Q\Gamma^*(t)g(0), \Gamma^*(t)g(0)]_E.$$

Note some further abuse of notation in (2.6), where $Q_t^{1/2} g(s)$ is regarded as an element of H_{t+s}.

Proof. Let $f : [0, \infty) \to H_t$ be defined by $f(s) = Q_t^{1/2} g(s)$. Since $Q_t^{1/2}$ is unitary as an operator from E_t to H_t, f is differentiable at 0 with derivative $f'(0) = Q_t^{1/2} g'(0)$. Let $Q := BB^*$. By Theorem 2.5,

$$\phi(s) := \|i_{t,t+s} f(s)\|^2_{H_{t+s}} = \|Q_t^{1/2} g(s)\|^2_{H_{t+s}}$$

is differentiable at 0 with derivative

$$\begin{aligned} \phi'(0) &= 2[f'(0), f(0)]_{H_t} - \|B^*\Sigma^*(t)f(0)\|^2_H \\ &= 2[Q_t^{1/2} g'(0), Q_t^{1/2} g(0)]_{H_t} - [Q\Gamma^*(t)g(0), \Gamma^*(t)g(0)]_E \\ &= 2[g'(0), g(0)]_E - [Q\Gamma^*(t)g(0), \Gamma^*(t)g(0)]_E. \end{aligned} \tag{2.7}$$

In the second identity of (2.7) we used that $\Gamma^*(t)$ extends $S^*(t)Q_t^{-1/2}$ on E_t and that for all $h = Q_t y \in H_t$ we have

$$[B^*\Sigma^*(t)h, B^*\Sigma^*(t)h]_H = [Q\Sigma^*(t)i_t^* y, \Sigma^*(t)i_t^* y]_E = [Q^* S(t)y, S^*(t)y]_E,$$

recalling that we identify $Q_t y = i_t i_t^* y$ and $i_t^* y$. In the third identity of (2.7) we used that $Q_t^{1/2}$ is unitary from E_t to H_t. ∎

3. Null controllability with vanishing energy. Following Priola and Zabczyk [10] we call the pair (A, B) *null controllable with vanishing energy* if for all $\varepsilon > 0$ and $x \in E$ there exists a time $t > 0$ and a control $u \in L^2(0, t; H)$ with $y^{u,x}(t) = 0$ and $\|u\|_{L^2(0,t;H)} < \varepsilon$. Clearly, null controllability with vanishing energy implies null controllability with bounded energy.

THEOREM 3.1. *If the pair (A, B) is null controllable with vanishing energy, then it is null controllable in finite time.*

Proof. For $n = 1, 2, \ldots$, let E_n denote the set of all $x \in E$ for which there exists a control $u \in L^2(0, n; H)$ with $y^{u,x}(n) = 0$ and $\|u\|_{L^2(0,n;H)} \leqslant 1$. Notice that $\bigcup_{n \geqslant 1} E_n = E$.

We claim that each E_n is closed. To see this, fix $n \geqslant 1$ and let $\lim_{k\to\infty} x_k = x$ in E with all $x_k \in E_n$. We must check that $x \in E_n$. For each k we choose a

control $u_k \in L^2(0,n;H)$ with $y^{u_k,x_k}(n) = 0$ and $\|u_k\|_{L^2(0,n;H)} \leqslant 1$. After passing to a subsequence, we may assume that there exists a control $u \in L^2(0,n;H)$ with $\|u\|_{L^2(0,n;H)} \leqslant 1$ such that $\lim_{k\to\infty} u_k = u$ weakly in $L^2(0,n;H)$. Then for all $x^* \in E^*$ we have

$$\begin{aligned}\langle y^{u,x}(n), x^*\rangle &= \langle S(n)x, x^*\rangle + \int_0^n [u(s), B^*S^*(n-s)x^*]_H\, ds \\ &= \lim_{k\to\infty}\left(\langle S(n)x_k, x^*\rangle + \int_0^n [u_k(s), B^*S^*(n-s)x^*]_H\, ds\right) \\ &= \lim_{k\to\infty}\langle y^{u_k,x_k}(n), x^*\rangle = 0.\end{aligned}$$

Hence $y^{u,x}(n) = 0$ and $x \in E_n$.

By the Baire category theorem, at least one E_{n_0} has a nonempty interior. Fix an arbitrary x_0 in the interior of E_{n_0} and consider the set $E_{n_0} - x_0$. This is a neighborhood of 0 consisting of elements that can be steered to 0 in time n_0. By linearity it follows that every $x \in E$ can be steered to 0 in time n_0. This means that the pair (A,B) is null controllable in time n_0. □

Recall that if (A,B) is null controllable in time t_0, then for all $t \geqslant t_0$ the square norm $\|\Sigma(t)x\|^2_{H_t}$ is the minimal energy to steer from x to 0 in time t. Hence the following observation is a straightforward consequence of (2.3) and the above theorem.

COROLLARY 3.2. *The following assertions are equivalent:*

1. *The pair (A,B) is null controllable with vanishing energy.*
2. *The pair (A,B) is null controllable in finite time and $\lim_{t\to\infty}\|\Sigma(t)x\|_{H_t} = 0$ for all $x \in E$.*

We proceed with two simple examples of systems that are null controllable with vanishing energy.

Example 3.3. If (A,B) is null controllable in finite time and the semigroup S generated by A is *strongly stable*, i.e., if $\lim_{t\to\infty} S(t)x = 0$ for all $x \in E$, then (A,B) is null controllable with vanishing energy. Indeed, if (A,B) is null controllable in time t_0, then for all $t \geqslant t_0$ we have

$$\|\Sigma(t)x\|_{H_t} = \|i_{t_0,t}\Sigma(t_0)S(t-t_0)x\|_{H_t} \leqslant \|\Sigma(t_0)\|\,\|S(t-t_0)x\|.$$

Example 3.4. The range of B is a Hilbert space with norm

$$\|Bh\|_{\operatorname{range} B} = \inf\bigl\{\|h'\|_H :\ Bh' = Bh\bigr\}.$$

With this norm, the range of B equals the reproducing kernel Hilbert space of the operator BB^*. Accordingly we shall denote the range of B by H_{BB^*}. If S restricts to a C_0-semigroup S_B on H_{BB^*}, then it follows from [6, Theorem 3.5] that the reachable spaces H_t for the pair (A,B) coincide with the reproducing kernel space of the operators $R_t \in \mathcal{L}(H_{BB^*})$ defined by

$$R_t h := \int_0^t S_B(s)S_B^*(s)h\, ds \qquad (h \in H_{BB^*}),$$

and the pair (S_B, I_B) is null controllable for all times $t > 0$. Here I_B denotes the identity operator on H_{BB^*}. It follows from the same reference that for all $h \in \operatorname{range} B$ and $t > 0$ we have an estimate

$$\|\Sigma_B(t)h\|^2_{H_t} \leqslant \frac{1}{t^2}\int_0^t \|S_B(s)h\|^2_{H_{BB^*}}\, ds \qquad (h \in H_{BB^*}).$$

Here, $\Sigma_B(t)$ denotes $S_B(t)$, regarded as an operator from H_{BB^*} into H_t. In particular the pair (S_B, I_B) is null controllable with vanishing energy if the semigroup S_B is uniformly bounded on H_{BB^*}.

In [10], under the assumption that E is a Hilbert space it was shown by control theoretic methods that a pair (A, B) which is null controllable in finite time is null controllable with vanishing energy if and only if the algebraic Riccati equation

$$XA + A^*X - XBB^*X = 0 \tag{3.1}$$

admits $X = 0$ as its only positive self-adjoint solution. A *solution* of (3.1) is a bounded operator $X \in \mathcal{L}(E)$ such that

$$\langle XAx, y\rangle + \langle Xx, Ay\rangle - \langle XBB^*Xx, y\rangle = 0 \quad \text{for all } x, y \in D(A). \tag{3.2}$$

In this identity the brackets denote the scalar product of E.

In this section we shall prove an extension of this result to Banach spaces E. It shares with [10] the strategy of first solving a differential Riccati equation and obtaining the final characterization from a maximality argument, but both steps are accomplished in a completely different way. In the Banach space setting, a *solution* of (3.1) is a bounded operator $X \in \mathcal{L}(E, E^*)$ such that (3.2) holds for all $x, y \in D(A)$; this time the brackets denote the duality pairing between E^* and E. The notions of positivity and self-adjointness extend as follows: we call $X \in \mathcal{L}(E, E^*)$ *positive* if $\langle Xx, x\rangle \geqslant 0$ for all $x \in E$ and *symmetric* if $\langle Xx, y\rangle = \langle Xy, x\rangle$ for all $x, y \in E$.

We begin with a result which states that the operator function $t \mapsto \Sigma^*(t)\Sigma(t)$ solves, in some appropriate sense, the differential Riccati equation

$$\frac{d}{dt}X(t) = X(t)A + A^*X(t) - X(t)BB^*X(t)$$

on the interval $[t_0, \infty)$.

In the Hilbert space literature, existence of a solution is usually derived from a fixed point argument. Here, we obtain it as a direct consequence of Theorem 2.5.

PROPOSITION 3.5. *Let the pair (A, B) be null controllable in time t_0. For all $x, y \in D(A)$ the function $t \mapsto \langle \Sigma^*(t)\Sigma(t)x, y\rangle$ is differentiable on the interval $[t_0, \infty)$, with derivative*

$$\begin{aligned}\frac{d}{dt}&\langle \Sigma^*(t)\Sigma(t)x, y\rangle \\ &= \langle \Sigma^*(t)\Sigma(t)Ax, y\rangle + \langle \Sigma^*(t)\Sigma(t)x, Ay\rangle - \langle BB^*\Sigma^*(t)\Sigma(t)x, \Sigma^*(t)\Sigma(t)y\rangle.\end{aligned}$$

Proof. Since both BB^* and $\Sigma^*(t)\Sigma(t)$ are symmetric operators, by polarization it suffices to prove that for all $x \in D(A)$ and $t \geqslant t_0$ we have

$$\frac{d}{dt}\langle \Sigma^*(t)\Sigma(t)x, x\rangle = 2\langle \Sigma^*(t)\Sigma(t)Ax, x\rangle - \langle BB^*\Sigma^*(t)\Sigma(t)x, \Sigma^*(t)\Sigma(t)x\rangle.$$

For this, in turn, it suffices to prove right differentiability. Indeed, by Lemma 2.3 the functions $\langle \Sigma^*(t)\Sigma(t)x, x\rangle$ and $2\langle \Sigma^*(t)\Sigma(t)Ax, x\rangle - \langle BB^*\Sigma^*(t)\Sigma(t)x, \Sigma^*(t)\Sigma(t)x\rangle$ are continuous functions of $t \in [t_0, \infty)$, and by elementary calculus a continuous function that is right differentiable with continuous right derivative is differentiable; cf. [13].

Fix $x \in D(A)$ and $t \geqslant t_0$. By Theorem 2.5 applied to $f(s) = \Sigma(t)S(s)x$ we have

$$\begin{aligned}\lim_{s\downarrow 0} \frac{1}{s}\Big(\langle \Sigma^*(t+s)\Sigma(t+s)x, x\rangle - \langle \Sigma^*(t)\Sigma(t)x, x\rangle\Big)\\ &= \lim_{s\downarrow 0} \frac{1}{s}\Big(\|i_{t,t+s}\Sigma(t)S(s)x\|^2_{H_{t+s}} - \|\Sigma(t)x\|^2_{H_t}\Big)\\ &= 2[\Sigma(t)Ax, \Sigma(t)x]_{H_t} - \|B^*\Sigma^*(t)\Sigma(t)x\|^2_H\\ &= 2\langle \Sigma^*(t)\Sigma(t)Ax, x\rangle - \langle BB^*\Sigma^*(t)\Sigma(t)x, \Sigma^*(t)\Sigma(t)x\rangle. \qquad \square\end{aligned}$$

Remark 3.6. In the special case where E is a Hilbert space, instead of using Theorem 2.5 we could apply Corollary 2.6 to the E_t-valued function $g(s) := \Gamma(t)S(s)$; note that $Q_t^{1/2} g(s) = \Sigma(t)S(s)x = f(s)$.

From Proposition 3.5 we obtain the following.

PROPOSITION 3.7. *Let the pair (A, B) be null controllable in time t_0. For all $x, y \in E$ the limit $\lim_{t\to\infty}\langle \Sigma^*(t)\Sigma(t)x, y\rangle$ exists, and the operator $X \in \mathcal{L}(E, E^*)$ defined by*

$$\langle Xx, y\rangle := \lim_{t\to\infty} \langle \Sigma^*(t)\Sigma(t)x, y\rangle \tag{3.3}$$

defines a positive symmetric solution of the algebraic Riccati equation

$$XA + A^*X - XBB^*X = 0.$$

Proof. For all $x \in E$ we have $\langle \Sigma^*(t)\Sigma(t)x, x\rangle = \|\Sigma(t)x\|^2_{H_t}$, which is a nonincreasing function of $t \geqslant t_0$. In particular, for all $x \in E$ the limit $\lim_{t\to\infty}\langle \Sigma^*(t)\Sigma(t)x, x\rangle$ exists. Since each $\Sigma^*(t)\Sigma(t)$ is positive and symmetric, by polarization it follows that for all $x, y \in E$ the limit $\lim_{t\to\infty}\langle \Sigma^*(t)\Sigma(t)x, y\rangle$ exists, and then (3.3) defines a positive and symmetric operator X.

Since $t \mapsto \Sigma^*(t)\Sigma(t)$ solves the differential Riccati equation, a standard argument implies that X solves the algebraic Riccati equation. $\square$

Our next aim is to show that the weak operator limit $X = \lim_{t\to\infty} \Sigma^*(t)\Sigma(t)$ is in fact the *maximal* symmetric solution of the algebraic Riccati equation. More precisely we have the following.

THEOREM 3.8. *Let the pair (A, B) be null controllable at time $t_0 > 0$. If Y is a symmetric solution of the algebraic Riccati equation, then for all $x \in E$ we have $\langle Yx, x\rangle \leqslant \langle Xx, x\rangle$.*

Proof. Fix $t \geqslant t_0$ and $x \in E$, and let $u \in L^2(0, t; H)$ be any control steering x to 0 in time t:

$$y^{u,x}(t) = S(t)x + \int_0^t S(t-s)Bu(s)\, ds = 0.$$

We will show that the function $f_u : [0, t] \to \mathbb{R}$ defined by

$$f_u(s) := \int_0^s \|u(r)\|^2_H\, dr + \langle Y y^{u,x}(s), y^{u,x}(s)\rangle$$

is nondecreasing. To prove this we shall show that f_u is almost everywhere differentiable with nonnegative derivative.

Let us first consider the function $g_u(s) := \langle Y y^{u,x}(s), y^{u,x}(s)\rangle$. In order to show that g_u is differentiable we introduce a regularization operator as follows. For $\lambda > 0$ large enough, put $E_\lambda := \lambda(\lambda - A)^{-1}$ and define

$$g_{u,\lambda}(s) := \langle Y E_\lambda y^{u,x}(s), E_\lambda y^{u,x}(s)\rangle.$$

Then, by the symmetry of Y and the fact that this operator solves the algebraic Riccati equation,

$$\begin{aligned}
g'_{u,\lambda}(s) &= 2\left\langle YE_\lambda y^{u,x}(s), \frac{d}{ds}E_\lambda y^{u,x}(s)\right\rangle \\
&= 2\left\langle YE_\lambda y^{u,x}(s), \frac{d}{ds}\left(S(s)E_\lambda x + \int_0^s S(s-r)E_\lambda Bu(r)\,dr\right)\right\rangle \\
&= 2\left\langle YE_\lambda y^{u,x}(s), A\left(S(s)E_\lambda x + \int_0^s S(s-r)E_\lambda Bu(r)\,dr\right) + E_\lambda Bu(s)\right\rangle \\
&= \langle YBB^*YE_\lambda y^{u,x}(s), E_\lambda y^{u,x}(s)\rangle + 2\langle YE_\lambda y^{u,x}(s), E_\lambda Bu(s)\rangle \\
&=: G_{u,\lambda}(s).
\end{aligned}$$

From $\lim_{\lambda\to\infty} E_\lambda = I$ strongly we have $\lim_{\lambda\to\infty} g_{u,\lambda} = g_u$ and

$$\lim_{\lambda\to\infty} G_{u,\lambda} = \langle YBB^*Yy^{u,x}(s), y^{u,x}(s)\rangle + 2\langle Yy^{u,x}(s), Bu(s)\rangle$$

uniformly on $[0,t]$ (notice that $y^{u,x}$ is continuous on $[0,t]$). The closedness of the first derivative now implies that g_u is differentiable, with derivative

$$g'_u(s) = \langle YBB^*Yy^{u,x}(s), y^{u,x}(s)\rangle + 2\langle Yy^{u,x}(s), Bu(s)\rangle.$$

It follows that f_u is almost everywhere differentiable, with derivative

$$\begin{aligned}
f'_u(s) &= \|u(s)\|_H^2 + \langle YBB^*Yy^{u,x}(s), y^{u,x}(s)\rangle + 2\langle Yy^{u,x}(s), Bu(s)\rangle \\
&= \|u(s)\|_H^2 + \|B^*Yy^{u,x}(s)\|_H^2 + 2[B^*Yy^{u,x}(s), u(s)]_H \\
&\geqslant \|u(s)\|_H^2 + \|B^*Yy^{u,x}(s)\|_H^2 - 2\|B^*Yy^{u,x}(s)\|_H\|u(s)\|_H \\
&= \big(\|u(s)\|_H - \|B^*Yy^{u,x}(s)\|_H\big)^2,
\end{aligned}$$

which is nonnegative.

By what has been shown so far, we have

$$\begin{aligned}
\|u\|^2_{L^2(0,t;H)} &= \int_0^t \|u(r)\|_H^2\,dr = \int_0^t \|u(r)\|_H^2\,dr + \langle Yy^{u,x}(t), y^{u,x}(t)\rangle \\
&= f_u(t) \geqslant f_u(0) = \langle Yy^{u,x}(0), y^{u,x}(0)\rangle = \langle Yx, x\rangle.
\end{aligned}$$

Taking the infimum over all admissible controls we obtain

$$\|\Sigma(t)x\|^2_{H_t} \geqslant \langle Yx, x\rangle.$$

Finally, letting $t\to\infty$, this gives

$$\langle Xx, x\rangle = \lim_{t\to\infty} \|\Sigma(t)x\|^2_{H_t} \geqslant \langle Yx, x\rangle. \qquad \square$$

The preceding two results may now be combined to prove the following characterization of null controllability with vanishing energy, which extends the corresponding Hilbert space result of [10] to Banach spaces.

THEOREM 3.9. *The following assertions are equivalent:*

1. *The pair (A,B) is null controllable with vanishing energy.*
2. *The pair (A,B) is null controllable in finite time and the only positive symmetric solution of the algebraic Riccati equation $XA + A^*X - XBB^*X = 0$ is the trivial solution $X = 0$.*

Proof. We will use Corollary 3.2.

(1)⇒(2): Let Y be any positive symmetric solution of the algebraic Riccati equation. Then for all $x \in E$ we have

$$0 \leqslant \langle Yx, x\rangle \leqslant \langle Xx, x\rangle = \lim_{t\to\infty} \|\Sigma(t)x\|_{H_t}^2 = 0,$$

which implies that $Y = 0$.

(2)⇒(1): Since $X = \lim_{t\to\infty} \Sigma^*(t)\Sigma(t)$ is a positive symmetric solution of the algebraic Riccati equation, it follows that $\lim_{t\to\infty} \|\Sigma(t)x\|_{H_t}^2 = \langle Xx, x\rangle = 0$ for all $x \in E$. □

Under additional spectral assumptions (which are satisfied, e.g., if S is eventually compact), it is shown in [10] that the pair (A, B) is null controllable with vanishing energy if and only if $\sup\{\operatorname{Re}\lambda : \lambda \in \sigma(A)\} \leqslant 0$. This result is applied in [11], where it is used to obtain necessary and sufficient conditions for the validity of Liouville's theorem for the Ornstein–Uhlenbeck operator associated with the pair (A, B).

As an application of Theorem 3.9 we give a sufficient condition for null controllability with vanishing energy in the symmetric case.

THEOREM 3.10. *Let the pair (A, B) be null controllable at time $t_0 > 0$. Assume furthermore that*

- *(nondegeneracy) B has dense range,*
- *(BB^*-symmetry) $S(t)BB^* = BB^*S^*(t)$ for all $t \geqslant 0$.*

If the limit $Q_\infty := \lim_{t\to\infty} Q_t$ exists in the weak operator topology, then (A, B) is null controllable with vanishing energy.

Without any nondegeneracy condition on B, the assumptions of the theorem imply that S restricts to a strongly stable C_0-semigroup of contractions S_B on the range of B [6, Theorem 4.5]. By Examples 3.3 and 3.4, the pair (S_B, I_B) is null controllable with vanishing energy.

Proof. We shall use the fact that $Q_\infty := \lim_{t\to\infty} Q_t$ exists in the weak operator topology if and only if there exists a positive symmetric solution in $\mathcal{L}(E^*, E)$ of the Lyapunov equation

$$AY + YA^* + BB^* = 0$$

and that in this case Q_∞ is the minimal positive symmetric solution of this equation [6, Theorem 4.4]. In this context, a bounded operator $Y \in \mathcal{L}(E^*, E)$ is called *positive* if $\langle Yx, x\rangle \geqslant 0$ for all $x \in E$ and *symmetric* if $\langle Yx, y\rangle = \langle Yy, x\rangle$ for all $x, y \in E$.

Assume now that $X \in \mathcal{L}(E, E^*)$ is a positive symmetric solution of the algebraic Riccati equation. We have to show that $X = 0$.

Since B is assumed to have dense range, it is an easy consequence of the Hahn–Banach theorem that BB^* is injective and has dense range as well. From this it follows that Q_∞ is injective and has dense range [5, Lemma 5.2].

By the same argument as in the proof of [6, Theorem 4.5], the assumption $S(t)BB^* = BB^*S(t)$ implies that the semigroup S_t on H_t is self-adjoint for all $t \geqslant t_0$. Moreover, for all $x \in D(A)$ we have $\Sigma(t)x \in D(A_t)$ and $A_t\Sigma(t)x = \Sigma(t)Ax$. Similarly, for all $h \in D(A_t^*)$ we have $\Sigma^*(t)h \in D(A^*)$ and $A^*\Sigma^*(t)h = \Sigma^*(t)A_t^*h$. Using these facts, for all $x, y \in D(A)$ we obtain

$$\begin{aligned}\langle Xx, Ay\rangle &= \lim_{t\to\infty} \langle \Sigma^*(t)\Sigma(t)x, Ay\rangle = \lim_{t\to\infty} \langle \Sigma^*(t)A_t^*\Sigma(t)x, y\rangle \\ &= \lim_{t\to\infty} \langle \Sigma^*(t)A_t\Sigma(t)x, y\rangle = \lim_{t\to\infty} \langle \Sigma^*(t)\Sigma(t)Ax, y\rangle = \langle XAx, y\rangle.\end{aligned}$$

It follows that $Xx \in D(A^*)$ and $A^*Xx = XAx$. Thus, $A^*X = XA$. Similarly one proves that $AQ_\infty = Q_\infty A^*$. As X and Q_∞ are symmetric and solve the algebraic Riccati equation and the Lyapunov equation, respectively, for all $x^*, y^* \in D(A^*)$ we obtain

$$\begin{aligned}
0 &= \langle A^*XQ_\infty x^*, Q_\infty y^*\rangle + \langle XAQ_\infty x^*, Q_\infty y^*\rangle - \langle XBB^*XQ_\infty x^*, Q_\infty y^*\rangle \\
&= \langle XQ_\infty A^* x^*, Q_\infty y^*\rangle + \langle XAQ_\infty y^*, Q_\infty x^*\rangle - \langle XBB^*XQ_\infty x^*, Q_\infty y^*\rangle \\
&= -\langle XBB^* x^*, Q_\infty y^*\rangle - \langle XBB^*XQ_\infty x^*, Q_\infty y^*\rangle \\
&= -\langle Q_\infty XBB^* x^*, y^*\rangle - \langle Q_\infty XBB^*XQ_\infty x^*, y^*\rangle.
\end{aligned}$$

Thus,

$$\langle Q_\infty XBB^*(I + XQ_\infty)x^*, y^*\rangle = 0 \tag{3.4}$$

for all $x^*, y^* \in D(A^*)$. Since $D(A^*)$ is weak*-dense, it follows that

$$Q_\infty XBB^*(I + XQ_\infty)x^* = 0 \tag{3.5}$$

for all $x^* \in D(A^*)$. Furthermore, by the symmetry of Q_∞, X, and BB^*, from (3.4) we obtain

$$\langle (I + Q_\infty X)BB^*XQ_\infty y^*, x^*\rangle = 0$$

for all $x^*, y^* \in D(A^*)$. Since $D(A^*)$ is weak*-dense, it follows that

$$(I + Q_\infty X)BB^*XQ_\infty y^* = 0 \tag{3.6}$$

for all $y^* \in D(A^*)$. Taking $y^* = x^*$ and subtracting (3.5) and (3.6), we find

$$BB^*XQ_\infty x^* = Q_\infty XBB^* x^*$$

for all $x^* \in D(A^*)$. Hence, by (3.6),

$$(I + Q_\infty X)Q_\infty XBB^* x^* = 0$$

for all $x^* \in D(A^*)$. Since $D(A^*)$ is weak*-dense and BB^* is weak*-to-weakly continuous and has weakly dense range, this implies that

$$(I + Q_\infty X)Q_\infty X = 0$$

or, equivalently, $P(I - P) = 0$, where $P := -Q_\infty X$. Thus, P is a projection in E.

For any $x \in \ker P$ we have $Q_\infty Xx = 0$ and therefore $Xx = 0$ by the injectivity of Q_∞.

For any $x \in \ker(I - P)$ we have $-Q_\infty Xx = x$ and therefore

$$0 \leqslant \langle Xx, x\rangle = -\langle Xx, Q_\infty Xx\rangle = -\langle Q_\infty Xx, Xx\rangle \leqslant 0$$

by the positivity of Q_∞. It follows that $\langle Q_\infty Xx, Xx\rangle = \|i_\infty^* Xx\|_{H_\infty}^2 = 0$, where $i_\infty : H_\infty \hookrightarrow E$ denotes the reproducing kernel Hilbert space of Q_∞. Since $Q_\infty = i_\infty \circ i_\infty^*$ is injective, i_∞^* is injective, and we conclude that $Xx = 0$.

Combining the facts just proved, we obtain that $Xx = 0$ for all $x \in E$, i.e., $X = 0$. □

It is worthwhile to point out that Theorem 3.10 is not covered by Example 3.3, since the existence of Q_∞ does not imply strong stability of the semigroup S.

Example 3.11. Let $E = \mathbb{R}^2$ and $S(t) = \left(\begin{smallmatrix} e^{-t} & 0 \\ 0 & 1 \end{smallmatrix}\right)$. The semigroup S is not strongly stable. Taking $H = \mathbb{R}$ and $Bh = (h, 0)$, the limit $Q_\infty = \lim_{t\to\infty} Q_t$ exists: we have

$$\lim_{t\to\infty} Q_t = \lim_{t\to\infty} \int_0^t \begin{pmatrix} e^{-2s} & 0 \\ 0 & 0 \end{pmatrix} ds = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}.$$

Let us finally observe that in Theorem 3.10 the condition on existence of Q_∞ is not a necessary one (take $E = H = \mathbb{R}$, $B = I$, and $S(t) = I$), nor can it be dropped (take $E = H = \mathbb{R}$, $B = I$, and $S(t) = e^t I$).

Acknowledgment. The author thanks Professors Jerzy Zabczyk and Ben Goldys for stimulating discussions and kind hospitality.

REFERENCES

[1] V. Barbu and G. Da Prato, *Hamilton-Jacobi Equations in Hilbert Spaces*, Research Notes in Math., Pitman, Boston, MA, 1986.

[2] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite-Dimensional Systems, Vol.* 1, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 1992.

[3] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite-Dimensional Systems, Vol.* 2, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 1993.

[4] R. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[5] B. Goldys, F. Gozzi, and J. M. A. M. van Neerven, *On closability of directional gradients*, Potential Anal., 18 (2003), pp. 289–310.

[6] B. Goldys and J. M. A. M. van Neerven, *Transition semigroups of Banach space valued Ornstein-Uhlenbeck processes*, Acta Appl. Math., 76 (2003), pp. 283–330.

[7] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories.* I, Encyclopedia Math. Appl. 74, Cambridge University Press, Cambridge, UK, 2000.

[8] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories.* II, Encyclopedia Math. Appl. 75, Cambridge University Press, Cambridge, UK, 2000.

[9] J. M. A. M. van Neerven, *Nonsymmetric Ornstein-Uhlenbeck semigroups in Banach spaces*, J. Funct. Anal., 155 (1998), pp. 495–535.

[10] E. Priola and J. Zabczyk, *Null controllability with vanishing energy*, SIAM J. Control Optim., 42 (2003), pp. 1013–1032.

[11] E. Priola and J. Zabczyk, *Liouville Theorems in Finite and Infinite Dimensions*, Preprint 9/2003, Scuola Normale Superiore, Pisa, 2003.

[12] M. Sîrbu, *A Riccati equation approach to the null controllability of linear systems*, Commun. Appl. Anal., 6 (2002), pp. 163–177.

[13] J. Zabczyk, *Mathematical Control Theory: An Introduction*, Birkhäuser Boston, Boston, MA, 1992.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1328–1353

BOUNDS AND ASYMPTOTIC APPROXIMATIONS FOR UTILITY PRICES WHEN VOLATILITY IS RANDOM*

RONNIE SIRCAR† AND THALEIA ZARIPHOPOULOU‡

Abstract. This paper is a contribution to the valuation of derivative securities in a stochastic volatility framework, which is a central problem in financial mathematics. The derivatives to be priced are of European type with the payoff depending on both the stock and the volatility. The valuation approach uses utility-based criteria under the assumption of exponential risk preferences. This methodology yields the indifference prices as solutions to second order quasilinear PDEs. Two sets of price bounds are derived that highlight the important ingredients of the utility approach, namely, nonlinear pricing rules with dynamic certainty equivalent characteristics, and pricing measures depending on correlation and the Sharpe ratio of the traded asset. The problem is further analyzed by asymptotic methods in the limit of the volatility being a fast mean-reverting process. The analysis relates the traditional market-selected volatility risk premium approach and the preference-based valuation techniques.

Key words. financial mathematics, derivative pricing, stochastic volatility, utility indifference pricing

AMS subject classifications. 91B28, 91B70, 93E20, 34E05

DOI. 10.1137/S0363012903409253

1. Introduction.

We study the utility indifference pricing mechanism for European derivative contracts in financial markets with uncertain volatility. As is well known, in such incomplete markets, there are many possible no-arbitrage pricing (or "risk-neutral") measures and typically an interval of arbitrage-free option prices. The traditional pricing methodology is that the market selects a pricing probability measure that is reflected in the prices of liquidly traded derivative contracts (for example, at-the-money call options). Indifference pricing is an alternative mechanism whereby a no-arbitrage price is selected according to investment optimality criteria of a risk-averse investor. Our analysis, using bounds and asymptotic approximations, sheds some light on the relation between the two. Specifically, Theorem 3.2 shows that the nonlinear utility pricing rule lies between a linear no-arbitrage pricing rule and an insurance-type certainty equivalent pricing rule.

Stochastic volatility models are popular because they capture the deviation of stock price data from the Black–Scholes geometric Brownian motion model in a parsimonious way. They were originally introduced in the late 1980's by Hull and White [22] and others for option pricing. Much of their success derives from their predicted option prices exhibiting the implied volatility skew that is observed in many options markets. See [17], for example, for details.

However, a market with stochastic volatility is incomplete in that volatility is a source of uncertainty that is not traded. Therefore, enforcement of *no arbitrage* does not lead to a unique derivative pricing rule. The usual way to "close" the model is

*Received by the editors April 21, 2002; accepted for publication (in revised form) April 9, 2004; published electronically January 5, 2005.

http://www.siam.org/journals/sicon/43-4/40925.html

†Department of Operations Research & Financial Engineering, Princeton University, E-Quad, Princeton, NJ 08544 (sircar@princeton.edu). The research of this author was partially supported by NSF grant SES-0111499.

‡Departments of Mathematics and Management Science and Information Systems, The University of Texas at Austin, Austin, TX 78712 (zariphop@math.utexas.edu). The research of this author was partially supported by NSF grants DMS-0102909 and DMS-0091946.

to assume the market chooses a pricing measure which is implicit in the prices of liquidly traded options. The indifference pricing mechanism is an alternative in which the price is uniquely (and endogenously) determined at the cost of depending on the preferences of the pricer. It has been studied in various incomplete market problems, for example, when there are transaction costs [6, 11, 21] or nontraded assets [10, 31], and under exponential utilities [33].

Our analysis is presented as follows. In section 2, we describe the mechanism in the context of a standard stochastic volatility model and characterize the indifference price in terms of solutions of related Hamilton–Jacobi–Bellman (HJB) equations. We derive a quasilinear PDE (2.32) that the pricing function satisfies. In addition, a specific measure, $\mathcal{Q}$ in Definition 2.7, emerges as a natural "prior" pricing measure, and the indifference price can be characterized as a worst-case expected payoff, penalized by relative entropy with respect to this prior (section 2.1).

In section 3, we derive two sets of bounds for the indifference price by analysis of the associated HJB equations. Section 4 presents asymptotic approximations that relate the indifference price to a particular *no-arbitrage* price. Finally, section 5 concludes and lists some remaining questions about the mechanism for future investigation.

2. Indifference prices. We assume a dynamic market setting with two assets, a riskless bond B, and a stock S. The stock price is modeled as a diffusion process satisfying

$$dS_s = \mu S_s\, ds + \sigma(Y_s, s) S_s\, dW_s^1, \qquad s \geq 0, \tag{2.1}$$

with $\mu > 0$. The volatility coefficient of the stock is driven by the stochastic factor $Y \in \mathbb{R}$ which is modeled as a correlated diffusion satisfying

$$dY_s = b(Y_s, s)\, ds + a(Y_s, s)\, (\rho\, dW_s^1 + \rho' dW_s^2), \tag{2.2}$$

with $\rho \in (-1, 1)$ the correlation coefficient and $\rho' = \sqrt{1 - \rho^2}$.

The processes W^1 and W^2 are independent standard Brownian motions defined on a probability space $(\Omega, \mathcal{F}, (\mathcal{F}_s), \mathbb{P})$, where $\mathcal{F}_s$ is the augmented σ-algebra generated by $((W_u^1, W_u^2);\, 0 \leq u \leq s)$. We assume that a riskless bond with maturity T is available for trading, yielding constant interest rate $r = 0$. The case $r \neq 0$ can be treated using standard discounting arguments and it is not presented herein. The derivative to be priced is of European type with payoff $g(S_T, Y_T)$ at expiration T. We make the following assumptions throughout.

Assumption 1.

1. *The volatility function $\sigma(\cdot)$ and the diffusion coefficient $a(\cdot, \cdot)$ are smooth and bounded above and below away from zero.*
2. *The drift coefficient $b(\cdot, \cdot)$ in (2.2) is Lipschitz continuous on $\mathbb{R} \times [0, T]$.*
3. *The payoff function $g(\cdot, \cdot)$ is smooth and bounded.*

Under these assumptions, (2.1) and (2.2) have a unique solution with $S_s \geq 0$ $\mathbb{P}$-a.s., $s \geq 0$ a.e. The assumption on the payoff excludes put options (whose payoff has discontinuous first derivative) and call options (which are unbounded). Handling these issues will require regularization methods (see, for example, [19]) which we do not address in this paper.

The utility-based valuation method relies on the comparison of maximal expected utilities corresponding to investment opportunities with and without the derivative. In both settings, trading takes place between the bond and the stock, and the objective

is to maximize the terminal utility of wealth. The investor starts, at time $t \geq 0$, with initial endowment x and dynamically rebalances his portfolio allocations, say, π_s^0 and π_s, representing the amounts invested at time $s \geq t$ in the bond and the stock accounts. It is assumed that no intermediate consumption or infusion of extraneous funds is allowed. The total current wealth satisfies

$$X_s = \pi_s^0 + \pi_s, \quad t \leq s \leq T, \tag{2.3}$$

and thus solves the state controlled diffusion equation

$$\begin{cases} dX_s = \mu\pi_s\, ds + \sigma(Y_s, s)\pi_s\, dW_s^1, & t \leq s \leq T, \\ X_t = x. \end{cases} \tag{2.4}$$

The above equation can be easily derived from (2.1) and the budget constraint (2.3) (see Merton [30]). Note that because the coefficients in (2.1) are linear in S, the latter does not appear explicitly in (2.4). Moreover, the budget constraint results in eliminating the first control variable π_s^0. The single control variable π_s is called admissible if it is $\mathcal{F}_s$-measurable and satisfies the integrability constraint $E\int_t^T \sigma(Y_s, s)^2\pi_s^2\, ds < +\infty$. The set of admissible policies is denoted by $\mathcal{A}$.

The next task is to introduce and analyze the three fundamental optimal investment problems via which the indifference prices for the writer and the buyer of the derivative will be constructed. Throughout the analysis, it is assumed that the individual preferences are modeled via an exponential utility function

$$U(x) = -e^{-\gamma x}, \qquad x \in \mathbb{R}, \tag{2.5}$$

with risk-aversion parameter $\gamma > 0$, and that they remain the same, independently of whether the derivative is written, bought, or not traded at all. The first is the classical Merton portfolio optimization problem, appropriately modified to accommodate stochastic volatility. Its value function is

$$V(x, y, t) = \sup_{\mathcal{A}} E(-e^{-\gamma X_T} \mid X_t = x,\ Y_t = y), \tag{2.6}$$

where X and Y solve (2.4) and (2.2), respectively. The investor seeks to maximize his terminal expected utility.

If the derivative with payoff $g(S_T, Y_T)$ is written, the *writer's* value function is

$$u^w(x, S, y, t) = \sup_{\mathcal{A}} E(-e^{-\gamma(X_T - g(S_T, Y_T))} \mid X_t = x,\ S_t = S,\ Y_t = y), \tag{2.7}$$

and if the derivative is bought, the *buyer's* value function is

$$u^b(x, S, y, t) = \sup_{\mathcal{A}} E(-e^{-\gamma(X_T + g(S_T, Y_T))} \mid X_t = x,\ S_t = S,\ Y_t = y). \tag{2.8}$$

It is immediate that

$$u^w(x, S, y, t; g) = u^b(x, S, y, t; -g). \tag{2.9}$$

The payoffs in (2.7) and (2.8) reflect, respectively, the obligation of the writer and the compensation of the buyer at expiration T.

A fundamental assumption is that both the writer and the buyer optimize over the same set of admissible strategies . Moreover, the traditional no-bankruptcy constraint $X_s \geq 0$ a.e. $t \leq s \leq T$ is not imposed herein due to the fact that exponential

utilities may allow for negative wealth levels. Imposing such constraints affects significantly the nature of the indifference prices and, in most cases, yields prices that are considerably high and not applicable. For models with transaction costs, such issues were studied by Constantinides and Zariphopoulou [6, 7].

We next review the definition of indifference prices (see Hodges and Neuberger [21]). The *writer's indifference price* of the European claim $g(S_T, Y_T)$ is defined as the amount $h^w = h^w(x, S, y, t)$, such that the writer is indifferent to the following two scenarios: optimize the utility payoff without writing the derivative and optimize the utility payoff with the *liability* $g(S_T, Y_T)$ at expiration, but with an initial *compensation* $h^w(x, S, y, t)$ at the time of inscription t. Similarly, the *indifference buyer's price* of the European claim $g(S_T, Y_T)$ is defined as the amount $h^b = h^b(x, S, y, t)$ such that the buyer is indifferent to the following two scenarios: optimize the utility payoff without buying the derivative and optimize the utility payoff with the *payoff* $g(S_T, Y_T)$ at expiration, but with the initial *cost* $h^b(x, S, y, t)$ at the time of inscription t.

DEFINITION 2.1. *The indifference prices h^w and h^b are defined by*

$$V(x, y, t) = u^w\Big(x + h^w(x, S, y, t), S, y, t\Big), \tag{2.10}$$

$$V(x, y, t) = u^b\Big(x - h^b(x, S, y, t), S, y, t\Big). \tag{2.11}$$

The above definition allows for derivative prices that depend on the investor's wealth, as reflected in their x-argument. Such dependence might look like an undesirable feature given the wealth-independent prices that arbitrage-free theory yields in complete markets. As the calculations below show, the choice of exponential utility leads to wealth-independent prices, at least for the case of European claims and in the absence of trading constraints. Wealth independence, however, might not hold for other choices of risk preferences and/or trading constraints. In such situations, universality may be achieved by relaxing the notion of indifference price to *reservation prices*. The latter prices are defined as wealth-independent price bounds for which the price equalities (2.10) and (2.11) hold as inequalities (see, for example, [6, 7]).

The value functions V, u^w, and u^b, whose arguments will yield the indifference prices, may be studied via their associated HJB equations. Although these equations are fully nonlinear, the convenience of the exponential utility function allows us to construct classical solutions after an initial separation of variables. It follows from a standard viscosity solution argument that these solutions coincide with the respective value functions.

To facilitate the presentation, we introduce the following operators and Hamiltonians:

$$\begin{aligned}\mathcal{A}^{(S,y)}u = {} & \frac{1}{2}\sigma(y,t)^2S^2u_{SS} + \rho\sigma(y,t)a(y,t)Su_{Sy} \\ & + \frac{1}{2}a(y,t)^2u_{yy} + \mu Su_S + b(y,t)u_y,\end{aligned} \tag{2.12}$$

$$\mathcal{A}^{(y)}u = \frac{1}{2}a(y,t)^2u_{yy} + b(y,t)u_y, \tag{2.13}$$

$$\begin{aligned}\mathcal{H}^{(S,y)}(u_{xx}, u_{xy}, u_{xS}, u_x) = {} & \max_{\pi}\bigg(\frac{1}{2}\sigma(y,t)^2\pi^2u_{xx} + \pi(\rho\sigma(y,t)a(y,t)u_{xy} \\ & + \sigma(y,t)^2Su_{xS} + \mu u_x)\bigg),\end{aligned} \tag{2.14}$$

$$\mathcal{H}^{(y)}(u_{xx}, u_{xy}, u_x) = \max_{\pi}\left(\frac{1}{2}\sigma(y,t)^2\pi^2u_{xx} + \pi(\rho\sigma(y,t)a(y,t)u_{xy} + \mu u_x)\right). \tag{2.15}$$

Note that $\mathcal{A}^{(S,y)}$ is the infinitesimal generator of the Markov process (S, Y), and $\mathcal{A}^{(y)}$ is the infinitesimal generator of Y, which is a Markov process by itself in our stochastic volatility models.

The HJB equation associated with the value function u^w defined in (2.7) is

$$\begin{aligned} &u_t + \mathcal{A}^{(S,y)}u + \mathcal{H}^{(S,y)}(u_{xx}, u_{xy}, u_{xS}, u_x) = 0, \\ &u(x, S, y, T) = -e^{-\gamma(x-g(S,y))}, \end{aligned} \tag{2.16}$$

on $D = \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R} \times [0, T]$. The relevant PDE for u^b, defined in (2.8), is the same as (2.16), with the sign of g changed in the terminal condition. For V, defined in (2.6), we simply set g to zero and remove the S-derivatives from the equation (i.e., $\mathcal{A}^{(S,y)}$ and $\mathcal{H}^{(S,y)}$ are replaced by $\mathcal{A}^{(y)}$ and $\mathcal{H}^{(y)}$, respectively).

In the theorems below, we produce a closed form expression for the value function V and we provide regularity results for $u^{w,b}$. The proofs are based on the construction of a candidate solution that is actually smooth and therefore a classical solution of the HJB equation. We readily identify it also as the unique viscosity solution and, from there, by a standard argument, with the value function. The intermediate step is required because classical verification theorems require polynomial growth (in x) restrictions on the value functions, which do not hold with exponential utility.

For convenience, we introduce

$$\mathcal{L}^{(S,y)}u = \mathcal{A}^{(S,y)}u - \rho\frac{\mu}{\sigma(y,t)}a(y,t)u_y, \tag{2.17}$$

$$\mathcal{L}^{(y)}u = \mathcal{A}^{(y)}u - \rho\frac{\mu}{\sigma(y,t)}a(y,t)u_y, \tag{2.18}$$

$$\mathcal{M}(G_S, G_y, G) = \frac{1}{2}\sigma(y,t)^2S^2\frac{G_S^2}{G} + \rho\sigma(y,t)a(y,t)S\frac{G_SG_y}{G} + \frac{1}{2}\rho^2a(y,t)^2\frac{G_y^2}{G}. \tag{2.19}$$

THEOREM 2.2. *The value function V is given by*

$$V(x, y, t) = -e^{-\gamma x}f(y, t)^{\frac{1}{1-\rho^2}}, \tag{2.20}$$

where f solves

$$f_t + \mathcal{L}^{(y)}f = \frac{1}{2}(1-\rho^2)\frac{\mu^2}{\sigma(y,t)^2}f \tag{2.21}$$

in $y \in \mathbb{R}$, $t < T$, with $f(y, T) = 1$ for $y \in \mathbb{R}$.

Proof. We first consider a candidate solution of the form $\tilde{V}(x, y, t) = -e^{-\gamma x}F(y, t)$. This form is suggested by the scaling properties of the exponential utility. We recall that V solves the HJB equation (2.16) for $g = 0$ which reduces to

$$V_t + \mathcal{H}^{(y)}(V_{xx}, V_{xy}, V_x) + \mathcal{A}^{(y)}u = 0, \tag{2.22}$$

with $V(x, y, T) = -e^{-\gamma x}$. Evaluating this equation at the candidate solution $\tilde{V}$ yields that F must satisfy the quasilinear equation

$$F_t + \mathcal{L}^{(y)}F = \frac{1}{2}\frac{\mu^2}{\sigma(y,t)^2}F + \frac{1}{2}\rho^2a(y,t)^2\frac{F_y^2}{F}, \tag{2.23}$$

with $F(y, T) = 1$. A power transformation $F = f^\delta$ for $\delta = \frac{1}{1-\rho^2}$ yields that f must solve the linear PDE (2.21), with $f(y, T) = 1$. The fact that the linear equation

(2.21) has a unique smooth and bounded solution follows from Assumption 1 on the coefficients (see [26, Theorem 2.9.10], for example). We then deduce that the candidate $\tilde{V}$ satisfies the HJB equation (2.22), with terminal value $-e^{-\gamma x}$, and it is also smooth. Therefore, it is a classical solution of (2.22). To conclude, we use uniqueness results for viscosity solutions of the HJB equation. This approach has become by now familiar in incomplete market models (see [34] for an overview).

Following the arguments of Theorems 4.1 and 4.2 in [13], we deduce that the value function V defined in (2.6) is the *unique* viscosity solution of (2.22). Uniqueness holds in the class of functions that are concave and of exponential growth in the wealth argument and are uniformly bounded in the variable y. (We note that in the model analyzed in [13], market incompleteness was generated by stochastic labor income modeled as a correlated diffusion. The two associated HJB equations have similar nonlinearities, and the technical arguments work under rather minor modifications.) We next observe that the candidate $\tilde{V}$ is smooth and therefore a viscosity solution of (2.22). Moreover, the assumptions on the model coefficients yield that it belongs to the class of viscosity solutions in which uniqueness holds. Therefore, $\tilde{V} \equiv V$ and the result follows. Note that classical verification results (for instance, [14, Theorem III.8.1]) require more stringent polynomial growth conditions on the candidate solution, a requirement that is bypassed by the viscosity arguments. □

DEFINITION 2.3. *Let $\mathbb{P}$ be the historical measure. We define an equivalent measure $\widetilde{\mathbb{P}}$ by*

$$\frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}} = \exp\left(- \int_0^T \frac{\mu}{\sigma(Y_s, s)}\, dW_s^1 - \frac{1}{2} \int_0^T \frac{\mu^2}{\sigma(Y_s, s)^2}\, ds \right).$$

By Girsanov's theorem, the dynamics of (S, Y) under $\widetilde{\mathbb{P}}$ are

$$dS_s = \sigma(Y_s, s) S_s\, d\widetilde{W}_s^1, \tag{2.24}$$

$$dY_s = \left(b(Y_s, s) - \rho \frac{\mu}{\sigma(Y_s, s)} a(Y_s, s) \right) ds + a(Y_s, s)\, (\rho\, d\widetilde{W}_s^1 + \rho' d\widetilde{W}_s^2), \tag{2.25}$$

where $\widetilde{W}_s^1 = W_s^1 + \int_0^s \frac{\mu}{\sigma(Y_u, u)}\, du$ and $\widetilde{W}_s^2 = W_s^2$ are independent $\widetilde{\mathbb{P}}$-Brownian motions. The measure $\widetilde{\mathbb{P}}$ is often known as the minimal martingale measure [16].

From the formula (2.20) for V, the Feynman–Kac representation of the solution to (2.21), and the definition of the measure $\widetilde{\mathbb{P}}$, we obtain the following proposition.

PROPOSITION 2.4. *The solution f of* (2.21) *admits the probabilistic representation*

$$f(y, t) = \mathbb{E}_{\widetilde{\mathbb{P}}} \left(e^{- \int_t^T \frac{\mu^2 (1 - \rho^2)}{2\sigma(Y_s, s)^2}\, ds} \mid Y_t = y \right), \tag{2.26}$$

where the process Y satisfies (2.25). *The value function V is then given by*

$$V(x, y, t) = -e^{-\gamma x} \left(\mathbb{E}_{\widetilde{\mathbb{P}}} \left(e^{- \int_t^T \frac{\mu^2 (1 - \rho^2)}{2\sigma(Y_s, s)^2}\, ds} \mid Y_t = y \right) \right)^{\frac{1}{1 - \rho^2}}. \tag{2.27}$$

THEOREM 2.5. *The writer's value function u^w is given by*

$$u^w(x, S, y, t) = -e^{-\gamma x} G(S, y, t), \tag{2.28}$$

where $G \in C^{2,2,1}(\mathbb{R}^+ \times \mathbb{R} \times [0,T])$ *is the unique bounded solution of the quasilinear equation*

$$G_t + \mathcal{L}^{(S,y)} G = \frac{1}{2} \frac{\mu^2}{\sigma(y,t)^2} G + \mathcal{M}(G_S, G_y, G), \tag{2.29}$$

with $G(S,y,T) = e^{\gamma g(S,y)}$.

Proof. We look at a candidate solution of the form $\tilde{u}(x,S,y,t) = -e^{-\gamma x} G(S,y,t)$. Straightforward calculations in (2.16) imply that G must satisfy the quasilinear equation (2.29). Note that G must be positive as u is negative.

A logarithmic transformation $G = e^{\phi}$ gives that ϕ should solve a quasilinear equation with quadratic nonlinearity, namely,

$$\phi_t + \mathcal{L}^{(S,y)} \phi + \frac{1}{2} a(y)^2 (1-\rho^2) \phi_y^2 = \frac{\mu^2}{2\sigma(y)^2}, \tag{2.30}$$

with $\phi(S,y,T) = \gamma g(S,y)$. Equation (2.30) is the familiar HJB equation of a *quadratic cost* stochastic control problem (see Fleming and Rishel [15, section VI.5]). Under Assumption 1, we obtain that ϕ is bounded, $\phi \in C^{2,2,1}(\mathbb{R}^+ \times \mathbb{R} \times [0,T])$, and that it is the unique solution in this class (see, for example, Ladyzenskaja, Solonnikov, and Uralceva [27], Fleming and Rishel [15], or Pham [32]). This in turn yields the same properties for G. Following similar arguments as in the proof of Theorem 2.2 to identify $\tilde{u}$ as the unique viscosity solution of (2.16) and therefore the value function, we conclude that $u^w = \tilde{u}$. □

We note that even though a simple power transformation can linearize the reduced equation (2.23) in the one dimensional case, quasilinear equations of the form (2.29) cannot be linearized in higher dimensional settings unless the nonlinearity is a quadratic in ∇G, where ∇ denotes the gradient with respect to the spatial variables, and there are no cross-derivative terms. Of course under a logarithmic transformation, one may reduce (2.29) to (2.30).

Before we construct the indifference prices, we introduce some convenient notation.

LEMMA 2.1. *Let*

$$L(y,t) = \frac{1}{\rho'} a(y,t) \frac{f_y(y,t)}{f(y,t)}, \tag{2.31}$$

with f *given in* (2.26). *Then under Assumption* 1, L *is smooth, and bounded.*

Proof. From (2.21), f is positive, smooth, and bounded for fixed $t < T$ under Assumption 1. To establish that $f_y(y,t)$ is also smooth and bounded, it suffices to differentiate (2.21) with respect to y and use the relevant probabilistic representation of f_y. □

THEOREM 2.6. (i) *The writer's indifference price* h^w *is the unique* $C^{2,2,1}(\mathbb{R}^+ \times \mathbb{R} \times [0,T])$ *bounded solution of the pricing equation*

$$h^w_t + \mathcal{L}^{(S,y)} h^w + \rho' a(y,t) L(y,t) h^w_y + \frac{1}{2} \gamma (1-\rho^2) a(y,t)^2 (h^w_y)^2 = 0 \tag{2.32}$$

with $h^w(S,y,T) = g(S,y)$.

(ii) *The buyer's indifference price satisfies*

$$h^w(S,y,t;g) = -h^b(S,y,t;-g) \tag{2.33}$$

and solves

$$h_t^b + \mathcal{L}^{(S,y)} h^b + \rho' a(y,t) L(y,t) h_y^b - \frac{1}{2}\gamma(1-\rho^2) a(y,t)^2 (h_y^b)^2 = 0 \tag{2.34}$$

with $h^b(S,y,T) = g(S,y)$.

Proof. We discuss only the arguments for h^w. To this end, we first observe that the pricing equality (2.10) together with the representations (2.20) and (2.28) of the value functions V and u^w yield that h^w is given by

$$h^w(S,y,t) = \frac{1}{\gamma} \ln \frac{G(S,y,t)}{f(y,t)^{1/(1-\rho^2)}}, \tag{2.35}$$

and as such is independent of x. Direct substitution shows that h^w solves the claimed quasilinear equation (2.32). It also satisfies the terminal condition $h^w(S,y,T) = g(S,y)$. In (2.32), the coefficient of the additional h_y^w term is smooth and bounded by Lemma 2.1. The uniqueness and regularity results for h^w follow from an appropriate adaptation of Theorem 4.1 in Pham [32], which utilizes that (2.32), like (2.30) for ϕ, is the HJB equation of a quadratic cost control problem. The parity property (2.33) follows from the definition of the indifference prices and the properties of the value functions u^w and u^b. □

The indifference price equation (2.32) indicates that a new measure, defined below, emerges from the utility-based valuation.

Definition 2.7. *We define* $\mathcal{Q}$ *by*

$$\frac{d\mathcal{Q}}{d\mathbb{P}} = \exp\left(- \int_0^T \frac{\mu}{\sigma(Y_s,s)}\, dW_s^1 + \int_0^T L(Y_s,s)\, dW_s^2 \right.$$
$$\left. - \frac{1}{2}\int_0^T \left(\frac{\mu^2}{\sigma(Y_s,s)^2} + L(Y_s,s)^2 \right) ds \right).$$

In the language of stochastic volatility models, the function $-L$ is a particular market price of volatility risk (and $\mathcal{Q}$ a particular equivalent martingale measure), as described in section 4.1.1. In fact, $\mathcal{Q}$ is the minimal relative entropy martingale measure, as we discuss in section 2.1.

It is worth observing that the price equation (2.32) does not have a zeroth order term or drift terms in S. It does not have a nonlinear term involving h_S either. This is an immediate consequence of the assumptions of zero interest rate and that the stock is tradeable. We also observe that the only place where the risk-aversion coefficient γ appears is in front of the non-linear term that directly reflects market incompleteness. The latter has also a fixed sign with respect to γ which, in turn, yields the following intuitive result.

Theorem 2.8. *The writer's (resp., buyer's) indifference price is nondecreasing (resp., nonincreasing) with respect to the risk-aversion parameter* γ. *As* $\gamma \to 0$, *the writer's and buyer's indifference prices satisfy*

$$\lim_{\gamma \downarrow 0} h^{w,b}(S,y,t) = \mathbb{E}_{\mathcal{Q}}(g(S_T,Y_T) \mid S_t = S, Y_t = y), \tag{2.36}$$

where the measure $\mathcal{Q}$ *is defined above.*

Proof. We denote by h^{γ_1} and h^{γ_2} the writer's indifference prices corresponding to risk-aversion coefficients γ_1 and γ_2 with $\gamma_1 < \gamma_2$. Straightforward calculations show

that h^{γ_1} is a subsolution of the indifference pricing equation that h^{γ_2} satisfies and that $h^{\gamma_1}(S,y,T) = h^{\gamma_2}(S,y,T)$. Using classical comparison results for (2.32), we conclude.

We observe that as $\gamma \to 0$, the indifference pricing equation (2.32) formally converges to the linear equation

$$h^0_t + \mathcal{L}^{(S,y)} h^0 + \rho' a(y,t) L(y,t) h^0_y = 0 \tag{2.37}$$

with terminal condition $h^0(S,y,T) = g(S,y)$. Under Assumption 1, the latter has a unique smooth (and thus viscosity) solution given by the Feynman–Kac formula

$$h^0(S,y,t) = \mathbb{E}_{\mathcal{Q}}(g(S_T, Y_T) \mid S_t = S, Y_t = y). \tag{2.38}$$

The stability properties of viscosity solutions (see Lions [29, Proposition I.3]) yield that $\{h^\gamma(S,y,t)\}$ converges, along subsequences, to the viscosity solution of (2.37) and, by uniqueness, we conclude.

The analogous result for the buyer's price follows from similar calculations. □

Expressions such as (2.38) have also been obtained in other utility-based pricing approaches [9, 25, 24]. It is well known that the indifference price (buyer's or writer's) of $\alpha > 0$ derivative contracts with bounded payoff g, written $h(\alpha;\gamma)$ as a function of the quantity and risk-aversion parameter, satisfies

$$h(\alpha;\gamma) = \alpha h(1;\alpha\gamma),$$

as is clear in the current context from the PDEs (2.32) and (2.34) by replacing g by αg in the terminal condition and making the change of variable $h = \alpha h'$. Therefore, taking the limit of zero risk-aversion is analogous to taking the limit of the price per unit $h(\alpha;\gamma)/\alpha$ as α goes to zero (with $\gamma > 0$ fixed). This is how the "fair" price is defined in [9], and the measure $\mathcal{Q}$ that arises in its characterization is labeled the neutral pricing measure in [24]. In the next section, we also point out that $\mathcal{Q}$ is in fact the minimal relative entropy measure.

2.1. Interpretation via relative entropy penalization. One way to interpret the utility-based valuation mechanism is in terms of relative entropy penalization. This is a specific example of the well-known connection between exponential utility and entropy as discussed, for example, in [12, 33].

Recall that $\mathcal{Q}$ denotes the probability measure under which the dynamics of (S,Y) are

$$\begin{aligned} dS_s &= \sigma(Y_s,s) S_s \, dW^{\mathcal{Q}(1)}_s, \\ dY_s &= \left(b(Y_s,s) - \rho \frac{\mu}{\sigma(Y_s,s)} a(Y_s,s) + \rho' a(Y_s,s) L(Y_s,s) \right) ds \\ &\quad + a(Y_s,s) \left(\rho \, dW^{\mathcal{Q}(1)}_s + \rho' dW^{\mathcal{Q}(2)}_s \right), \end{aligned} \tag{2.39}$$

where $W^{\mathcal{Q}(1)}_s = W^1_s + \int_0^s \frac{\mu}{\sigma(Y_u,u)} \, du$ and $W^{\mathcal{Q}(2)}_s = W^2_s - \int_0^s L(Y_u,u) \, du$ are independent $\mathcal{Q}$-Brownian motions. Note that $\mathcal{Q}$ is already a "risk-neutral" martingale measure because S is a $\mathcal{Q}$-martingale.

Then, let $\mathbb{P}^{(\lambda)}$ be any *equivalent local martingale measure*, which, in this context, is parameterized by an adapted process, λ, say, with $\int_0^T \lambda_s^2 \, ds < \infty$ a.s., such that

$$\frac{d\mathbb{P}^{(\lambda)}}{d\mathcal{Q}} = \exp\left(- \int_0^T \lambda_s \, dW^{\mathcal{Q}(2)}_s - \frac{1}{2} \int_0^T \lambda_s^2 \, ds \right).$$

That is, defining $W_s^{\lambda(2)} = W_s^{\mathcal{Q}(2)} + \int_0^s \lambda_u\, du$, $(W^{\mathcal{Q}(1)}, W^{\lambda(2)})$ are independent Brownian motions under $\mathbb{P}^{(\lambda)}$, and the dynamics of (S, Y) are described by (2.39) and

$$dY_s = \left(b(Y_s, s) - \rho \frac{\mu}{\sigma(Y_s, s)} a(Y_s, s) + \rho' a(Y_s, s)(L(Y_s, s) - \lambda_s) \right) ds \\ + a(Y_s, s)\,(\rho\, dW_s^{\mathcal{Q}(1)} + \rho' dW_s^{\lambda(2)}).$$

We can think of $\mathcal{Q}$ as a *prior* risk-neutral measure, and we define the relative entropy $H_t(\mathbb{P}^{(\lambda)} \mid \mathcal{Q})$ between the conditional laws on the process $\{(S_s, Y_s), t \le s \le T\}$ starting at the same point (S, y) at time t as follows. First, let (ξ_s) denote the Radon–Nikodym process

$$\xi_s = \mathbb{E}_{\mathcal{Q}} \left(\frac{d\mathbb{P}^{(\lambda)}}{d\mathcal{Q}} \mid \mathcal{F}_s \right).$$

We then define

$$H_t(\mathbb{P}^{(\lambda)} \mid \mathcal{Q}) = \mathbb{E}_{\mathbb{P}^{(\lambda)}} \left(\ln(\xi_T/\xi_t) \mid \mathcal{F}_t \right).$$

By direct calculation, this is a quadratic penalization on the "additional" volatility risk premium λ:

$$H_t(\mathbb{P}^{(\lambda)} \mid \mathcal{Q}) = \frac{1}{2} \mathbb{E}_{\mathbb{P}^{(\lambda)}} \left(\int_t^T \lambda_s^2\, ds \mid \mathcal{F}_t \right). \tag{2.40}$$

We denote by $\mathcal{M}_f$ the set of λ with

$$\mathbb{E}_{\mathbb{P}^{(\lambda)}} \left(\int_0^T \lambda_s^2\, ds \right) < \infty,$$

which guarantees finiteness of the relative entropies $H_t(\mathbb{P}^{(\lambda)} \mid \mathcal{Q})$.

Therefore, we can interpret the writer's indifference pricing mechanism as choosing a measure which tries to maximize the derivative's expected payout but is constrained from deviating too far from the prior in terms of relative entropy:

$$h^w = \sup_{\lambda \in \mathcal{M}_f} \left[\mathbb{E}_{\mathbb{P}^{(\lambda)}} \left(g(S_T, Y_T) \mid \mathcal{F}_t \right) - \frac{1}{\gamma} H_t(\mathbb{P}^{(\lambda)} \mid \mathcal{Q}) \right]. \tag{2.41}$$

This is because the HJB equation associated with this stochastic control problem is (2.32), as follows from using the formula (2.40). Notice that the upshot of the utility mechanism is to identify the prior $\mathcal{Q}$ which does not depend on the risk-aversion γ or the claim g being priced. The relative entropy arises naturally as the dual of the exponential utility, and γ^{-1} weights the penalty term.

In [12], the writer's indifference price (at time $t = 0$) is characterized as

$$h^w = \sup_{Q \in P_f} \left[\mathbb{E}_Q(g(S_T, Y_T)) - \frac{1}{\gamma} H_0(Q \mid \mathbb{P}) \right] - \sup_{Q \in P_f} \left[-\frac{1}{\gamma} H_0(Q \mid \mathbb{P}) \right] \tag{2.42}$$

(their equation (5.6)) under quite general conditions. The set P_f consists of measures Q that are absolutely continuous with respect to $\mathbb{P}$, such that the wealth process X is a $(Q, \mathcal{F})$-local martingale, and $H_0(Q \mid \mathbb{P}) < \infty$. In (2.42), the indifference price

is given as the difference between the solutions of two optimization problems, with $\mathbb{P}$ as the prior measure. Our expression (2.41) gives the indifference price under our diffusion stochastic volatility models as the solution of a single optimization problem with prior (risk-neutral) measure $\mathcal{Q}$ (which arises from the solution of the Merton problem).

In fact, it is easy to show from the associated HJB equations that $\mathcal{Q}$ is the minimal relative entropy martingale measure (see Fritelli [20]) solution of

$$\sup_{Q \in P_f} \left(-H_0(Q \mid \mathbb{P})\right)$$

and that

$$H_0(\mathcal{Q} \mid \mathbb{P}) = -\frac{1}{1-\rho^2} \ln f \mid_{t=0},$$
$$\sup_{Q \in P_f} \left[\mathbb{E}_Q(g(S_T, Y_T)) - \frac{1}{\gamma} H_0(Q \mid \mathbb{P})\right] = \frac{1}{\gamma} \ln G \mid_{t=0},$$

which connects (at time $t = 0$) our expression (2.35) with (2.42), which is taken from [12].

Relative entropy minimization has been extensively used for calibration from market data, with additional constraints that some benchmark derivative contracts are priced exactly [2, 4]. The prior measure $\mathcal{Q}$ will also emerge later in section 3.1 in obtaining bounds for the indifference prices.

3. Indifference price spreads. The indifference pricing equations (2.32) and (2.34) do not in general have explicit solutions. Given that risk aversion is taken into account in the utility-based valuation, one would naturally expect to recover indifference prices in terms of the so-called certainty equivalent pricing rule. This is in fact the classical risk-based pricing device used in the de facto incomplete insurance market (see, for example, Bowers et al. [3]). However, as simple calculations show, indifference prices do not correspond to straightforward generalizations of certainty equivalents given that they result from an interplay between dynamic optimization among investment opportunities and risk monitoring. We note that classical arbitrage-free financial markets use a risk-neutral measure and incomplete insurance markets use the historical one.

In what follows we aim at addressing some of the above issues by looking at price bounds. We derive bounds on the indifference prices that involve characteristics of linear and nonlinear prices, namely, expected payoffs, certainty equivalents, and related pricing measures.

PROPOSITION 3.1. *Let $\widetilde{\mathbb{P}}$ be the minimal martingale measure described in Definition* 2.3*, and define*

$$R(y,t) = \frac{1}{2\gamma} \left(\frac{\mu}{\sigma(y,t)}\right)^2,$$

with $\mu/\sigma(y,t)$ being the (time-varying) Sharpe ratio of the traded stock, and

$$\zeta(S,y,t) = \frac{1}{\gamma(1-\rho^2)} \ln \mathbb{E}_{\widetilde{\mathbb{P}}} \left(e^{-\gamma(1-\rho^2)\int_t^T R(Y_s,s)\,ds} \mid S_t = S,\, Y_t = y\right).$$

(i) *The writer's indifference price h^w satisfies*

$$h^w(S,y,t) \le \frac{1}{\gamma}\ln \mathbb{E}_{\widetilde{\mathbb{P}}}\left(e^{\gamma\left(g(S_T,Y_T)-\int_t^T R(Y_s,s)\,ds\right)} \mid S_t = S, Y_t = y\right) - \zeta(S,y,t), \tag{3.1}$$

$$h^w(S,y,t) \ge \mathbb{E}_{\widetilde{P}}\left(g(S_T,Y_T) - \int_t^T R(Y_s,s)\,ds \mid S_t = S,\, Y_t = y\right) - \zeta(S,y,t). \tag{3.2}$$

(ii) *The buyer's indifference price h^b satisfies*

$$h^b(S,y,t) \le \mathbb{E}_{\widetilde{\mathbb{P}}}\left(g(S_T,Y_T) + \int_t^T R(Y_s,s)\,ds \mid S_t = S, Y_t = y\right) + \zeta(S,y,t),$$

$$h^b(S,y,t) \ge -\frac{1}{\gamma}\ln \mathbb{E}_{\widetilde{\mathbb{P}}}\left(e^{-\gamma\left(g(S_T,Y_T)+\int_t^T R(Y_s,s)\,ds\right)} \mid S_t = S, Y_t = y\right) + \zeta(S,y,t).$$

Proof. We first present the arguments for the derivation of the lower bound (3.2). We recall that

$$h(S,y,t) = \frac{1}{\gamma}\ln\frac{G(S,y,t)}{V(y,t)} = \frac{1}{\gamma}\ln\frac{G(S,y,t)}{f(y,t)^{1/(1-\rho^2)}} \tag{3.3}$$

and that G solves (2.29). Because $\rho^2 \le 1$ and $G > 0$, we easily conclude that G is a supersolution of

$$\bar{G}_t + \mathcal{L}^{(S,y)}\bar{G} = \frac{1}{2}\frac{\mu^2}{\sigma(y,t)^2}\bar{G} + \frac{1}{2}a^2(y,t)\frac{\bar{G}_y^2}{\bar{G}} + \rho\sigma(y,t)a(y,t)S\frac{\bar{G}_S\bar{G}_y}{\bar{G}} + \frac{1}{2}\sigma^2(y,t)S^2\frac{\bar{G}_S^2}{\bar{G}} \tag{3.4}$$

and, thus,

$$G(S,y,t) \ge \bar{G}(S,y,t). \tag{3.5}$$

The solution $\bar{G}$ to (3.4) can be derived via an exponential transformation $\bar{G} = e^{\bar{\phi}}$, with $\bar{\phi}$ solving

$$\bar{\phi}_t + \mathcal{L}^{(S,y)}\bar{\phi} = \frac{\mu^2}{2\sigma^2(y,t)}, \qquad \bar{\phi}(S,y,T) = \gamma g(S,y). \tag{3.6}$$

Under the boundedness assumptions on the coefficients and payoff, (3.6) has a unique classical solution, and $\bar{\phi}$ has the probabilistic representation

$$\bar{\phi}(S,y,t) = \gamma\mathbb{E}_{\widetilde{P}}\left(g(S_T,Y_T) - \int_t^T R(Y_s,s)\,ds \mid S_t = S, Y_t = y\right). \tag{3.7}$$

Combining $G \ge e^{\bar{\phi}}$ with (3.3) and (2.26) gives the lower bound (3.2) for the writer's indifference price h^w.

Following similar arguments, we can derive the upper bound for the indifference price. In fact, from (2.19), $\mathcal{M}(G_S, G_y, G) \ge 0$, so (2.29) yields

$$G_t + \mathcal{L}^{(S,y)}G \ge \frac{\mu^2}{2\sigma^2(y,t)}G.$$

Therefore, G is a subsolution to

$$\hat{G}_t + \mathcal{L}^{(S,y)}\hat{G} = \frac{\mu^2}{2\sigma^2(y,t)}\hat{G}, \qquad \hat{G}(S,y,T) = e^{\gamma g(S,y)}. \tag{3.8}$$

This in turn yields $G(S,y,t) \leq \hat{G}(S,y,t)$, and, in view of the probabilistic representation of the solution of (3.8), gives

$$G(S,y,t) \leq \mathbb{E}^{\widetilde{\mathbb{P}}}_{\mathbb{P}}\left(e^{\gamma\left(g(S_T,Y_T)-\int_t^T R(Y_s,s)\,ds\right)} \mid S_t = S, Y_t = y\right).$$

The upper bound (3.1) follows easily. Part(ii) can be derived from the parity formula (2.33). □

3.1. Alternative bounds. We continue with the derivation of alternative reservation prices. We stress that the bounds derived below have rather natural and desirable properties. The lower bound is given by an *arbitrage-free–type* price of the payoff g. This price corresponds to the limiting case $\gamma \to 0$ described in Theorem 2.8. The upper bound is given in terms of a *certainty-equivalent–type* price of the payoff g. It corresponds to what the writer would charge under a pricing device based entirely on static certainty equivalent valuation without taking into account dynamic rebalancing and optimal investments. It is important to observe that all bounds are expressed in terms of the same measure $\mathcal{Q}$.

THEOREM 3.2. *Let $\mathcal{Q}$ be the measure introduced in Definition* 2.7.

(i) *The writer's indifference price satisfies*

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\Big(g(S_T,Y_T) \mid S_t = S, Y_t = y\Big) &\leq h^w(S,y,t)\\ &\leq \frac{1}{\gamma}\ln \mathbb{E}_{\mathcal{Q}}\Big(e^{\gamma g(S_T,Y_T)} \mid S_t = S, Y_t = y\Big).\end{aligned} \tag{3.9}$$

(ii) *The buyer's indifference price satisfies*

$$\begin{aligned}-\frac{1}{\gamma}\ln \mathbb{E}_{\mathcal{Q}}\Big(e^{-\gamma g(S_T,Y_T)} \mid S_t = S, Y_t = y\Big) &\leq h^b(S,y,t)\\ &\leq \mathbb{E}_{\mathcal{Q}}\Big(g(S_T,Y_T) \mid S_t = S, Y_t = y\Big).\end{aligned}$$

Proof. We first construct appropriate sub- and supersolutions of (2.29). In particular, we look for sub- and supersolutions of the separable form $M(y,t)N(S,y,t)$. Inserting the above function in (2.29) yields

$$\begin{cases} N\left(M_t + \mathcal{L}^{(y)}M - \frac{1}{2}\rho^2 a^2(y,t)\frac{M_y^2}{M} - \frac{\mu^2}{2\sigma^2(y,t)}M\right)\\ \quad + M\left(N_t + \mathcal{L}^{(S,y)}N + (1-\rho^2)a^2(y,t)\frac{M_y}{M}N_y - \mathcal{M}(N_S,N_y,N)\right) = 0, \end{cases}$$

where $\mathcal{M}$ was defined in (2.19), and with $M(y,T)N(S,y,T) = e^{\gamma g(S,y)}$. Next, we choose M and N to solve, respectively,

$$M_t + \mathcal{L}^{(y)}M = \frac{\mu^2}{2\sigma^2(y,t)}M + \frac{1}{2}\rho^2 a^2(y,t)\frac{M_y^2}{M}, \tag{3.10}$$

with $M(y,T)=1$, and

$$N_t+\mathcal{L}^{(S,y)}N+(1-\rho^2)a^2(y,t)\frac{M_y}{M}N_y=\mathcal{M}(N_S,N_y,N), \tag{3.11}$$

with $N(S,y,T)=e^{\gamma g(S,y)}$. We observe that M and F solve the same equations, (3.10) and (2.23), and satisfy the same terminal condition. By uniqueness, we deduce that $M\equiv F$ or, equivalently,

$$M(y,t)=f(y,t)^{1/(1-\rho^2)}, \tag{3.12}$$

with f solving (2.21). Therefore, we can write the solution of (2.29) as $G(S,y,t)=f(y,t)^{1/(1-\rho^2)}N(S,y,t)$, with N solving (3.11). It is worth observing that N solves the quasilinear equation (3.11) that is similar to (2.29) but with two modifications; namely, (3.11) does not have a potential term, and, also, its first derivative coefficient contains the extra term $(1-\rho^2)a^2(y,t)\frac{M_y(y,t)}{M(y,t)}$.

Because of (3.12), we have

$$\frac{M_y(y,t)}{M(y,t)}=\frac{1}{1-\rho^2}\frac{f_y(y,t)}{f(y,t)}\ , \tag{3.13}$$

which is smooth and bounded as observed in Lemma 2.1. Therefore, (3.11) can be written as

$$\begin{cases} N_t+\mathcal{L}^{(S,y)}N+\rho' a(y,t)L(y,t)N_y=\mathcal{M}(N_S,N_y,N),\\ N(S,y,T)=e^{\gamma g(S,y)}, \end{cases} \tag{3.14}$$

where $L(y,t)$ was introduced in (2.31). Observing that $\mathcal{M}(N_S,N_y,N)\geq 0$ and that $\rho^2\leq 1$, by arguments similar to the ones used in the derivation of the lower and upper bounds, (3.2) and (3.1), respectively, we readily deduce that

$$\underline{N}(S,y,t)\leq N(S,y,t)\leq \bar{N}(S,y,t), \tag{3.15}$$

where $\underline{N}$ and $\bar{N}$ solve, respectively,

$$\bar{N}_t+\mathcal{L}^{(S,y)}\bar{N}+\rho' a(y,t)L(y,t)\bar{N}_y=0, \tag{3.16}$$

with $\bar{N}(S,y,T)=e^{\gamma g(S,y)}$, and

$$\begin{aligned}\underline{N}_t+\mathcal{L}^{(S,y)}\underline{N}+\rho' a(y,t)L(y,t)\underline{N}_y=\frac{1}{2}a^2(y,t)\frac{\underline{N}_y^2}{\underline{N}}+\rho a(y,t)\sigma(y,t)S\frac{\underline{N}_S\underline{N}_y}{\underline{N}}\\ +\frac{1}{2}\sigma^2(y,t)S^2\underline{N}_{SS},\end{aligned} \tag{3.17}$$

with $\underline{N}(S,y,T)=e^{\gamma g(S,y)}$. The linear equation (3.16) has a unique classical solution under Assumption 1, and the Feynman–Kac formula yields

$$\bar{N}(S,y,t)=\mathbb{E}_{\mathcal{Q}}\Big(e^{\gamma g(S_T,Y_T)}\mid S_t=S,Y_t=y\Big), \tag{3.18}$$

with $\mathcal{Q}$ given in Definition 2.7. Using an exponential transformation $\underline{N}(S,y,t)=e^{k(S,y,t)}$ gives that k must solve

$$k_t+\mathcal{L}^{(S,y)}k+\rho' a(y,t)L(y,t)k_y=0,$$

with $k(S,y,T) = \gamma g(S,y)$, which has a unique classical solution under our assumptions. The same then follows for (3.17). The probabilistic representation for k is

$$k(S,y,t) = \mathbb{E}_{\mathcal{Q}}\Big(\gamma g(S_T, Y_T) \mid S_t = S, Y_t = y\Big).$$

Using (3.12) and (3.3), we deduce

$$\frac{1}{\gamma}\ln \underline{N}(S,y,t) \le h^w(S,y,t) \le \frac{1}{\gamma}\ln \bar{N}(S,y,t),$$

which yields the desired upper and lower bounds (3.9). □

From these bounds, we easily obtain the following bounds on the *price spread*.

PROPOSITION 3.3. *The price spread $h^w - h^b$ is bounded by*

$$0 \le h^w - h^b \le \frac{1}{\gamma}\ln\left[\mathbb{E}_{\mathcal{Q}}\Big(e^{-\gamma g(S_T,Y_T)} \mid S_t = S, Y_t = y\Big)\mathbb{E}_{\mathcal{Q}}\Big(e^{\gamma g(S_T,Y_T)} \mid S_t = S, Y_t = y\Big)\right].$$

4. Fast mean-reverting stochastic volatility. We now study the indifference price using asymptotic approximations. In this section, we assume the European contract is a claim on S_T only and not on Y_T. That is, $g = g(S_T)$, as is usually the case.

4.1. Stochastic volatility framework. For clarity of exposition, we take the volatility-driving process (Y_t) to be an Ornstein–Uhlenbeck (OU) process, namely,

$$dY_t = \alpha(m - Y_t)\,dt + \beta\,(\rho\,dW_t^1 + \rho' dW_t^2),$$

where α is the rate of mean-reversion, m the long-run mean, and β the volatility of the volatility factor Y, which we shall call the "v-vol." In terms of the previous notation, we have $b(y) = \alpha(m-y)$ and $a(y) = \beta$. The process admits a unique invariant distribution, $\mathcal{N}(m,\nu^2)$, where $\nu^2 = \beta^2/(2\alpha)$. We also define at this stage the density of this distribution $\Phi(y)$ and the average $\langle\cdot\rangle$ with respect to this density:

$$\langle\chi\rangle = \int \chi\Phi.$$

In particular, we denote by $\bar{\sigma}$ the long-run (root-mean-square) volatility

$$\bar{\sigma} = \sqrt{\langle\sigma^2\rangle}. \tag{4.1}$$

The utility-based pricing equation (2.32) describes the indifference price h as the solution of a quasilinear differential equation depending on the risk-aversion parameter γ, the level of the volatility driving process y, and the parameters of this model α, β, m, ρ as well as the function $\sigma(\cdot)$. Given a fully specified model and parameters estimated from data, we could compute the utility price by numerically discretizing (2.32). Another approach is to use asymptotic approximations that are exact in some limit in which the problem simplifies. These can be used in certain parameter ranges which may be valid in some markets. They also give a deeper analytical understanding of the pricing mechanism and its relationship to modeling assumptions.

The limit we focus on here is fast mean-reversion of the volatility process, meaning that α is large. We write

$$\alpha = 1/\varepsilon, \qquad 0 < \varepsilon << 1,$$

and we are interested in the limit $\varepsilon \downarrow 0$ with the variance of the invariant distribution ν^2 fixed. This implies the scaling $\beta = \sqrt{2}\,\nu/\sqrt{\varepsilon}$. The choice of scaling is natural, because it allows us to pick up the effects of both mean-reversion and v-vol in the correction (first order) term of the approximation (4.10) below.

Evidence of a rapidly mean-reverting volatility factor in the S&P 500 is presented in the empirical study [18] of high-frequency data. Another recent empirical study [1] has found evidence of a fast volatility scale in exchange rate dynamics. Chernov et al. [5] propose and give evidence from data for two-factor stochastic volatility models in which one factor mean-reverts on a short time-scale. For present purposes, the method of this section can be regarded as yielding an approximation whose validity will depend on specific market conditions, in particular over time horizons when other slower factors can be considered effectively constant. The extension to incorporate slower scales using mixed singular and regular perturbation techniques is the subject of future investigation.

This method was previously used for *no-arbitrage* derivative pricing and hedging problems in Fouque, Papanicolaou, and Sircar [17]. The arguments were extended for stochastic control problems in [23]. We summarize the main findings from the former for *no-arbitrage* pricing and hedging European claims in order to compare the analogous results for the indifference pricing mechanism.

4.1.1. No-arbitrage pricing of European claims. Let $P(S, y, t)$ be the pricing function for a European claim with payoff $g(S_T)$. By *no-arbitrage* arguments, this price is given by

$$P(S, y, t) = \mathbb{E}^{(\lambda_m)}\{g(S_T) \mid S_t = S, Y_t = y\}, \tag{4.2}$$

where the expectation is taken with respect to the equivalent martingale measure $\mathbb{P}^{(\lambda_m)}$, and λ_m is the market price of volatility risk. We assume that $\lambda_m = \lambda_m(Y_t)$ in that it is a bounded function of Y_t only and therefore that (S, Y) is also a Markov process under $\mathbb{P}^{(\lambda_m)}$, which justifies the notation in (4.2). This premium is implicit in the prices of liquidly traded options or the market-set implied volatility skew. Under the measure $\mathbb{P}^{(\lambda_m)}$, the dynamics of (S, Y) can be written

$$\begin{aligned} dS_t &= \sigma(Y_t) S_t \, dW_t^\star, \\ dY_t &= \left[\frac{1}{\varepsilon}(m - Y_t) - \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}} \left(\rho \frac{\mu}{\sigma(Y_t)} - \rho' \lambda_m(Y_t) \right) \right] dt + \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}} \left(\rho \, dW_t^\star + \rho' dZ_t^\star \right), \end{aligned}$$

where $(W_t^\star)$ and $(Z_t^\star)$ are independent $\mathbb{P}^{(\lambda_m)}$-Brownian motions.

We make the following assumption.

ASSUMPTION 2. *The market price of volatility risk function $\lambda_m(\cdot)$ is smooth and bounded.*

The analysis of [17, Chapter 5] leads to the following approximation for P in the limit of fast mean-reversion:

$$P(S, y, t) \approx P^{(0)}(S, t) + \widetilde{P^{(1)}}(S, t), \tag{4.3}$$

where $P^{(0)}(S, t)$ is the Black–Scholes pricing function for the claim using the long-run average volatility parameter $\bar{\sigma}$ which is related to the original stochastic volatility model through (4.1). In other words, $P^{(0)}(S, t)$ solves the Black–Scholes PDE problem

$$\begin{aligned} \mathcal{L}_{BS}(\bar{\sigma}) P^{(0)} &= 0; \qquad t < T, \\ P^{(0)}(S, T) &= g(S), \end{aligned} \tag{4.4}$$

where

$$\mathcal{L}_{BS}(\bar{\sigma}) = \frac{\partial}{\partial t} + \frac{1}{2}\bar{\sigma}^2 S^2 \frac{\partial^2}{\partial S^2}, \tag{4.5}$$

the Black–Scholes differential operator at volatility level $\bar{\sigma}$. Under Assumptions 1 and 2, $P^{(0)}$ is smooth and bounded with bounded derivatives.

The correction term $\widetilde{P^{(1)}}$ accounting for stochastic volatility effects is given by

$$\widetilde{P^{(1)}}(S,t) = -(T-t)\left(V_2 S^2 P^{(0)}_{SS} + V_3 S^3 P^{(0)}_{SSS}\right) \tag{4.6}$$

for some constants V_2 and V_3 related to the original parameters and the functions σ and λ_m by formulas given below. In the case of smooth payoff g, we have the following convergence result, shown in [17]. We use the order notation

$$f(\varepsilon) = \mathcal{O}(g(\varepsilon)), \qquad \text{as } \varepsilon \downarrow 0 \qquad \Rightarrow \qquad \lim_{\varepsilon \downarrow 0} \frac{f(\varepsilon)}{g(\varepsilon)} = c,$$

for some constant c independent of ε, and $f(\varepsilon) = o(g(\varepsilon))$ if $c = 0$.

PROPOSITION 4.1. *Under Assumptions* 1 *and* 2*, for a fixed point* (S, y, t)*,*

$$|P(S,y,t) - (P^{(0)}(S,t) + \widetilde{P^{(1)}}(S,t))| = \mathcal{O}(\varepsilon). \tag{4.7}$$

For the case of call and put options when the payoff is only $\mathcal{C}^0$, the following convergence result is proved in [19] using a regularization technique.

PROPOSITION 4.2. *Under Assumptions* 1 *and* 2*, for a fixed point* (S, y, t)*,*

$$|P(S,y,t) - (P^{(0)}(S,t) + \widetilde{P^{(1)}}(S,t))| = o(\varepsilon|\log\varepsilon|^{1+p})$$

for any $p > 0$*.*

In Theorem 4.3 below, we prove the analogue of (4.7) for smooth and bounded g and the indifference pricing mechanism.

We note the following points about the approximation (4.3).

- To this level, namely, zeroth plus first order of approximation, the price is insensitive to the present level of the stochastic volatility process $\sigma(Y_t)$.
- The group parameters V_2 and V_3 are related to the original model as follows:

$$V_2 = \frac{\nu}{\sqrt{2\alpha}}\left(2\rho\langle\sigma\psi_1'\rangle - \left\langle\left(\frac{\mu\rho}{\sigma} + \rho'\lambda_m\right)\psi_1'\right\rangle\right), \tag{4.8}$$

$$V_3 = \frac{\rho\nu}{\sqrt{2\alpha}}\langle\sigma\psi_1'\rangle, \tag{4.9}$$

  where $\nu^2 = \beta^2/(2\alpha)$ and $\psi_1(y)$ is a solution of the Poisson equation (4.23) below. In practice, these relations are not used and V_2 and V_3 are estimated directly from the market-implied volatility skew and then can be used for pricing American and exotic claims to the same order of approximation.
- The formulas (4.8) and (4.9) show that V_2 and V_3 are of order $1/\sqrt{\alpha}$ and so are small under the assumption of fast mean-reversion. Moreover, V_3 is zero when $\rho = 0$ and in the case of nonzero correlation, the third-derivative term describes the leverage effect. In the case of equities, ρ is typically negative and returns distributions are asymmetric with a fatter left tail. The second-derivative term contains effects due to the extra kurtosis of stochastic volatility models over geometric Brownian motion, and the market price of volatility risk λ_m.

- Finally, the approximation is *robust* in the sense that it does not depend on specification of σ or λ_m within the class of functions described in Assumptions 1 and 2.

4.2. Approximation of indifference prices and interpretation. The main result of this section is the following theorem.

THEOREM 4.3. *Let h denote either the writer's or the buyer's indifference price, defined in* (2.10) *and* (2.11). *Under Assumption* 1, *for a fixed point* (S, y, t),

$$|h(S,y,t) - (P^{(0)}(S,t) - (T-t)(V_3 S^3 P^{(0)}_{SSS} + V_2^{(0)} S^2 P^{(0)}_{SS}))| = \mathcal{O}(\varepsilon). \tag{4.10}$$

Here, V_3 is defined in (4.9) *and $V_2^{(0)}$ denotes V_2 defined in* (4.8), *with $\lambda_m = 0$.*

Comparing with the *no-arbitrage* fast mean-reverting stochastic volatility approximation (4.6), we see from (4.8) that to this order of approximation, *the (writer's or buyer's) utility indifference price is exactly the no-arbitrage price in which there is zero risk premium from the second Brownian motion ($\lambda_m(\cdot) \equiv 0$). In particular, the risk-aversion coefficient γ does not appear in these first two terms of the approximation.*

The intuition for this is best understood from the relative entropy formulation of the indifference pricing mechanism discussed in section 2.1. Given a prior risk-neutral measure $\mathcal{Q}$ under which the volatility is *fast mean-reverting*, the utility pricer does not use his freedom to deviate from this belief up to the level of accuracy we have computed. In particular, he would have to choose a very large volatility risk premium λ in (2.41) for the asymptotics to lead to a different approximation in the first two terms, and this is penalized heavily by the entropy.

4.3. Expansions for indifference prices. To produce an asymptotic expansion for the indifference price, one may either analyze the indifference price equation (2.32) directly or, alternatively, approximate the value functions V and u involved in the pricing mechanism and, subsequently, approximate h via (2.10). We choose to proceed in the latter way because it also gives, as an intermediate output, useful results for the optimization problems that are interconnected with the utility-based prices.

We recall that the value functions u and V of the writer and the plain investor, respectively, can be written $u(x,S,y,t) = -e^{-\gamma x} G(S,y,t)$, with G solving (2.29), and $V(x,y,t) = -e^{-\gamma x} F(y,t)$, with F solving (2.23), and that the indifference price is given by

$$h(S,y,t) = \frac{1}{\gamma} \ln \frac{G(S,y,t)}{F(y,t)}. \tag{4.11}$$

It is convenient to work with the logarithmic transformations of G and F. To this end, we set

$$G = e^{\phi}, \qquad F = e^{\psi}, \tag{4.12}$$

where $\phi(S,y,t)$ solves (2.30) and $\psi(y,t)$ solves

$$\psi_t + \mathcal{L}^{(y)}\psi + \frac{1}{2} a(y)^2 (1-\rho^2) \psi_y^2 - \frac{\mu^2}{2\sigma(y)^2} = 0, \tag{4.13}$$

with $\psi(y,T) = 0$. In the latter case, we could as easily work with f, which solves the linear equation (2.21), but since we will need the expansion for ϕ, it is simpler to

construct that and then obtain the expansion for ψ by setting $g \equiv 0$ and removing the S dependence.

The first two terms of the asymptotic expansions of ϕ and ψ in powers of $\sqrt{\varepsilon}$ are summarized in the following proposition, which is proven in the following sections.

PROPOSITION 4.4. *Under Assumption* 1*, for a fixed point* (S, y, t) *the following hold.*

(i) *The first two terms of the expansion for* ϕ*, solution of* (2.30)*, lead to the approximation*

$$|\phi(S,y,t) - (\phi^{(0)}(S,t) + \widetilde{\phi^{(1)}}(S,t))| = \mathcal{O}(\varepsilon), \tag{4.14}$$

where

$$\phi^{(0)}(S,t) = \gamma P^{(0)}(S,t) - \frac{\mu^2}{2\sigma_\star^2}(T-t), \tag{4.15}$$

$$\widetilde{\phi^{(1)}}(S,t) = -\gamma(T-t)\left[V_3 S^3 P^{(0)}_{SSS}(S,t) + V_2^{(0)} S^2 P^{(0)}_{SS}(S,t) + \frac{\mu^3 C_4}{\gamma}\right]. \tag{4.16}$$

Here, V_3 *is defined in* (4.9)*,* $V_2^{(0)}$ *denotes* V_2 *defined in* (4.8)*, with* $\lambda_m = 0$*,* C_4 *is a market constant,* $C_4 = \frac{\rho\nu}{\sqrt{2\alpha}}\langle \frac{\psi_2'}{\sigma}\rangle$*, with* ψ_2 *defined below in* (4.24)*, and*

$$\bar{\sigma}^2 = \langle \sigma^2 \rangle, \qquad \frac{1}{\sigma_\star^2} = \left\langle \frac{1}{\sigma^2} \right\rangle. \tag{4.17}$$

(ii) *The first two terms of the expansion for* ψ*, solution of* (4.13)*, lead to the approximation*

$$|\psi(S,y,t) - (\psi^{(0)}(S,t) + \widetilde{\psi^{(1)}}(S,t))| = \mathcal{O}(\varepsilon), \tag{4.18}$$

where

$$\psi^{(0)}(S,t) = -\frac{\mu^2}{2\sigma_\star^2}(T-t), \qquad \widetilde{\psi^{(1)}}(S,t) = -(T-t)\mu^3 C_4.$$

Before we give the proof of Proposition 4.4, we introduce some convenient notation.

4.4. Operator notation. We write (2.30) in the compact form

$$\mathcal{L}^\varepsilon \phi + \frac{\nu^2}{\varepsilon}(1-\rho^2)\phi_y^2 = \frac{\mu^2}{2\sigma(y)^2}, \tag{4.19}$$

where we define

$$\begin{aligned}
\mathcal{L}^\varepsilon &= \frac{1}{\varepsilon}\mathcal{L}_0 + \frac{1}{\sqrt{\varepsilon}}\mathcal{L}_1 + \mathcal{L}_2,\\
\mathcal{L}_0 &= \nu^2 \frac{\partial^2}{\partial y^2} + (m-y)\frac{\partial}{\partial y},\\
\mathcal{L}_1 &= \sqrt{2}\,\rho\nu\left(\sigma(y) S \frac{\partial^2}{\partial S \partial y} - \frac{\mu}{\sigma(y)}\frac{\partial}{\partial y}\right),\\
\mathcal{L}_2 &= \frac{\partial}{\partial t} + \frac{1}{2}\sigma(y)^2 S^2 \frac{\partial^2}{\partial S^2}.
\end{aligned} \tag{4.20}$$

We notice that

1. $\mathcal{L}_0$ is the infinitesimal generator of the OU process with unit rate of mean-reversion;
2. $\mathcal{L}_1$ takes derivatives in y and so kills any function that does not depend on y; and
3. $\mathcal{L}_2 = \mathcal{L}_{BS}(\sigma(y))$, where $\mathcal{L}_{BS}(\cdot)$ is the Black–Scholes differential operator as a function of the volatility level defined in (4.5).

4.5. Proof of Proposition 4.4. We first describe an expansion of the form

$$\begin{aligned}\phi(S,y,t) = \phi^{(0)}(S,y,t) + \sqrt{\varepsilon}\phi^{(1)}(S,y,t) + \varepsilon\phi^{(2)}(S,y,t)\\ + \varepsilon^{3/2}\phi^{(2)}(S,y,t) - Z^{\varepsilon}(S,y,t),\end{aligned} \tag{4.21}$$

and write an equation for the *error term* Z^{ε}.

We define $\phi^{(0)}$ and $\sqrt{\varepsilon}\phi^{(1)} = \widetilde{\phi^{(1)}}$ by (4.15) and (4.16), respectively. In particular, $\phi^{(0)}$ and $\phi^{(1)}$ do not depend on y, and, by the smoothness assumptions on g, they are also smooth.

We define $\phi^{(2)}$ by

$$\phi^{(2)}(S,y,t) = -\frac{1}{2}\psi_1(y)S^2\phi^{(0)}_{SS} + \frac{\mu^2}{2}\psi_2(y), \tag{4.22}$$

where ψ_1 and ψ_2 are solutions of the Poisson equations

$$\mathcal{L}_0\psi_1 = \sigma(y)^2 - \bar{\sigma}^2, \tag{4.23}$$

$$\mathcal{L}_0\psi_2 = \frac{1}{\sigma(y)^2} - \frac{1}{\sigma_\star^2}. \tag{4.24}$$

The average volatilities $\bar{\sigma}$ and $\sigma_\star$ are defined in (4.17) and are finite as σ is bounded. The expressions (4.15), (4.16), and (4.22) are motivated by a formal expansion. Finally, we define $\phi^{(3)}(S,y,t)$ by

$$\phi^{(3)}(S,y,t) = \frac{\rho\nu}{\sqrt{2}}\left[\psi_3(y)(S^3\phi^{(0)}_{SSS} + 2S^2\phi^{(0)}_{SS}) - \mu\psi_4(y)S^2\phi^{(0)}_{SS} - \mu\psi_5(y)\right], \tag{4.25}$$

where ψ_3, ψ_4, and ψ_5 are solutions of the Poisson equations

$$\mathcal{L}_0\psi_3 = \sigma(y)\psi_1'(y) - \langle\sigma\psi_1'\rangle, \tag{4.26}$$

$$\mathcal{L}_0\psi_4 = \frac{\psi_1'(y)}{\sigma(y)} - \left\langle\frac{\psi_1'}{\sigma}\right\rangle, \tag{4.27}$$

$$\mathcal{L}_0\psi_5 = \frac{\psi_2'(y)}{\sigma(y)} - \left\langle\frac{\psi_2'}{\sigma}\right\rangle. \tag{4.28}$$

Here, $\phi^{(3)}$ has been chosen to be a solution of

$$\mathcal{L}_0\phi^{(3)} + \mathcal{L}_1\phi^{(2)} + \mathcal{L}_2\phi^{(1)} = 0,$$

which is a Poisson equation (in y). Its solvability condition (see [17, section 5.2.2]) is

$$\langle\mathcal{L}_2\phi^{(1)} + \mathcal{L}_1\phi^{(2)}\rangle = 0, \tag{4.29}$$

which is satisfied because that is how $\phi^{(1)}$ and $\phi^{(2)}$ were chosen.

LEMMA 4.1. *The functions $\phi^{(2)}$ and $\phi^{(3)}$ are smooth as functions of S and t and can be chosen to be at most logarithmically growing in y at infinity.*

Proof. As shown in [19, Appendix C], the boundedness assumption on σ implies that we can choose ψ_1 and ψ_2 to be at most logarithmically growing at infinity:

$$|\psi_{1,2}(y)| \leq c(1 + \ln(1 + |y|))$$

for some constant c. In particular, the derivatives $\psi'_{1,2}$ are bounded, and the right-hand sides of (4.26)–(4.28) are therefore bounded. Then we can also choose ψ_3, ψ_4, and ψ_5 to be at most logarithmically growing at infinity. Since $\phi^{(0)}(S,t)$ and $\phi^{(1)}(S,t)$ are smooth, the result follows from the explicit expressions (4.22) and (4.25). ∎

We assume hereon that $\phi^{(2)}$ and $\phi^{(3)}$ *are* chosen to be at most logarithmically growing in y at infinity.

Inserting the expansion (4.21) into the PDE (2.30), we obtain the following equation and terminal condition for Z^ε:

$$\mathcal{L}^\varepsilon Z^\varepsilon + 2\nu^2(1-\rho^2)(\phi_y^{(2)} + \sqrt{\varepsilon}\phi_y^{(3)})Z_y^\varepsilon - \frac{\nu^2}{\varepsilon}(1-\rho^2)(Z_y^\varepsilon)^2 = \varepsilon J, \tag{4.30}$$

$$Z^\varepsilon(S,y,T) = \varepsilon K(S,y). \tag{4.31}$$

In this calculation, we have used the fact that $\phi^{(0)}$, $\phi^{(1)}$, $\phi^{(2)}$, and $\phi^{(3)}$ satisfy

$$\mathcal{L}_0\phi^{(0)} + \nu^2(1-\rho^2)(\phi_y^{(0)})^2 = 0, \qquad \mathcal{L}_0\phi^{(1)} + \mathcal{L}_1\phi^{(0)} = 0,$$

$$\mathcal{L}_0\phi^{(2)} + \mathcal{L}_1\phi^{(1)} + \mathcal{L}_2\phi^{(0)} = \frac{\mu^2}{2\sigma(y)^2}, \tag{4.32}$$

as well as the solvability condition (4.29) for (4.32).

In (4.30) and (4.31), the source term J and terminal data K are given by

$$J = \mathcal{L}_1\phi^{(3)} + \mathcal{L}_2\phi^{(2)} + \sqrt{\varepsilon}\mathcal{L}_2\phi^{(3)} + \nu^2(1-\rho^2)\left(\phi_y^{(2)} + \sqrt{\varepsilon}\phi_y^{(3)}\right)^2, \tag{4.33}$$

$$K = \phi^{(2)}(S,y,T) + \sqrt{\varepsilon}\phi^{(3)}(S,y,T). \tag{4.34}$$

From Lemma 4.1, it follows that J and K are smooth and at most logarithmically growing in y at infinity. Notice that Z^ε is the unique classical solution of a quasilinear parabolic PDE problem. Defining

$$\theta(S,y,t) = 2\nu^2(1-\rho^2)\left(\phi_y^{(2)} + \sqrt{\varepsilon}\phi_y^{(3)}\right), \tag{4.35}$$

we can write (4.30) as

$$\hat{\mathcal{L}}^\varepsilon Z^\varepsilon = \varepsilon J + \frac{\nu^2}{\varepsilon}(1-\rho^2)(Z_y^\varepsilon)^2, \tag{4.36}$$

where $\hat{\mathcal{L}}^\varepsilon = \mathcal{L}^\varepsilon + \theta(S,y,t)\frac{\partial}{\partial y}$ is a linear parabolic operator. Notice that θ is bounded.

LEMMA 4.2. *Let $\bar{Z}^\varepsilon(S,y,t)$ be the unique classical solution of the linear PDE problem*

$$\begin{aligned} \hat{\mathcal{L}}^\varepsilon \bar{Z}^\varepsilon &= \varepsilon J, \\ \bar{Z}^\varepsilon(S,y,T) &= \varepsilon K(S,y). \end{aligned} \tag{4.37}$$

Then $Z^\varepsilon \leq \bar{Z}^\varepsilon$.

Proof. The result follows from the nonnegativity of the nonlinear term in (4.36) and classical comparison results for linear parabolic equations. □

LEMMA 4.3. *Define* $\underline{Z}^\varepsilon(S,y,t)$ *by* $\underline{Z}^\varepsilon = -\ln \bar{q}^\varepsilon$, *where* $\bar{q}^\varepsilon(S,y,t)$ *is the unique classical solution of the linear PDE problem*

$$\begin{aligned} \hat{\mathcal{L}}^\varepsilon \bar{q}^\varepsilon + \varepsilon J \bar{q}^\varepsilon &= 0, \\ \bar{q}^\varepsilon(S,y,T) &= e^{-\varepsilon K(S,y)}. \end{aligned} \tag{4.38}$$

Then $Z^\varepsilon \geq \underline{Z}^\varepsilon$.

Proof. We first observe that $q^\varepsilon = e^{-Z^\varepsilon}$ satisfies

$$\hat{\mathcal{L}}^\varepsilon q^\varepsilon + \varepsilon J q^\varepsilon = \frac{1}{2q^\varepsilon}\left(\frac{\rho\nu\sqrt{2}}{\sqrt{\varepsilon}} q_y^\varepsilon + \sigma(y) S q_S^\varepsilon\right)^2, \tag{4.39}$$

with terminal condition

$$q^\varepsilon(S,y,T) = e^{-\varepsilon K(S,y)}.$$

By the nonnegativity of the nonlinear term on the right-hand side of (4.39) and classical comparison results for linear parabolic equations, the result follows. □

We next need to show that the upper and lower bounds go to zero with ε.

LEMMA 4.4. *At fixed* (S,y,t),

$$\bar{Z}(S,y,t) = \mathcal{O}(\varepsilon).$$

Proof. We can write the probabilistic representation of (4.37),

$$\bar{Z}^\varepsilon(S,y,t) = \mathbb{E}^\star\left(\varepsilon K(\hat{S}_T, \hat{Y}_T) - \varepsilon \int_t^T J(\hat{S}_u, \hat{Y}_u, u)\, du \mid \hat{S}_t = S, \hat{Y}_t = y\right), \tag{4.40}$$

in terms of the processes $(\hat{S}, \hat{Y})$ defined by

$$\begin{aligned} d\hat{S} &= \sigma(\hat{Y})\hat{S}\left(\rho\, d\hat{B}^1 + \rho' d\hat{B}^2\right), \\ d\hat{Y} &= \frac{1}{\varepsilon}\left((m - \hat{Y}) - \sqrt{2\varepsilon}\frac{\rho\nu\mu}{\sigma(\hat{Y})} + \varepsilon\theta(\hat{S}, \hat{Y}, t)\right) dt + \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\, d\hat{B}^1, \end{aligned} \tag{4.41}$$

on some probability space $(\hat{\Omega}, \hat{\mathcal{F}}, (\hat{\mathcal{F}}_t), \hat{\mathbb{P}}^\star)$ where $\hat{B}^1$ and $\hat{B}^2$ are independent Brownian motions and where $\mathbb{E}^\star$ denotes expectation with respect to $\hat{\mathbb{P}}^\star$. This is because the infinitesimal generator of $(\hat{S}, \hat{Y})$ is $\hat{\mathcal{L}}^\varepsilon$.

We recall that $J(S,y,t)$ and $K(S,y)$ are at most logarithmically growing as functions of y and are bounded as functions of (S,t). Then the expectation (4.40) can be bounded by a combination of terms of the form $\varepsilon\mathbb{E}^\star\{\chi(\hat{Y}_u) \mid \hat{S}_t = S, \hat{Y}_t = y\}$, with $t < u \leq T$, for some logarithmically growing functions χ. Using Lemma 4.6 below, for ε sufficiently small, the terms $\mathbb{E}^\star\{\chi(\hat{Y}_u) \mid \hat{S}_t = S, \hat{Y}_t = y\}$ are bounded independent of ε. The result follows from (4.40). □

LEMMA 4.5. *At fixed* (S,y,t),

$$\underline{Z}(S,y,t) = \mathcal{O}(\varepsilon).$$

Proof. The solution of (4.38) has the probabilistic representation, via the Feynman–Kac formula

$$\bar{q}^{\varepsilon}(S,y,t) = \mathbb{E}^{\star}\left\{\exp\left(-\varepsilon K(\hat{S}_T,\hat{Y}_T) + \varepsilon\int_t^T J(\hat{S}_u,\hat{Y}_u,u)\,du\right) \mid \hat{S}_t = S, \hat{Y}_t = y\right\},$$

where $(\hat{S},\hat{Y})$ were defined in the previous lemma. From the properties of J and K, this can be bounded above by

$$\mathbb{E}^{\star}\left\{\exp\left(\varepsilon\chi(\hat{Y}_T,T) + \varepsilon\int_t^T \chi(\hat{Y}_u,u)\,du\right) \mid \hat{S}_t = S, \hat{Y}_t = y\right\}$$

for some functions χ at most logarithmically growing in $\hat{Y}$. From the ε-independent exponential moments of $\hat{Y}$ given in Lemma 4.6 below, it follows that

$$\bar{q}^{\varepsilon}(S,y,t) = 1 + \mathcal{O}(\varepsilon)$$

for fixed (S,y,t). The result follows from $\underline{Z}^{\varepsilon} = -\ln\bar{q}^{\varepsilon}$. □

LEMMA 4.6. *For $\hat{Y}$ defined in (4.41), there exist some $\varepsilon_0 > 0$ and a constant $C(t,T,v,y)$, independent of ε, such that for any $t \le s \le T$ and $0 < \varepsilon < \varepsilon_0$,*

$$\mathbb{E}^{\star}\{e^{v\hat{Y}_s} \mid \hat{S}_t = S, \hat{Y}_t = y\} < C(t,T,v,y).$$

Proof. The proof is a modification of [8, Proposition 2]. We first rewrite (4.41) as

$$d\hat{Y}_t = \left(\frac{1}{\varepsilon}(m - \hat{Y}_t) - \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\Lambda(\hat{S},\hat{Y},t)\right)dt + \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\,d\hat{B}^1_t,$$

with

$$\Lambda(S,y,t) = \frac{\rho\mu}{\sigma(y)} - \sqrt{\varepsilon}\frac{\theta(S,y,t)}{\nu\sqrt{2}},$$

where θ was defined in (4.35). Since θ is bounded, for any $\varepsilon_0 > 0$ and $0 < \varepsilon < \varepsilon_0$, Λ is bounded independent of ε.

Next, using Girsanov's theorem, we define an equivalent measure $\hat{\mathbb{P}}$ under which $\hat{Y}$ is a standard OU process. Introducing $\hat{W}^1_t = \hat{B}^1_t - \int_0^t \Lambda(\hat{S}_u,\hat{Y}_u,u)\,du$, we define $\hat{\mathbb{P}}$ by

$$\frac{d\hat{\mathbb{P}}^{\star}}{d\hat{\mathbb{P}}} = M_T^{(\Lambda)},$$

where

$$M_s^{(\Lambda)} = e^{-\int_t^s \Lambda(\hat{S}_u,\hat{Y}_u,u)d\hat{W}^1_u - \frac{1}{2}\int_t^s \Lambda(\hat{S}_u,\hat{Y}_u,u)^2 du}.$$

Then $\hat{W}^1$ is a $\hat{\mathbb{P}}$-Brownian motion and $M^{(\Lambda)}$ is a $(\hat{\mathbb{P}},(\hat{\mathcal{F}}_t))$-martingale.

Now we have

$$\mathbb{E}^{\star}\{e^{v\hat{Y}_s} \mid \hat{S}_t = S, \hat{Y}_t = y\} = \mathbb{E}\{e^{v\hat{Y}_s}M_s^{(\Lambda)} \mid \hat{S}_t = S, \hat{Y}_t = y\}, \tag{4.42}$$

where $\mathbb{E}$ denotes the expectation under $\hat{\mathbb{P}}$. We rewrite (4.42) as

$$\mathbb{E}\{e^{v\hat{Y}_s} M_s^{(\Lambda)} \mid \hat{S}_t = S, \hat{Y}_t = y\} = \mathbb{E}\left\{ e^{v\hat{Y}_s} e^{\frac{1}{2}\int_t^s \Lambda(\hat{S}_u, \hat{Y}_u, u)^2 du} \sqrt{M_s^{(2\Lambda)}} \mid \hat{S}_t = S, \hat{Y}_t = y \right\},$$

and, using the Cauchy–Schwarz inequality, we deduce that

$$\mathbb{E}^\star\{e^{v\hat{Y}_s} \mid \hat{S}_t = S, \hat{Y}_t = y\} \le \sqrt{\mathbb{E}\left\{ e^{2v\hat{Y}_s} e^{\int_t^s \Lambda(\hat{S}_u, \hat{Y}_u, u)^2 du} \mid \hat{S}_t = S, \hat{Y}_t = y \right\}}, \tag{4.43}$$

since $M^{(2\Lambda)}$ is a martingale with expected value equal to one. Therefore,

$$\mathbb{E}^\star\{e^{v\hat{Y}_s} \mid \hat{S}_t = S, \hat{Y}_t = y\} \le e^{\frac{1}{2}(s-t)||\Lambda||_\infty^2} \sqrt{\mathbb{E}\left\{ e^{2v\hat{Y}_s} \mid \hat{S}_t = S, \hat{Y}_t = y \right\}}. \tag{4.44}$$

Under $\hat{\mathbb{P}}$, the dynamics of $\hat{Y}$ are given by

$$d\hat{Y}_t = \frac{1}{\varepsilon}(m - \hat{Y}_t)\, dt + \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\, d\hat{W}_t^1.$$

That is, it is an autonomous standard OU process. From [8, Lemma 2], there exists a constant $c(v, y)$ such that for any $t \le s \le T$ and $\varepsilon > 0$ we have

$$\mathbb{E}\{e^{v\hat{Y}_s} \mid \hat{S}_t = S, \hat{Y}_t = y\} \le c(v, y).$$

The result follows by setting

$$C(t, T, v, y) = e^{\frac{1}{2}(T-t)||\lambda||_\infty^2} \sqrt{c(2v, y)}. \qquad \square$$

Combining these results, we have

$$\mathcal{O}(\varepsilon) = -\ln \bar{q}^\varepsilon(S, y, t) \le Z^\varepsilon(S, y, t) \le \bar{Z}^\varepsilon(S, y, t) = \mathcal{O}(\varepsilon)$$

for fixed (S, y, t). It follows that

$$Z^\varepsilon(S, y, t) \to 0 \qquad \text{as } \varepsilon \downarrow 0,$$

and part (i) of Proposition 4.4 follows from (4.21).

The proof of convergence of the approximation (4.18) for the solution ψ of (4.13) is a simpler version of the preceding, setting $g \equiv 0$ and thereby removing the S-dependences.

4.6. Proof of Theorem 4.3. From (4.11) and (4.12), we have that

$$h(S, y, t) = \frac{1}{\gamma}(\phi(S, y, t) - \psi(S, y, t)).$$

Using (4.14) and (4.18) and the triangle inequality trivially leads to (4.10), completing the proof.

5. Conclusions. The practical implications of the preceding analysis are pricing spreads in section 3 that can be tuned by the pricer's risk aversion; robust asymptotic approximations in section 4 that do not require precise specification of a stochastic volatility model; and analytical results for the problem of optimizing a portfolio consisting of dynamic positions in a liquid underlying asset combined with a static position in a derivative security. Such a problem is important in incomplete markets when investors would like to use derivatives to "trade volatility" indirectly but cannot rebalance frequently because of high transaction costs. Our results herein are applied to this problem in work in preparation.

Many questions about the utility pricing mechanism remain for future investigation: the implications for hedging derivative risk; the effect on implied volatilities, especially the term-structure, or variation with time-to-maturity; the multidimensional problem, meaning both the case of multifactor stochastic volatility models (where perhaps one factor is slow and another is fast mean-reverting), and the case of options on a basket of stocks. Mathematically, the main challenge is in extending the present analysis to the important cases of unbounded and nonsmooth payoffs.

REFERENCES

[1] S. Alizadeh, M. Brandt, and F. Diebold, *Range-based estimation of stochastic volatility models*, J. Finance, 57 (2002), pp. 1047–1092.

[2] M. Avellaneda, *The minimum-entropy algorithm and related methods for calibrating asset-pricing models*, Doc. Math., extra volume III (1998), pp. 545–563.

[3] N.L. Bowers, H.U. Gerber, J.C. Hickman, and D.A. Jones, *Actuarial Mathematics*, 2nd ed., Society of Actuaries, Schaumburg, IL, 1997.

[4] R. Carmona and L. Xu, *Calibrating Arbitrage-Free Stochastic Volatility Models by Relative Entropy Method*, Technical report, CEOR, Princeton University, Princeton, NJ, 1997.

[5] M. Chernov, R. Gallant, E. Ghysels, and G. Tauchen, *Alternative models for stock price dynamics*, J. Econometrics, 116 (2003), pp. 225–257.

[6] G.M. Constantinides and T. Zariphopoulou, *Bounds on prices of contingent claims in an intertemporal economy with proportional transaction costs and general preferences*, Finance Stoch., 3 (1999), pp. 345–369.

[7] G.M. Constantinides and T. Zariphopoulou, *Bounds on derivative prices in an intertemporal setting with proportional transaction costs and multiple securities*, Math. Finance, 11 (2001), pp. 331–346.

[8] P. Cotton, J.-P. Fouque, G. Papanicolaou, and K.R. Sircar, *Stochastic volatility corrections for interest rate derivatives*, Math. Finance, 14 (2004), pp. 173–200.

[9] M.H.A. Davis, *Option pricing in incomplete markets*, in Mathematics of Derivative Securities, M.A.H. Dempster and S.R. Pliska, eds., Cambridge University Press, Cambridge, UK, 1997, pp. 216–226.

[10] M.H.A. Davis, *Optimal Hedging with Basis Risk*, Technical report, Vienna University of Technology, Vienna, Austria, 2000.

[11] M.H.A. Davis, V. G. Panas, and T. Zariphopoulou, *European option pricing with transaction costs*, SIAM J. Control Optim., 31 (1993), pp. 470–493.

[12] F. Delbaen, P. Grandits, T. Rheinlander, D. Samperi, M. Schweizer, and C. Striker, *Exponential hedging and entropic penalties*, Math. Finance, 12 (2002), pp. 99–123.

[13] D. Duffie and T. Zariphopoulou, *Optimal investment with undiversifiable income risk*, Math. Finance, 3 (1993), pp. 135–148.

[14] W.H. Fleming and H.M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[15] W.H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[16] H. Foellmer and M. Schweizer, *Hedging of contingent claims under incomplete information*, in Applied Stochastic Analysis, M.H.A. Davis and R.J. Elliott, eds., Gordon and Breach, London, 1990.

[17] J.-P. Fouque, G. Papanicolaou, and K.R. Sircar, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, Cambridge, UK, 2000.

[18] J.-P. Fouque, G. Papanicolaou, K.R. Sircar, and K. Solna, *Short time-scale in S&P* 500 *volatility*, J. Computat. Finance, 6 (2003), pp. 1–23.
[19] J.-P. Fouque, G. Papanicolaou, K.R. Sircar, and K. Solna, *Singular perturbations in option pricing*, SIAM J. Appl. Math., 63 (2003), pp. 1648–1665.
[20] M. Fritelli, *The minimal entropy martingale measure and the valuation problem in incomplete markets*, Math. Finance, 10 (2000), pp. 39–52.
[21] S. Hodges and A. Neuberger, *Optimal replication of contingent claims under transaction costs*, Rev. Futures Markets, 8 (1989), pp. 222–239.
[22] J. Hull and A. White, *The pricing of options on assets with stochastic volatilities*, J. Finance, 42 (1987), pp. 281–300.
[23] M. Jonsson and K.R. Sircar, *Partial hedging in a stochastic volatility environment*, Math. Finance, 12 (2002), pp. 375–409.
[24] J. Kallsen, *Utility-based derivative pricing in incomplete markets*, in Mathematical Finance Finance Bachelier Congress 2000, Springer-Verlag, New York, 2000, pp. 313–338.
[25] I. Karatzas and S. Kou, *On the pricing of contingent claims under constraints*, Ann. Appl. Probab., 6 (1996), pp. 321–369.
[26] N.V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.
[27] O.A. Ladyzenskaja, V.A. Solonnikov, and N.N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.
[28] P.-L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. Part* 1: *The dynamic programming principle and applications*, Comm. Partial Differential Equations, 8 (1983), pp. 1101–74.
[29] P.-L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. Part* 2: *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–76.
[30] R. C. Merton, *Lifetime portfolio selection under uncertainty: The continuous-time case*, Rev. Econom. Statist., 51 (1969), pp. 247–257.
[31] M. Musiela and T. Zariphopoulou, *An example of indifference prices under exponential preferences*, Finance Stoch., 8 (2004), pp. 229–239.
[32] H. Pham, *Smooth solutions to optimal investment models with stochastic volatilities and portfolio constraints*, Appl. Math. Optim., 46 (2002), pp. 55–78.
[33] R. Rouge and N. ElKaroui, *Pricing via utility maximization and entropy*, Math. Finance, 10 (2000), pp. 259–276.
[34] T. Zariphopoulou, *Stochastic control methods in asset pricing*, in Handbook of Stochastic Analysis and Applications, D. Kannan and V. Lakshmikanathan, eds., Marcel Dekker, New York, 2001.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1354–1372

NEUMANN BOUNDARY CONTROL OF HYPERBOLIC EQUATIONS WITH POINTWISE STATE CONSTRAINTS*

BORIS S. MORDUKHOVICH† AND JEAN-PIERRE RAYMOND‡

Abstract. We consider optimal control problems for hyperbolic equations with controls in Neumann boundary conditions with pointwise constraints on the control and state functions. Focusing on the multidimensional wave equation with a nonlinear term, we derive new necessary optimality conditions in the form of a pointwise Pontryagin maximum principle for the state-constrained problem under consideration. Our approach is based on modern methods of variational analysis that allow us to obtain refined necessary optimality conditions with no convexity assumptions on integrands in the minimizing cost functional.

Key words. optimal control, wave equation, Neumann boundary controls, state constraints, necessary optimality conditions, Pontryagin maximum principle

AMS subject classifications. 49K20, 93C20, 35L20

DOI. 10.1137/S0363012903431177

1. Introduction.

This paper is concerned with optimal control problems for hyperbolic equations with controls in Neumann boundary conditions in the presence of *pointwise* constraints on the control and state functions. It is well known that *state-constrained* control problems are among the most challenging and difficult in dynamic optimization; see, e.g., [2, 3, 4, 7, 9, 17, 22, 25] and their references for state-constrained problems governed by parabolic and elliptic partial differential equations. To the best of our knowledge, such optimal control problems have not been studied yet for semilinear *hyperbolic equations* with controls in *Neumann boundary conditions*, which is the objective of this paper. For problems with no state constraints we refer to [18], where an optimal control problem for a semilinear hyperbolic equation with controls in Neumann boundary conditions has been studied.

Let Ω be an open bounded domain in $\mathbb{R}^N$, with a boundary Γ of class C^2, and let T be a positive time. We mainly pay attention to the following optimal control problem governed by the *semilinear wave equation*: minimize

$$J(y,u)=\int_\Omega f(x,y(T))dx+\int_Q g(x,t,y)dxdt+\int_\Sigma h(s,t,u)dsdt$$

over admissible pairs (y,u) satisfying

$$\left\{\begin{array}{ll} y_{tt}-\Delta y+\Phi(\cdot,y)=0 & \text{in } Q:=\Omega\times(0,T),\\ \partial_\nu y=u & \text{on } \Sigma:=\Gamma\times(0,T),\\ y(0)=y_0,\quad y_t(0)=y_1 & \text{in } \Omega, \end{array}\right. \tag{1.1}$$

under the pointwise constraints on control and state functions

$$u\in U_{ad}\subset L^2(\Sigma),\qquad y\in\mathcal{C}\subset C([0,T];L^2(\Omega)).$$

*Received by the editors July 7, 2003; accepted for publication (in revised form) March 4, 2004; published electronically January 5, 2005.

http://www.siam.org/journals/sicon/43-4/43117.html

†Department of Mathematics, Wayne State University, Detroit, MI 48202 (boris@math.wayne.edu). Research of this author was partly supported by the National Science Foundation under grants DMS-0072179 and DMS-0304989.

‡Laboratoire MIP, Université Paul Sabatier, 31062 Toulouse Cedex 4, France (raymond@mip.ups-tlse.fr).

We denote this problem by (P) and write it as

$$\text{(P)} \qquad \inf\{J(y,u) \mid (y,u) \text{ satisfies (1.1)},\ u \in U_{ad},\ y \in \mathcal{C}\}.$$

Assumptions on the nonlinear function Φ, as well as on the integrands f, g, and h, are presented and discussed in section 2. The initial state $(y_0, y_1) \in H^1(\Omega) \times L^2(\Omega)$ is fixed. Note that the main constructions and results of the paper can be extended to hyperbolic equations governed by more general *strongly elliptic operators* in (1.1)—not just by the Laplacian Δ—with *time-independent and regular* coefficients.

As mentioned above, we are not familiar with any publications concerning state-constrained control problems with Neumann boundary controls for hyperbolic equations. Some results for *distributed controls* in state-constrained hyperbolic systems are obtained in [7, 9, 28, 29]. Our preceding paper [21] deals with necessary optimality conditions for state-constrained problems governed by the wave equation with *Dirichlet* boundary controls.

It has been well recognized that Neumann and Dirichlet boundary conditions are essentially different in both parabolic and hyperbolic dynamical settings. While for parabolic equations the Dirichlet boundary value problem is considerably more difficult than the Neumann one, this is not the case for the hyperbolic dynamics; see, e.g., [11, 12, 13, 15, 17, 22, 24] and the references therein. On the contrary, the fundamental *regularity theory* for hyperbolic equations with Dirichlet boundary conditions has come first; cf. [11] and [13]. The sharp regularity results developed by Lasiecka and Triggiani for hyperbolic equations with Neumann boundary conditions [13] play a crucial role in this paper.

The main goal of this paper is to establish *necessary optimality conditions* for the state-constrained Neumann boundary control problem (P), which will be derived under rather mild and natural assumptions. In contrast to the Dirichlet case [21], where the dynamics is linear and the initial condition is in $L^2(\Omega) \times H^{-1}(\Omega)$, here we consider the *nonlinear dynamics* in (1.1) and we require *stronger* regularity assumptions on the initial state: $(y_0, y_1) \in H^1(\Omega) \times L^2(\Omega)$ in (1.1). On the other hand, the Neumann case provides *more regularity* of the corresponding solutions to the boundary value problem in (1.1), which eventually allows us to entirely *avoid the convexity assumptions*—on the integrands in the cost functional $J(y,u)$—that play a crucial role in the Dirichlet control problem considered in [21]. Moreover, in this paper we are able to establish necessary optimality conditions for (P) in the *pointwise form* of the Pontryagin maximum principle in contrast to the weaker integral form of [21].

Our approach to deriving necessary optimality conditions in the Neumann problem (P) is completely different from the one in [21] developed for the Dirichlet case. Instead of reducing the original problem to an abstract optimization problem and then using a suitable version of the Lagrange multiplier rule as in [21], we now employ *perturbation methods of modern variational analysis* involving *penalizing* state constraints and then passing to the limit from necessary optimality conditions in unconstrained approximating problems. In the case of optimal control problems governed by ordinary differential systems with even nonsmooth data this approach was developed in the 1970s; see, e.g., [5, 20, 27]. For problems governed by partial differential equations the situation is more complicated, and the first results have been obtained in the 1990s for bounded controls [2, 4]; see also [17] and the references therein. As mentioned in [9, page 595], versions of the maximum principle for unbounded control operators in the case of problems governed by partial differential equations were discovered by Fattorini [8] and independently by Raymond and Zidani in [25]. Based on

Ekeland's variational principle [6] and the approach developed in [25], we derive necessary optimality conditions for the original state-constrained hyperbolic problem (P) in the pointwise form of the maximum (actually minimum) principle of Pontryagin's type. Note that the approach developed in this paper allows us to obtain necessary optimality conditions for a more general version of problem (P), where the integrand h also depends on the *state variable* y. We are not going to pursue this issue here for simplicity.

The rest of the paper is organized as follows. In section 2 we present and discuss the *basic assumptions* used throughout the paper and then formulate the *main result* giving necessary conditions for optimal solutions to (P). Section 3 is devoted to the proper definitions of solutions and the subsequent analysis of the *state system* (1.1) and in the corresponding *adjoint system* appearing in the necessary optimality conditions.

Section 4 contains preparatory material allowing us to derive in the concluding section necessary optimality conditions in the pointwise maximum principle form for approximating problems with no state constraints. Namely, we obtain the so-called *increment formula* for the minimizing functional with respect to *diffuse/needle variations* of controls. This technique, which is well known for the case of ordinary differential equations (see, e.g., [10, 20]), requires a more delicate analysis in the case of partial differential equations. Following the constructions developed in [25] for the control of parabolic equations, we obtain an increment formula for approximating hyperbolic problems based on suitable *Taylor expansions* of the problem data with respect to diffuse perturbations of reference controls.

In the final section 5 we give the proof of the main result of this paper that involves the three major steps of variational analysis: (a) *perturbation* of the original state-constrained problem by a family of approximating problems with *no state constraints* by using Ekeland's variational principle in an appropriate metric space, (b) deriving necessary optimality conditions in the *approximating problems* that provide *suboptimality conditions* for the original problem, and finally (c) *passing to the limit* from the approximating problems to obtain the desired necessary conditions for the reference optimal solution to the state-constrained problem (P).

2. Basic assumptions and statement of the main theorem. Throughout the paper we use standard notation. For the reader's convenience we recall that $\mathcal{M}([0,T];L^2(\Omega))$ is the space of measures on $[0,T]$ with values in $L^2(\Omega)$, which is the topological dual of $C([0,T];L^2(\Omega))$. The topological dual of

$$C_0(]0,T];L^2(\Omega)) := \big\{y \in C([0,T];L^2(\Omega)) \mid y(0)=0\big\}$$

is denoted by $\mathcal{M}_b(]0,T];L^2(\Omega))$, and, similarly, the topological dual of

$$C_0(]0,T[;L^2(\Omega)) = \big\{y \in C([0,T];L^2(\Omega)) \mid y(0)=0,\ y(T)=0\big\}$$

is denoted by $\mathcal{M}_b(]0,T[;L^2(\Omega))$. Note that, according to the usual notation, the space $C_0(]0,T];L^2(\Omega))$ (respectively $C_0(]0,T[;L^2(\Omega))$) consists of continuous mappings on the interval $]0,T]$ (respectively $]0,T[$), vanishing at infinity. In what follows we identify $]0,T]$ and $]0,T[$ with $(0,T]$ and $(0,T)$, respectively.

It is well known that every measure $\mu \in \mathcal{M}_b(]0,T];L^2(\Omega))$ can be identified with a measure $\tilde{\mu} \in \mathcal{M}([0,T];L^2(\Omega))$ such that $\tilde{\mu}(\{0\})=0$ and $\tilde{\mu}|_{]0,T]}=\mu$, where $\tilde{\mu}|_{]0,T]}$ denotes the restriction of $\tilde{\mu}$ to $]0,T]$. Therefore, if $y \in C([0,T];L^2(\Omega))$ and $\mu \in \mathcal{M}_b(]0,T];L^2(\Omega))$, we still use the notation

$$\langle y,\mu\rangle_{C([0,T];L^2(\Omega)),\mathcal{M}_b(]0,T];L^2(\Omega))} \quad \text{for} \quad \langle y,\tilde{\mu}\rangle_{C([0,T];L^2(\Omega)),\mathcal{M}([0,T];L^2(\Omega))}.$$

Since we have to deal with equations satisfied in the sense of distributions in Q (see, e.g., (2.4)), it is also convenient to identify $\mathcal{M}_b(]0,T];L^2(\Omega))$ with a subspace of $\mathcal{M}_b(\Omega\times]0,T])$; this identification follows from the continuous and dense imbedding $C_0(\Omega\times]0,T]) \hookrightarrow C_0(]0,T];L^2(\Omega))$. Therefore, if $\mu \in \mathcal{M}_b(]0,T];L^2(\Omega))$, the notation $\mu|_Q$—the restriction of μ to Q—is meaningful if μ is considered as a bounded measure on $\Omega\times]0,T] = \Omega\times(0,T]$, and so $\mu|_{\Omega\times\{T\}}$ stands for $\mu(\{T\})$. The same kind of notation is used in the paper in similar settings. For $z \in L^2(Q)$ we denote by z_t (respectively by z_{tt}) the derivative (respectively, the second derivative) of z in t in the sense of distributions in Q.

Given a Banach space Z, the duality pairing between Z and Z' is denoted by $\langle\cdot,\cdot\rangle_{Z,Z'}$. When there is no ambiguity, we sometimes write $\langle\cdot,\cdot\rangle$ instead of $\langle\cdot,\cdot\rangle_{Z,Z'}$. To emphasize a specific kind of regularity of solutions to the hyperbolic equations under consideration, we may write, e.g., that $(y,y_t) \in C([0,T];X)\times C([0,T];Y)$ is a solution to (1.1) instead of just indicating that y is a solution to this system.

For the definition of the space $BV([0,T];(H^1(\Omega))')$—the space of functions of bounded variation on $[0,T]$ with values in $(H^1(\Omega))'$—we refer to [1, 19]. If $p \in BV([0,T];(H^1(\Omega))')$, one can define $p(t^-)$ and $p(t^+)$ for every $t \in (0,T)$ and also $p(0^+)$ and $p(T^-)$, while the values $p(0)$ and $p(T)$ may be generally different from $p(0^+)$ and $p(T^-)$. There is a unique Radon measure on $[0,T]$ with values in $(H^1(\Omega))'$, denoted by d_tp, such that the restriction of d_tp to $(0,T)$ is the vector-valued distributional derivative of p in $(0,T)$ with $d_tp(\{0\}) = (p(0^+) - p(0))$ and $d_tp(\{T\}) = (p(T) - p(T^-))$. Moreover, identifying p with its representative right-hand side continuous in $(0,T)$, one has

$$p(0^+) = p(0) + d_tp(\{0\}) \quad \text{and} \quad p(t) = p(0) + d_tp([0,t]) \quad \text{for every } t \in]0,T].$$

Recall that if $\{p_n\}$ is a bounded sequence in $BV([0,T];(H^1(\Omega))')$, then there is a subsequence $\{p_{n_k}\}$ and a function $p \in BV([0,T];(H^1(\Omega))')$ such that

$$p_{n_k}(t) \to p(t) \quad \text{weakly in } (H^1(\Omega))' \quad \text{for almost every (a.e.) } t \in [0,T].$$

Note that this convergence may hold for every $t \in [0,T]$ if the above representative right-hand side continuous in $(0,T)$ is not specified; see [1, Theorem 3.5] and [19, Proposition 16.1]. In particular,

$$p_{n_k}(T) \to p(T) \quad \text{weakly in } (H^1(\Omega))'.$$

Now let us formulate the standing *basic assumptions* on the initial data of problem (P) that are needed throughout this paper.

(A1) For every $y \in \mathbb{R}$, $\Phi(\cdot,\cdot,y)$ is measurable in Q. For a.e. pair $(x,t) \in Q$, $\Phi(x,t,\cdot)$ is of class C^1 on $\mathbb{R}$. Moreover, one has

$$\Phi(\cdot,0) \in L^1(0,T;L^2(\Omega)), \qquad |\Phi'_y(x,t,y)| \le M \quad \text{in } Q\times\mathbb{R}, \tag{2.1}$$

where M is a positive constant.

(A2) For every $y \in \mathbb{R}$, $f(\cdot,y)$ is measurable in Ω with $f(\cdot,0)$ belonging to $L^1(\Omega)$. For a.e. $x \in \Omega$, $f(x,\cdot)$ is a function of class C^1 on $\mathbb{R}$. Moreover, there is a constant $C > 0$ such that

$$|f'_y(x,y)| \le C(1+|y|) \quad \text{whenever } (x,y) \in \Omega\times\mathbb{R}.$$

(A3) For every $y \in \mathbb{R}$, $g(\cdot,\cdot,y)$ is measurable in Q with $g(\cdot,0)$ belonging to $L^1(Q)$. For a.e. $(x,t) \in Q$, $g(x,t,\cdot)$ is of class C^1. Moreover, there is a constant $C > 0$ such that

$$|g'_y(x,t,y)| \le C(1+|y|) \quad \text{whenever } (x,t,y) \in Q \times \mathbb{R}.$$

(A4) For every $u \in \mathbb{R}$, $h(\cdot,\cdot,u)$ is measurable on Σ with $h(\cdot,0)$ belonging to $L^1(\Sigma)$. For a.e. $(s,t) \in \Sigma$, $h(s,t,\cdot)$ is of class C^1. Moreover, there is a constant $C > 0$ such that

$$|h'_u(s,t,u)| \le C(1+|u|) \quad \text{whenever } (s,t,u) \in \Sigma \times \mathbb{R}.$$

(A5) The state constraint set $\mathcal{C} \subset C([0,T];L^2(\Omega))$ is closed and convex with $\operatorname{int}\mathcal{C} \neq \emptyset$. We suppose that the function defined by $\hat{y}_0(x,t) := y_0(x)$ belongs to the interior of $\mathcal{C}$ (y_0 denotes the initial state).

(A6) The control set U_{ad} is given in the form

$$U_{ad} := \{u \in L^2(\Sigma) \mid u(s,t) \in K(s,t) \text{ a.e. on } \Sigma\},$$

where K is a measurable multifunction whose values are nonempty and closed subsets of $\mathbb{R}$.

Of course, we suppose as usual that the set of *feasible pairs* (y,u) to (P) is *nonempty*, i.e., there is $u \in U_{ad}$ such that $J(y_u,u) < \infty$ and $y_u \in \mathcal{C}$, where y_u is the weak solution of system (1.1) corresponding to u; see section 3.

Observe that the above basic assumptions do *not* impose any *convexity* requirements on the integrands in the cost functional with respect to either state or control variables, as well as on the control set U_{ad}. This is different from the setting of [21] for the corresponding Dirichlet problem. The reason is that the Neumann boundary value problem offers *more regularity* in comparison with the Dirichlet one and allows us to employ powerful variational methods to prove necessary optimality conditions that *do not rely on weak convergences*. These methods applied to the Dirichlet boundary value problem definitely require full convexity for the limiting procedures to end up with pointwise results. On the other hand, in this paper we do not establish any existence theorems for optimal solutions, in contrast to [21]. In fact, in the Neumann setting under consideration it would be enough to assume convexity only with *respect to control* variables to justify the existence of optimal solutions by the so-called direct method. The stronger convexity assumptions imposed in [21] with respect to *both state and control* variables are due to the lack of regularity in the Dirichlet setting and are needed not only for the existence of optimal solutions but also for the proof of necessary optimality conditions as given in [21].

To formulate our main result, let us define the *Hamiltonian* function

$$H(s,t,u,p,\lambda) := pu + \lambda h(s,t,u)$$

for the control problem (P). The following theorem gives necessary conditions for optimal solutions to (P) that are a version of the *Pontryagin maximum principle* in pointwise form for the Neumann boundary control problem under consideration. Note that it is more convenient in our case to formulate this result with the *minimum* (not maximum) condition.

THEOREM 2.1 (pointwise necessary optimality conditions). *Let $(\bar{y},\bar{u})$ be an optimal solution to problem* (P) *satisfying assumptions* (A1)–(A6). *Then there exist $\lambda \in \mathbb{R}^+$, $\mu \in \mathcal{M}_b(]0,T];L^2(\Omega))$, and a measurable subset $\widetilde{\Sigma} \subset \Sigma$ such that*

$\mathcal{L}^N(\Sigma \setminus \widetilde{\Sigma}) = 0$,

$$(\lambda, \mu) \neq 0, \qquad \langle \mu, z - \bar{y} \rangle \leq 0 \quad \textit{for all } z \in \mathcal{C}, \quad \textit{and} \tag{2.2}$$

$$H(s,t,\bar{u}(s,t),p(s,t),\lambda) = \min_{u \in K(s,t)} H(s,t,u,p(s,t),\lambda) \quad \textit{for all } (s,t) \in \widetilde{\Sigma}, \tag{2.3}$$

where $\mathcal{L}^N$ denotes the N-dimensional Lebesgue measure, and p is the corresponding solution to the adjoint system

$$\begin{cases} p_{tt} - \Delta p + \Phi'_y(\cdot,\bar{y})p = \lambda g'_y(x,t,\bar{y}) + \mu|_Q & \textit{in } Q, \\ \partial_\nu p = 0 & \textit{on } \Sigma, \\ p(T) = y_0, \quad p_t(T) = -\lambda f'_y(x,\bar{y}(T)) - \mu|_{\Omega\times\{T\}} & \textit{in } \Omega. \end{cases} \tag{2.4}$$

The proof of Theorem 2.1 is conducted in section 5. The definitions of solutions to the state and adjoint systems in this theorem are given and discussed in the next section.

3. Analysis of the state and adjoint systems. Let us start with the Neumann boundary value problem for the *linear wave equation*

$$\begin{cases} y_{tt} - \Delta y = \phi & \text{in } Q, \\ \partial_\nu y = u & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } \Omega. \end{cases} \tag{3.1}$$

The following fundamental *regularity* result is established by Lasiecka and Triggiani [13]. Note that this result involves the space $C^1([0,T];(H^{1/2}(\Omega))')$ for weak solutions to (3.1) as stated in [14].

LEMMA 3.1 (basic regularity). *Assume that $(\phi, u, y_0, y_1) \in L^1(0,T;L^2(\Omega)) \times L^2(\Sigma) \times H^1(\Omega) \times L^2(\Omega)$ and let $y(\phi,u,y_0,y_1) \in C([0,T];L^2(\Omega)) \cap C^1([0,T];(H^1(\Omega))')$ be the unique weak solution to the linear Neumann boundary value problem* (3.1). *Then the mapping $u \mapsto y(0,u,0,0)$ is bounded from $L^2(\Sigma)$ to $C([0,T];H^{1/2}(\Omega)) \cap C^1([0,T];(H^{1/2}(\Omega))')$, and it is also bounded from $L^2(\Sigma)$ to $H^{3/5-\varepsilon}(Q)$ for all $\varepsilon > 0$. Furthermore, the mapping $(\phi, y_0, y_1) \mapsto y(\phi,0,y_0,y_1)$ is bounded from $L^1(0,T;L^2(\Omega)) \times H^1(\Omega) \times L^2(\Omega)$ to $C([0,T];H^1(\Omega)) \cap C^1([0,T];L^2(\Omega))$.*

Next we consider the Neumann boundary value problem for the linear wave equation with possibly *nonsmooth data*:

$$\begin{cases} y_{tt} - \Delta y + ay = \phi & \text{in } Q, \\ \partial_\nu y = u & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } \Omega, \end{cases} \tag{3.2}$$

where the nonsmooth coefficient $a(x,t)$ belongs to $L^\infty(Q)$. The following estimate of solutions to the homogeneous problem in (3.2) is needed in what follows.

LEMMA 3.2 (solution estimate for the homogeneous Neumann problem). *Assume that $u = 0$ and that $(\phi, y_0, y_1) \in L^1(0,T;L^2(\Omega)) \times H^1(\Omega) \times L^2(\Omega)$. Equation* (3.2) *admits a unique weak solution in $C([0,T];L^2(\Omega)) \cap C^1([0,T];(H^1(\Omega))')$. This solution satisfies the estimate*

$$\|y\|_{C([0,T];H^1(\Omega))} + \|y_t\|_{C([0,T];L^2(\Omega))} \leq C(\|\phi\|_{L^1(0,T;L^2(\Omega))} + \|y_0\|_{H^1(\Omega)} + \|y_1\|_{L^2(\Omega)}),$$

where the constant C may depend on $\|a\|_{L^\infty(Q)}$ and $\|\phi\|_{L^1(0,T;L^2(\Omega))}$, but it is invariant with respect to all $a(x,t)$ having the same $L^\infty(Q)$-norm.

Proof. The proof is standard. It is sufficient to multiply the first equation in (3.2) by y_t, to integrate it over Ω, and then to use Gronwall's lemma (see, for example, [18, page 184]). ∎

LEMMA 3.3 (compactness of the solution operator). *Assume that* $(\phi, y_0, y_1) = (0,0,0)$ *and that* $u \in L^2(\Sigma)$. *Equation* (3.2) *admits a unique weak solution* $y(u)$ *in* $C([0,T];L^2(\Omega)) \cap C^1([0,T];(H^1(\Omega))')$. *The mapping* $u \mapsto (y(u), y_t(u))$ *is a bounded operator from* $L^2(\Sigma)$ *into* $C([0,T];H^{1/2}(\Omega)) \times C([0,T];(H^{1/2}(\Omega))')$, *and the mapping* $u \mapsto y(u)$ *is a compact operator from* $L^2(\Sigma)$ *into* $C([0,T];L^2(\Omega))$.

Proof. The existence and uniqueness of the corresponding solution to (3.2) can be deduced from the well-known result for (3.1) by using a fixed point method in $L^2(0,\bar{t};L^2(\Omega))$ as $\bar{t}$ is sufficiently small and then by iterating the process n times with $n\bar{t} > T$; cf. [23] for more details. Moreover, in this way we get the estimate

$$\|y\|_{C([0,T];H^{1/2}(\Omega))} + \|y_t\|_{C([0,T];(H^{1/2}(\Omega))')} \le C\|u\|_{L^2(\Sigma)},$$

where C depends on an upper bound for the *norm* $\|a\|_{L^\infty(Q)}$ but not on $a(\cdot)$ itself. Now the compactness result follows from [26, Corollary 5]. ∎

Our next goal is to study the Neumann boundary value problem (1.1), which is labelled as the *state system* for convenience. We first recall the notion of *weak solutions* to the Neumann problem in (1.1) that is appropriate for the purposes of this paper.

DEFINITION 3.4 (weak solutions to the state system). *A function* $(y, y_t) \in C([0,T];L^2(\Omega)) \times C([0,T];(H^1(\Omega))')$ *is a* weak solution *to system* (1.1) *if*

$$\begin{aligned} &\int_Q -\Phi(\cdot,y)z\,dxdt \\ &\quad = \int_Q y\varphi\,dxdt - \langle y_t(0), z(0)\rangle_{(H^1(\Omega))',H^1(\Omega)} + \int_\Omega y(0)z_t(0)\,dx + \int_\Sigma zu\,dsdt, \end{aligned} \tag{3.3}$$

for all $\varphi \in L^1(0,T;L^2(\Omega))$, *where* z *solves the homogeneous Neumann boundary value problem*

$$\begin{cases} z_{tt} - \Delta z = \varphi & \text{in } Q, \\ \partial_\nu z = 0 & \text{on } \Sigma, \\ z(T) = 0, \quad z_t(T) = 0 & \text{in } \Omega. \end{cases} \tag{3.4}$$

The advantage of the above definition is that it allows us to establish the existence, uniqueness, and regularity of weak solutions to the original state system under the standing assumptions made in section 2.

THEOREM 3.5 (existence, uniqueness, and regularity of weak solutions to the state system). *For every* $(u, y_0, y_1) \in L^2(\Sigma) \times H^1(\Omega) \times L^2(\Omega)$ *the state system* (1.1) *admits a unique weak solution* (y, y_t) *in* $C([0,T];L^2(\Omega)) \times C([0,T];(H^1(\Omega))')$. *This solution belongs to* $C([0,T];H^{1/2}(\Omega)) \times C([0,T];(H^{1/2}(\Omega))')$ *and satisfies the estimate*

$$\|y\|_{C([0,T];H^{1/2}(\Omega))} + \|y_t\|_{C([0,T];(H^{1/2}(\Omega))')} \le C(\|u\|_{L^2(\Sigma)} + \|y_0\|_{H^1(\Omega)} + \|y_1\|_{L^2(\Omega)} + 1)$$

with some constant $C > 0$. *Moreover, the mapping* $(u, y_0, y_1) \mapsto y$ *is continuous from* $(u, y_0, y_1) \in L^2(\Sigma) \times H^1(\Omega) \times L^2(\Omega)$ *into* $C([0,T];H^{1/2}(\Omega)) \cap C^1([0,T];(H^{1/2}(\Omega))')$.

Proof. The existence of solutions in the space $C([0,\bar{t}];L^2(\Omega)) \cap C^1([0,\bar{t}];(H^1(\Omega))')$ with $\bar{t}$ sufficiently small can be obtained by a standard fixed point method. Then assumption (A1) and the estimates in Lemmas 3.2 and 3.3 allow us to ensure the existence of solutions in the space given in the theorem. The proof of uniqueness is also

standard and is omitted for brevity. The estimate of (y, y_t) in $C([0,T]; H^{1/2}(\Omega)) \cap C^1([0,T]; (H^{1/2}(\Omega))')$ follows from the estimate of y in $C([0,T]; L^2(\Omega))$ due to the basic Lemma 3.1. To justify the continuity of the mapping $(u, y_0, y_1) \mapsto y$ from $(u, y_0, y_1) \in L^2(\Sigma) \times H^1(\Omega) \times L^2(\Omega)$ into $C([0,T]; H^{1/2}(\Omega)) \cap C^1([0,T]; (H^{1/2}(\Omega))')$, we again use assumption (A1) and the corresponding estimates for the linearized system (3.2) presented in Lemmas 3.2 and 3.3. □

Next we consider the *adjoint system* given by

$$\begin{cases} p_{tt} - \Delta p + ap = \mu|_Q & \text{in } Q, \\ \partial_\nu p = 0 & \text{on } \Sigma, \\ p(T) = 0, \quad p_t(T) = -\mu|_{\Omega \times \{T\}} & \text{in } \Omega, \end{cases} \tag{3.5}$$

where $\mu \in \mathcal{M}_b(]0,T]; L^2(\Omega))$, where $\mu|_Q$ and $\mu|_{\Omega\times\{T\}}$ denote the restriction of μ to Q and to $\Omega \times \{T\}$, respectively, and where $a \in L^\infty(Q)$.

In order to introduce and justify an appropriate definition of solutions to the adjoint system (3.5), we need the following lemma that is certainly of independent interest.

LEMMA 3.6 (divergence formula). *The space*

$$W := \{\vec{V} \in (L^2(Q))^{N+1} \mid \operatorname{div}(\vec{V}) \in \mathcal{M}_b(]0,T[; L^2(\Omega))\}$$

endowed with the norm $\|\vec{V}\|_W := \|\vec{V}\|_{(L^2(Q))^{N+1}} + \|\operatorname{div}(\vec{V})\|_{\mathcal{M}_b(]0,T[;L^2(\Omega))}$ *is a Banach space. There exists a unique continuous operator* γ_{ν_Q} *from* W *into* $H^{-1/2}(\partial Q)$ *satisfying*

$$\gamma_{\nu_Q}(\vec{V}) = \gamma_0(\vec{V}) \cdot \nu_Q$$

for every $\vec{V} \in (C^1(\overline{Q}))^{N+1}$ *and such that the divergence formula*

$$\begin{aligned} \int_Q \vec{V} \cdot \nabla\phi + \langle \phi, \operatorname{div}(\vec{V})\rangle_{C([0,T];L^2(\Omega)),\mathcal{M}_b(]0,T[;L^2(\Omega))} \\ = \langle \gamma_{\nu_Q}(\vec{V}), \gamma_0(\phi)\rangle_{H^{-1/2}(\partial Q), H^{1/2}(\partial Q)} \end{aligned} \tag{3.6}$$

holds for all $\phi \in H^1(Q)$.

Proof. It is easy to see that the space W is Banach. Let Λ be a continuous extension operator from $H^{1/2}(\partial Q)$ into $H^1(Q)$ that is a bounded linear operator from $H^{1/2}(\partial Q)$ into $H^1(Q)$ satisfying

$$\gamma_0 \Lambda\varphi = \varphi \quad \text{for all } \varphi \in H^{1/2}(\partial Q).$$

Taking $\vec{V} \in (C^1(\overline{Q}))^{N+1}$, observe that the functional

$$\varphi \longmapsto \int_Q \vec{V} \cdot \nabla \Lambda\varphi + \langle \Lambda\varphi, \operatorname{div}(\vec{V})\rangle_{C([0,T];L^2(\Omega)),\mathcal{M}_b(]0,T[;L^2(\Omega))}$$

is linear and bounded on $H^{1/2}(\partial Q)$. Denoting this functional by $\gamma_{\nu_Q}(\vec{V})$, we directly verify that

$$\gamma_{\nu_Q}(\vec{V}) = \gamma_0(\vec{V}) \cdot \nu_Q,$$

and that the divergence formula (3.6) is satisfied. This means that $\gamma_{\nu_Q}(\vec{V})$ does not depend on the extension operator Λ. Furthermore, one has

$$\left| \int_Q \vec{V} \cdot \nabla \Lambda\varphi + \langle \Lambda\varphi, \operatorname{div}(\vec{V}) \rangle_{C([0,T];L^2(\Omega)),\mathcal{M}_b(]0,T[;L^2(\Omega))} \right| \le C\|\varphi\|_{H^{1/2}(\partial Q)} \|\vec{V}\|_W,$$

which implies that

$$\|\gamma_{\nu_Q}(\vec{V})\|_{H^{-1/2}(\partial Q)} \le C\|\vec{V}\|_W \quad \text{for all } \vec{V} \in (C^1(\overline{Q}))^{N+1}.$$

Since $(C^1(\overline{Q}))^{N+1}$ is dense in W, the proof is complete. □

Next take $(p, p_t) \in L^2(0,T;H^1(\Omega)) \times L^2(0,T;L^2(\Omega))$ and assume that the combination $p_{tt} - \Delta p$, calculated in the sense of distributions on Q, belongs to $\mathcal{M}_b(]0,T[;L^2(\Omega))$. Employing Lemma 3.6, we define the *normal trace* on ∂Q of the vectorfield $(-\nabla p, p_t)$ as an element of $H^{-1/2}(\partial Q)$. Then one has the estimate

$$\begin{aligned} &\|\gamma_{\nu_Q}(-\nabla p, p_t)\|_{H^{-1/2}(\partial Q)} \\ &\quad \le C(\|p\|_{L^2(0,T;H^1(\Omega))} + \|p_t\|_{L^2(Q)} + \|p_{tt} - \Delta p\|_{\mathcal{M}_b(]0,T[;L^2(\Omega))}), \end{aligned}$$

where the constant $C > 0$ is independent of p. Since $\Omega \times \{0\}$ is an open subset in ∂Q, the restriction of $\gamma_{\nu_Q}(-\nabla p, p_t)$ to $\Omega \times \{0\}$ belongs to $H^{-1/2}(\Omega)$. Thus we get

$$\gamma_{\nu_Q}(-\nabla p, p_t)|_{\Omega\times\{0\}} = p_t(0) \in H^{-1/2}(\Omega).$$

Note that this result can be improved. We are going to show in Theorem 3.8 that a properly defined solution p to (3.5) actually has the property $p_t(0) \in L^2(\Omega)$.

Now we are ready to introduce an appropriate notion of *weak solutions* to the adjoint system (3.5) and justify their basic properties needed in what follows.

DEFINITION 3.7 (weak solutions to the adjoint system). *A function $p \in L^\infty(0,T;L^2(\Omega))$ is a* weak solution *to* (3.5) *if*

$$\langle y(\varphi), \mu \rangle_{C([0,T];L^2(\Omega))\times\mathcal{M}_b(]0,T];L^2(\Omega))} - \int_Q p\varphi \, dxdt = 0 \tag{3.7}$$

for all $\varphi \in L^1(0,T;L^2(\Omega))$, where $y(\varphi)$ is the solution to

$$\begin{cases} y_{tt} - \Delta y + ay = \varphi & \text{in } Q, \\ \partial_\nu y = 0 & \text{on } \Sigma, \\ y(0) = 0, \quad y_t(0) = 0 & \text{in } \Omega. \end{cases} \tag{3.8}$$

The next theorem establishes the existence, uniqueness, and regularity of weak solutions to the adjoint system under the standing assumptions made.

THEOREM 3.8 (existence, uniqueness, and regularity of weak solutions to the adjoint system). *The adjoint system* (3.5) *admits a unique weak solution $(p, p_t) \in L^\infty(0,T;H^1(\Omega)) \times L^\infty(0,T;L^2(\Omega))$. This solution satisfies $p_t \in BV([0,T];(H^1(\Omega))')$, $p \in C_w([0,T];H^1(\Omega))$, and*

$$p_t(\tau) \in L^2(\Omega) \quad \text{whenever } \tau \in \{t \in [0,T] \mid \mu(\{t\}) = 0\},$$

which implies that $p_t(0) \in L^2(\Omega)$ ($C_w([0,T];H^1(\Omega))$ denotes the space of continuous functions from $[0,T]$ into $H^1(\Omega)$ endowed with its weak topology). Moreover, one has the estimate

$$\|p\|_{L^\infty(0,T;H^1(\Omega))} + \|p_t\|_{L^\infty(0,T;L^2(\Omega))} \le C\|\mu\|_{\mathcal{M}_b(]0,T];L^2(\Omega))}, \tag{3.9}$$

where C depends on $\|a\|_{L^\infty(Q)}$ but is invariant with respect to the functions $a(x,t)$ having the same norm in the space $L^\infty(Q)$.

Proof. Observe that $p=0$ when the pair $(p,p_t)\in L^\infty(0,T;H^1(\Omega))\times L^\infty(0,T;L^2(\Omega))$ satisfies (3.7) with $\mu=0$. This implies that the adjoint system (3.5) *cannot admit more than one* weak solution. To prove the *existence* of a weak solution, we develop an approximation procedure. First build a sequence $\{\mu_n\}\subset L^1(0,T;L^2(\Omega))$ satisfying

$$\begin{aligned}\lim_{n\to\infty}\int_Q y\mu_n\,dxdt &= \langle y,\mu|_{]0,T[}\rangle_{C([0,T];L^2(\Omega)),\mathcal{M}_b(]0,T[;L^2(\Omega))}\\ &\text{for all } y\in C([0,T];L^2(\Omega)),\quad \text{and}\\ &\|\mu_n\|_{L^1(0,T;L^2(\Omega))}=\|\mu|_{]0,T[}\|_{\mathcal{M}_b(]0,T[;L^2(\Omega))}.\end{aligned}\tag{3.10}$$

To define μ_n, we follow the construction in the appendix of [24]. Let $\bar\mu$ be the extension of $\mu|_{]0,T[}$ by zero to $\mathbb{R}$, let $\{\rho_n\}$ be a sequence of nonnegative symmetric mollifiers on $\mathbb{R}$ with their supports in $(-1/n,1/n)$, and let S_0 and S_T be the functions on $\mathbb{R}$ defined by $S_0(t):=-t$ and $S_T(t):=2T-t$. Given $n\ge 2$, we set

$$\bar\mu_n(A):=(\bar\mu*\rho_n)(A)+(\bar\mu*\rho_n)(S_0(A))+(\bar\mu*\rho_n)(S_T(A))$$

for every Borel subset A in $\mathbb{R}$, and then construct the desired measure by

$$\mu_n:=\frac{\|\mu\|_{\mathcal{M}_b(]0,T[;L^2(\Omega))}}{\|\bar\mu_n|_{]0,T[}\|_{\mathcal{M}_b(]0,T[;L^2(\Omega))}}\bar\mu_n|_{]0,T[}.$$

Following [24, appendix], one can verify both relations formulated in (3.10).

Considering now the unique solution p_n to the system

$$\begin{cases} p_{tt}-\Delta p+ap=\mu_n & \text{in } Q,\\ \partial_\nu p=0 & \text{on } \Sigma,\\ p(T)=0,\quad p_t(T)=-\mu|_{\Omega\times\{T\}} & \text{in } \Omega\end{cases}\tag{3.11}$$

and applying Lemma 3.2, we get the estimate

$$\begin{aligned}&\|p_n\|_{L^\infty(0,T;H^1(\Omega))}+\|p_{nt}\|_{L^\infty(0,T;L^2(\Omega))}+\|p_n(0)\|_{H^1(\Omega)}+\|p_{nt}(0)\|_{L^2(\Omega)}\\ &\quad\le C\|\mu\|_{\mathcal{M}_b(]0,T];L^2(\Omega))}\end{aligned}\tag{3.12}$$

with a constant $C>0$ independent of n, where p_{nt} stands for the derivative of p_n with respect to t in $(0,T)$ in the sense of vector-valued distributions. Denoting by p_{ntt} the corresponding derivative of p_{nt} with respect to t in $(0,T)$ and using (3.11), we arrive at

$$p_{ntt}=\pi_n+\mu_n\in L^\infty(0,T;(H^1(\Omega))')+\mathcal{M}_b(]0,T[;L^2(\Omega))\subset\mathcal{M}_b(]0,T[;(H^1(\Omega))'),$$

where the operator π_n is defined by

$$\langle\pi_n,y\rangle_{L^\infty(0,T;(H^1(\Omega))'),L^1(0,T;H^1(\Omega))}:=\int_Q(\nabla p_n\cdot\nabla y-ap_ny)\,dxdt.$$

Therefore, in addition to (3.12), the sequences $\{p_{ntt}\}$ and $\{p_{nt}\}$ are bounded in the spaces $\mathcal{M}_b(]0,T[;(H^1(\Omega))')$ and $BV([0,T];(H^1(\Omega))')$, respectively. Observing that $\mathcal{M}_b(]0,T[;(H^1(\Omega))')$ is the dual of a *separable* Banach space, we select weak* convergent *subsequences* of the above sequences. The same *sequential compactness* property

holds also for the space $BV([0,T];(H^1(\Omega))')$; see section 2. In this way we find $p \in L^\infty(0,T;H^1(\Omega))$ with $p_t \in L^\infty(0,T;L^2(\Omega)) \cap BV([0,T];(H^1(\Omega))')$ and a subsequence $\{p_n\}$ converging to p in the weak* topology of $L^\infty(0,T;H^1(\Omega))$ and such that the corresponding subsequence $\{p_{nt}\}$ converges weak* in $L^\infty(0,T;L^2(\Omega))$ to p_t. Furthermore, since $\gamma_{\nu_Q}(-\nabla p_n, p_{nt})$ is bounded in $L^2(\partial Q)$, we can also deduce that the sequence of $\gamma_{\nu_Q}(-\nabla p_n, p_{nt})$ converges to $\gamma_{\nu_Q}(-\nabla p, p_t)$ in the weak topology of $L^2(\partial Q)$. Taking into account the relations

$$\gamma_{\nu_Q}(-\nabla p_n, p_{nt})|_{\Omega\times\{T\}} = \mu|_{\Omega\times\{T\}} \quad \text{and} \quad \gamma_{\nu_Q}(-\nabla p_n, p_{nt})|_\Sigma = 0,$$

one gets that $\gamma_{\nu_Q}(-\nabla p, p_t)|_\Sigma = -\partial_\nu p = 0$ and that

$$\gamma_{\nu_Q}(-\nabla p_n, p_{nt})|_{\Omega\times\{0\}} = p_{nt}(0) \rightharpoonup \gamma_{\nu_Q}(-\nabla p, p_t)|_{\Omega\times\{0\}} = p_t(0)$$

in the weak topology of $L^2(\Omega)$. Finally, by passing to the limit in the equality

$$\langle y(\varphi), \mu_n\rangle_{C([0,T];L^2(\Omega)),\mathcal{M}_b(]0,T];L^2(\Omega))} - \int_Q p_n\varphi = 0,$$

where $y(\varphi)$ is the solution of (3.8), we conclude that (p, p_t) is the desired weak solution to the adjoint system (3.5). The proof of the theorem is complete. ∎

The last result of this section gives a useful Green-type relationship between the corresponding solutions of the (linearized) state and adjoint systems.

THEOREM 3.9 (Green formula). *Assume that* $(\phi, y_0, y_1) = (0,0,0)$ *and that* $u \in L^2(\Sigma)$, *let* y *be the corresponding weak solution to system* (3.2), *and let* p *satisfy* (3.5). *Then*

$$\langle y, \mu\rangle_{C([0,T];L^2(\Omega)),\mathcal{M}_b(]0,T];L^2(\Omega))} - \int_Q p\varphi = \int_\Sigma up\, ds\, dt. \tag{3.13}$$

Proof. This formula can be proved for the pair (y, p_n), where p_n is the solution to the approximating adjoint system (3.11). Passing there to the limit as $n \to \infty$, we obtain the desired Green formula (3.13) as formulated in the theorem. ∎

4. Diffuse perturbations and increment formula. As mentioned in section 1, our approach to deriving necessary optimality conditions in the original state-constrained problem (P) includes an approximation procedure to penalize the state constraints. In this way we arrive at a family of Neumann boundary control problems for hyperbolic equations with pointwise (or hard) constraints on the control variable but with *no state constraints*. Although the latter approximating problems are essentially easier than the initial state-constrained problem (P), they still require a delicate variational analysis. As is well known in the control theory for ordinary differential equations, a key element in obtaining maximum-type conditions for problems with hard constraints on control but not on state variables is the so-called *increment formula* for the minimizing cost functional with respect to *needle variations* of reference controls; see, e.g., [10, 20]. In this section we obtain some counterparts of such results for the hyperbolic control problems under consideration, by using the so-called diffuse perturbations first introduced in [16] and then developed in [2, 3, 4, 25]; see also the references therein. Here we follow the construction developed in [25].

Given a reference control $\bar u \in U_{ad}$, an admissible control $u \in U_{ad}$, and a number $\rho \in (0,1)$, a *diffuse perturbation* of $\bar u$ is defined by

$$u_\rho(s,t) := \begin{cases} \bar u(s,t) & \text{on } \Sigma \setminus E_\rho, \\ u(s,t) & \text{on } E_\rho, \end{cases} \tag{4.1}$$

where E_ρ is a measurable subset of Σ. The next theorem can be viewed as an increment formula for the cost functional $J(y,u)$ with respect to diffuse perturbations of the reference control. Note that it also contains the corresponding Taylor expansion for state trajectory of (1.1), which is an essential ingredient of the increment formula.

THEOREM 4.1 (increment formula). *Given arbitrary controls $\bar{u}, u \in U_{ad}$ and a number $\rho \in (0,1)$, we consider the diffuse perturbation defined in* (4.1) *and the weak solutions $\bar{y}$ and y_ρ of system* (1.1) *corresponding to $\bar{u}$ and u_ρ, respectively. Then there exists a measurable subset $E_\rho \subset \Sigma$ such that the following hold:*

$$\mathcal{L}^N(E_\rho) = \rho \mathcal{L}^N(\Sigma), \tag{4.2}$$

$$\int_{E_\rho} (h(s,t,\bar{u}) - h(s,t,u))\, dsdt = \rho \int_\Sigma (h(s,t,\bar{u}) - h(s,t,u))\, dsdt, \tag{4.3}$$

$$y_\rho = \bar{y} + \rho z + \rho r_\rho \quad \text{with } \lim_{\rho \to 0} \|r_\rho\|_{C([0,T];L^2(\Omega))} = 0, \tag{4.4}$$

$$J(y_\rho, u_\rho) = J(\bar{y},\bar{u}) + \rho \Delta J + o(\rho) \quad \text{with } \Delta J := J'_y(\bar{y},\bar{u})z + J(\bar{y},u) - J(\bar{y},\bar{u}), \tag{4.5}$$

where z is the weak solution to the system

$$\begin{cases} z_{tt} - \Delta z + \Phi'_y(\cdot,\bar{y})z = 0 & \text{in } Q, \\ \partial_\nu z = \bar{u} - u & \text{on } \Sigma, \\ z(0) = 0, \quad z_t(0) = 0 & \text{in } \Omega. \end{cases} \tag{4.6}$$

The proof of the theorem given below relies on the following technical lemma, which follows from [25, Lemma 4.1].

LEMMA 4.2 (diffuse perturbations). *Let $\bar{u}, u \in U_{ad}$. For every $\rho \in (0,1)$ there is a sequence of measurable subsets E^n_ρ in Σ such that*

$$\mathcal{L}^N(E^n_\rho) = \rho \mathcal{L}^N(\Sigma), \tag{4.7}$$

$$\int_{E^n_\rho} (h(s,t,\bar{u}) - h(s,t,u))\, dsdt = \rho \int_\Sigma (h(s,t,\bar{u}) - h(s,t,u))\, dsdt, \tag{4.8}$$

$$\frac{1}{\rho}\chi_{E^n_\rho} \rightharpoonup 1 \quad \text{weak}^* \text{ in } L^\infty(\Sigma) \text{ as } n \to \infty, \tag{4.9}$$

where χ_E stands for the characteristic function of the set E.

Proof of Theorem 4.1. The existence of the subsets E_ρ satisfying (4.2) and (4.3) is an easy consequence of Lemma 4.2. The main issue is to justify the Taylor expansion (4.4) for the trajectories y_ρ of (1.1) corresponding to the diffuse control perturbations. One clearly sees that (4.4) and (4.3) imply the increment formula (4.5) due to the construction of diffuse perturbations.

To prove (4.4), we pick a number $\rho \in (0,1)$, take the sets E^n_ρ from Lemma 4.2, and build the diffuse control perturbations

$$u^n_\rho(s,t) := \begin{cases} \bar{u}(s,t) & \text{on } \Sigma \setminus E^n_\rho, \\ u(s,t) & \text{on } E^n_\rho. \end{cases}$$

Let y^n_ρ be the solution of (1.1) corresponding to u^n_ρ and let z be the (unique) weak solution of (4.6). It is easy to see that for all n the function $\xi^n_\rho := (y^n_\rho - \bar{y})/\rho - z$ is the unique weak solution to the system

$$\begin{cases} \xi_{tt} - \Delta \xi + a^n_\rho \xi = f^n_\rho & \text{in } Q, \\ \partial_\nu \xi = w^n_\rho & \text{on } \Sigma, \\ \xi(0) = 0, \quad \xi_t(0) = 0 & \text{in } \Omega, \end{cases}$$

with the following data:

$$a_\rho^n := \int_0^1 \Phi_y'(\cdot, \bar{y}+\theta(y_\rho^n-\bar{y}))\, d\theta, \quad f_\rho^n := (\Phi_y'(\cdot,\bar{y})-a_\rho^n)z, \quad w_\rho^n := \left(1-\frac{1}{\rho}\chi_{E_\rho^n}\right)(u-\bar{u}).$$

Denote by $\xi_\rho^{n,1}$ the solution to

$$\begin{cases} \xi_{tt} - \Delta\xi + a_\rho^n \xi = f_\rho^n & \text{in } Q, \\ \partial_\nu \xi = 0 & \text{on } \Sigma, \\ \xi(0) = 0, \quad \xi_t(0) = 0 & \text{in } \Omega, \end{cases}$$

by $\xi_\rho^{n,2}$ the solution to

$$\begin{cases} \xi_{tt} - \Delta\xi + a_\rho^n \xi = 0 & \text{in } Q, \\ \partial_\nu \xi = w_\rho^n & \text{on } \Sigma, \\ \xi(0) = 0, \quad \xi_t(0) = 0 & \text{in } \Omega, \end{cases}$$

and by ζ_ρ^n the solution to

$$\begin{cases} \zeta_{tt} - \Delta\zeta + a\zeta = 0 & \text{in } Q, \\ \partial_\nu \zeta = w_\rho^n & \text{on } \Sigma, \\ \zeta(0) = 0, \quad \zeta_t(0) = 0 & \text{in } \Omega, \end{cases}$$

where $a(x,t) := \Phi_y'(x,t,\bar{y}(x,t))$. One clearly has

$$\begin{gathered} (\xi_\rho^{n,2}-\zeta_\rho^n)_{tt} - \Delta(\xi_\rho^{n,2}-\zeta_\rho^n) + a_\rho^n(\xi_\rho^{n,2}-\zeta_\rho^n) = (a-a_\rho^n)\zeta_\rho^n \quad \text{in } Q, \\ \partial_\nu(\xi_\rho^{n,2}-\zeta_\rho^n) = 0 \quad \text{on } \Sigma, \\ (\xi_\rho^{n,2}-\zeta_\rho^n)(0) = 0, \qquad (\xi_\rho^{n,2}-\zeta_\rho^n)_t(0) = 0 \quad \text{in } \Omega. \end{gathered}$$

By Lemma 3.2, we find a constant $C > 0$, independent of n and ρ, ensuring the following estimates for all $n = 1, 2, \ldots$ and $0 < \rho < 1$:

$$\begin{aligned} \|\xi_\rho^{n,2}-\zeta_\rho^n\|_{C([0,T];L^2(\Omega))} &\le C\|a-a_\rho^n\|_{L^1(0,T;L^{2N}(\Omega))}\|\zeta_\rho^n\|_{L^\infty(0,T;L^{2N/(N-1)}(\Omega))} \\ &\le C\|a-a_\rho^n\|_{L^1(0,T;L^{2N}(\Omega))}\|\zeta_\rho^n\|_{L^\infty(0,T;H^{1/2}(\Omega))}, \end{aligned} \tag{4.10}$$

$$\|\xi_\rho^{n,1}\|_{C([0,T];L^2(\Omega))} \le C\|f_\rho^n\|_{L^1(0,T;L^2(\Omega))}, \tag{4.11}$$

where $\|\zeta_\rho^n\|_{L^\infty(0,T;L^{2N/(N-1)}(\Omega))}$ are uniformly bounded due to Lemma 3.1. Taking (4.9) into account, we conclude that for all $0 < \rho < 1$ the sequence of w_ρ^n converges to zero in the weak topology of $L^2(\Sigma)$ and, by Lemma 3.3, the sequence of ζ_ρ^n converges to zero in $C([0,T];L^2(\Omega))$. Thus there is an integer $n(\rho)$ such that

$$\|\zeta_\rho^{n(\rho)}\|_{C([0,T];L^2(\Omega))} \le \rho \quad \text{for all } 0 < \rho < 1. \tag{4.12}$$

Observe further that $u_\rho^{n(\rho)}$ converge to $\bar{u}$ in $L^2(\Sigma)$ as $\rho \downarrow 0$. It follows now from Theorem 3.5 that $y_\rho^{n(\rho)}$ converge to $\bar{y}$ in $C([0,T];L^2(\Omega))$ as $\rho \downarrow 0$. Invoking assumption (A1), one has that $f_\rho^{n(\rho)}$ converge to zero in $L^1(0,T;L^2(\Omega))$ and that $(a-a_\rho^{n(\rho)})$ converge to zero in $L^1(0,T;L^{2N}(\Omega))$ as $\rho \downarrow 0$. This, together with (4.10)–(4.12), implies that

$$\begin{aligned} \lim_{\rho\to 0}\|\xi_\rho^{n(\rho)}\|_{C([0,T];L^2(\Omega))} \le \lim_{\rho\to 0}(&\|\xi_\rho^{n(\rho),1}\|_{C([0,T];L^2(\Omega))} + \|\xi_\rho^{n(\rho),2} \\ &-\zeta_\rho^{n(\rho)}\|_{C([0,T];L^2(\Omega))} + \|\zeta_\rho^{n(\rho)}\|_{C([0,T];L^2(\Omega))}) = 0. \end{aligned}$$

Finally, setting $E_\rho := E_\rho^{n(\rho)}$, $u_\rho := u_\rho^{n(\rho)}$, and $\frac{1}{\rho}r_\rho := \xi_\rho^{n(\rho)}$, we end the proof of the theorem. ∎

5. Proof of necessary optimality conditions. As mentioned, in the proof of our main theorem we are going to use *Ekeland's variational principle* [6], which is one of the most powerful tools of nonlinear analysis and is especially important in applications of variational methods. In the framework of deriving necessary optimality conditions for the state-constrained problem (P), Ekeland's variational principle allows us to perform an efficient *strong approximation* of the given optimal solution to the original problem by some functions that happen to be optimal solutions to *perturbed* optimal control problems with *no state constraints*. To accomplish this procedure, we first describe a complete metric space and a lower semicontinuous functional, which are suitable for the application of Ekeland's principle to our problem.

Given $\bar{u} \in U_{ad}$ and a fixed positive number k, we define the set

$$U_{ad}(\bar{u}, k) := \{u \in U_{ad} \mid |u(s,t) - \bar{u}(s,t)| \le k \text{ for a.e. } (s,t) \in \Sigma\}$$

and endow this set with the metric, which goes back to Ekeland's seminal paper [6],

$$d(v,u) := \mathcal{L}^N(\{(s,t) \mid v(s,t) \ne u(s,t)\}),$$

where $\mathcal{L}^N(\Omega)$ denotes as before the N-dimensional Lebesgue measure of $\Omega \subset \mathbb{R}^N$. Observe that if $\{u_n\} \subset U_{ad}(\bar{u}, k)$ and $u \in U_{ad}(\bar{u}, k)$ are such that $\lim_{n\to\infty} d(u_n, u) = 0$, then the sequence $\{u_n\}$ *strongly* converges to u in the norm of $L^2(\Sigma)$. The next result provides more information about this space and about the cost functional of (P) on it, where y_u stands for the weak solution of (1.1) corresponding to u.

LEMMA 5.1 (proper setting for Ekeland's principle). *The metric space* $(U_{ad}(\bar{u}, k), d)$ *is complete, and the mapping* $u \mapsto (y_u, J(y_u, u))$ *is continuous from* $(U_{ad}(\bar{u}, k), d)$ *into* $C([0,T]; L^2(\Omega)) \times \mathbb{R}$.

Proof. The completeness of the space $(U_{ad}(\bar{u}, k), d)$ is a well-known fact; cf. [6, 25]. Let us prove the continuity statement of the lemma based on the regularity of weak solutions to the state system (1.1) established in section 3.

Take $\{u_n\} \subset U_{ad}(\bar{u}, k)$ and $u \in U_{ad}(\bar{u}, k)$ such that the control sequence $\{u_n\}$ converges to u in the above d-metric as $n \to \infty$. Denote by y and by y_n the weak solutions of (1.1) corresponding to u and to u_n, respectively. Since $u_n \to u$ strongly in $L^2(\Sigma)$, the corresponding trajectories y_n converge to y in $C([0,T]; L^2(\Omega))$ by Theorem 3.5. Furthermore, it follows from the estimates in assumptions (A2)–(A4) that the sequence of values $J(y_n, u_n)$ converges to $J(y,u)$ as $n \to \infty$, which ensures the desired continuity. □

Now using the classical results in the geometry of Banach spaces presented, e.g., in [17, Chapter 2] (see Theorem 2.18 and Proposition 2.20 therein), we conclude by the separability of $C([0,T]; L^2(\Omega))$ that there is an equivalent norm $|\cdot|_{C([0,T];L^2(\Omega))}$ on this space such that it is Gâteaux differentiable at any nonzero point and its dual norm on $\mathcal{M}([0,T]; L^2(\Omega))$—denoted by $|\cdot|_{\mathcal{M}([0,T];L^2(\Omega))}$—is *strictly convex*. Given the constraint set $\mathcal{C} \subset C([0,T]; L^2(\Omega))$ in the original problem (P), we define the *distance function*

$$d_{\mathcal{C}}(x) := \inf_{z \in \mathcal{C}} |x - z|_{C([0,T];L^2(\Omega))} \tag{5.1}$$

via the new norm $|\cdot|_{C([0,T];L^2(\Omega))}$ on $C([0,T]; L^2(\Omega))$. Since $\mathcal{C}$ is convex, the distance function (5.1) is also convex, and it is Lipschitz continuous on $C([0,T]; L^2(\Omega))$ with rank 1. As is well known,

$$|\xi|_{\mathcal{M}([0,T];L^2(\Omega))} \le 1 \quad \text{whenever } \xi \in \partial d_{\mathcal{C}}(x) \text{ with } x \in \mathcal{C};$$

moreover, one has

$$|\xi|_{\mathcal{M}([0,T];L^2(\Omega))} = 1 \quad \text{for every } \xi \in \partial d_{\mathcal{C}}(x) \text{ and } x \notin \mathcal{C},$$

where $\partial d_{\mathcal{C}}$ stands for the *subdifferential* of convex analysis. Taking into account that the dual norm $|\cdot|_{\mathcal{M}([0,T];L^2(\Omega))}$ is strictly convex on $\mathcal{M}([0,T];L^2(\Omega))$, we conclude that the subdifferential $\partial d_{\mathcal{C}}(x)$ is a *singleton*, and hence $d_{\mathcal{C}}$ is *Gâteaux differentiable* at x for every $x \notin \mathcal{C}$.

Let $(\bar{y},\bar{u})$ be an optimal solution to the original problem (P). Using the distance function (5.1), we define the *penalized functional* by

$$J_k(y,u) := \left[\left(J(y,u) - J(\bar{y},\bar{u}) + \frac{1}{k^2}\right)^+\right]^2 + d^2_{\mathcal{C}}(y), \quad k = 1,2,\dots,$$

where J is the cost functional in (P). Since $J_k(\bar{y},\bar{u}) = k^{-4}$, one has that

$$J_k(\bar{y},\bar{u}) < \inf\{J_k(y,u) \mid u \in U_{ad}(\bar{u},k^{1/3}),\ (y,u) \text{ satisfies (1.1)}\} + \frac{1}{k^2},$$

for all k, i.e., $(\bar{y},\bar{u})$ is a $\frac{1}{k^2}$-*optimal solution* to the penalized problem.

Notice that the functional J_k is *smooth* at points where it *does not vanish*, in the sense that it is Gâteaux differentiable at such points; cf. [20] in the case of control systems governed by ordinary differential equations. This follows from the construction of J_k, assumptions (A2)–(A4), and the above property of (5.1). Ekeland's principle allows us to *strongly* approximate $(\bar{y},\bar{u})$ by a pair (y_k,u_k) satisfying (1.1) in such a way that (y_k,u_k) is an *exact solution* to some *perturbed* optimal control problem for system (1.1) with the same control constraints and *no state constraints*.

Proof of Theorem 2.1. We divide the proof of this theorem into the three major steps.

Step 1. *Approximating problems via Ekeland's principle.* Given an optimal solution $(\bar{y},\bar{u})$ to the original problem (P), we fix a natural number $k = 1,2,\dots$ and get from Lemma 5.1 that the metric space $(U_{ad}(\bar{u},k^{1/3}),d)$ is complete, and that the functional $u \longmapsto J_k(y_u,u)$ is lower semicontinuous (even continuous) on this space. By the Ekeland variational principle [6] we find an admissible control u_k satisfying

$$\begin{aligned} &u_k \in U_{ad}(\bar{u},k^{1/3}), \qquad d(u_k,\bar{u}) \le \frac{1}{k}, \quad \text{and} \\ &J_k(y_k,u_k) \le J_k(y_u,u) + \frac{1}{k}d(u_k,u) \quad \text{for all } u \in U_{ad}(\bar{u},k^{1/3}), \end{aligned} \tag{5.2}$$

where y_k and y_u are the weak solutions of (1.1) corresponding to u_k and u, respectively. The latter means that, for all natural numbers k, u_k is an *optimal solution* to the *perturbed problem*

$$\inf\left\{J_k(y,u) + \frac{1}{k} \,\middle|\, u \in U_{ad}(\bar{u},k^{1/3}),\ (y,u) \text{ satisfies (1.1)}\right\}. \tag{P_k}$$

Step 2. *Necessary conditions in approximating problems.* First take an arbitrary $u_0 \in U_{ad}$ and construct the following modification of $\bar{u}$ feasible to (P_k) by

$$u_{0k}(s,t) := \begin{cases} u_0(s,t) & \text{if } |u_0(s,t) - \bar{u}(s,t)| \le k^{1/3}, \\ \bar{u}(s,t) & \text{otherwise.} \end{cases} \tag{5.3}$$

Then, given any $0 \le \rho < 1$, we define *diffuse perturbations* of the optimal control u_k in (P_k) as

$$u_\rho^k(s,t) := \begin{cases} u_k(s,t) & \text{on } \Sigma \setminus E_\rho^k, \\ u_{0k}(s,t) & \text{on } E_\rho^k. \end{cases} \tag{5.4}$$

Theorem 4.1 ensures the existence of measurable sets $E_\rho^k \subset \Sigma$ for which $\mathcal{L}^N(E_\rho^k) = \rho\mathcal{L}^N(\Sigma)$,

$$y_\rho^k = y_k + \rho z_k + \rho r_\rho^k, \qquad \lim_{\rho\to 0} \|r_\rho^k\|_{C([0,T];L^2(\Omega))} = 0, \quad \text{and} \tag{5.5}$$

$$J(y_\rho^k, u_\rho^k) = J(y_k, u_k) + \rho\Delta J^k + o(\rho), \tag{5.6}$$

where y_ρ^k is the weak solution of (1.1) corresponding to u_ρ^k, where z_k is the weak solution to

$$\begin{cases} z_{tt} - \Delta z + \Phi_y'(\cdot, y_k)z = 0 & \text{in } Q, \\ \partial_\nu z = u_k - u_{0k} & \text{on } \Sigma, \\ z(0) = 0, \quad z_t(0) = 0 & \text{in } \Omega, \end{cases}$$

and where ΔJ^k is defined by

$$\Delta J^k := \int_Q g_y'(\cdot, y_k)z_k \, dxdt + \int_\Omega f_y'(\cdot, y_k(T))z_k \, dx + \int_\Sigma (h(\cdot, u_{0k}) - h(\cdot, u_k)) \, dsdt.$$

Since each u_ρ^k is clearly feasible for (P_k), from (5.2) and the construction of the metric d, we deduce that

$$\lim_{\rho\to 0} \frac{J_k(y_k, u_k) - J_k(y_\rho^k, u_\rho^k)}{\rho} \le \frac{1}{k}\mathcal{L}^N(\Sigma). \tag{5.7}$$

Observe that $J_k(y_k, u_k) \ne 0$ for all k due to the optimality of u_k in (P_k) and the structure of J_k. Hence J_k is *Gâteaux differentiable* at (y_k, u_k) by the discussion above. Then it easily follows from (5.6) and (5.7) that

$$-\lambda_k\Delta J^k - \langle \mu_k, z_k \rangle \le \frac{1}{k}\mathcal{L}^N(\Sigma), \tag{5.8}$$

where the multipliers λ_k and μ_k are computed by

$$\lambda_k := \frac{(J(y_k, u_k) - J(\bar{y}, \bar{u}) + \frac{1}{k^2})^+}{J_k(y_k, u_k)}, \qquad \mu_k := \begin{cases} \dfrac{d_{\mathcal{C}}(y_k)\nabla d_{\mathcal{C}}(y_k)}{J_k(y_k, u_k)} & \text{if } y_k \notin \mathcal{C}, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that $\mu_k \in \mathcal{M}([0,T]; L^2(\Omega))$. Now let p_k be the (unique) weak solution to the *adjoint system*

$$\begin{cases} p_{tt} - \Delta p + \Phi_y'(\cdot, y_k)p = \lambda_k g_y'(\cdot, y_k) + \mu_k|_Q & \text{in } Q, \\ \partial_\nu p = 0 & \text{on } \Sigma, \\ p(T) = 0, \quad p_t(T) = -\lambda_k f_y'(\cdot, y_k(T)) - \mu_k|_{\Omega\times\{T\}} & \text{in } \Omega, \end{cases} \tag{5.9}$$

where $\mu_k|_Q$ and $\mu_k|_{\Omega\times\{T\}}$ are the restrictions of μ_k to Q and $\Omega\times\{T\}$, respectively. Employing the Green formula in Theorem 3.9, we have

$$\begin{aligned}\lambda_k\int_Q g'_y(x,t,y_k)z_k\,dxdt+\lambda_k\int_\Omega f'_y(x,y_k(T))z_k(T)\,dx+\langle\mu_k,z_k\rangle\\ =\int_Q p_k(z_{ktt}-\Delta z_k+\Phi'_y(\cdot,y_k)z_k)\,dxdt+\int_\Sigma p_k\partial_\nu z_k\,dsdt\\ =\int_\Sigma p_k(u_k-u_{0k})\,dsdt.\end{aligned}$$

The latter implies, by (5.8) and the definition of ΔJ^k, that

$$\int_\Sigma(\lambda_k h(s,t,u_k)+p_k u_k)\,dsdt\le\int_\Sigma(\lambda_k h(s,t,u_{0k})+p_k u_{0k})\,dsdt+\frac{1}{k}\mathcal{L}^N(\Sigma) \tag{5.10}$$

for every $k=1,2,\dots$, which gives necessary optimality conditions for the solutions u_k to the approximating problems (P_k).

Step 3. *Passing to the limit.* To conclude the proof of the theorem, we need to pass to the limit in the above relations for the optimal solutions u_k to (P_k) as $k\to\infty$. First observe that

$$\lambda_k^2+|\mu_k|^2_{\mathcal{M}([0,T];L^2(\Omega))}=1\quad\text{for all } k=1,2,\dots.$$

Invoking basic functional analysis, we find an element $(\lambda,\bar\mu)\in\mathbb{R}\times\mathcal{M}([0,T];L^2(\Omega))$, with $\lambda\ge0$, and a subsequence of (λ_k,μ_k), still indexed by k, such that

$$\lambda_k\to\lambda\quad\text{in }\mathbb{R}\quad\text{and}\quad\mu_k\rightharpoonup\bar\mu\quad\text{weak}^*\text{ in }\mathcal{M}([0,T];L^2(\Omega)).$$

Furthermore, Theorem 3.8 ensures the estimate

$$\begin{aligned}&\|p_k\|_{L^\infty(0,T;H^1(\Omega))}+\|p_{kt}\|_{L^\infty(0,T;L^2(\Omega))}\\ &\quad\le C(\|\mu\|_{\mathcal{M}([0,T];L^2(\Omega))}+\|g'_y(\cdot,y_k)\|_{L^1(0,T;L^2(\Omega))}+\|f'_y(\cdot,y_k(T))\|_{L^2(\Omega)}).\end{aligned}$$

Since the sequences $\{\lambda_k\}\subset\mathbb{R}$, $\{\mu_k\}\subset\mathcal{M}([0,T];L^2(\Omega))$, $\{y_k\}\subset C([0,T];L^2(\Omega))$, and $\{u_k\}\subset L^2(\Sigma)$ are bounded, the sequence $\{(p_k,p_{kt})\}$ is bounded in $L^\infty(0,T;H^1(\Omega))\times L^\infty(0,T;L^2(\Omega))$. Then there is a subsequence of $\{(p_k,p_{kt})\}$ converging to some (p,p_t) in the weak* topology of $L^\infty(0,T;H^1(\Omega))\times L^\infty(0,T;L^2(\Omega))$ and a subsequence of $\{y_k\}$ converging to some $\bar y\in L^\infty(0,T;L^2(\Omega))$ in the weak* topology of $L^\infty(0,T;L^2(\Omega))$. We already know that $\{u_k\}$ tends to $\bar u$ in $L^2(\Sigma)$. Employing standard arguments, we prove that $(\bar y,\bar y_t)$ is the solution of (1.1) corresponding to $\bar u$, and that (p,p_t) is the (unique) weak solution of (2.4) corresponding to $\bar y$.

We choose $(\lambda,\mu)=(\lambda,\bar\mu|_{]0,T]})$ as the multipliers for the necessary optimality conditions stated in Theorem 2.1. Taking into account assumption (A5) on the convexity and nonemptiness interiority of the set $\mathcal{C}$, one has the necessary condition (2.2) for the limiting multipliers (λ,μ). In particular, let us verify that $(\lambda,\mu)\ne0$. Suppose the contrary, which gives

$$\lim_{k\to\infty}|\mu_k|^2_{\mathcal{M}([0,T];L^2(\Omega))}=1. \tag{5.11}$$

By assumption (A5) we have $\hat y_0\in\operatorname{int}\mathcal{C}$. Thus there exists a ball $B(\hat y_0,\rho)$ in $C([0,T];L^2(\Omega))$ centered at $\hat y_0$ with radius $\rho>0$ such that $B(\hat y_0,\rho)\subset\mathcal{C}$. Using (5.11) and taking any $k=1,2,\dots$, we find $z_k\in B(0,\rho)$ satisfying

$$\langle z_k,\mu_k\rangle_{C([0,T];L^2(\Omega)),\mathcal{M}([0,T];L^2(\Omega))}=\frac{\rho}{2}|\mu_k|_{\mathcal{M}([0,T];L^2(\Omega))}.$$

Since $\hat{y}_0 + z_k \in \mathcal{C}$, one has by definition of μ_k that

$$\langle \hat{y}_0 + z_k - y_k, \mu_k \rangle_{C([0,T];L^2(\Omega)),\mathcal{M}([0,T];L^2(\Omega))} \leq 0 \quad \text{for all } k = 1, 2, \dots .$$

Passing to the limit as $k \to \infty$, we get

$$\frac{\rho}{2} + \langle \hat{y}_0 - \bar{y}, \bar{\mu} \rangle_{C([0,T];L^2(\Omega)),\mathcal{M}([0,T];L^2(\Omega))} \leq 0.$$

Remember that $\bar{y}(x,0) = \hat{y}_0(x,0)$ and $\mu = \bar{\mu}|_{]0,T]}$; therefore

$$\langle \hat{y}_0 - \bar{y}, \bar{\mu} \rangle_{C([0,T];L^2(\Omega)),\mathcal{M}([0,T];L^2(\Omega))} = \langle \hat{y}_0 - \bar{y}, \mu \rangle_{C([0,T];L^2(\Omega)),\mathcal{M}_b(]0,T];L^2(\Omega))},$$

which implies that

$$\langle \hat{y}_0 - \bar{y}, \mu \rangle_{C([0,T];L^2(\Omega)),\mathcal{M}_b(]0,T];L^2(\Omega))} \leq -\frac{\rho}{2} < 0.$$

The latter contradicts the assumption on $(\lambda, \mu) = 0$ and thus justifies the nontriviality condition in the theorem.

It remains to verify the minimum condition (2.3). To do this, recall that $u_k \to \bar{u}$ strongly in $L^2(\Sigma)$. Passing to the limit as $k \to \infty$ in (5.10), we get

$$\int_\Sigma (\lambda h(s,t,\bar{u}) + p\bar{u})\, dsdt \leq \int_\Sigma (\lambda h(s,t,u_0) + pu_0)\, dsdt \quad \text{for all } u_0 \in U_{ad}. \tag{5.12}$$

Finally, taking into account the structure of U_{ad} in (A6) and employing the standard arguments (see, e.g., [25, section 5.2]), we derive the pointwise condition (2.3) from the integral one in (5.12). □

Acknowledgment. The authors are indebted to anonymous referees for their valuable suggestions and remarks that allowed them to improve the original presentation.

REFERENCES

[1] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, 2nd ed., D. Reidel, Dordrecht, The Netherlands, 1986.

[2] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.

[3] E. Casas, J.-P. Raymond, and H. Zidani, *Pontryagin's principle for local solutions of control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1182–1203.

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

[5] F. H. Clarke, *The maximum principle under minimal hypotheses*, SIAM J. Control Optim., 14 (1976), pp. 1078–1091.

[6] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[7] H. O. Fattorini, *Optimal control problems with state constraints for semilinear distributed-parameter systems*, J. Optim. Theory Appl., 88 (1996), pp. 25–59.

[8] H. O. Fattorini, *Nonlinear infinite-dimensional optimal control problems with state constraints and unbounded control sets*, Rend. Istit. Mat. Univ. Trieste, 28 (1996), suppl., pp. 127–146 (1997).

[9] H. O. Fattorini, *Infinite-Dimensional Optimization and Control Theory*, Cambridge University Press, Cambridge, UK, 1999.

[10] R. Gabasov and F. M. Kirillova, *Qualitative Theory of Optimal Processes*, Marcel Dekker, New York, 1976.

[11] I. Lasiecka, J.-L. Lions, and R. Triggiani, *Nonhomogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl. (9), 65 (1986), pp. 149–192.
[12] I. Lasiecka and R. Triggiani, *Dirichlet boundary control problem for parabolic equations with quadratic cost: Analyticity and Riccati's feedback synthesis*, SIAM J. Control Optim., 21 (1983), pp. 41–67.
[13] I. Lasiecka and R. Triggiani, *Sharp regularity theory for second order hyperbolic equations of Neumann type. Part* I*: L_2 nonhomogeneous data*, Ann. Mat. Pura Appl. (4), 157 (1990), pp. 285–367.
[14] I. Lasiecka and R. Triggiani, *Regularity theory of hyperbolic equations with nonhomogeneous Neumann boundary conditions.* II*. General boundary data*, J. Differential Equations, 94 (1991), pp. 112–164.
[15] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 2000.
[16] X. J. Li and Y. Yao, *Maximum principle of distributed parameter systems with time lags*, in Proceedings of the Conference on Control Theory of Distributed Parameter Systems and Applications, Lecture Notes in Control and Inform. Sci. 75, F. Kappel and K. Kunish, eds., Springer-Verlag, Berlin, 1985, pp. 410–427.
[17] X. J. Li and J. Yong, *Optimal Control Theory for Infinite-Dimensional Systems*, Systems Control Found. Appl., Birkhäuser Boston, Boston, 1995.
[18] J.-L. Lions, *Contrôle des Systèmes Distribués Singuliers*, Gauthier-Villars, Paris, 1983.
[19] J.-J. Moreau, *Bounded variation in time*, in Topics in Nonsmooth Mechanics, J. J. Moreau, P. D. Panagiotopoulos, and G. Strang, eds., Birkhäuser, Basel, 1988, pp. 1–74.
[20] B. S. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[21] B. S. Mordukhovich and J.-P. Raymond, *Dirichlet boundary control of hyperbolic equations in the presence of state constraints*, Appl. Math. Optim., 49 (2004), pp. 145–157.
[22] B. S. Mordukhovich and K. Zhang, *Minimax control of parabolic systems with Dirichlet boundary conditions and state constraints*, Appl. Math. Optim., 36 (1997), pp. 323–360.
[23] P. A. Nguyen and J.-P. Raymond, *Control problems for convection-diffusion equations with a control localized on manifolds*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 417–448.
[24] J.-P. Raymond, *Nonlinear boundary control of semilinear parabolic problems with pointwise state constraints*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 341–370.
[25] J. P. Raymond and H. Zidani, *Pontryagin's principle for state-constrained control problems governed by parabolic equations with unbounded controls*, SIAM J. Control Optim., 36 (1998), pp. 1853–1879.
[26] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
[27] R. Vinter, *Optimal Control*, Systems Control. Found. Appl., Birkhäuser Boston, Boston, 2000.
[28] L. W. White, *Control of a hyperbolic problem with pointwise stress constraints*, J. Optim. Theory Appl., 41 (1983), pp. 359–369.
[29] L. W. White, *Distributed control of a hyperbolic problem with control and stress constraints*, J. Math. Anal. Appl., 106 (1985), pp. 41–53.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1373–1399

OPTIMAL CONTROL OF UNCERTAIN SYSTEMS WITH INCOMPLETE INFORMATION FOR THE DISTURBANCES*

MARC QUINCAMPOIX† AND VLADIMIR M. VELIOV‡

Abstract. We investigate the problem of optimization of a terminal cost function for a system depending on a control, and on two disturbances for which a priori set membership is known. The disturbances are of different natures: One becomes known to the controller at the current time (we called it observable) while the other remains unknown. No state measurements are available. The problem can be viewed as a differential game of min-max type where the controller aims at minimization of the objective function by a strategy which depends only on the observable disturbance. Since the state of the system is not exactly known due to the presence of an unobservable disturbance, we reformulate the problem through a set-valued dynamics describing the evolution of the current set estimation of the state. To reduce the complexity of the problem, we pass to a suboptimal problem where the evolution of the state estimation is restricted to a prescribed collection of sets. The main result of the paper is a characterization of the value function of this problem through a Hamilton–Jacobi inequality in terms of Dini derivatives, which implies a convergent scheme for numerical computations. As necessary auxiliary tools, we provide new results on evolution and viability of tubes in a given collection of sets that may be of independent interest.

Key words. uncertain systems, differential games, optimal control, viability theory, reachable sets

AMS subject classifications. 90D25, 49J24, 49K35, 28B20

DOI. 10.1137/S0363012903420863

1. Introduction.

We consider the system

$$\dot{x} = f(x, u, y, v), \qquad x(0) = e \in E_0, \tag{1}$$

where $x \in \mathbf{R}^n$ is the state, $u \in U$ is the control, and $y \in Y$ and $v \in V(y)$ are disturbances (U, Y, and $V(y)$ are given subsets of finite dimensional spaces $E_0 \subset \mathbf{R}^n$). The main concern of the paper is the optimal control problem where the controller wants to minimize (by choosing u) the cost

$$g(T, x(T)) \tag{2}$$

against the worst case of disturbances y and v and initial state $e \in E_0$. We distinguish two types of disturbance:

- *observable uncertainty* y, for which the current realization $y(t) \in Y$ becomes known to the controller;
- *unobservable uncertainty* $v \in V(y)$, for which the realization of $v(t) \in V(y(t))$ remains unknown.

*Received by the editors January 9, 2003; accepted for publication (in revised form) January 2, 2004; published electronically January 5, 2005. This research was partially supported by the Austrian Science Foundation under contract 14060-OEK and by the European Community's Human Potential Programme under contract HPRN-CT-2002-00281 (Evolution Equations).

http://www.siam.org/journals/sicon/43-4/42086.html

†Laboratoire de Mathématiques, Unité CNRS FRE 2218, Université de Bretagne Occidentale, 6 Avenue Victor Le Gorgeu, BP 809, 29285 Brest, France (Marc.Quincampoix@univ-brest.fr).

‡Institute for Econometrics, Operations Research and Systems Theory, Vienna University of Technology, Argentinierstrasse 8, A-1040 Vienna, Austria (vveliov@eos.tuwien.ac.at), and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria.

Thus we consider a min-max problem or a differential game where the second player wants to maximize (by choosing e, y, and v) the cost (2) while the first player—the controller—wants to minimize it (by choosing u). The information available to the controller implies that the control u should be considered in a feedback form which may depend on the current and the past values of y, but not on v.

For every given open-loop control $u(\cdot)$ and observable uncertainty $y(\cdot)$, the unobservable uncertainty $v(\cdot)$ gives rise to a differential inclusion

$$\dot{x} \in f(x, u(t), y(t), V(y(t))), \qquad x(0) \in E_0, \tag{3}$$

whose solution is a time-dependent tube providing the deterministic estimation of the trajectory. Notice that a (set-valued) tube starting from E_0 is involved even in the case of precisely known initial state $x(0)$, due to the unobservable uncertainty.

Let us formulate the problem in a more precise way. Let $\mathcal{U}_{[t,\theta]}$ be the set of all open-loop admissible controls on the interval $[t, \theta]$, that is, the measurable functions with values in U. Similarly, $\mathcal{Y}_{[t,\theta]}$ denotes the set of all measurable selections of Y on $[t, \theta]$, and $\mathcal{V}_{[t,\theta]}(y(\cdot))$ denotes the set of all measurable selections of the mapping $V(y(\cdot))$ (for a given $y(\cdot)$) on the same interval. The suppositions formulated in section 4 will imply that for any $t \in [0, T)$, $e \in \mathbf{R}^n$, $u \in \mathcal{U}_{[t,T]}$, $y \in \mathcal{Y}_{[t,T]}$, and $v \in \mathcal{V}_{[t,T]}(y)$ system (1) has a unique solution on $[t, T]$ starting from e, denoted by $x[t, e; u, y, v](\cdot)$.

To make use of the dynamic programming principle, we consider problem (1), (2) also for an arbitrary initial time t and initial set E, instead of the fixed $t = 0$ and $E = E_0$. The optimal control for initial time t and initial compact set E is sought as a *nonanticipative strategy* (called also a Varaiya–Roxin–Elliot–Kalton strategy [17]; cf. the definition in section 4), $\alpha : \mathcal{Y}_{[t,T]} \mapsto \mathcal{U}_{[t,T]}$. The guaranteed result obtained by using the strategy α for initial data (t, E) is

$$I(t, E; \alpha) \stackrel{def}{=} \sup\{g(T, x[t, e; \alpha(y), y, v](T));\ e \in E,\ y \in \mathcal{Y}_{[t,T]},\ v \in \mathcal{V}_{[t,T]}(y)\}.$$

Then

$$I(t, E) \stackrel{def}{=} \inf_{\alpha} I(t, E; \alpha) \tag{4}$$

is the minimal guaranteed (lower) value that can be achieved starting from the set E at time t.

As previous works indicate (see [7, 8, 6]), the optimal control problem in the case of incomplete/inexact information is qualitatively different compared with the perfect information case, since the corresponding Hamilton–Jacobi–Isaacs (HJI) equation for the value function becomes, essentially, infinite dimensional. It is shown in [8] that the problem with incomplete measurement can be reformulated as a problem with complete information in the *information space*, which, however, is infinite dimensional.[1] In contrast, in the present paper we reformulate the problem as such with complete information, but for a dynamic system in the *estimation space*, which, in principle, is also infinite dimensional, but in some cases can be equivalently replaced by a finite dimensional one. In this case the corresponding HJI equation is finite dimensional. (A finite dimensional HJI equation was derived in [30] for a specific game on the plane where the state information is incomplete by employing the certainty equivalence principle [7, 9].) If this is not the case, one can still formulate a "suboptimal"

[1] We stress the fact that the *state* is partially measured in these papers, while we suppose that the *disturbance* in the equation is partially measured, while state measurements are not available.

version of the problem by restricting the consideration to an estimation space, which is a finitely parametrized collection of sets. To develop the technique for passing to such a "suboptimal" problem and its analysis is one of our main goals.

To pass to a finitely parametrized estimation we recall in section 2 the concept of solution tubes in a given collection of compact sets developed in [29]. We restrict the presentation to more special collections than in [29] which are more convenient for the present paper, but we add some necessary new facts. In particular, we prove in section 2.2 the continuous dependence of the solution tubes on data, which is essential for the subsequent analysis. For $y(\cdot)$ fixed, (3) can be considered as a controlled differential inclusion whose solution tube takes values in the given collection. Therefore the next step is to develop the viability theory for collections of sets and controlled differential inclusions. In section 3 we extend some of our results from [28] to the case of controlled inclusions and to the concept of *viability with a target* introduced in [27], which is essential for the control problem considered here. To maintain the flow of the paper some technical proofs are postponed until the last section.

With the appropriate viability theory at hand, we cope with problem (2), (1) in section 4. We can consider the end time T as fixed but, in fact, in section 4 the end time is determined by a terminal condition, which, if certainly reached (that is, reached for all possible realizations of the unobservable uncertainty), then the control process terminates and the performance index is evaluated. This formulation includes minimal time problems.

Our main result extends that of [10], the latter concerning the "classical" case where unobservable uncertainty is not present and the initial state is precisely known. We prove that the value function of the problem is the unique minimal Dini supersolution of the respective HJI equation. The epigraph of the value function is characterized also as the maximal set that for every $y \in Y$ is a viability domain (in a specified collection of sets) for an auxiliary controlled inclusion depending on y. This makes it possible to apply an appropriate modification of the viability kernel algorithm (see [31, 11]) for numerical calculation. The applicability of such a modification is, in principle, shown in [25] (where, however, a discrete-time problem in a somewhat different setting is considered) including numerical schemes and examples. In the last part of section 4 we also discuss the suppositions and possible extensions.

We mention that the material in sections 2 and 3 is aimed at solving the control problem formulated above, but the results presented there could also be useful for other control/estimation problems for continuous-time uncertain systems.

2. Solution tubes in a collection of sets. The reachable set of a differential inclusion (the latter interpreted as an uncertain system, as in (3)) is the minimal guaranteed estimation of the current state. Therefore, to calculate reachable sets is a cornerstone of the deterministic estimation and control of uncertain systems (see, e.g., [21]) and a lot of work has been done toward developing numerical approximation methods (see, e.g., the surveys [15, 24]). Since the geometry of the reachable sets could be rather complicated, specific subclasses of sets are usually used as approximation tools: boxes, polyhedral sets, ellipsoids (see [12, 22, 13, 23]), box or polyhedral complexes (see [31, 19, 11, 20]), etc. In some cases convergence results are obtained, but usually, to achieve a good approximation, one has to use rather complex approximating sets. On the other hand, in problems of control of uncertain systems and differential games, where the state estimation is just an auxiliary tool, one has to employ only fairly simple sets. (The associated HJI equation, for example, has the dimension of the state estimators, and therefore the latter should not be too large.)

In such cases the issue of *approximation* is not that relevant. A different problem arises: to obtain inclusions of the reachable set in sets from a prescribed collection $\mathcal{E}$, that is, to replace the solution tube $X(t)$ of the differential inclusion by a tube $E(t)$ with values in $\mathcal{E}$. In doing this, one has to ensure at least $X(t) \subset E(t)$, but two more properties are also desirable: (i) the Markov property of the evolution of $E(\cdot)$, which, together with $X(t) \subset E(t)$, requires invariance of the tube $E(t)$ with respect to the differential inclusion, and (ii) minimality.

In this section we modify some results from our paper [29] (where more bibliographical data and examples are provided) and establish some new ones needed in the subsequent sections.

2.1. Definitions and main suppositions. We shall use the following notation: $\mathcal{B}$ is the Euclidean unit ball in $\mathbf{R}^n$, $\mathrm{comp}(\mathbf{R}^n)$ is the set of all nonempty compact subsets of $\mathbf{R}^n$, $\mathrm{dist}(X,Y) \stackrel{def}{=} \sup_{x\in X} \inf_{y\in Y} |x-y|$ is the distance from $X \in \mathrm{comp}(\mathbf{R}^n)$ to $Y \in \mathrm{comp}(\mathbf{R}^n)$, $|X| = \mathrm{dist}(X,\{0\})$, and $H(X,Y) = \max\{\mathrm{dist}(X,Y), \mathrm{dist}(Y,X)\}$ is the Hausdorff distance between X and Y. Multiplication of a set with a scalar and summation of sets are understood in the usual (Minkowski) sense. For a set X, $f(X)$ stays for $\{f(x);\ x \in X\}$. For a given closed $X \subset \mathbf{R}^n$, the set of all Hausdorff continuous mappings $X \mapsto \mathrm{comp}(\mathbf{R}^n)$ is a complete metric cone (with respect to the Minkowski operations), which will be denoted by $C(X; \mathrm{comp}(\mathbf{R}^n))$.

Definition. A set-valued map $E(\cdot) : [0,T] \Rightarrow \mathbf{R}^n$ is called a *tube*. Throughout this paper it would be convenient to use the term "tube" only for mappings $E(\cdot)$ that have nonempty compact values and a closed graph. A tube is *Lipschitz continuous* if there is a constant L such that

$$H(E(s), E(t)) \le L|t-s| \quad \text{for every } s,t \in [0,T].$$

We consider a differential inclusion

$$\dot{x} \in F(x,t), \quad x \in \mathbf{R}^n,\ t \in [0,T], \tag{5}$$

supposing the following.

Condition A. $F : \mathbf{R}^n \times [0,T] \Rightarrow \mathbf{R}^n$ is a set-valued mapping with nonempty convex, compact values, measurable in t for every fixed x, and locally Lipschitz continuous in x uniformly with respect to t. Moreover, F satisfies the linear growth condition

$$|F(x,t)| \le a(1+|x|) \quad \forall x \in \mathbf{R}^n,\ t \in [0,T].$$

As usual, a solution to (5) is any absolutely continuous function that satisfies (5) for a.e. t. Given a set of initial states $E_0 \subset \mathbf{R}^n$, the solution tube of (5) on $[0,T]$ is defined as

$$X(s) \stackrel{def}{=} X[0,E_0](s) \stackrel{def}{=} \{x(s);\ x(\cdot) - \text{solution of (5) on } [0,s] \quad \text{with } x(0) \in E_0\}.$$

This is the unique *minimal* tube that starts from E_0 at $t=0$ and is *invariant* with respect to (5), the latter meaning that

$$\forall s \in [0,T) \quad \forall\, x(\cdot) - \text{solution of (5) on } [s,T] \quad \text{with } x(s) \in X(s) \\ \Rightarrow x(t) \in X(t) \quad \forall t \in [s,T].$$

Here and below, "minimal," when applied to sets, means "minimal with respect to inclusion"; when applied to tubes, minimality is meant with respect to the partial ordering in which $E_1(\cdot) \prec E_2(\cdot)$ if and only if $E_1(t) \subset E_2(t)$ for every $t \in [0,T]$.

Definition. Let $\mathcal{E}$ be a given collection of compact sets in $\mathbf{R}^n$. The tube $E(\cdot) : [0,T] \mapsto \mathrm{comp}(\mathbf{R}^n)$ is called the *solution tube of* (5) *in the collection* $\mathcal{E}$ if and only if $E(\cdot)$ is a minimal invariant tube with values in $\mathcal{E}$.

The above definition meets the requirements (i) and (ii) formulated in the preamble of this section and extends the usual concept of a solution tube. Clearly, $X[0,E_0](\cdot)$ is the unique solution tube in the collection $\mathcal{E} = \mathrm{comp}(\mathbf{R}^n)$, starting from E_0. To ensure the existence of a solution tube in a more general collection $\mathcal{E}$ we introduce the following conditions for $\mathcal{E}$.

Condition B.1. The collection $\mathcal{E}$ consists of nonempty compact sets and is closed in the Hausdorff metric. For every compact Z there is some $E \in \mathcal{E}$ containing Z.

Condition B.2. There exists a constant $L_{\mathcal{E}}$ such that for each $\varepsilon > 0$ and each $E \in \mathcal{E}$ there exists $E' \in \mathcal{E}$ for which $E + \varepsilon\mathcal{B} \subset E' \subset E + \varepsilon L_{\mathcal{E}}\mathcal{B}$.

Obviously Conditions B.1 and B.2, together with the Zorn lemma, imply that for every $Z \in \mathrm{comp}(\mathbf{R}^n)$ there exists a minimal element of $\mathcal{E}$ containing Z.

Condition B.3. For every $Z \in \mathrm{comp}(\mathbf{R}^n)$ there is a unique minimal element of $\mathcal{E}$ containing Z.

The last condition is not necessary for many of the considerations below, including the existence, but it is convenient in the context of the optimal control problem investigated in the present paper.

2.2. Existence and continuity of the solution tubes.

THEOREM 2.1 (adapted from [29]). *Suppose that Conditions* A *and* B.1–B.3 *are fulfilled. Then for every* $E_0 \in \mathcal{E}$ *inclusion* (5) *has a unique solution tube in* $\mathcal{E}$ *starting from* E_0*, it satisfies the growth estimation*

$$E(t) \subset E_0 + (1 + |E_0|)(e^{2a(L_{\mathcal{E}}+1)t} - 1)\mathcal{B}, \tag{6}$$

and is Lipschitz continuous with Lipschitz constant $2a(1 + L_{\mathcal{E}})(1 + |E_0|)e^{2a(1+L_{\mathcal{E}})T}$.

COROLLARY 2.2. *For every* $E_0 \in \mathcal{E}$ *inclusion* (5) *has a unique solution tube in* $\mathcal{E}$ *on* $[0, +\infty)$*, starting from* E_0*. The solution tube is contained in every invariant tube of* (5) *starting from* E_0 *and taking values in* $\mathcal{E}$.

In the next sections we shall need the following continuity property of the solution tubes. Let us consider a sequence of differential inclusions of the form

$$\dot{x} \in F_0^k(x,s) + B^k(x,s)u_k(s), \tag{7}$$

where $F_0^k : \mathbf{R}^n \times [t_k, T] \Rightarrow \mathbf{R}^n$, $u_k : [t_k, T] \mapsto U$, U is a subset of a finite dimensional space, and B^k is a matrix function with appropriate size.

PROPOSITION 2.3. *Let the following conditions be fulfilled:*

(i) *For every* $k = 0, 1, \ldots$, *the mapping* $F^k(x,s) \stackrel{def}{=} F_0^k(x,s) + B^k(x,s)u_k(s)$ *satisfies Condition* A *with the growth constant* a *independent of* k; B^k *are measurable in* s*, locally Lipschitz continuous in* x*, uniformly in* s; U *is compact; the functions* $u_k(\cdot)$ *are measurable; and the collection* $\mathcal{E}$ *satisfies Condition* B.

(ii) $t_k \in [0,T]$ *and* $E_0^k \in \mathcal{E}$, $k = 0, 1, \ldots$, *are such that* $\lim_{k\to+\infty} t_k = t_0$, $\liminf_{k\to+\infty} \mathrm{dist}(E_0^0, E_0^k) = 0$ *and* $u_k(\cdot)$ *converges* L_1*-weakly to* $u_0(\cdot)$*; for every compact set* Z

$$\liminf_{k\to+\infty} \int_0^T \sup_{x\in Z} [\mathrm{dist}(F_0^0(x,s), F_0^k(x,s)) + |B^k(x,s) - B^0(x,s)|]\,\mathrm{d}s = 0.$$

Let $E_k(\cdot) : [t_k, T] \mapsto \mathcal{E}$ *be the solution tube in* $\mathcal{E}$ *of* (7), $k = 0, 1, \ldots$, *starting from* E_0^k *at time* t_k*. Then there is a subsequence of* $\{E_k(\cdot)\}$ *that converges uniformly to*

some tube $E(\cdot) : [t_0, T] \mapsto \mathcal{E}$. *Every such limit tube* $E(\cdot)$ *is Lipschitz continuous, and* $E_0(t) \subset E(t)$ *for all* $t \in [t_0, T]$.

Proof. To avoid the obvious extension/restriction technicalities needed to cope strictly with the case $t_k \neq t_0$, we assume that $t_k = t_0$ in the proof below. Since the growth constant a is the same for all k, from Theorem 2.1 it follows that the tubes $E_k(\cdot)$ are equi-Lipschitz, and therefore, also equibounded. According to the Arzelà–Ascoli theorem applied to the space $C([t, T]; \mathcal{E})$, there is a subsequence uniformly converging to a tube $E(\cdot)$. To avoid multiple indexes, we suppose that the whole sequence is convergent and that the two "liminf" in supposition (ii) of the proposition are "lim." We shall prove that $E(\cdot)$ is an invariant tube for (7) with $k = 0$, which implies the claim of the proposition according to Corollary 2.2.

To prove the invariance of $E(\cdot)$, we take an arbitrary $\tau \in [t_0, T)$ and an arbitrary trajectory $x(\cdot)$ of (7) with $k = 0$, starting from a point $x_0^0 \in E(\tau)$. Then we define the Carathéodory selection

$$\psi_k(x, s) = \mathcal{P}_{F_0^k(x,s)}(\dot{x}(s) - B^0(x(s), s)u_0(s))$$

of F_0^k (where $\mathcal{P}_Z(y)$ is the projection of y on the convex compact set Z) and consider the equation

$$\dot{x}_k(s) = \psi_k(x_k(s), s) + B^k(x_k(s), s)u_k(s), \qquad x_k(\tau) = x_k^0,$$

where x_k^0 is chosen from $E_k(\tau)$ in such a way that $|x_0^0 - x_k^0| \leq \operatorname{dist}(E(\tau), E_k(\tau)) \stackrel{def}{=} \rho_k$. Thus $\rho_k \to 0$. Moreover, obviously

$$\begin{aligned}
|\psi_k(x_k(s), s) - (\dot{x}(s) - B^0(x(s), s)u_0(s))| &\leq \operatorname{dist}(F_0^0(x(s), s), F_0^k(x_k(s), s)) \\
&\leq H(F_0^0(x(s), s), F_0^0(x_k(s), s)) + \operatorname{dist}(F_0^0(x_k(s), s), F_0^k(x_k(s), s)) \\
&= H(F_0^0(x(s), s), F_0^0(x_k(s), s)) + \gamma_k(s),
\end{aligned}$$

where $\gamma_k \stackrel{def}{=} \int_\tau^T \gamma_k(s)\,\mathrm{d}s \to 0$ due to (ii) and the uniform boundedness of $x_k(\cdot)$. For the same reason,

$$\delta_k \stackrel{def}{=} \int_\tau^T |B^k(x_k(s), s) - B^0(x_k(s), s)|\,\mathrm{d}s \to 0.$$

Then we have

$$\begin{aligned}
&|x_k(s) - x(s)| \\
&\quad \leq \rho_k + \gamma_k + \int_\tau^s H(F_0^0(x(\theta), \theta), F_0^0(x_k(\theta), \theta))\,\mathrm{d}\theta \\
&\qquad + \left| \int_\tau^s [B^k(x_k(\theta), \theta)u_k(\theta) - B^0(x(\theta), \theta)u_0(\theta)]\,\mathrm{d}\theta \right| \\
&\quad \leq \rho_k + \gamma_k + \int_\tau^s L|x_k(\theta) - x(\theta)|\,\mathrm{d}\theta + \delta_k + \int_\tau^s L|x_k(\theta) - x(\theta)||U|\,\mathrm{d}\theta \\
&\qquad + \left| \int_\tau^s [B^0(x(\theta), \theta)(u_k(\theta) - u_0(\theta))]\,\mathrm{d}\theta \right|,
\end{aligned}$$

where $|U| \stackrel{def}{=} \operatorname{dist}(U, \{0\})$, L is the Lipschitz constant of $F_0^0(\cdot, s)$, and $B^0(\cdot, s)$ in a compact set Z that contains all $x_k(s)$. Since the last term is continuous and tends to

zero with k, the Gronwall inequality implies that $x_k(s) \to x(s)$. From the invariance of $E_k(\cdot)$ we have $x_k(s) \in E_k(s)$ for every $s \in [\tau, T]$, and therefore $x(s) \in E(s)$. This proves the invariance of $E(\cdot)$, and thus the claim of the theorem thanks to Corollary 2.2. □

Remark 2.4. In the case of single-valued F_0^k and E_0, the above theorem holds without the supposition that the second summand in (7) is linear in u. An example showing that Proposition 2.3 is false for nonaffine inclusions (even with single-valued F_0^k and convex compact-valued $f_1^k(x, s, U)$) is given in the next subsection.

2.3. Particular cases and examples. Every collection of closed sets $\mathcal{E}$ can be represented as consisting of sublevel sets of a parametric family of Lipschitz functions. Namely, if we define

$$\varphi(E, x) = \operatorname{dist}(x, E), \quad E \in \mathcal{E},\ x \in \mathbf{R}^n,$$

and denote $P = \mathcal{E}$, then obviously

$$E \in \mathcal{E} \Leftrightarrow \exists p \in P :\ E = E(p), \quad \text{where } E(p) \stackrel{def}{=} \{x \in \mathbf{R}^n;\ \varphi(p, x) \le 0\}. \tag{8}$$

For practical reasons, however, we are mainly interested in collections $\mathcal{E}$ that admit a parameterization as in (8) with a set P being a subset of a finite dimensional space, as in the examples below, which are especially convenient and easy for calculation.

(a) Let us fix a finite or countable subset $L = \{l_1, l_2, \dots\}$ of the unit sphere $\partial\mathcal{B} \subset \mathbf{R}^n$ such that the convex cone spanned by L coincides with $\mathbf{R}^n$. With every $Z \in \operatorname{comp}(\mathbf{R}^n)$ we associate the sequence of numbers $p(Z) = (p_1, p_2, \dots)$, where $p_i = \max_{z \in Z} \langle l_i, z\rangle$. Denote $P \stackrel{def}{=} \{p(Z);\ Z \in \operatorname{comp}(\mathbf{R}^n)\}$.

It is easy to check that the corresponding collection $\mathcal{E}_L \stackrel{def}{=} \{E(p);\ p \in P\}$ satisfies Conditions B.1–B.3. It may consist of all convex compact subsets of $\mathbf{R}^n$ (if L is dense in $\partial\mathcal{B}$), of all "boxes" (if $L = \{\pm e_i\}_{i=1}^n$ with $\{e_i\}_i$—an orthogonal basis in $\mathbf{R}^n$), and of all polyhedrons with given normal vectors to the faces.

(b) The following is an alternative collection $\mathcal{E}$ that satisfies Conditions B.1–B.3. Let us fix the points $z_1, \dots, z_N \in \mathbf{R}^n$, $N \ge 1$, and define

$$\mathcal{E} \stackrel{def}{=} \left\{ \bigcap_{i=1,\dots,N} (z_i + s_i\mathcal{B}),\ s = (s_1, \dots, s_N) \in S \right\},$$

where S is the set of those s for which the intersection is nonempty. For every $Z \in \operatorname{comp}(\mathbf{R}^n)$, the unique minimal element from $\mathcal{E}$ that contains Z has

$$s_i = \operatorname{dist}(Z, z_i).$$

(c) Similarly, as in the collection in (b), instead of translated sublevel sets of the Euclidean norm, one may use sublevel sets of a collection of other functions defined in $\mathbf{R}^n$, caring that Conditions B.1–B.3 are satisfied. Such a collection may involve nonconvex sets and may be "adapted" to a given differential inclusion in such a way that reachable sets initiated from elements of the collection are well approximated (locally in time) by other elements of the collection.

Example 2.5 (proof of the negative claim in Remark 2.4). Consider the system

$$\begin{aligned} &\dot{x}_1 = u_1, \quad \dot{x}_2 = u_1^2 + (1 + x_3^2)u_2 - 1, \quad \dot{x}_3 = 0, \\ &U = \{(u_1, u_2);\ -1 \le u_1 \le 1,\ 0 \le u_2 \le 1 - u_1^2\}. \end{aligned} \tag{9}$$

Notice that the right-hand side of the corresponding differential inclusion is convex and compact for every x. Let us consider the collection $\mathcal{E} = \mathrm{comp}(\mathbf{R}^n)$ and the initial element $E_0 = (0, 0, [-1, 1])$. Define the sequence u_1^k on $[0, 1]$ such that u_1^k switches alternatively between -1 and 1 at times $t = i/k$, $i = 0, 1, \dots, k-1$, and define $u_2^k \equiv 0$. Then $E_k(t) = (x_1^k(t), 0, [-1, 1])$ with $x_1^k(t) = \int_0^t u_1^k(s)\,\mathrm{d}s$ is a solution tube of (9). Obviously $x_1^k(\cdot) \to 0$ and $E_k(\cdot) \to E_0$ (since u_1^k converges weakly to zero, $(u_1^k)^2 \equiv 1$, and $x_2(0) = 0$). The last set, however, is neither a solution tube nor contains such a tube starting from E_0. Indeed, if E_0 is a solution tube, then u_1 must be identically zero. Then the solution tube of (9) becomes

$$\left\{(0, (1 + x_3^2) \int_0^t u_2(s)\,\mathrm{d}s - t, x_3);\ x_3 \in [-1, 1]\right\},$$

and the second component cannot be identically zero independently of $x_3 \in [-1, 1]$, no matter what the control u_2 is. This proves the negative claim in Remark 2.4. The solution tube $E(t) = (0, 0, [-1, 1])$ is, in fact, realized by the Young measure $\mu(t) = 0.5\delta_{-1} + 0.5\delta_1$, where δ_u is the atomic measure concentrated at u. Any control $u(t)$, however, that realizes the same trajectory as μ has to depend on its initial state x_3, which is "unknown" in the context of the solution tubes.

3. Viability theory for collections of sets and solution tubes. The theory developed in the previous section extends the usual notion of trajectory of a control system to the notion of a solution tube (in a collection of sets) of an *uncertain* control system (controlled differential inclusion). Then the problem arises of how to build a viability theory for solution tubes similarly to the way it has been done for usual trajectories [2].

General theories for set-valued dynamical systems (in metric spaces) were developed in [3, 26, 4]. In particular, the viability theory was extended to more general dynamics in metric spaces in [16, 18, 4]. The dynamics of the solution tubes in a given collection of sets, however, does not fit within the above-mentioned framework. A relevant notion for contingency and viability for collections of sets was developed in [28], but it concerns standard (certain) control systems, where only the initial state is uncertain. Below we extend this theory in two directions: (i) to uncertain control systems, and (ii) to the concept of "viability with a target" introduced in [27].

3.1. Viability domains and kernels. Let a collection of compact sets $\mathcal{E}$ in $\mathbf{R}^n$ be fixed and let $Z \in \mathrm{comp}(\mathbf{R}^n)$.

Definition. A mapping $L(\cdot) \in C(Z, \mathrm{comp}(\mathbf{R}^n))$ is called a (continuous) contingent field to $\mathcal{E}$ at Z if

$$\liminf_{h \to 0+} \inf_{\tilde{E} \in \mathcal{E}} \sup_{x \in Z} \mathrm{dist}\left(L(x), \frac{\tilde{E} - x}{h}\right) = 0.$$

The set of all contingent fields to $\mathcal{E}$ at Z will be denoted by $\mathcal{T}_{\mathcal{E}}(Z)$. The following lemma is a straightforward consequence of the definition.

LEMMA 3.1. *The (set-valued) mapping $L(\cdot) \in C(Z; \mathrm{comp}(\mathbf{R}^n))$ belongs to $\mathcal{T}_{\mathcal{E}}(Z)$ if and only if there are sequences $h_k \to 0+$, $\gamma_k \to 0$, and $E_k \in \mathcal{E}$ such that*

$$(I + h_k L(\cdot))(Z) \subset E_k + h_k \gamma_k \mathcal{B} \quad \forall k$$

(here and below, I is the identity mapping).

In the particular case of the collection $\mathcal{E}_K = \{\{x\};\ x \in K\}$ of all single points from a given set $K \subset \mathbf{R}^n$, the set $\mathcal{T}_{\mathcal{E}}(Z)$ is nonempty if and only if $Z = \{x\}$ is a singleton and every contingent field is single-valued. In fact, the contingent field at $\{x\}$ is defined at the single point x only, and therefore can be identified with a vector $l \in \mathbf{R}^n$. It is straightforward that $l \in \mathcal{T}_{\mathcal{E}_K}(\{x\})$ if and only if $l \in T_K(x)$, where $T_K(x)$ is the usual (Bouligand) contingent cone[2] to K at x (see, e.g., [5]). In the case of collections that do not consist only of singletons, we use the term *contingent field* (in contrast to "contingent vector") to stress that *mappings* defined on Z are considered "tangent" objects, rather than constant vectors only. All properties of the contingent fields (similar to those for contingent vectors) obtained in [28, Proposition 1] in the case of single-valued fields apply also to set-valued fields.

Below $\mathcal{M}$ will be another (possibly empty) collection of compact sets in $\mathbf{R}^n$ that will be interpreted later as a "target." Let $\mathcal{L}$ be a family of mappings from $C(\mathbf{R}^n; \mathrm{comp}(\mathbf{R}^n))$. Then $\mathcal{L}_{|Z}$ will denote the set of restrictions of the mappings from $\mathcal{L}$ to the set $Z \in \mathrm{comp}(\mathbf{R}^n)$.

Definition. The collection $\mathcal{E}$ is called a *viability domain with target* $\mathcal{M}$ for $\mathcal{L}$ if

$$\mathcal{T}_{\mathcal{E}}(E) \cap \mathcal{L}_{|E} \neq \emptyset \quad \forall E \in \mathrm{cl}(\mathcal{E}) \setminus \mathcal{M}.$$

Obviously $\mathcal{E}$ is a viability domain if and only if $\mathrm{cl}(\mathcal{E})$ is a viability domain.

The following proposition plays a key role in the whole theory.[3] It claims that the sequences γ_k in Lemma 3.1 can be chosen to a certain extent independent of the particular $\mathcal{E}$, E, and L. The idea of the proof is the same as that of Lemma 2 in [28], obtained in the case of a family of single-valued fields $\mathcal{L}$ and $\mathcal{M} = \emptyset$. However, the suppositions in [28] are too restrictive in the present context, so the proof below is technically rather different from that in [28].

We shall use the notation

$$\rho(E, \mathcal{M}) \stackrel{def}{=} \inf_{M \in \mathcal{M}} \mathrm{dist}(E, M).$$

Definition. The subset $\mathcal{E}'$ of $\mathcal{E}$ is *inclusion-complete* (in $\mathcal{E}$) if the inclusions $E \in \mathcal{E}'$, $E' \subset E$, and $E' \in \mathcal{E}$ together imply $E' \in \mathcal{E}'$.

PROPOSITION 3.2. *Let the closed collections $\mathcal{E}$ and $\mathcal{M}$ of compact sets and the closed convex family $\mathcal{L} \subset C(\mathbf{R}^n; \mathrm{comp}(\mathbf{R}^n))$ be given. Suppose that $\mathcal{E}$ satisfies Conditions* B.1–B.3 *and that for every $Z \in \mathrm{comp}(\mathbf{R}^n)$, the family $\mathcal{L}_{|Z}$ is equi-Lipschitz and uniformly bounded.*

Then for every number R there exist positive numbers β, c, and C such that for every $h \in (0, \beta]$, for every $\mathcal{E}' \subset \mathcal{E}$ which is a closed inclusion-complete viability domain with target $\mathcal{M}$ for $\mathcal{L}$, and for every $E \in \mathcal{E}'$, for which $E \subset R\mathcal{B}$ and $\rho(E, \mathcal{M}) > ch$, there exist $L(\cdot) \in \mathcal{L}$ and $\tilde{E} \in \mathcal{E}'$, with $\tilde{E} \subset (R+1)\mathcal{B}$, such that

$$(I + hL(\cdot))(E) \subset \tilde{E} + Ch^2\mathcal{B}.$$

The proof is given in section 5.

[2] $T_K(x)$ is defined as the set of those $l \in \mathbf{R}^n$ for which $\liminf_{h \to 0+} \frac{1}{h} \mathrm{dist}(x + hl, K) = 0$.

[3] To our knowledge the proposition below is new also in the "classical" case of a collection of singletons and single-valued fields (in fact, vectors). Its claim seems to provide one deep reason for the Euler discretizability of Hamilton–Jacobi equations associated with optimal control, despite the fact that the derivatives in these equations do not exist in the classical sense and the solution can be even discontinuous.

If a collection $\mathcal{E} \subset \mathrm{comp}(\mathbf{R}^n)$ is not a viability domain for $\mathcal{L}$, then it may happen that it contains a collection $\mathcal{E}' \subset \mathcal{E}$ which is a viability domain.

THEOREM 3.3. *Let $\mathcal{E}$, $\mathcal{M}$, and $\mathcal{L}$ be as in Proposition* 3.2. *Then there exists a (possibly empty) closed collection $\mathcal{E}_0 \subset \mathcal{E}$ which is a viability domain for $\mathcal{L}$ with target $\mathcal{M}$ and which contains every other viability domain in $\mathcal{E}$ with target $\mathcal{M}$.*

Definition. The (possibly empty) collection $\mathcal{E}_0 \subset \mathcal{E}$ obtained in the above theorem is called the *viability kernel of $\mathcal{E}$ with target $\mathcal{M}$* for $\mathcal{L}$ and will be denoted by $\mathrm{Viab}_{\mathcal{L}}(\mathcal{E}; \mathcal{M})$.

Proof. If no subset of $\mathcal{E}$ is a viability domain (below, we skip mentioning $\mathcal{L}$ and $\mathcal{M}$), then $\mathrm{Viab}_{\mathcal{L}}(\mathcal{E}; \mathcal{M}) = \emptyset$. Otherwise we denote

$$\mathcal{E}_0 \stackrel{def}{=} \mathrm{cl} \cup \mathcal{E}',$$

where the union is taken with respect to all viability domains included in $\mathcal{E}$. However, the collection $\mathcal{E}_0$ would not change if we take in the above union only the closed inclusion-complete viability domains $\mathcal{E}'$. Indeed, the viability property is invariant with respect to taking closure or completing a collection $\mathcal{E}'$ by adding to it all subsets of elements of $\mathcal{E}'$ that belong to $\mathcal{E}$.

We shall prove that $\mathcal{E}_0$ is a viability domain with target $\mathcal{M}$. Take an arbitrary $E_0 \in \mathcal{E}_0 \backslash \mathcal{M}$. Let us denote $R = \mathrm{dist}(E_0, 0)+1$ and let β, c, and C be the corresponding numbers from Proposition 3.2. Let $h \in (0, 1]$ be arbitrary, but such that $2ch \leq \rho(E_0, \mathcal{M})$.

By the definition of $\mathcal{E}_0$ there is a viability domain $\mathcal{E}'_h \subset \mathcal{E}_0$ and $E_h \in \mathcal{E}'_h$ such that $H(E_h, E_0) \leq h \min\{1, ch\}$. We have $\mathrm{dist}(E_h, 0) \leq \mathrm{dist}(E_0, 0) + 1 \leq R$ and $\rho(E_h, \mathcal{M}) \geq \rho(E_0, \mathcal{M}) - ch \geq ch$. Then we can apply Proposition 3.2 to obtain $L_h(\cdot) \in \mathcal{L}$ and $\tilde{E}_h \in \mathcal{E}'_h$ such that

$$(I + hL_h(\cdot))(E_h) \subset \tilde{E}_h + Ch^2\mathcal{B}.$$

Let D be a Lipschitz constant of the mappings from $\mathcal{L}_{|Y}$, where $Y = E_0 + \mathcal{B}$. Then

$$\begin{aligned}(I + hL_h(\cdot))(E_0) &\subset (I + hL_h(\cdot))(E_h) + (1 + D)ch^2\mathcal{B} \subset \tilde{E}_h + [(1 + D)c + C]h^2\mathcal{B} \\ &= \tilde{E}_h + C_1h^2\mathcal{B}\end{aligned}$$

with $C_1 = (1 + D)c + C$. Since L_h are equi-Lipschitz and uniformly bounded on Y, there is a sequence $h_k \to 0$ such that L_{h_k} converges to some $L \in \mathcal{L}_{|Y}$ in $C(Y; \mathrm{comp}(\mathbf{R}^n))$. Then

$$(I + h_kL(\cdot))(E_0) \subset \tilde{E}_{h_k} + C_1h^2\mathcal{B} + h_k\|L(\cdot) - L_{h_k}(\cdot)\|_{C(Y;\mathrm{comp}(\mathbf{R}^n))}\mathcal{B}.$$

Since $\tilde{E}_{h_k} \in \mathcal{E}'_{h_k} \subset \mathcal{E}_0$, the last relation implies that $L(\cdot) \in \mathcal{T}_{\hat{\mathcal{E}}}(\hat{E})$ according to Lemma 3.1. Thus $\mathcal{E}_0$ is a viability domain with target $\mathcal{M}$. □

The following lemma is another direct consequence of Proposition 3.2.

LEMMA 3.4. *Let $\mathcal{E}$, $\mathcal{M}$, and $\mathcal{L}$ be as in Proposition* 3.2*, but with a compact family $\mathcal{L}$. Let $\mathcal{E}_k \in \mathcal{E}$, $k = 1, 2, \ldots$, be a sequence of closed inclusion-complete viability domains with target $\mathcal{M}$ for $\mathcal{L}$. Then the Kuratowski upper limit of $\{\mathcal{E}_k\}$ is a viability domain with target $\mathcal{M}$ for $\mathcal{L}$.*

3.2. Existence of a viable solution tube. In this subsection we address the key issue of the viability theory specified below in Theorem 3.5 in the present framework of collections of sets and solution tubes: the existence of a solution tube in a viability domain.

We consider a specific family of mappings $\mathcal{L} \subset C(\mathbf{R}^n; \mathrm{comp}(\mathbf{R}^n))$ having the affine form

$$\mathcal{L} = \{F_0(\cdot) + B(\cdot)u;\ u \in U\}, \tag{10}$$

where $F_0 \in C(\mathbf{R}^n; \mathrm{comp}(\mathbf{R}^n))$, $U \subset \mathbf{R}^r$, and $B(\cdot)$ is an $(n \times r)$-matrix function. Below in this section we suppose that

(11) $\mathcal{E} \subset \mathrm{comp}(\mathbf{R}^n)$ is a collection satisfying Conditions B.1–B.3;

(12) is convex and compact, $B(\cdot)$ is locally Lipschitz continuous;

(13) $F_u(x) \stackrel{def}{=} F_0(x) + B(x)u$ satisfies Condition A uniformly in $u \in U$.

With the family $\mathcal{L}$ we associate the controlled differential inclusion

$$\dot{x} \in F_0(x) + B(x)u(t), \quad u(t) \in U. \tag{14}$$

Let us denote by $\mathcal{U}_{[s,\tau]}$ the set of all measurable $u(\cdot) : [s, \tau] \mapsto U$. According to Corollary 2.2, for every $s \geq 0$, every $E \in \mathcal{E}$, and every $u(\cdot) \in \mathcal{U}_{[s,+\infty)}$, inclusion (14) has a unique solution tube $E_{u(\cdot)}[s, E](\cdot)$ in the collection $\mathcal{E}$ on $[0, +\infty)$, starting from E at time s. Similarly, as before, the solution tube in $\mathrm{comp}(\mathbf{R}^n)$ will be denoted by $X_{u(\cdot)}[s, E](\cdot)$.

The following definition adapts the terminology from [2].

Definition. Let $\mathcal{E}'$ and $\mathcal{M}$ be given closed subsets of $\mathcal{E}$. The collection $\mathcal{E}'$ enjoys the *viability property* with target $\mathcal{M}$ with respect to (14) if and only if for every $E \in \mathcal{E}' \setminus \mathcal{M}$ there exists $u(\cdot) \in \mathcal{U}_{[0,+\infty)}$ such that the solution tube $E_{u(\cdot)}(\cdot) = E_{u(\cdot)}[0, E](\cdot)$ either satisfies $E_{u(\cdot)}(t) \in \mathcal{E}'$ for all $t \geq 0$, or there is $T > 0$ such that $E_{u(\cdot)}(t) \in \mathcal{E}'$ on $[0, T]$ and $E_{u(\cdot)}(T) \in \mathcal{M}$.

THEOREM 3.5. *Let Conditions* (11)–(13) *hold. The closed inclusion-complete collection $\mathcal{E}' \subset \mathcal{E}$ enjoys the viability property with target $\mathcal{M}$ with respect to* (14) *if and only if it is a viability domain with target $\mathcal{M}$.*

The proof of this result is rather technical; therefore it is postponed to section 5.

Remark 3.6. We mention that in the "classical" viability theory (that is, with single-valued F_0) affineness with respect to u is not required. Without asserting that Theorem 3.5 does not hold under weaker assumptions for the dependence of F_u on u, we point out that in the case of a nonlinear dependence on u, some undesirable irregularities may take place. For example, in the "classical" case the intersection of a monotone decreasing sequence of viability domains is a viability domain, too. This is not the case for Example 2.5 in section 2.3. The collection $\mathcal{E}'_\alpha = \{([-a, a], 0, [-1, 1]);\ a \in [0, \alpha]\} \subset \mathcal{E} \stackrel{def}{=} \mathrm{comp}(\mathbf{R}^3)$ is easily seen to be a viability domain (with $\emptyset$ as a target) if $\alpha > 0$. On the other hand, it was shown there that $\mathcal{E}'_0 \stackrel{def}{=} \cap_\alpha \mathcal{E}'_\alpha = \{(0, 0, [-1, 1]\}$ fails to enjoy the viability property. We shall comment further on this effect in section 4.3.

3.3. Dini derivative and viability. In this subsection, $\mathcal{E}$ will always be a closed collection of nonempty compact sets in $\mathbf{R}^n$ and $J : \mathcal{E} \to \mathbf{R}$ will be a lower semicontinuous function (abbreviated as l.s.c.). Let $E \in \mathcal{E}$ be fixed and let $F : E \mapsto \mathrm{comp}(\mathbf{R}^n)$ be a set-valued field on E.

Definition. We define the lower Dini derivative of J at E in the direction of the field F as

$$D^-_{\mathcal{E}} J(E; F) \stackrel{def}{=} \liminf_{h,\delta \to 0+} \inf \left\{ \frac{J(E') - J(E)}{h};\ E' \in \mathcal{E},\ (I + hF)(E) \subset E' + h\delta\mathcal{B} \right\}.$$

Remark 3.7. If J is Lipschitz and the collection $\mathcal{E}$ satisfies Condition B.1, then one can equivalently take $\delta = 0$ in the above definition.

For $\mathcal{E}$ and J as above, we define as usual

$$\operatorname{epi} J \stackrel{def}{=} \{(E,\nu);\ E \in \mathcal{E},\ \nu \in \mathbf{R} :\ J(E) \le \nu\} \subset \mathcal{E} \times \mathbf{R},$$

which is also a closed collection of compact sets.

The following lemma follows easily from the definitions.

LEMMA 3.8. *For every $E \in \mathcal{E}$ and $F : E \mapsto \operatorname{comp}(\mathbf{R}^n)$, the following two conditions are equivalent:*

(i) $D^-_{\mathcal{E}} J(E;F) \le 0$;

(ii) $(F,0) \in \mathcal{T}_{\operatorname{epi} J}(E, J(E))$.

Let $\mathcal{M} \subset \mathcal{E}$ be closed and let $\mathcal{L}$ be a closed family of mappings $L : \mathbf{R}^n \Rightarrow \mathbf{R}^n$ such that, for every $Z \in \operatorname{comp}(\mathbf{R}^n)$, the family $\mathcal{L}_{|Z}$ is equi-Lipschitz and uniformly bounded. As a consequence of the above lemma, we obtain that

$$\inf_{L \in \mathcal{L}} D^-_{\mathcal{E}} J(E; L_{|E}) \le 0 \quad \forall E \in \mathcal{E} \setminus \mathcal{M}$$

is a necessary condition for the collection $\operatorname{epi} J$ to be a viability domain for $\mathcal{L} \times \{0\}$ with a target $\mathcal{M} \times \mathbf{R}$. The next lemma implies that the "inf" in the above inequality is attained.

LEMMA 3.9. *For every fixed $E \in \mathcal{E}$, the mapping*

$$L \to D^-_{\mathcal{E}} J(E;L) \tag{15}$$

defined in the space $C(E; \operatorname{comp}(\mathbf{R}^n))$ of continuous set-valued fields on E is l.s.c.

Proof. Let $L_k \to L$, that is, $\rho_k \stackrel{def}{=} \sup_{x \in E} H(L(x), L_k(x)) \to 0$. There exist sequences h_k, δ_k, and $E'_k \in \mathcal{E}$ such that $0 < h_k, \delta_k \le 1/k$,

$$(I + h_k L_k)(E) \subset E'_k + h_k \delta_k \mathcal{B} \quad \text{and} \quad \frac{J(E'_k) - J(E)}{h_k} \le D^-_{\mathcal{E}} J(E; L_k) + \frac{1}{k}. \tag{16}$$

Then

$$(I + h_k L)(E) \subset E'_k + h_k \rho_k \mathcal{B} + h_k \delta_k \mathcal{B}.$$

From here and the second relation in (16),

$$D^-_{\mathcal{E}} J(E;L) \le \liminf_{k \to +\infty} \left(D^-_{\mathcal{E}} J(E;L_k) + \frac{1}{k} \right),$$

which proves the lemma. ∎

COROLLARY 3.10. $\operatorname{epi} J$ *is a viability domain with a target $\mathcal{M} \times \mathbf{R}$ for $\mathcal{L} \times \{0\}$ if and only if*

$$\min_{L \in \mathcal{L}} D^-_{\mathcal{E}} J(E; L_{|E}) \le 0 \quad \forall E \in \mathcal{E} \setminus \mathcal{M}.$$

Notice that we have replaced "inf" with "min" in the above inequality, meaning that the infimum is attained.

4. Optimal control in the presence of observable and unobservable uncertainties. We consider the minimization problem for (1), (2) formulated in the introduction. In addition we introduce a set $M \subset [0,T] \times \mathbf{R}^n$, which will determine the termination time of the control process, as described below. The following suppositions will hold for the rest of the section.

Condition C.1. U, Y, and $\bar{V}$ are compact subsets of finite dimensional vector spaces, U is convex, and the mapping $y \to V(y) \subset \bar{V}$ is compact valued and Lipschitz continuous.

Condition C.2. The function $f : \mathbf{R}^n \times U \times Y \times \bar{V} \mapsto \mathbf{R}^n$ has the form

$$f(x,u,y,v) = f_0(x,y,v) + B(x,y)u,$$

where f_0 and B are continuous, locally Lipschitz in x uniformly with respect to the other variables. The sets $f_0(x,y,V(y))$ are convex. f has linear growth with respect to x, uniformly in u,y,v.

The admissible sets $\mathcal{U}_{[t,\theta]}$, $\mathcal{Y}_{[t,\theta]}$, and $\mathcal{V}_{[t,\theta]}(y(\cdot))$ are defined in the introduction. We recall the notion of *nonanticipative strategy* on $[t,\theta]$. This is any mapping $\alpha : \mathcal{Y}_{[t,\theta]} \mapsto \mathcal{U}_{[t,\theta]}$ that satisfies the nonanticipativity condition

$$\begin{aligned}&\forall\, y_1, y_2 \in \mathcal{Y}_{[t,\theta]},\ \forall\, \tau \in (t,\theta]\\ &\text{if } y_1(s) = y_2(s) \quad \text{for a.e. } s \in [t,\tau], \text{ then } \alpha(y_1)(s) = \alpha(y_2)(s) \quad \text{for a.e. } s \in [t,\tau].\end{aligned}$$

Let $\mathcal{A}_{[t,\theta]}$ denote the set of all such strategies on $[t,\theta]$.

Let us consider first the case of a fixed-end time T discussed in the introduction. With the system (1) and any fixed $u \in \mathcal{U}_{[t,T]}$ and $y \in \mathcal{Y}_{[t,T]}$, we associate the differential inclusion

$$\dot{x} \in f(x,u(s),y(s),V(y(s))). \tag{17}$$

Similarly, as before, we denote by $X_{u,y}[t,E](\cdot)$ the solution tube of (17) in $\mathrm{comp}(\mathbf{R}^n)$, starting from the set E at time t. Moreover, for a compact set $Z \subset \mathbf{R}^n$ we define $G : \mathrm{comp}(\mathbf{R}^n) \mapsto \mathbf{R}$ as

$$G(T,Z) = \sup_{z \in Z} g(T,z).$$

Then, obviously, definition (4) is equivalent to

$$I(t,E) = \inf_{\alpha \in \mathcal{A}_{[t,T]}} \sup_{y \in \mathcal{Y}_{[t,T]}} G(T, X_{\alpha(y),y}[t,E](T)). \tag{18}$$

In this formulation of the original problem there is no unobservable uncertainty. We passed to a problem with complete information (here y is an observable disturbance) but over the solution tubes to differential inclusion (17).

Because of the complexity of the problem so obtained, we can restrict the consideration to the solution tubes of (17) in a given collection of sets $\mathcal{E}$ instead of the whole $\mathrm{comp}(\mathbf{R}^n)$. Thus we come up with the more general problem, formulated below in the case of a target that determines the termination time. We also suppose the following condition.

Condition C.3. The collection $\mathcal{E}$ satisfies Conditions B.1–B.3. The set $M \subset [0,+\infty) \times \mathbf{R}^n$ is closed and contains $t \times \mathbf{R}^n$ for every $t \geq T$. The function $g(\cdot) : M \mapsto \mathbf{R}$ is l.s.c. Moreover, the following property holds: For every $(t,x_0) \in M$ with $t < T$,

for every $s \in (t,T]$, $y \in \mathcal{Y}_{[t,T]}$, and $u \in \mathcal{U}_{[t,T]}$, the inclusion (17) has a trajectory $x(\cdot)$ with $x(t) = x_0$ for which $g(s,x(s)) \geq g(t,x_0)$.

Formulation of the general problem. Let Conditions C.1–C.3 be satisfied. For any fixed $u \in \mathcal{U}_{[t,T]}$ and $y \in \mathcal{Y}_{[t,T]}$, we denote by $E_{u,y}[t,E](\cdot)$ the solution tube in the collection $\mathcal{E}$ starting from $E \in \mathcal{E}$ at time t. We define also the target collection $\mathcal{M} \stackrel{def}{=} \{(t,E);\ E \in \mathcal{E},\ (t,E) \subset M\}$, which is nonempty since $(T,\mathcal{E}) \subset \mathcal{M}$, closed, and inclusion complete in $\mathcal{E}$. For a given tube $E(\cdot) \in C([0,T];\mathcal{E})$ the termination time is determined as

$$T(E(\cdot)) = \min\{t \geq 0;\ (t,E(t)) \in \mathcal{M}\}.$$

We consider the following minimization problem for the initial pair $(t,E) \notin \mathcal{M}$:

$$\begin{aligned} &I_{\mathcal{E}}(t,E) \stackrel{def}{=} \inf_{\alpha \in \mathcal{A}_{[t,T]}} \sup_{y \in \mathcal{Y}_{[t,T]}} G(\tau, E(\tau)), \\ &\text{where } E(\cdot) \stackrel{def}{=} E_{\alpha(y),y}[t,E](\cdot),\ \tau = T(E(\cdot)). \end{aligned} \tag{19}$$

If $(t,E) \in \mathcal{M}$, then by definition $I_{\mathcal{E}}(t,E) = G(t,E)$. The role of the last part of Condition C.3 is to ensure that once the target is achieved, the continuation of the process would not lead to a better guaranteed result. This supposition is obviously fulfilled in the case $\mathcal{M} = \{T\} \times \mathcal{E}$ (fixed-end time) or $G(t,E) = t$ (minimal time problem). Obviously we have $I_{\mathcal{E}}(t,E) \geq I_{\mathrm{comp}(\mathbf{R}^n)}(t,E)$. The difference between these two values is the price for the simplification we make by passing to the collection $\mathcal{E}$. In some cases, however, equality may take place even for simple subcollections of $\mathrm{comp}(\mathbf{R}^n)$, as shown by a discrete-time example in [25].

4.1. Some basic properties. The proof of the following dynamic programming principle adapts standard arguments.

PROPOSITION 4.1. *Under Conditions* C.1–C.3*, for every $E \in \mathcal{E}$, $t \in [0,T)$, and $s \in (t,T]$,*

$$I_{\mathcal{E}}(t,E) = \inf_{\alpha \in \mathcal{A}_{[t,s]}} \sup_{y \in \mathcal{Y}_{[t,s]}} I_{\mathcal{E}}(\tau, E(\tau)),$$

$$\text{where } E(\cdot) \stackrel{def}{=} E_{\alpha(y),y}[t,E](\cdot),\ \tau \stackrel{def}{=} \min\{s, T(E(\cdot))\}.$$

PROPOSITION 4.2. *Under Conditions* C.1–C.3*, the value function $I_{\mathcal{E}} : [0,T] \times \mathcal{E}$ is l.s.c.*

PROPOSITION 4.3. *Under Conditions* C.1–C.3*, for every $(t,E) \in [0,T) \times \mathcal{E}$ there is $\alpha \in \mathcal{A}_{[t,T]}$ for which the infimum in* (19) *is attained (that is an optimal strategy).*

Proof. The proof of the second proposition[4] follows from the proof of Proposition 4.2, applied for $(t_k,E_k) \equiv (t,E)$.

Let the sequence $\{(t_k,E_k)\}$, with $t_k \in [0,T]$ and $E_k \in \mathcal{E}$, converge to (t,E). By the definition of $I_{\mathcal{E}}$, there is $\alpha_k \in \mathcal{A}_{[t_k,T]}$ such that

$$\sup_{y \in \mathcal{Y}_{[t_k,T]}} \{G(\tau,E(\tau));\ E(\cdot) = E_{\alpha_k(y),y}[t_k,E_k](\cdot),\ \tau = T(E(\cdot))\} \leq I_{\mathcal{E}}(t_k,E_k) + \frac{1}{k}. \tag{20}$$

[4]The existence of an optimal strategy can be obtained also by a slight modification of part 3 of the proof of Theorems 4.5 and 4.7 below. However, we give here a direct proof (not involving the viability theory), which applies also to Proposition 4.2.

We fix an infinite subset $\mathbf{N}_0 \subset \mathbf{N}$ such that $\lim_{k\in\mathbf{N}_0} I_{\mathcal{E}}(t_k, E_k) = \liminf_{k\in\mathbf{N}} I_{\mathcal{E}}(t_k, E_k)$. We fix also some $\bar{y} \in Y$. For an arbitrary $y \in \mathcal{Y}_{[t,T]}$ define $y_k \in \mathcal{Y}_{[t_k,T]}$ either by restricting y to $[t_k, T]$ (if $t < t_k$) or by concatenating $\bar{y}$ on $[t_k, t)$ (if $t > t_k$). Then we denote by $S(y)$ the set of all L_2-weak limits of the sequence $\{\alpha_k(y_k)\}_{k\in\mathbf{N}_0}$. Below we utilize the following form of the Cardaliaguet–Plaskacz lemma (see [10]).

LEMMA 4.4. *Let $S : \mathcal{Y}_{[t,T]} \Rightarrow U_{[t,T]}$ be a set-valued mapping with nonempty L_2-weakly compact images, having the following* nonanticipativity *property:*

For every $\theta \in (t, T]$, if $y', y'' \in \mathcal{Y}_{[t,T]}$ and $y'(s) = y''(s)$ almost everywhere in $[t, \theta]$, then, for every $u' \in S(y')$, there exists $u'' \in S(y'')$ such that $u'(s) = u'(s)$ almost everywhere in $[t, \theta]$.

Then there exists $\alpha \in \mathcal{A}_{[t,T]}$ such that $\alpha(y) \in S(y)$ for every $y \in \mathcal{Y}_{[t,T]}$.

Clearly our set $S(y)$ is L_2-weakly closed and thus also L_2-weakly compact, since it is contained in the L_2-weakly compact set $\mathcal{U}_{[t,T]}$. To prove the nonanticipativity of S we take y', y'', and u' as in Lemma 4.4. Denoting by y'_k and y''_k the corresponding extensions/restrictions defined above, we have $\alpha_k(y'_k)(s) = \alpha_k(y''_k)(s)$ for a.e. $s \in [t, \theta]$, since α_k is nonanticipative. If $\mathbf{N}' \subset \mathbf{N}_0$ is the subsequence of $\alpha_k(y'_k)$ convergent to u', then we define u'' as an arbitrary L_2-weak limit of $\{\alpha_k(y''_k)\}_{\mathbf{N}'}$. Then $u'' \in S(y'')$ and clearly $u''(s) = u'(s)$ a.e. in $[t, \theta]$.

From Lemma 4.4 we obtain an $\alpha(\cdot) \in \mathcal{A}_{[t,T]}$ which is a selection of S. In particular, for every $y \in \mathcal{Y}_{[t,T]}$, we fix an infinite subset $\mathbf{N}^y \subset \mathbf{N}_0$ such that $\{\alpha_k(y_k)\}_{k\in\mathbf{N}^y}$ converges L_2-weakly to $\alpha(y)$.

Let us fix an arbitrary $y \in \mathcal{Y}_{[t,T]}$. For the corresponding restricted/extended $y_k \in \mathcal{Y}_{[t_k,T]}$ we have, according to Proposition 2.3, that solution tubes $E_{\alpha_k(y_k),y_k}[t_k, E_k](\cdot)$ converge uniformly along some subsequence $\mathbf{N}' \subset \mathbf{N}^y$ to a tube $E(\cdot)$ in $\mathcal{E}$, and $E_{\alpha(y),y}[t, E](s) \subset E(s)$. Let $E_k(\cdot) \stackrel{def}{=} E_{\alpha_k(y_k),y_k}[t_k, E_k](\cdot)$, $\tau_k \stackrel{def}{=} T(E_k(\cdot))$. Denote $\tau_0 = \liminf_{k\in\mathbf{N}'} \tau_k$. Then from the closedness and the inclusion-completeness of $\mathcal{M}$, we have that $\hat{\tau} \stackrel{def}{=} T(E_{\alpha(y),y}[t, E](\cdot)) \le \tau_0$. From the lower semicontinuity and inclusion-monotonicity of G, and then by the choice of $\mathbf{N}^y \supset \mathbf{N}'$ and (20),

$$\begin{aligned} G(\tau_0, E_{\alpha(y),y}[t, E](\tau_0)) &\le G(\tau_0, E(\tau_0)) \le \limsup_{k\in\mathbf{N}'} G(\tau_k, E_k(\tau_k)) \\ &\le \lim_{k\in\mathbf{N}'} \left(I_{\mathcal{E}}(t_k, E_k) + \frac{1}{k} \right) = \liminf_{k\to+\infty} I_{\mathcal{E}}(t_k, E_k). \end{aligned}$$

Since $\hat{\tau} \le \tau_0$, the last part of Condition C.3 and the above inequality imply

$$G(\hat{\tau}, E_{\alpha(y),y}[t, E](\hat{\tau})) \le G(\tau_0, E_{\alpha(y),y}[t, E](\tau_0)) \le \liminf_{k\to+\infty} I_{\mathcal{E}}(t_k, E_k).$$

Since $y \in \mathcal{Y}_{[t,T]}$ was chosen arbitrarily, this completes the proof. □

4.2. Characterizations of the value function. The first theorem below gives a characterization of the value function as the unique minimal Dini supersolution of the associated HJI equation. Define the extended closed collection of compact sets $\hat{\mathcal{E}}$ in $\mathbf{R} \times \mathbf{R}^n$ as

$$\hat{\mathcal{E}} \stackrel{def}{=} \{(t, E);\ t \ge 0,\ E \in \mathcal{E}\}.$$

THEOREM 4.5. *Under Conditions* C.1–C.3*, the value function $I_{\mathcal{E}}$ is the unique minimal l.s.c. solution of the differential inequality*

$$\sup_{y\in Y}\min_{u\in U} D^-_{\hat{\mathcal{E}}} J(t,E;(1,f(\cdot,u,y,V(y)))) \le 0 \quad \forall\,(t,E)\in \hat{\mathcal{E}}\setminus\mathcal{M}, \tag{21}$$

with the side condition

$$J(t,E)\ge G(t,E) \quad \forall\,(t,E)\in\mathcal{M}. \tag{22}$$

That is,

$$I_{\mathcal{E}}(t,E)=\min\{J(t,E);\ J \text{ – } l.s.c.\ solution\ of\ (21),\ (22)\}.$$

Remark 4.6. We can write "min" instead of "inf" in (21) thanks to Lemma 3.9.

In parallel with Theorem 4.5 we shall prove also the next theorem, which extends the result in [10], the latter concerning the case of complete information (the disturbance v is not present). It shows that one can apply for calculation of $I_{\mathcal{E}}$ an extension, similar to that in [27], of the viability kernel algorithm [31, 11].

THEOREM 4.7. *Under Conditions* C.1–C.3*, denote*

$$\mathcal{E}^* \stackrel{def}{=} \hat{\mathcal{E}}\times\mathbf{R}, \qquad \mathcal{M}^* \stackrel{def}{=} \{(t,E,\nu);\ (t,E)\in\mathcal{M},\ \nu\ge G(t,E)\}.$$

Define $\mathcal{E}^*_1=\mathcal{E}^*$*, and recursively,*

$$\mathcal{E}^*_{k+1}=\bigcap_{y\in Y}\mathrm{Viab}_{\mathcal{L}_y}(\mathcal{E}^*_k;\mathcal{M}^*), \tag{23}$$

where $\mathcal{L}_y \stackrel{def}{=} \{(1,f(\cdot,u,y,V(y)),0);\ u\in U\}$*. Then*

$$\mathrm{epi}\, I_{\mathcal{E}}=\mathcal{E}_\infty \stackrel{def}{=} \bigcap_k \mathcal{E}^*_k.$$

Proof. We prove Theorems 4.5 and 4.7 in parallel.

1. First we shall prove that $I_{\mathcal{E}}$ is an l.s.c. solution of (21), (22). The function $I_{\mathcal{E}}$ is l.s.c. according to Proposition 4.2 and satisfies (22) as an equality by definition. To prove that it satisfies (21) we shall employ the dynamic programming principle in Proposition 4.1. Fix $(t,E)\in\hat{\mathcal{E}}\setminus\mathcal{M}$ and an arbitrary $y\in Y$. Since $\mathcal{M}$ is closed, there exists a sequence $h_k\to 0+$ such that for the sequence $t_k \stackrel{def}{=} t+h_k$ we have $(t_k,E_{u,y}[t,E](t_k))\notin\mathcal{M}$ for every $u\in\mathcal{U}_{[t,t_k]}$ and $y\in\mathcal{Y}_{[t,t_k]}$. Then $T(E_{u,y}[t,E](\cdot))>t_k$. According to Proposition 4.1 there is $\alpha_k\in\mathcal{A}_{[t,t_k]}$ such that

$$I_{\mathcal{E}}(t,E)\ge \sup_{y\in\mathcal{Y}_{[t,t_k]}} I_{\mathcal{E}}(t_k,E_{\alpha(y),y}[t,E](t_k))-\frac{h_k}{k}.$$

For $y\in Y$, denote $u_k(\cdot)\stackrel{def}{=}\alpha_k(y)(\cdot)$ (here the function y is identified with the constant value y), and $E_k(\cdot)=E_{u_k,y}[t,E](\cdot)$. We have

$$I_{\mathcal{E}}(t,E)\ge I_{\mathcal{E}}(t_k,E_k(t_k))-\frac{h_k}{k}. \tag{24}$$

Denote $\hat{E}_k\stackrel{def}{=}(t_k,E_k(t_k))\in\hat{\mathcal{E}}$ and

$$\bar{u}_k=\frac{1}{h_k}\int_t^{t_k} u_k(\tau)\,\mathrm{d}\tau.$$

Since U is compact we may assume that for some $\bar{u} \in U$ and for a subsequence, we have $\beta_k \overset{def}{=} |\bar{u}_k - \bar{u}| \to 0$. Thanks to the structural and convexity assumptions in C.1 and C.2, one can prove in a standard way that there is a constant C such that

$$H(X_{u_k(\cdot),y}[t,E](t_k), (I + h_k f(\cdot, \bar{u}, y, V))(E)) \leq Ch_k(h_k + \beta_k).$$

Since $X_{u_k(\cdot),y}[t,E](t_k) \subset E_k(t_k)$, we obtain

$$(I + h_k(1, f(\cdot, \bar{u}, y, V)))(t,E) \subset \hat{E}_k + Ch_k(h_k + \beta_k)\mathcal{B}.$$

This together with (24) implies

$$D^-_{\hat{\mathcal{E}}} J(t,E;(1, f(\cdot, \bar{u}, y, V))) \leq 0.$$

Then

$$\min_{u \in U} D^-_{\hat{\mathcal{E}}} J(t,E;(1, f(\cdot, u, y, V))) \leq 0,$$

which gives (21), since $y \in Y$ was arbitrarily chosen.

2. Now we shall prove that every l.s.c. solution J of (21), (22) satisfies the inclusion

$$\operatorname{epi} J \subset \mathcal{E}^*_\infty \overset{def}{=} \bigcap_k \mathcal{E}^*_k. \tag{25}$$

Thanks to (21), (22), and the definition of $\mathcal{M}^*$, we obtain from Corollary 3.10 that $\operatorname{epi} J$ is a viability domain with target $\mathcal{M}^*$ for each family $\mathcal{L}_y$ with $y \in Y$. Thus

$$\operatorname{epi} J \subset \bigcap_{y \in Y} \operatorname{Viab}_{\mathcal{L}_y}(\mathcal{E}^*_1; \mathcal{M}^*) = \mathcal{E}^*_2.$$

Repeating the same argument, we obtain that $\operatorname{epi} J \subset \mathcal{E}^*_k$ for every $k \geq 1$; hence (25).

3. It remains to prove that $\mathcal{E}^*_\infty \subset \operatorname{epi} I_{\mathcal{E}}$, which, in view of point 2, implies also the minimality of $I_{\mathcal{E}}$. To do this we consider the auxiliary differential inclusion

$$\begin{pmatrix} \dot{t} \\ \dot{x} \\ \dot{\nu} \end{pmatrix} \in \begin{pmatrix} 1 \\ f(x, u(s), y(s), V(y(s))) \\ 0 \end{pmatrix}, \tag{26}$$

which corresponds to the set-valued fields $\mathcal{L}_y$ involved in Theorem 4.7. Moreover, we fix an arbitrary $(t, E, \nu) \in \mathcal{E}^*_\infty$.

3.1. We shall prove that for every $y(\cdot) \in \mathcal{Y}_{[t,T]}$ there exists $u(\cdot) \in \mathcal{U}_{[t,T]}$ such that the corresponding solution tube $s \to (s, E(s), \nu)$ of (26), starting from (t, E, ν) at time $s = t$, satisfies

$$(s, E(s), \nu) \in \mathcal{E}^*_\infty \quad \text{as long as } s < T(E(\cdot)) \text{ or } \nu < G(s, E(s)). \tag{27}$$

Since for a fixed $y \in Y$ we have $\mathcal{E}_{k+1} \subset \operatorname{Viab}_{\mathcal{L}_y}(\mathcal{E}_k; \mathcal{M}^*) \subset \mathcal{E}_k$, one can write

$$\mathcal{E}_\infty = \bigcap_k \operatorname{Viab}_{\mathcal{L}_y}(\mathcal{E}_k; \mathcal{M}^*).$$

Since $\operatorname{Viab}_{\mathcal{L}_y}(\mathcal{E}_k; \mathcal{M}^*)$ is a viability domain with target $\mathcal{M}^*$ for $\mathcal{L}_y$, Lemma 3.4 implies the same for $\mathcal{E}^*_\infty$. Thus, for every fixed $y \in Y$ the collection $\mathcal{E}^*_\infty$ is a viability domain

with target $\mathcal{M}^*$ for (26). Then for the fixed $(t, E, \nu) \in \mathcal{E}^*_\infty$ and for every piecewise constant function $y(\cdot) \in \mathcal{Y}_{[t,T]}$ there exists $u(\cdot) \in \mathcal{U}_{[t,T]}$ such that the corresponding solution tube $(s, E(s), \nu)$ of (26) belongs to $\mathcal{E}^*_\infty$, as long as the target $\mathcal{M}^*$ is not reached. With the definition of the set $\mathcal{M}^*$ in mind, we obtain (27).

For any $y(\cdot) \in \mathcal{Y}_{[t,T]}$ we denote by $S(y(\cdot))$ the set of all $u(\cdot) \in \mathcal{U}_{[t,T]}$ for which (27) is satisfied. We have already proved that $S(y(\cdot))$ is nonempty for all piecewise constant $y(\cdot) \in \mathcal{Y}_{[t,T]}$. We shall prove that the mapping S has a closed graph with respect to the L_2-weak topology in $\mathcal{Y}_{[t,T]} \times \mathcal{U}_{[t,T]}$.

Let $(y_k, u_k) \to (y_0, u_0)$ weakly, and let $(\cdot, E_k(\cdot), \nu)$ be the solution tube of (26) corresponding to $u_k(\cdot)$ and $y_k(\cdot)$. Thanks to Conditions C.1–C.3, we may apply Proposition 2.3 and obtain that the solution tube $(\cdot, E_0(\cdot), \nu)$ corresponding to $y_0(\cdot)$ and $u_0(\cdot)$ is contained in a uniform limit of a subsequence of $(\cdot, E_k(\cdot), \nu)$. Denote this limit by $(\cdot, E(\cdot), \nu)$. Let s be such that

$$s < T(E_0(\cdot)) \quad \text{or} \quad \nu < G(s, E_0(s)).$$

From the inclusion-completeness of $\mathcal{M}$ and the inclusion monotonicity of G, we obtain that the same inclusion holds also for $E(s)$. From the closedness of $\mathcal{M}$ and the lower semicontinuity of G, we conclude that the above inequality is satisfied also by $E_k(s)$. Then from (27) (applied to $E_k(s)$) and the closedness of $\mathcal{E}^*_0$, we have $(s, E(s), \nu) \in \mathcal{E}^*_0$.

Moreover, if a set $\mathcal{E}^*_k$ is inclusion-complete in $\mathcal{E}^*$, then $\mathrm{Viab}_{\mathcal{L}_y}(\mathcal{E}^*_k; \mathcal{M}^*)$ is inclusion-complete, too, and also an arbitrary intersection of inclusion-complete collections is inclusion-complete. This implies inductively that $\mathcal{E}^*_\infty$ is inclusion-complete in $\mathcal{E}^*$. Then the inclusion $(s, E(s), \nu) \in \mathcal{E}^*_\infty$ obtained above, together with $E_0(s) \subset E(s)$ and $(s, E_0(s), \nu) \in \mathcal{E}^*$, imply $(s, E_0(s), \nu) \in \mathcal{E}^*_\infty$ on $[t, T(E_0)]$. Thus (27) is satisfied by u_0 and $(y, u_0) \in \operatorname{graph} S$. This proves the closedness of graph S in the L_2-weak topology in $\mathcal{Y}_{[t,T]} \times \mathcal{U}_{[t,T]}$.

From the closedness of graph S, together the compactness of $\mathcal{U}[t, T]$ and the proven fact that $S(y) \neq \emptyset$ for the piecewise constant functions y, we obtain that $S(y) \neq \emptyset$ for every $y \in \mathcal{Y}_{[t,T]}$.

3.2. To apply Lemma 4.4 to the mapping S, we first notice that $S(y(\cdot))$ is L_2-weakly closed (thanks to closedness of graph S) and therefore it is L_2-weakly compact, since $S(y(\cdot)) \subset \mathcal{U}_{[t,T]}$. Thanks to the claim in part 3.1, it is straightforward to prove that the mapping $S(\cdot)$ satisfies the nonanticipativity property in Lemma 4.4. Then there exists a nonanticipative selection $\alpha(\cdot)$ of $S(\cdot)$.

3.3. Let $y \in \mathcal{Y}_{[t,T]}$ be arbitrary, and let $(\cdot, E_y(\cdot), \nu)$ be the solution tube of (26) corresponding to $u(\cdot) \overset{def}{=} \alpha(y)(\cdot)$ and $y(\cdot)$. Since $u(\cdot) \in S(y(\cdot))$, we have that (27) is fulfilled for E_y. The set of those s for which

$$s < T(E_y(\cdot)) \quad \text{or} \quad \nu < G(s, E_y(s))$$

is bounded by T due to the definition of $\mathcal{M}$, and its supremum, $\bar{\tau}_y$, satisfies

$$\bar{\tau}_y \geq T(E_y(\cdot)) \quad \text{and} \quad \nu \geq G(\bar{\tau}_y, E_y(\bar{\tau}_y)).$$

The first inequality and the last part of Condition C.3 give

$$G(\tau_y, E_y(\tau_y)) \leq G(\bar{\tau}_y, E_y(\bar{\tau}_y)),$$

where $\tau_y \overset{def}{=} T(E_y(\cdot))$. Then

$$I_{\mathcal{E}}(t, E) \leq \sup_{y \in \mathcal{Y}_{[t,T]}} G(\tau_y, E_y(\tau_y)) \leq \sup_{y \in \mathcal{Y}_{[t,T]}} G(\bar{\tau}_y, E_y(\bar{\tau}_y)) \leq \nu.$$

This means $(t, E, \nu) \in \operatorname{epi} I_{\mathcal{E}}$, and proves $\mathcal{E}^*_\infty \subset \operatorname{epi} I_{\mathcal{E}}$. □

Remark 4.8 (on the implementation of Theorems 4.5 and 4.7). The last theorem shows that finding the value $I_{\mathcal{E}}$ is equivalent to the determination of suitable viability kernels. Notice that the viability kernels in (23) involve the complete information systems $\mathcal{L}_y$ only. The latter systems, however, are set-valued and the numerical algorithms developed in [31, 11, 27] (the last of which concerns also the case of viability with a target) are not directly applicable. At a conceptual level, we may summarize the algorithm for finding the viability kernel $\operatorname{Viab}_{\mathcal{L}}(\mathcal{E}; \mathcal{M})$ of a family of set-valued fields $\mathcal{L}$, with target $\mathcal{M}$ in the collection $\mathcal{E}$ (see subsection 3.1) as follows. For $h > 0$ and a natural number k, we define recursively

$$\mathcal{E}^h_{i+1} := \{E \in \mathcal{E}^h_i \setminus \mathcal{M};\ \exists L \in \mathcal{L}, \exists \widetilde{E} \in \mathcal{E}^h_i,\ (I + hL)(E) \subset \widetilde{E} + ch^2\mathcal{B}\} \cup \mathcal{M}, \quad \mathcal{E}^h_0 = \mathcal{E},$$

where $c \geq \lambda M/2$, and λ and M are local Lipschitz constant and bound for all $L \in \mathcal{L}$, respectively. Then

$$\operatorname{Viab}_{\mathcal{L}}(\mathcal{E}; \mathcal{M}) = \lim_{h \to 0+} \bigcap_i \mathcal{E}^h_i.$$

The above procedure should be applied to the family $\mathcal{L}_y = \{(1, f(\cdot, u, y, V(y)), 0);\ u \in U\}$ in (23) for each $y \in Y$.

The state-space discretization which certainly should be involved in a constructive procedure for approximating the viability kernel, as well as the corresponding convergence analysis, will not be discussed in the present paper.[5] Clearly, the algorithm has a high computational complexity, and to make it implementable, one has to choose a simple finitely parameterized collection $\mathcal{E}$.

We mention that Theorem 4.5 can also be used as a base for a numerical solution. The possibility of convergent time discretization is provided by the uniformity result in Proposition 3.2. In [25] we demonstrate the solvability of a discrete-time Hamilton–Jacobi–Bellman inequality of similar type, where sets of intervals were used for the collection $\mathcal{E}$. Essentially, the approach based on Theorem 4.5 coincides with a particular realization of the viability kernel algorithm.

4.3. Final discussions. 1. Since the control problem considered above includes minimal time problems, clearly the value function can be discontinuous, unless appropriate controllability conditions are fulfilled. If g is locally Lipschitz and the problem with fixed-end time is considered, then $I_{\mathcal{E}}$ is Lipschitz, provided that the solution tubes in $\mathcal{E}$ depend in a certain Lipschitz-like way on the data. Such dependence is established for some special collections in [29, Theorem 2].

2. Even in the case of a Lipschitz continuous value function, it may happen that (21) is not satisfied as an equality. To show this we adapt an example from [10]:

$$\dot{x} = u - y, \quad u, y \in [-1, 1], \quad g(x) = -|x|.$$

Here we may take $\mathcal{E} = \mathbf{R}^1$, since there are no unobservable uncertainties. From the consideration in [10], it follows that $I_{\mathcal{E}}(t, x) = -|x|$. The expression in the left-hand side of (21) equals zero for $x \neq 0$. For $x = 0$, however,

$$\begin{aligned} \sup_{y \in [-1,1]} \inf_{u \in [-1,1]} D^- I_{\mathcal{E}}(t, x; u - y) &= \sup_{y \in [-1,1]} \inf_{u \in [-1,1]} -|u - y| \\ &= \sup_{y \in [-1,1]} -(1 + |y|) = -1 < 0. \end{aligned}$$

[5] We refer to [31] for a state-space discretization technique and to [28] for a convergence analysis that could also be useful in the present context.

3. The consideration of the differential game in section 4 involves two conditions that need some more discussion. The first is Condition B.3, which significantly restricts the class of collections $\mathcal{E}$. In our opinion, the above theory can be developed also under weaker conditions for $\mathcal{E}$ (cf. [29]) which do not imply uniqueness of the solution tubes (in particular, including collections of ellipsoids). Such a consideration, however, would involve another "inf" (over the set of solution tubes, in the case of nonuniqueness) in the formulation (19) of the problem, and would bring essential technical complications that we choose to avoid in the present exposition.

The second "questionable" assumption is that equation (1) has the affine structure required in C.2. This assumption is, at least to a certain extent, essential. In particular, without affineness the existence of optimal strategy (Proposition 4.3) fails, even if the set $f(x, U, y, v)$ is convex. A counterexample can easily be constructed for the system in Example 2.5, section 2.3.

5. Proofs. In the proof of Proposition 3.2, we shall use the following lemma.

LEMMA 5.1. *Let the collection $\mathcal{E}$ satisfy Conditions* B.1–B.3*, and let $\mathcal{E}' \subset \mathcal{E}$ be inclusion-complete. Then for every $Z \in \mathrm{comp}(\mathbf{R}^n)$ and $E \in \mathcal{E}'$ there exists $E^* \in \mathcal{E}'$ such that*

$$\mathrm{dist}(Z, E^*) \le \mathrm{dist}(Z, E) \quad \textit{and} \quad \mathrm{dist}(E^*, Z) \le (L_{\mathcal{E}} + 1) d_H(Z, \mathcal{E}) + L_{\mathcal{E}} \,\mathrm{dist}(Z, E),$$

where

$$d_H(Z, \mathcal{E}) = \inf_{E \in \mathcal{E}} H(Z, E).$$

Proof. Let $E_0' \in \mathcal{E}$ be such that $H(Z, E_0') = d_H(Z, \mathcal{E})$. From Condition B.2 (with $\varepsilon = d_H(Z, \mathcal{E})$) there is $E_0 \in \mathcal{E}$ such that

$$Z \subset E_0' + d_H(Z, \mathcal{E})\mathcal{B} \subset E_0 \subset E_0' + L_{\mathcal{E}} d_H(Z, \mathcal{E})\mathcal{B}.$$

Then

$$Z \subset E_0 \subset E_0' + L_{\mathcal{E}} d_H(Z, \mathcal{E})\mathcal{B} \subset Z + (L_{\mathcal{E}} + 1) d_H(Z, \mathcal{E})\mathcal{B}. \tag{28}$$

Denote $Z' = \mathcal{P}_E(Z)$ (as before, $\mathcal{P}_E$ is the projection mapping over E). Since $Z' \subset Z + \mathrm{dist}(Z, E)\mathcal{B} \subset E_0 + \mathrm{dist}(Z, E)\mathcal{B}$, according to Condition B.2 there is $\tilde{E} \in \mathcal{E}$ such that

$$Z' \subset E_0 + \mathrm{dist}(Z, E)\mathcal{B} \subset \tilde{E} \subset E_0 + L_{\mathcal{E}} \,\mathrm{dist}(Z, E)\mathcal{B}. \tag{29}$$

We define $E^* = E \cap \tilde{E}$. Condition B.3 implies that the intersection of elements of $\mathcal{E}$ is also an element of $\mathcal{E}$. Hence $E^* \in \mathcal{E}$, but also $E^* \subset E \in \mathcal{E}'$, and the inclusion-completeness of $\mathcal{E}'$ implies $E^* \in \mathcal{E}'$.

Since Z' is contained both in E and in $\tilde{E}$, we have also $Z' \subset E^*$. Thus

$$\mathrm{dist}(Z, E^*) \le \mathrm{dist}(Z, Z') = \mathrm{dist}(Z, E).$$

Moreover, from (29) and (28),

$$\begin{aligned} \mathrm{dist}(E^*, Z) &\le \mathrm{dist}(\tilde{E}, Z) \le \mathrm{dist}(E_0, Z) + L_{\mathcal{E}} \,\mathrm{dist}(Z, E) \\ &\le (L_{\mathcal{E}} + 1) d_H(Z, \mathcal{E}) + L_{\mathcal{E}} \,\mathrm{dist}(Z, E). \end{aligned}$$

□

Now we prove Proposition 3.2.

Proof. 1. Let us fix the number R. Let L and M be a Lipschitz constant and a bound of $\mathcal{L}_{|(R+2)\mathcal{B}}$. We define

$$C = e^L(3 + LM) + M, \qquad c = M + 3, \qquad \beta = [M(L_{\mathcal{E}} + 1) + L_{\mathcal{E}} + 1]^{-1}. \tag{30}$$

2. Let us fix an arbitrary $\mathcal{E}'$ as in the formulation of the proposition. We fix also an arbitrary $h \in (0, \beta]$. According to Lemma 3.1, for every $E \in \mathcal{E}' \setminus \mathcal{M}$ there exist some $L_E(\cdot) \in \mathcal{L}$, $\sigma(E) \in (0, h^2]$ and $E' \in \mathcal{E}'$ such that

$$(I + \sigma(E)L_E(\cdot))(E) \subset E' + h\sigma(E)\mathcal{B}. \tag{31}$$

If $E \in \mathcal{M} \cap \mathcal{E}'$, we define $\sigma(E) = 0$.

For $Z \subset \mathbf{R}^n$ we denote similarly, as above,

$$\rho(Z, \mathcal{E}') = \inf_{E' \in \mathcal{E}'} \operatorname{dist}(Z, E').$$

For an arbitrary $Z \in \operatorname{comp}(\mathbf{R}^n)$, consider

$$\begin{aligned} \alpha(Z) = \sup\{\sigma(E);\ E \in \mathcal{E}',\ \operatorname{dist}(Z, E) \le \rho(Z, \mathcal{E}') + h\sigma(E), \quad \text{and} \\ \operatorname{dist}(E, Z) \le (L_{\mathcal{E}} + 1)d_H(Z, \mathcal{E}') + L_{\mathcal{E}}\rho(Z, \mathcal{E}') + \sigma(E)\}. \end{aligned} \tag{32}$$

First we shall prove that the set in the braces is nonempty. Let $E_k \in \mathcal{E}'$ be such that $\operatorname{dist}(Z, E_k) \le \rho(Z, \mathcal{E}') + 1/k$, $k = 1, 2, \ldots$. According to Lemma 5.1, there are $E_k^* \in \mathcal{E}'$ such that

$$\begin{aligned} \operatorname{dist}(Z, E_k^*) &\le \operatorname{dist}(Z, E_k) \le \rho(Z, \mathcal{E}') + \frac{1}{k}, \\ \operatorname{dist}(E_k^*, Z) &\le (L_{\mathcal{E}} + 1)d_H(Z, \mathcal{E}) + L_{\mathcal{E}} \operatorname{dist}(Z, E_k) \\ &\le (L_{\mathcal{E}} + 1)d_H(Z, \mathcal{E}) + L_{\mathcal{E}} \left(\rho(Z, \mathcal{E}') + \frac{1}{k} \right). \end{aligned}$$

The last inequality implies that the sequence E_k^* is bounded; therefore it has a convergent subsequence (indexed again by k) with limit E^* which belongs to $\mathcal{E}'$ since the latter is closed. Passing to a limit in the above two inequalities, we obtain (since $d_H(Z, \mathcal{E}) \le d_H(Z, \mathcal{E}')$)

$$\operatorname{dist}(Z, E^*) \le \rho(Z, \mathcal{E}') \quad \text{and} \quad \operatorname{dist}(E^*, Z) \le (L_{\mathcal{E}} + 1)d_H(Z, \mathcal{E}') + L_{\mathcal{E}}\rho(Z, \mathcal{E}').$$

Thus $\sigma(E^*)$ belongs to the set in the right-hand side of (32). Moreover, if $E^* \notin \mathcal{M}$, then $\alpha(Z) \ge \sigma(E^*) > 0$.

Then for every compact Z there exists $\mathcal{F}(Z) \in \mathcal{E}'$ such that

$$\operatorname{dist}(Z, \mathcal{F}(Z)) \le \rho(Z, \mathcal{E}') + h\sigma(\mathcal{F}(Z)), \tag{33}$$

$$\operatorname{dist}(\mathcal{F}(Z), Z) \le (L_{\mathcal{E}} + 1)d_H(Z, \mathcal{E}') + L_{\mathcal{E}}\rho(Z, \mathcal{E}') + \sigma(\mathcal{F}(Z)), \tag{34}$$

$$\sigma(\mathcal{F}(Z)) \ge \frac{1}{2}\alpha(Z). \tag{35}$$

3. Let us fix an arbitrary $E_0 \in \mathcal{E}'$ such that $E_0 \subset R\mathcal{B}$ and $\rho(E_0, \mathcal{M}) > ch$. We define a sequence $E_0, E_1, \ldots$ of elements of $\mathcal{E}'$ by

$$\sigma_k = \sigma(E_k), \quad Z_k = (I + \sigma_k L_{E_k}(\cdot))(E_k), \quad E_{k+1} = \mathcal{F}(Z_k).$$

We terminate the sequence at σ_{N-1}, Z_{N-1}, E_N if at least one of the following conditions fails to be fulfilled:

(i) $\sum_{i=0}^{N}\sigma_i \le h$;
(ii) $E_N \subset (R+1)\mathcal{B}$;
(iii) $E_N \notin \mathcal{M}$.

From (31) and from the definition of Z_k, we obtain, respectively,

$$\rho(Z_k,\mathcal{E}') \le \rho((I+\sigma_k L_{E_k}(\cdot))(E_k),\mathcal{E}') \le h\sigma_k, \tag{36}$$

$$H(Z_k,E_k) \le M\sigma_k, \quad d_H(Z_k,\mathcal{E}') \le M\sigma_k. \tag{37}$$

We shall prove that (ii) and (iii) cannot fail earlier than (i). Suppose that $E_i \subset (R+1)\mathcal{B}$ for all $i=0,\dots,k<N$, and let (i) be fulfilled for N. Then from (34), (36), and (37) we have

$$\begin{aligned} E_{k+1} = \mathcal{F}(Z_k) &\subset Z_k + (L_{\mathcal{E}}+1)d_H(Z_k,\mathcal{E}')\mathcal{B} + L_{\mathcal{E}}\rho(Z_k,\mathcal{E}')\mathcal{B} + \sigma_{k+1}\mathcal{B} \\ &\subset E_k + \sigma_k M\mathcal{B} + (L_{\mathcal{E}}+1)\sigma_k M\mathcal{B} + L_{\mathcal{E}}h\sigma_k\mathcal{B} + \sigma_{k+1}\mathcal{B} \\ &\subset E_k + [(L_{\mathcal{E}}+2)M + L_{\mathcal{E}}]\sigma_k\mathcal{B} + \sigma_{k+1}\mathcal{B}. \end{aligned}$$

From here and (i), we obtain, inductively,

$$E_{k+1} \subset E_0 + [(L_{\mathcal{E}}+2)M + L_{\mathcal{E}} + 1]\sum_{i=0}^{N}\sigma_i\mathcal{B} \subset E_0 + \frac{h}{\beta}\mathcal{B} \subset (R+1)\mathcal{B}. \tag{38}$$

Hence (ii) is satisfied as long as (i) holds.

Using (33) and (36), we obtain

$$Z_k \subset \mathcal{F}(Z_k) + (\rho(Z_k,\mathcal{E}') + h\sigma_{k+1})\mathcal{B} \subset E_{k+1} + (\sigma_k + \sigma_{k+1})h\mathcal{B}. \tag{39}$$

From (37) and (39),

$$E_k \subset Z_k + M\sigma_k\mathcal{B} \subset E_{k+1} + ((M+1)\sigma_k + \sigma_{k+1})\mathcal{B}$$

and, inductively,

$$E_0 \subset E_N + (M+2)\sum_{i=0}^{N}\sigma_i\mathcal{B} \subset E_N + h(M+2)\mathcal{B}.$$

If we assume that (iii) fails at N, then

$$\rho(E_0,\mathcal{M}) \le \rho(E_N,\mathcal{M}) + \operatorname{dist}(E_0,E_N) \le h(M+2) + \rho(E_N,\mathcal{M}) \le ch. \tag{40}$$

This contradicts the choice of E_0 with $\rho(E_0,\mathcal{M}) > ch$; therefore (iii) is fulfilled as long as (i) holds.

4. We shall prove that there is a smallest number N for which (i) fails. If such N does not exist, then $\sum_{i=0}^{\infty}\sigma_i$ is finite. In particular, $\sigma_k \to 0$, and according to (35), $\alpha(Z_k) \le 2\sigma_{k+1} \to 0$. Since E_k and Z_k are uniformly bounded (in the sets $(R+1)\mathcal{B}$ and $(R+2)\mathcal{B}$, correspondingly), there is a convergent subsequence (indexed again by k) $E_k \to \bar{E} \in \mathcal{E}'$ and $Z_k \to \bar{Z} \in \operatorname{comp}(\mathbf{R}^n)$. Since, from (40), we have

$$\rho(E_k,\mathcal{M}) \ge \rho(E_0,\mathcal{M}) - (M+2)h \ge (c-(M+2))h = h,$$

we obtain $\rho(\bar{E},\mathcal{M}) \ge h$; therefore $\sigma(\bar{E}) > 0$. Since $\alpha(Z_k) \to 0$, we have $\alpha(Z_k) \le \sigma(\bar{E})$ for all sufficiently large k. Hence from the definition of $\alpha(Z_k)$, we obtain that for all sufficiently large k at least one of the following relations holds:

$$\operatorname{dist}(Z_k,\bar{E}) > \rho(Z_k,\mathcal{E}') + h\sigma(\bar{E}),$$
$$\operatorname{dist}(\bar{E},Z_k) > (L_{\mathcal{E}}+1)d_H(Z_k,\mathcal{E}') + L_{\mathcal{E}}\rho(Z_k,\mathcal{E}') + \sigma(\bar{E}).$$

Passing to a limit, we obtain that $\operatorname{dist}(\bar{Z},\bar{E}) \geq h\sigma(\bar{E}) > 0$ or $\operatorname{dist}(\bar{E},\bar{Z}) \geq \sigma(\bar{E}) > 0$. On the other hand, passing to the limit in the inequality $H(Z_k, E_k) \leq \sigma_k M$, we obtain $\bar{E} = \bar{Z}$. This contradiction proves the existence of a smallest number N for which (i) fails.

5. If N is the number defined above, then

$$\sum_{i=0}^{N-1} \sigma_i \leq h, \quad h < \sum_{i=0}^{N} \sigma_i \leq h + h^2. \tag{41}$$

From (39) we have

$$Z_k = (I + \sigma_k L_{E_k}(\cdot))(E_k) \subset E_{k+1} + h(\sigma_k + \sigma_{k+1})\mathcal{B}.$$

Now we shall apply Lemma 4 of [28] for the sets E_k and $S = R\mathcal{B}$, and for $\delta = 1$. The conditions are fulfilled since $M\sum_{i=0}^{N-1} \sigma_i \leq Mh \leq 1$ (see (30)). The lemma gives

$$\begin{aligned}
&\left(I + \sum_{i=0}^{N-1} \sigma_i L_{E_i}(\cdot)\right)(E_0) \\
&\quad \subset E_N + e^{L\sum_{i=0}^{N-1}\sigma_i}\left(h\left(\sum_{i=0}^{N-1}\sigma_i + \sum_{i=1}^{N}\sigma_i\right) + LM\left(\sum_{i=0}^{N-1}\sigma_i\right)^2\right)\mathcal{B} \\
&\quad \subset E_N + h^2 e^L(3 + LM)\mathcal{B}.
\end{aligned}$$

Denote

$$L_h(\cdot) = \frac{1}{\sum_{i=0}^{N-1}\sigma_i} \sum_{i=0}^{N-1} \sigma_i L_{E_i}(\cdot) \in \mathcal{L}.$$

Then

$$\left(I + \left(\sum_{i=0}^{N-1}\sigma_i\right) L_h(\cdot)\right)(E_0) \subset E_N + h^2 e^L(3 + LM)\mathcal{B}.$$

Using (41), we obtain

$$\begin{aligned}
(I + hL_h(\cdot))(E_0) &\subset E_N + h^2 e^L(3 + LM)\mathcal{B} + \left(h - \sum_{i=0}^{N-1}\sigma_i\right) M\mathcal{B} \\
&\subset E_N + (h^2 e^L(3 + LM) + \sigma_N M)\mathcal{B} \\
&\subset E_N + (h^2 e^L(3 + LM) + h^2 M)\mathcal{B} \\
&= E_N + Ch^2\mathcal{B},
\end{aligned}$$

according to the definition of C. Notice that $E_N \subset (R+1)\mathcal{B}$ according to (38). The proof is complete since $E_N \in \mathcal{E}'$. □

The proof of Theorem 3.5 follows.

Proof. 1. Let $\mathcal{E}'$ enjoy the viability property with target $\mathcal{M}$. Let us fix an arbitrary $E \in \mathcal{E}' \setminus \mathcal{M}$. There exists $u(\cdot) \in \mathcal{U}_{[0,+\infty)}$ such that $E(t) \stackrel{def}{=} E_{u(\cdot)[0,E]}(t) \in \mathcal{E}'$ on some interval $[0,T]$ with $T > 0$. Then

$$X_{u(\cdot)}[0,E](t) \subset E(t) \in \mathcal{E}' \quad \forall t \in [0,T]. \tag{42}$$

Let us define $\bar{u} \in U$ as an arbitrary element from the set

$$\operatorname*{Limsup}_{h\to 0+} \frac{1}{h}\int_0^h u(s)\,\mathrm{d}s$$

and let $h_k \to 0+$ be the subsequence corresponding to $\bar{u}$. In a standard way, we obtain that

$$H(X_{\bar{u}}[0,E](h_k),\ (I + h_k F_{\bar{u}}(\cdot))(E)) \le C_1 h_k^2,$$

where C_1 is independent of k. On the other hand, it is also standard to prove (thanks to the convexity of $F_0(x)$ and the linear dependence on u) that

$$H(X_{u(\cdot)}[0,E](h_k), X_{\bar{u}}[0,E](h_k)) \le C_2 h_k^2 + C_3 h_k \left|\bar{u} - \frac{1}{h_k}\int_0^{h_k} u(s)\,\mathrm{d}s\right| = C_2 h_k^2 + h_k \gamma_k$$

with $\gamma_k \to 0$. Then

$$(I + h_k F_{\bar{u}}(\cdot))(E) \subset X_{u(\cdot)}[0,E](h_k) + h_k[(C_1 + C_2)h_k + \gamma_k]\mathcal{B},$$

which, together with (42) and Lemma 3.1, implies that $F_{\bar{u}}(\cdot) \in \mathcal{T}_{\mathcal{E}'}(E)$. Thus $\mathcal{E}'$ is a viability domain with target $\mathcal{M}$.

2. Now let $\mathcal{E}' \subset \mathcal{E}$ be an inclusion-complete viability domain with target $\mathcal{M}$. We fix an arbitrary $E_0 \in \mathcal{E}'$. According to Theorem 2.1, there is a number $R > 1$ such that $E_{u(\cdot)}[0,E_0](t) \subset (R-1)\mathcal{B}$ for every $u(\cdot) \in \mathcal{U}_{[0,1]}$ and $t \in [0,1]$. By a standard argument there exist constants S and $h^* > 0$ (depending only on the growth constant a and on the Lipschitz constant of F_u on $R\mathcal{B}$) such that

$$H(X_{u(\cdot)}[0,Z](h),\ (I + hF_{u(\cdot)}(\cdot))(Z)) \le Sh^2$$

for every $Z \subset R\mathcal{B}$ and $h \in [0,h^*]$. Moreover, the mapping $X_{u(\cdot)}[0,Z](\cdot)$ is Lipschitz continuous on $[0,h^*]$ with a Lipschitz constant L_x (depending on the data like S), provided that $Z \subset R\mathcal{B}$. We define also the constants $D \stackrel{def}{=} L_{\mathcal{E}}(L_x + 1)$ (where $L_{\mathcal{E}}$ is the constant from Condition B.2 for $\mathcal{E}$).

We apply Proposition 3.2 with the above number R and the fixed $\mathcal{E}'$. There exist positive β, c, and C such that for every $h \in (0,\beta]$ and for every $E \in \mathcal{E}'$, such that $E \subset R\mathcal{B}$ and $\rho(E,\mathcal{M}) \ge ch$, there exists $u = u[E,h] \in U$ and $\tilde{E} = \tilde{E}[E,h] \in \mathcal{E}'$ for which

$$(I + hF_u(\cdot))(E) \subset \tilde{E} + Ch^2\mathcal{B}.$$

Now we start proving the existence of a "viable" solution tube from E_0 with target $\mathcal{M}$. Clearly, if $E_0 \in \mathcal{M}$, there is nothing to prove. Below we shall let the positive number h vary between 0 and a number $\bar{h}$ chosen in such a way that $\bar{h} \le \beta$, $c\bar{h} < \rho(E_0,\mathcal{M})$, $\bar{h} \le h^*$, $2\bar{h} \le 1$, and $(C+S)\bar{h} \le 1$.

For a fixed (for the moment) $h \in (0,\bar{h}]$, we define

$$t_0 = 0, \quad t_0' = 0, \quad u_h(t_0') = \bar{u}, \quad Y_h(t_0') = E_0,$$

where $\bar{u}$ is an arbitrarily fixed element of U. Then we perform the following recursive construction starting with $i = 1$, and supposing that

(i) $t_{i-1}' + h + (C+S)h^2 \le 1$;

(ii) $E_{i-1} \subset R\mathcal{B}$;
(iii) $\rho(E_{i-1}, \mathcal{M}) \geq ch$;
(iv) $Y_h(t'_{i-1}) \subset E_{i-1} \subset \mathcal{E}'$.
Define

$$t_i = t'_{i-1} + h, \quad t'_i = t_i + (C+S)h^2, \quad u_h(t) = u[E_{i-1}, h], \quad Y_h(t) = X_{u_h}[t_{i-1}, E_{i-1}](t)$$

for $t \in [t'_{i-1}, t_i]$, which is possible thanks to (i)–(iii). Then, again using (ii),

$$Y_h(t_i) \subset (I + hF_{u_h}(\cdot))(E_{i-1}) + Sh^2\mathcal{B} \subset \tilde{E}[E_{i-1}, h] + h^2(C+S)\mathcal{B}. \tag{43}$$

Then we extend $u_h(\cdot)$ and the mapping Y_h for $t \in (t_i, t'_i]$ by the formula

$$u_h(t) = \bar{u}, \qquad Y_h(t) = \bigcup_{y \in Y_h(t_i)} \left(\frac{t'_i - t}{t'_i - t_i} y + \frac{t - t_i}{t'_i - t_i} \mathcal{P}_{\tilde{E}[E_{i-1}, h]}(y) \right),$$

where, as before, $\mathcal{P}$ is the projection operator. Obviously $Y_h(t'_i) = \mathcal{P}_{\tilde{E}[E_{i-1}, h]}(Y_h(t_i)) \subset \tilde{E}[E_{i-1}, h]$. It is easy to check that the mapping Y_h is Lipschitz continuous with constant 1 on $[t_i, t'_i]$, thanks to the inclusion (43) (this is, in fact, proven in [1], where the above "linear" interpolation was proposed). Thus Y_h is now Lipschitz on $[0, t'_i]$ with constant $L'_x = \max\{L_x, 1\}$, thanks to (ii). Moreover,

$$H(Y_h(t'_{i-1}), Y_h(t'_i)) \leq L_x h + (C+S)h^2.$$

Since, according to (iv), $Y_h(t'_{i-1}) \subset E_{i-1}$, Conditions B.2 and B.3 together imply that the minimal element E_i from $\mathcal{E}$ that contains $Y_h(t'_i)$ satisfies

$$Y_h(t'_i) \subset E_i \subset E_{i-1} + L_{\mathcal{E}}(L_x h + (C+S)h^2)\mathcal{B} \subset E_{i-1} + hD\mathcal{B}. \tag{44}$$

Since $Y_h(t'_i) \subset \tilde{E}[E_{i-1}, h] \in \mathcal{E}' \subset \mathcal{E}$, we have $E_i \subset \tilde{E}[E_{i-1}, h]$, and since the collection $\mathcal{E}'$ is inclusion-complete in $\mathcal{E}$, we obtain that $E_i \in \mathcal{E}'$. Thus we ensure (iv) for the next step.

This completes the description of the construction. Thanks to the suppositions for $\bar{h}$, one can perform at least one step. One can repeat the same procedure as long as the relations (i)–(iii) hold. Thus the procedure terminates when one of these relations is violated; let $i = N = N(h)$ be the last integer for which (i)–(iii) still hold (as obviously exists, since (i) is violated for $i > 1/h$). Inclusion (44) implies that either $t'_N \geq 1$, or $Nh \geq 1/D$, or $\rho(E_N, \mathcal{M}) \leq ch$. In each case, when $h \to 0$, the sequence $t'_{N(h)}$ has a strictly positive lower limit T, and either $T \geq T_0 \stackrel{def}{=} \min\{1, 1/D\}$ or $\rho(E^h_{N(h)}, \mathcal{M}) \to 0$ for a subsequence (here we add the superscript h to E_i to indicate the dependence on h).

The sequence $Y_h(\cdot)$ is equi-Lipschitz (with Lipschitz constant L'_x) and uniformly bounded; therefore there exists a subsequence (for which also $t'_{N(h)} \to T$) converging to some Lipschitz tube $Y(\cdot)$ on $[0, T]$. Because of (iv) and the closedness of $\mathcal{E}'$, we obtain that $Y(t) \subset \mathcal{E}'$ for every $t \in [0, T]$.

Since $\mathrm{meas}(\cup_i [t_i, t'_i]) \leq \mathrm{const}.h$ and on every $[t'_{i-1}, t_i]$ the tube Y_h is invariant with respect to (14) with $u(t) = u_h(t)$, one can prove that $Y(\cdot)$ is an invariant tube of (14) for some $u_0(\cdot) \in \mathcal{U}_{[0,T]}$ (which is an L_1-weak limit of a subsequence of $u_h(\cdot)$). This is proven in detail in [28, Lemma 7]. Thus we obtain that $Y(\cdot)$ is an invariant tube in $\mathcal{E}'$ starting from E_0; therefore it contains the solution tube $E_{u_0(\cdot)}(\cdot)$ in $\mathcal{E}$ starting

from E_0 (see Corollary 2.2). From the inclusion-completeness of $\mathcal{E}'$, we conclude that $E_{u_0(\cdot)}(\cdot)$ is a tube in $\mathcal{E}'$, and since it is the solution tube in the larger collection $\mathcal{E}$, it is the solution tube also in $\mathcal{E}'$.

If $T < 1$ and $E_{u_0(\cdot)}(T) \notin \mathcal{M}$, then by the definition of R, we see that $E_{u(\cdot)}[T, E_{u_0(\cdot)}(T)](t) \subset (R-1)\mathcal{B}$ for every $u(\cdot) \in \mathcal{U}_{[T,1]}$ and $t \in [T,1]$. Therefore the same construction can be repeated with the same constants. Thus in a finite number of steps, the solution tube $E_{u_0(\cdot)}(\cdot)$ will either satisfy $E_{u_0(\cdot)}(t) \in \mathcal{M}$ for some $t \in [0,1]$ or will be defined on $[0,1]$. This, together with the growth condition and the estimation in Theorem 2.1, implies the claim of the theorem. □

REFERENCES

[1] Z. Artstein, *Piecewise linear approximations of set-valued maps*, J. Approx. Theory, 56 (1989), pp. 41–47.

[2] J.-P. Aubin, *Viability Theory*, Birkhäuser, Boston, 1991.

[3] J.-P. Aubin, *Mutational equations in metric spaces*, Set-Valued Anal., 1 (1993), pp. 3–46.

[4] J.-P. Aubin, *Mutational and Morphological Analysis. Tools for Shape Evolution and Morphogenesis*, Systems Control Found. Appl., Birkhäuser, Boston, 1999.

[5] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Systems Control Found. Appl., Birkhäuser, Boston, 1990.

[6] J. S. Baras and N. S. Patel, *Robust control of set-valued discrete-time dynamical systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 61–75.

[7] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems*, Birkhäuser, Boston, 1995.

[8] J. S. Baras and M. R. James, *Partially observed differential games, infinite-dimensional Hamilton–Jacobi–Isaacs equations, and nonlinear H_∞ control*, SIAM J. Control Optim., 34 (1996), pp. 1342–1364.

[9] P. Bernhard, *A discrete-time min-max certainty equivalence principle*, Systems Control Lett., 24 (1995), pp. 229–234.

[10] P. Cardaliaguet, *A differential game with two players and one target*, SIAM J. Control Optim., 34 (1996), pp. 1441–1460.

[11] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Set-valued numerical analysis for optimal control and differential games*, in Stochastic and Differential Games, Ann. Internat. Soc. Dynam. Games 4, Birkhäuser, Boston, 1999, pp. 177–247.

[12] F. L. Chernousko, *Estimation of the Phase State of Dynamical Systems*, Nauka, Moscow, 1988 (in Russian).

[13] F. L. Chernousko and D. Ya. Rokityanskii, *Ellipsoidal bounds on reachable sets of dynamical systems with matrices subjected to uncertain perturbations*, J. Optim. Theory Appl., 104 (2000), pp. 1–19.

[14] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[15] A. Dontchev and F. Lempio, *Difference methods for differential inclusions: A survey*, SIAM Rev., 34 (1992), pp. 263–294.

[16] L. Doyen, *Filippov and invariance theorems for mutational equations for tubes*, Set-Valued Anal., 1 (1993), pp. 289–303.

[17] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[18] A. Gorre, *Evolutions of tubes under operability constraints*, J. Math. Anal. Appl., 216 (1997), pp. 1–22.

[19] G. Häckl, *Numerical approximation of reachable sets and control sets*, Random Comput. Dynam., 1 (1992–93), pp. 371–394.

[20] L. Jaulin, M. Kieffer, I. Braems, and E. Walter, *Guaranteed non-linear estimation using constraint propagation on sets*, Internat. J. Control, 74 (2001), pp. 1772–1782.

[21] A. B. Kurzhanski and T. F. Filippova, *On the theory of trajectory tubes—a mathematical formalism for uncertain dynamics, viability and control*, in Advances in Nonlinear Dynamics and Control: A Report from Russia, Progr. Systems Control Theory 17, Birkhäuser, Boston, 1993, pp. 122–188.

[22] A. B. Kurzhanski and I. Vályi, *Ellipsoidal Calculus for Estimation and Control*, Birkhäuser, Boston, 1997.

[23] A. B. Kurzhanski and P. Varaiya, *Ellipsoidal techniques for reachability analysis: Internal approximation*, Systems Control Lett., 41 (2000), pp. 201–211.

[24] F. Lempio and V. M. Veliov, *Discrete approximations to differential inclusions*, GAMM Mitt. Ges. Angew. Math. Mech., 21 (1998), pp. 101–135.

[25] R. Moitié, M. Quincampoix, and V. M. Veliov, *Optimal control of discrete-time uncertain systems with imperfect measurement*, IEEE Trans. Automat. Control, 47 (2002), pp. 1909–1914.

[26] A. I. Panasyuk, *Quasidifferential equations in a complete metric space under conditions of the Caratheodory type.* I, Differential Equations, 31 (1995), pp. 901–910 (translation from Differ. Uravn., 31 (1995), pp. 962–972).

[27] M. Quincampoix and V. M. Veliov, *Viability with a target: Theory and applications*, in Applications of Mathematics in Engineering, B. I. Cheshankov and M. D. Todorov, eds., Heron Press, Sofia, 1998, pp. 47–54.

[28] M. Quincampoix and V. M. Veliov, *Open-loop viable control under uncertain initial state information*, Set-Valued Anal., 7 (1999), pp. 55–87.

[29] M. Quincampoix and V. M. Veliov, *Solution tubes to differential equations within a collection of sets*, Control Cybernet., 31 (2002), pp. 847–862.

[30] A. Rapaport and P. Bernhard, *On a planar pursuit game with imperfect knowledge of a coordinate*, Automatique Productique Informatique Industrielle, 29 (1995), pp. 575–601 (in French).

[31] P. Saint-Pierre, *Approximation of the viability kernel*, Appl. Math. Optim., 29 (1994), pp. 187–209.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1400–1434

UNIQUE CONTINUATION AND CONTROL FOR THE HEAT EQUATION FROM AN OSCILLATING LOWER DIMENSIONAL MANIFOLD*

CARLOS CASTRO† AND ENRIQUE ZUAZUA‡

Abstract. We consider the linear heat equation with Dirichlet boundary conditions in a bounded domain of $\mathbb{R}^n$, $n \geq 1$, and with a control acting on a lower-dimensional time-dependent manifold of dimension $k \leq n-1$. We analyze the approximate controllability problem. This problem is equivalent to a suitable uniqueness or unique continuation property of solutions of the heat equation without control. More precisely, it consists of proving that the unique solution of the Dirichlet problem vanishing on the time-dependent manifold is identically zero. This uniqueness problem, however, does not fit in the class of classical Cauchy problems and therefore, the existing tools based on power series expansions, Carleman inequalities, and doubling properties do not seem to apply. We give sufficient conditions on the time-dependent manifold for this uniqueness property to hold. The techniques we employ combine the Fourier series representation and the time analyticity of solutions and allow us to reduce the problem to a uniqueness question for the eigenfunctions of the Laplacian. We then apply well-known results on the nodal sets of these eigenfunctions.

We also analyze the asymptotic behavior of the control when the time-oscillation of the manifold supporting the control increases. When the frequency of oscillation tends to infinity we prove that the controls converge to an approximate control for the same heat equation but on a manifold of dimension $k+1$ that is independent of time. This is done under suitable time-periodicity assumptions on the original manifold and confirms the fact that increasing time-oscillations of the support of the control increases the efficiency of the control mechanism.

Key words. approximate controllability, heat equation, lower-dimensional manifold

AMS subject classifications. 93C20, 35B37, 35B60

DOI. 10.1137/S0363012903430317

1. Introduction.

This paper is devoted to a study of the properties of approximate controllability for the linear, constant coefficient heat equation in a bounded domain of $\mathbb{R}^n$, $n \geq 1$, with Dirichlet boundary conditions and with a control acting on a time-dependent lower-dimensional manifold of dimension $k \leq n-1$.

The main novelty of the analysis carried out in this paper lies precisely in the fact that the control acts on a lower-dimensional manifold that moves in time.

The problem under consideration is rather natural. Indeed, in the ultimate goal of optimally controlling a given system with the minimal amount of control it is natural to consider controls located, for instance, in a single point or on a finite collection of points. This is the so-called pointwise control problem (see Lions [L3]). However, in that case the system may easily fail to be controllable. That is the case, for instance, when the support of the control is located on a nodal set of an eigenfunction of the Laplacian. As a consequence of that, the property of pointwise controllability, even if it holds on some geometric configurations (when the support is chosen in an appropriate way), is extremely sensitive to the location of the controller.

*Received by the editors June 24, 2003; accepted for publication (in revised form) May 11, 2004; published electronically January 27, 2005. Supported by grants BFM2002-03345 of the MCYT (Spain) and the TMR projects of the EU "Homogenization and Multiple Scales" and "New materials, adaptive systems and their nonlinearities: modelling, control and numerical simulations."

http://www.siam.org/journals/sicon/43-4/43031.html

† Dep. Matemática e Informática, ETSI Caminos, Canales y Puertos, Univ. Politécnica de Madrid, 28040 Madrid, Spain (ccastro@caminos.upm.es).

‡Dep. Matemáticas, Facultad de Ciencias, Univ. Autónoma de Madrid, 28049 Madrid, Spain (enrique.zuazua@uam.es).

On the contrary, when the control is located on an open subset of the domain, approximate controllability holds as a consequence of Holmgren's uniqueness theorem without any geometric restriction. In fact, in that case, a much stronger property, the so-called null-controllability property, holds (see [FI], [LR], or [FZ] among others).

The case where the control is supported on a manifold of dimension $k \leq n-1$ is an intermediate situation between the problem of pointwise control and the control on an open subset. However, when the support of the control is independent of time the situation is quite similar to the one encountered when dealing with the pointwise control problem: If the manifold is contained on the nodal set of some eigenfunction, the approximate controllability property fails.

Berggrem [B] pointed out that a possible way of enhancing the control property when the control was supported in a manifold of dimension $k \leq n-1$ was to make this support oscillate in time more and more. In [B] some numerical evidence of this fact for the one-dimensional (1-d) case was also given.

This paper is devoted to analytically investigating this issue. We consider, in particular, the following two problems.

Problem 1*:* To give sufficient conditions on a time-dependent manifold to ensure the approximate controllability of the heat equation.

Problem 2*:* To analyze the asymptotic limit of the controllers when the time-frequency of the oscillations of their supports tends to infinity.

Let us briefly describe the results we obtain and the techniques we employ. Concerning the first problem, using the time-analyticity of solutions of the heat equation and the Fourier development of solutions, under suitable assumptions on the time-evolution of the manifolds where the control is supported, the problem is reduced to a unique continuation question on the eigenfunctions of the Laplacian that we solve applying classical results on its nodal sets. At this point, it should be noted that we do not fully use the existing results on the size of the nodal sets of eigenfunctions ([DF], [L], [JL]) whose consequences in the context of approximate control of the heat equation remain to be investigated.

Once the heat equation is known to be approximately controllable (i.e., once Problem 1 is solved) the control can be characterized as the minimum of a suitable quadratic functional over the set of solutions of the adjoint heat equation as in [FPZ].

We then address Problem 2. More precisely, assuming that the manifold in which the control is located is time-periodic and that it satisfies the requirements of Problem 1 to guarantee approximate controllability, we investigate the behavior of the controls as the time-frequency (described by a parameter $1/\varepsilon$) tends to infinity. The control then undergoes a *homogenization* process (see [DN] and [Z1] for similar situations in different problems). In the limit we get controls distributed on a time-independent manifold of dimension $k+1$ modulated by a density factor varying along the manifold. This result does show indeed that time oscillations on the support of the control enhance the controllability properties.

The proof of this second result uses the characterization of controls as minima of suitable quadratic functionals, Γ-convergence arguments, and a careful analysis of the behavior as $\varepsilon \to 0$ of traces of solutions of the heat equation along rapidly oscillating manifolds.

Let us now state more precisely the problems we shall address and introduce the notation we shall employ.

2. Problem formulation. Let Ω be a bounded smooth domain of $\mathbb{R}^n$ ($n = 1, 2, 3$), $\partial\Omega$ its boundary, $T > 0$, $Q = \Omega \times (0, T)$, and $\Sigma = \partial\Omega \times (0, T)$. We consider the linear heat equation with an interior control $f(x, t)$ which acts in an open subset $\omega \subset \Omega$:

$$\begin{cases} u_t - \Delta u = f(x,t)\chi_\omega(x) & \text{in } Q, \\ u = 0 & \text{on } \Sigma, \\ u(x,0) = u^0(x) & \text{in } \Omega. \end{cases} \tag{2.1}$$

Here χ_ω represents the characteristic function of the region ω.

Given any $T > 0$ and any open subset $\omega \subset \Omega$, the following approximate controllability property is known to hold: *For any initial data $u^0 \in L^2(\Omega)$, any final data $u^1 \in L^2(\Omega)$, and any $\alpha > 0$ there exists a control $f(x, t) \in L^2(\omega \times (0, T))$ such that the solution u of system* (2.1) *satisfies*

$$\|u(x,T) - u^1\|_{L^2(\Omega)} \leq \alpha. \tag{2.2}$$

Moreover, it is known (see [FPZ]) that the optimal control (the one with minimal L^2-norm) can be obtained by minimizing a suitable continuous, convex, and coercive functional over the space of solutions of the following adjoint system endowed with the L^2-norm of its datum at $t = T$:

$$\begin{cases} -\varphi_t - \Delta\varphi = 0 & \text{in } \Omega \times (0,T), \\ \varphi(T) = \varphi^0 & \text{in } \Omega, \\ \varphi = 0 & \text{on } \partial\Omega \times (0,T). \end{cases} \tag{2.3}$$

On the other hand, it is well known that the approximate controllability property above is equivalent to the following uniqueness property of (2.3): If $\varphi = 0$ in $\omega \times (0, T)$, then $\varphi \equiv 0$. In this case, this uniqueness property does hold as a consequence of Holmgren's theorem.

This paper is devoted to analyzing these questions when the open subset ω of Ω is replaced by a lower dimensional continuous manifold $\gamma(t) \subset \Omega$. In fact, in view of the fact that, in practice, the support of the control needs to be very small compared to the total size of the domain Ω it is very natural to consider the control to be located in such lower-dimensional manifolds. When the manifold $\gamma \subset \Omega$ is independent of time the corresponding control system reads as follows:

$$\begin{cases} u_t - \Delta u = f(x,t)\delta_\gamma(x) & \text{in } Q, \\ u = 0 & \text{on } \Sigma, \\ u(x,0) = u^0(x) & \text{in } \Omega, \end{cases} \tag{2.4}$$

where $\delta_\gamma(x)$ represents the Dirac measure on γ. Here γ can be a point, a curve if $n \geq 2$, or a surface if $n = 3$, for instance. When γ is a point we consider it as a manifold of dimension zero.

It turns out that the approximate controllability property of system (2.4) depends on the location of γ. Indeed, the problem of approximate controllability for system (2.4) can be reduced to a uniqueness problem for the adjoint system (2.3):

$$\varphi = 0 \text{ on } \gamma \Rightarrow \quad \varphi \equiv 0.$$

Using the Fourier series representation of solutions of system (2.3) it can be shown that these properties hold if and only if the only eigenfunction of the Laplacian with

homogeneous Dirichlet boundary conditions and vanishing on γ is the identically zero one. In what follows, the manifolds γ for which this spectral property is satisfied will be referred to as *strategic manifolds*.

The property of γ being strategic is difficult to establish in practice since it is extremely unstable. For example, if $\Omega = (0,1)$ with $n = 1$, then $\gamma = x_0 \in \Omega$ is strategic if and only if it is irrational. In general, k-dimensional manifolds in Ω with $k < n$ are generically strategic. But, by the contrary, γ fails to be strategic if it is contained in a nodal set of any of the eigenfunctions of the Laplacian. Consequently, controllability properties over low-dimensional manifolds are hard to use in practice. At this point it is worth noting that the strategic property for a k-dimensional manifold is obviously more likely to hold when the dimension k is larger.

To overcome this difficulty one may consider controls supported on moving (in time) manifolds $\{\gamma(t)\}_{0\leq t\leq T}$.

The main advantage of moving controls is that it is easy to construct families $\{\gamma(t)\}_{0\leq t\leq T}$ for which the strategic property holds for $\gamma(t)$ a.e. in $t \in [0,T]$. For example, this is the case in the 1-d example above when we assume that the control is located at a point that moves continuously in time. In this case, $\gamma(t)$ is irrational, and therefore strategic, a.e. in $t \in [0,T]$. Therefore, the approximate controllability is likely to hold for such moving controls. A previous result in this direction is given in [K] where the approximate controllability for the 1-d case is proved when considering two pointwise moving controls that meet at a time $t_0 > 0$.

However, even if the system can be controlled from a one-parameter family of lower-dimensional manifolds $\{\gamma(t)\}_{0\leq t\leq T}$, the control is expected to be singular because it acts in a very small part of the domain. Indeed, as it was pointed out in [B] for the 1-d case, the control exhibits, in general, a highly oscillatory behavior in time.

To improve the efficiency of these moving controls, the possibility of increasing the time oscillations of the curve $\{\gamma(t)\}_{0\leq t\leq T}$ was suggested in [B]. More precisely, a highly oscillating periodic family of manifolds of the form $\{\gamma(t/\varepsilon)\}_{0\leq t\leq T}$ was considered, $\{\gamma(t)\}_{0\leq t}$ being a $2\pi-$periodic in time family of manifolds. We refer to these controls as *rapidly oscillating controllers*. As $\varepsilon \to 0$ the control acts at any point of the range of $\{\gamma(s)\}_{s\in[0,2\pi)}$ for an increasing number of times. In this way, as $\varepsilon \to 0$, the controls are likely to be close, in some sense, to a control acting on the open set ω defined as the interior set of $P = \ range\ \{\gamma(s)\}_{s\in[0,2\pi)}$ with respect to the relative topology, which, typically, contains a manifold of dimension $k+1$. As we mentioned above, controls acting on higher dimensional manifolds are likely to be more efficient. In the particular case $k = n-1$, the limit set (as $\varepsilon \to 0$) ω is an open subset of Ω and the limit system is approximately controllable. Thus, in general, rapidly oscillating controllers should provide a more efficient way to control the system.

This paper is devoted to rigorously proving that both ideas presented before are correct. First we provide several controllability results of the heat equation for a large class of lower-dimensional *moving controls* and second, we prove the convergence of *rapidly oscillating controllers* as $\varepsilon \to 0$ to a certain class of controllers distributed on a $k+1$-dimensional manifold.

The rest of this paper is divided into four more sections. In section 3 we give the main approximate controllability results in this paper. We distinguish the case of one space dimension and the multidimensional one. We also state the main results on the asymptotic behavior of the controls on rapidly oscillating control regions. Section 4 is devoted to analyzing the problem of unique continuation which is equivalent to the approximate controllability one. We distinguish again the 1-d and the multidimensional cases. In section 5 we prove the convergence results for the rapidly oscillating

controllers as the oscillation parameter ϵ goes to zero. Finally, in section 6 we provide some comments and extensions.

3. Main results. In this paper we restrict ourselves to the case where Ω is an open set of $\mathbb{R}^n$ in dimensions $n = 1, 2, 3$, but the techniques we employ and the results we get can be easily generalized to higher dimensions.

We consider system (2.4) with a moving control

$$\begin{cases} u_t - \Delta u = f(x,t)\delta_{\gamma(t)}(x) & \text{in } Q, \\ u = 0 & \text{on } \Sigma, \\ u(x,0) = u^0(x) & \text{in } \Omega. \end{cases} \tag{3.1}$$

We assume that $\gamma(t)$ satisfies the following hypotheses:

1. $\gamma(t)$ is a Lipschitz-continuous k-dimensional manifold in Ω with $0 \le k \le n-1$ for all $t \in [0,T]$.
2. The set $\gamma = \{\gamma(t)\}_{0\le t\le T}$ is a time-continuous family of manifolds in the sense of the following definition.

DEFINITION 3.1. *We say that $\{\gamma(t)\}_{0\le t\le T}$ is a C^s $(s \ge 0)$ (resp., analytic) family of k-dimensional manifolds when $\gamma(t)$ is a Lipschitz-continuous k-dimensional manifold for all $t \in [0,T]$ and there exists a finite family of functions $\{\psi_\alpha(x,t)\}_{\alpha=1}^A$, $A \ge 1$ being independent of t, such that*

1. *$\psi_\alpha : V_\alpha \times [0,T] \subset \mathbb{R}^k \times [0,T] \to \{\gamma(t)\}_{0\le t\le T}$, where V_α are compact sets of $\mathbb{R}^k$, for any $\alpha = 1, \dots, A$, and*

$$\{\gamma(t)\}_{0\le t\le T} \subset \bigcup_{\alpha=1}^{A} \psi_\alpha(V_\alpha, t) \quad \text{for all } 0 \le t \le T.$$

2. *$\psi_\alpha(y,t) \in C([0,T]; W^{1,\infty}(V_\alpha))$ for all $\alpha = 1, \dots, A$.*
3. *for any fixed $y \in V_\alpha$, $\psi_\alpha(y,t)$ is C^s (resp., analytic) in the time variable t.*

Remark 3.1. 1. Note that a C^s family of k-dimensional manifolds may be constituted by manifolds $\gamma(t)$ which are not C^s. More precisely, the condition of being of class C^s refers only to the regularity on the time variable.

2. The assumption on the Lipschitz-continuity in space of the manifolds $\gamma(t)$ is the minimal one to define the surface measure σ on $\gamma(t)$ (see [N, Chap. 4, sect. 7]). Thus, the measure of $\gamma(t)$, arising below, and the integral on $\gamma(t)$ are well defined for any $t \in [0,T]$.

3. The measure of $\gamma(t) \subset \Omega$ is finite for all $t \in [0,T]$. Moreover, hypothesis 2 in Definition 3.1 ensures the time-continuity of the total measure of $\gamma(t)$. Therefore the measure of $\gamma(t)$ is in fact uniformly bounded in $t \in [0,T]$.

4. The case $k = 0$ corresponds to that in which, for each $t \in [0,T]$, $\gamma(t)$ is reduced to a single point or a finite number of points. For instance, when $\gamma(t) = \{x(t)\}$ for all $t \in [0,T]$ the control in (3.1) takes the form $f(t)\delta_{x=\gamma(t)}$ and f is independent of x.

In Figure 1 we show some examples of time-continuous families of manifolds that satisfy the hypotheses in Definition 3.1.

The control $f(x,t)$ (resp., $f(t)$ if $k = 0$) in (3.1) is assumed to belong to $L^2(0,T; L^2(\gamma(t)))$ (resp., $L^2(0,T)$), i.e.,

$$\int_0^T \int_{\gamma(t)} |f(x,t)|^2 d\sigma\, dt < \infty \qquad \left(\text{resp.}, \int_0^T |f(t)|^2 dt < \infty\right). \tag{3.2}$$

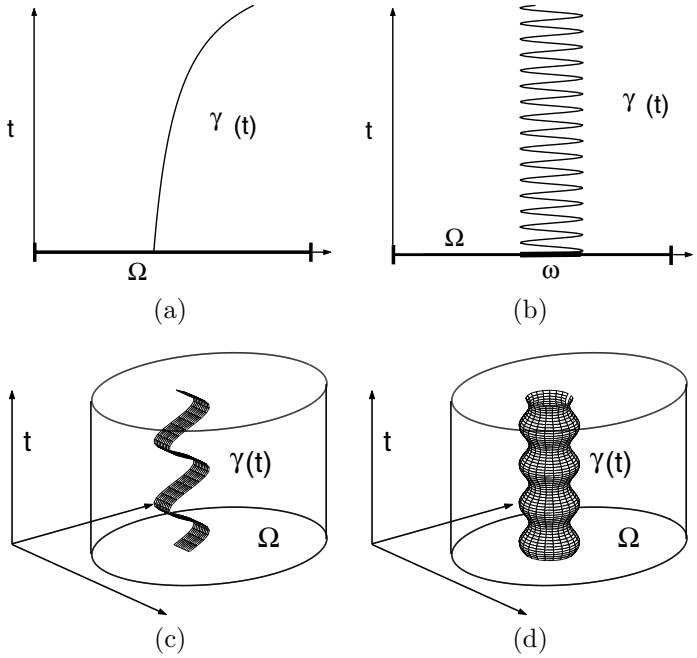


FIG. 1. *Examples of time-continuous families of manifolds that illustrate Definition* 3.1*:* (a) *and* (b)*, single point* ($k = 0$) *with nonregular and regular in time trajectories, respectively;* (c) *and* (d)*, curves* ($k = 1$) *with regular in time trajectories. Note that in case* (d) *the parametrization of* $\gamma(t)$*, for each* $t \in [0, T]$*, is a closed curve that requires at least two charts. Thus we have to consider* $A = 2$ *in Definition* 3.1.

Then, system (3.1) is well defined in different Sobolev spaces depending on the space dimension n and the dimension k of the manifold $\gamma(t)$. We introduce the spaces

$$H_{-1} = \begin{cases} H^{-1}(\Omega) & \text{if } n-k=1, \\ L^2(\Omega) & \text{if } n-k>1, \end{cases} \qquad H_0 = \begin{cases} L^2(\Omega) & \text{if } n-k=1, \\ H_0^1(\Omega) & \text{if } n-k>1, \end{cases}$$

$$H_1 = \begin{cases} H_0^1(\Omega) & \text{if } n-k=1, \\ H^2 \cap H_0^1(\Omega) & \text{if } n-k>1, \end{cases} \tag{3.3}$$

and we denote by H_i', $i = 1, 2, 3$, their duals. Recall that we are assuming the space dimension $n = 1, 2, 3$. Similar results can be proved for higher dimensions. But the choice of H_i, $i = 1, 2, 3$, has to be suitably modified for $n \geq 4$.

The term $f(x,t)\delta_{\gamma(t)}$ on the right-hand side of (3.1) clearly satisfies

$$f(x,t)\delta_{\gamma(t)} \in L^2(0,T;H_1'). \tag{3.4}$$

Indeed, if we denote by $\langle \cdot, \cdot \rangle$ the duality pairing between $L^2(0,T;H_1)$ and its dual we have

$$\langle f(x,t)\delta_{\gamma(t)}, \varphi \rangle = \begin{cases} \int_0^T \int_{\gamma(t)} f(x,t)\varphi(x,t)d\sigma\, dt & \text{if } k \geq 1, \\ \int_0^T f(t)\varphi(\gamma(t),t)dt, & \text{if } k = 0, \end{cases} \tag{3.5}$$

which is well defined under the assumptions above on $\gamma(t)$, since

$$\left|\int_0^T \int_{\gamma(t)} f(x,t)\varphi(x,t)d\sigma\, dt\right| \le \|f\|_{L^2(0,T;L^2(\gamma(t)))}\|\varphi\|_{L^2(0,T;L^2(\gamma(t)))} \quad \text{if } k \ge 1,$$
$$\left|\int_0^T f(t)\varphi(\gamma(t),t)dt\right| \le \|f\|_{L^2(0,T)}\|\varphi\|_{L^2(0,T;L^\infty(\Omega))}, \qquad \text{if } k = 0,$$

and

$$\|\varphi\|_{L^2(0,T;L^2(\gamma(t)))} \le \begin{cases} C(\gamma)\|\varphi\|_{L^2(0,T;L^\infty)} \le C(\gamma)\|\varphi\|_{L^2(0,T;H^2\cap H_0^1)} & \text{if } n=3 \text{ and } k=1,\\ C(\gamma)\|\varphi\|_{L^2(0,T;H_0^1(\Omega))} & \text{if } n=2,3 \text{ and } n-k=1, \end{cases}$$

$$\|\varphi\|_{L^2(0,T;L^\infty(\Omega))} \le \begin{cases} C(\gamma)\|\varphi\|_{L^2(0,T;H_0^1)} & \text{if } n=1,\\ C(\gamma)\|\varphi\|_{L^2(0,T;H^2\cap H_0^1)} & \text{if } n=2,3. \end{cases} \tag{3.6}$$

Here we have used the Sobolev embeddings

$$\begin{aligned} &H_0^1(\Omega) \subset C(\bar\Omega) && \text{if } n=1,\\ &H^2\cap H_0^1(\Omega) \subset C(\bar\Omega) && \text{if } n=2,3 \text{ and } n-k>1, \end{aligned}$$

and the trace theorem (see [N, Chap. 4, sect. 7]), in the case $n=2,3$ and $n-k=1$, guaranteeing that

$$\|\,\varphi|_{\gamma(t)}\,(\cdot,t)\|_{L^2(\gamma(t))} \le C(\gamma(t))\|\varphi(\cdot,t)\|_{H_0^1(\Omega)} \qquad \text{for } n=2,3 \text{ and } n-k=1.$$

The constant $C(\gamma(t))$ depends only on the measure of $\gamma(t)$. Thus, in view of the assumptions on γ, it can be chosen to be uniformly bounded in $t\in[0,T]$. Consequently (3.6) holds and therefore (3.4) holds too.

It is at this point where defining H_1 distinguishing the value of $n-k$ is needed. The same can be said about the assumptions we have made on the manifold γ.

We now define the weak solution of system (3.1) by transposition (see [LM]). Let $\psi\in L^2(0,T;H_{-1})$ and consider the adjoint heat equation

$$\begin{cases} -\varphi_t - \Delta\varphi = \psi & \text{in } Q,\\ \varphi = 0 & \text{on } \Sigma,\\ \varphi(x,T) = 0 & \text{in } \Omega. \end{cases} \tag{3.7}$$

Note that the unique solution of (3.7) belongs to the class

$$\varphi \in C([0,T];H_0)\cap L^2(0,T;H_1).$$

Multiplying the equation in (3.1) by φ and integrating by parts we obtain formally the following identity:

$$\int_0^T\int_{\gamma(t)} f\varphi d\sigma\, dt + \int_\Omega u^0\varphi(0)dx = \int_0^T\int_\Omega u\psi dx\, dt \quad \text{if } k\ge 1,$$
$$\int_0^T f(t)\varphi(\gamma(t),t)dt + \int_\Omega u^0\varphi(0)dx = \int_0^T\int_\Omega u\psi dx\, dt \quad \text{if } k=0.$$

This motivates the following definition: *We say that u is a solution of* (3.1), *in the sense of transposition, if*

$$\langle f\delta_{\gamma(t)},\varphi\rangle + \langle u^0,\varphi(0)\rangle_0 = \int_0^T \langle u,\psi\rangle_{-1}dt \qquad \forall\psi\in L^2(0,T;H_{-1}), \tag{3.8}$$

where $\langle\cdot,\cdot\rangle$ (resp., $\langle\cdot,\cdot\rangle_0$, $\langle\cdot,\cdot\rangle_{-1}$) is the duality pairing between $L^2(0,T;H_1)$ (resp., H_0, H_{-1}) and its dual space $L^2(0,T;H_1')$ (resp., H_0', H_{-1}).

It is easy to see that for any initial data $u^0\in H_0'$ there exists a unique solution of (3.1) in the sense of transposition in the class

$$u\in C\left([0,T];H_0'\right). \tag{3.9}$$

For example, if we consider the case $n-k>1$ (the case $n-k=1$ is even simpler), then $H_{-1}=L^2(\Omega)$ and $\psi\in L^2(0,T;L^2(\Omega))$. The solution φ of system (3.7) is in the class $\varphi\in L^2(0,T;H^2\cap H_0^1(\Omega))\cap C([0,T];H_0^1(\Omega))$, and then the map

$$\psi\to(\varphi(x,t),\varphi(x,0))$$

is linear and continuous from $L^2(0,T;L^2(\Omega))$ to $L^2(0,T;H^2\cap H_0^1(\Omega))\times H_0^1(\Omega)$. Therefore, for any initial data $u^0\in H^{-1}(\Omega)$, the left-hand side of (3.8) is linear and continuous from $\psi\in L^2(0,T;L^2(\Omega))$ to $\mathbb{R}$. Then, there exists a unique $u\in L^2(0,T;L^2(\Omega))$ satisfying (3.8), i.e., a solution of (3.1) in the sense of transposition. Moreover, it is easy to see that this solution u satisfies the first equation in system (3.1) in the sense of distributions, and then we deduce that $u_t\in L^2(0,T;(H^2\cap H_0^1(\Omega))')$. Consequently,

$$u\in C([0,T];H^{-1}(\Omega)),$$

and (3.9) holds for $n-k>1$.

We consider the following approximate controllability problem for system (3.1): *given u^0, $u^1\in H_0'$ and $\alpha>0$, to find a control $f\in L^2(0,T;L^2(\gamma(t)))$ such that the solution $u=u(x,t)$ of* (3.1) *satisfies*

$$\|u(T)-u^1\|_{H_0'}\le\alpha. \tag{3.10}$$

For the sake of clarity we divide the rest of this section into two subsections where we state separately the results for the 1-d case and those for higher space dimensions.

3.1. The one-dimensional case. We assume that $\Omega=(0,L)$ with $L>0$. Let $\gamma:[0,T]\to\Omega$ be any continuous curve. The following theorem holds.

THEOREM 3.1. *Let $\gamma(t):[0,T]\to\Omega$ be a nonconstant continuous curve satisfying at least one of the following conditions:*

1. *There exists an open subinterval $\mathcal{U}\subset[0,T]$ where γ is not analytic at any point $t\in\mathcal{U}$.*
2. *There exists $t_1\in(0,T)$ where $t\to\gamma(t)$ is not analytic and a subinterval $(t_1,t_2)\subset(0,T)$ where γ is analytic and can be extended analytically to a subinterval (t_0,t_2) with $t_1\in(t_0,t_2)$.*
3. *γ can be extended analytically to a curve $\bar\gamma(t):(-\infty,T]\to\mathbb{R}$ satisfying one of the following conditions:*
   (a) *$\bar\gamma(t)$ meets the boundary of Ω, i.e., there exists $t_0\in(-\infty,T]$ such that $\bar\gamma(t_0)\in\partial\Omega$.*
   (b) *The set of accumulation points of $\bar\gamma(t)$ as $t\to-\infty$ contains at least one point $x_0\in\Omega$ such that x_0/L is irrational.*

(c) *The set of accumulation points of* $\bar{\gamma}(t)$ *as* $t \to -\infty$ *is reduced to a unique* $x_0 \in \Omega$*, and there exists a sequence* $t_n \to -\infty$ *such that*

$$\lim_{t_n \to -\infty} (\bar{\gamma}(t_n) - x_0)^{-1} e^{t_n(\lambda_3 - \lambda_2)} = 0, \tag{3.11}$$

where $\lambda_3 - \lambda_2 = 5\pi^2/L^2$ *is the difference between the third and second eigenvalues of the Laplace operator in* Ω*.*

Then, for any $T > 0$*, system* (3.1) *is approximately controllable.*

Remark 3.2. The conditions on γ in Theorem 3.1 above do not characterize all possible curves for which approximate controllability holds. However, they may be considered sharp in different senses:

1. Condition 1 is sharp in the sense that if γ is analytic in one interval $(t_0, t_1) \subset [0, T]$, then the approximate controllability may fail. We can consider, for example, a curve $\gamma(t) = x_0$ with x_0 nonstrategic, i.e., x_0/L rational.
2. Condition 2 is sharp in the sense that if $\gamma : (t_1, t_2) \to \Omega$ is analytic but cannot be extended analytically to a subinterval (t_0, t_2) with $t_1 \in (t_0, t_2)$, then approximate controllability may fail (see example 1 in section 4.1 below).
3. Condition 3 is sharp in the sense that if $\gamma : [0, T] \to \Omega$ is analytic but cannot be extended analytically to $t \in (-\infty, T]$, then approximate controllability may fail (see example 2 in section 4.1 below).
4. Condition 3 is also sharp in the sense that if $\gamma : [0, T] \to \Omega$ is analytic and can be extended analytically to $\bar{\gamma} : (-\infty, T] \to \Omega$ in such a way that the set of accumulation points of $\bar{\gamma}(t)$ as $t \to -\infty$ is reduced to an unique nonstrategic point $x_0 \in \Omega$ that does not satisfy (3.11), then approximate controllability may fail (see example 3 in section 4.1 below).

We give now a number of examples showing that the conditions in Theorem 3.1 cover a large class of curves.

Examples.

1. *Weierstrass-type functions.* Let $\gamma(t) : [0, T] \to \Omega$ be a C^1 function with nowhere defined second derivative. Then $\gamma(t)$ is Lipschitz and satisfies condition 1 above. A function with this property can be constructed integrating the classical Weierstrass example of a continuous function that is nowhere differentiable (see [Pu, Chap. 4, sect. 7], for example).
2. *Piecewise analytic curves.* Let $\gamma_1, \gamma_2 : [0, T] \to \Omega$ be two analytic curves that meet at time t_0, i.e., there exists $t_0 \in (0, T)$, where $\gamma_1(t_0) = \gamma_2(t_0)$ and $\gamma_1'(t_0) \neq \gamma_2'(t_0)$. Then $\gamma : [0, T] \to \Omega$ defined as follows:

$$\gamma(t) = \begin{cases} \gamma_1(t) & \text{if } t \in [0, t_0], \\ \gamma_2(t) & \text{if } t \in [t_0, T] \end{cases}$$

   satisfies condition 2 in Theorem 3.1 (see Figure 1(a)).
3. *Constant velocity curves.* Let $\gamma(t) = x_0 + \alpha t$ with $x_0 \in \Omega$ (see Figure 2(a)). Then clearly γ satisfies condition 3(a) in Theorem 3.1.
4. *Periodic analytic curves.* Let $\gamma : \mathbb{R} \to \Omega$ be any nonconstant periodic analytic curve (see Figure 1(b)). In this case, the set of accumulation points of $\gamma(t)$ as $t \to -\infty$ is the range of γ over a period, which is an interval, and therefore it contains infinitely many points such that x_0/L is irrational. Then γ satisfies condition 3(b) in Theorem 3.1.

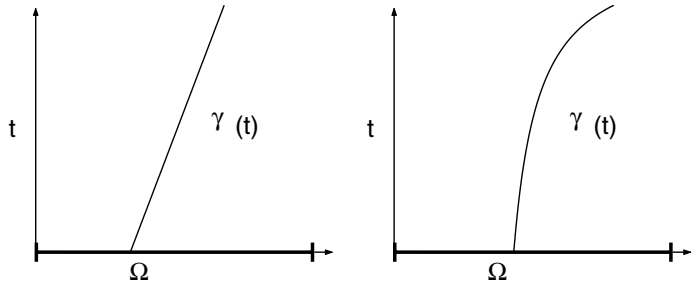


FIG. 2. *Trajectories of γ for which the approximate controllability property holds.*

5. *Curves that converge to a unique point as $t \to -\infty$ and satisfy* (3.11). For example, let $\gamma(t) = x_0 + \beta/t$ with $x_0 \in \Omega$ and let β be chosen in order to have $\gamma(t) \in \Omega$ for all $t \in [0, T]$. Then γ satisfies condition 3(c) in Theorem 3.1 (see Figure 2(b)).

3.2. The higher dimensional case. The situation is now more complex. We only consider nonconstant (in time) analytical families of k-dimensional manifolds $\{\gamma(t)\}_{0\le t\le T}$ satisfying, in particular, the following hypothesis:

$$\{\gamma(t)\}_{0\le t\le T} \quad \begin{array}{l}\text{can be extended to an analytical family } \{\bar\gamma(t)\}_{-\infty<t\le T}\\ \text{s.t. } \bar\gamma(t)\subset\Omega \quad \forall t\in(-\infty,T).\end{array} \tag{3.12}$$

Thus we are considering only the analogue of case 3 in the statement of Theorem 3.1.

In order to state our result let us introduce the eigenvalue problem associated to system (2.4):

$$\begin{cases} -\Delta w(x) = \lambda w(x), & x\in\Omega\\ w(x)=0, & x\in\partial\Omega. \end{cases} \tag{3.13}$$

The associated eigenvalues will be denoted by

$$0<\lambda_1<\lambda_2<\cdots<\lambda_j<\cdots$$

each one with finite multiplicity $l(j)\ge 1$.

We also introduce the following sets associated to the families $\{\gamma(t)\}_{0\le t\le T}$ satisfying (3.12).

DEFINITION 3.2. *Let $\{\gamma(t)\}_{0\le t\le T}$ be a family satisfying* (3.12). *We define the set of "accumulation points" of $\{\gamma(t)\}_{0\le t\le T}$ as*

$$P=\left\{x\in\overline{\Omega}\ s.t.\ \exists t_n\to-\infty,\ x_n\in\bar\gamma(t_n)\ with\ x_n\to x\right\}, \tag{3.14}$$

and for each $x\in P$, the set of "accumulation directions"

$$D_x=\left\{ v\in S^{n-1}\ s.t. \quad \begin{array}{c}\exists t_n\to-\infty,\ x_n\in\bar\gamma(t_n)\ with\ x_n\to x,\ \frac{x_n-x}{\|x_n-x\|}\to v\\ and\ \lim_{t_n\to-\infty}\|x_n-x\|^{-1}e^{ct_n}=0\ \forall c>0\end{array}\right\}. \tag{3.15}$$

Remark 3.3. Observe that D_x contains only those directions v for which $\|x_n - x\|$ converges to zero more slowly than any exponential as $t_n \to -\infty$. As indicated in Remark 3.2, in the 1-d case approximate controllability fails for some curves that converge to a unique accumulation point more rapidly than a certain exponential.

The following result reduces the approximate controllability property of system (3.1) to a suitable unique continuation property for the eigenfunctions of (3.13).

THEOREM 3.2. *Assume that* $\{\gamma(t)\}_{0\le t\le T}$ *satisfies the hypothesis* (3.12)*, and consider the sets* P *and* D_x *introduced in Definition* 3.2*. Assume also that the following spectral unique continuation property holds for the eigenfunctions of* (3.13)*:*

$$\begin{aligned} &\textit{The only eigenfunction } w \textit{ of } (3.13) \textit{ that satisfies}\\ &w(x_0+\mu v)=0,\ \forall x_0\in P,\ \forall v\in D_{x_0},\ \textit{and } \forall \mu\in\mathbb{R}\ \text{s.t. } x_0+\bar\mu v\in\bar\Omega,\\ &\forall \bar\mu\in[0,\mu],\ \textit{is the trivial one.} \end{aligned} \tag{3.16}$$

Then, for any $T>0$ *system* (2.4) *is approximately controllable.*

We observe that there are two special situations where the spectral unique continuation property (3.16) holds trivially:

1. P is not included in the zero set of any of the eigenfunctions of (3.13).
2. P is included in the zero set of some of the eigenfunctions of (3.13) but there exists $x_0\in P$ for which D_{x_0} contains an open set of S^{n-1}. Indeed, in this case if

$$w(x_0+\mu v)=0,\ \forall x_0\in P,\ \forall v\in D_{x_0},\ \text{and } \forall\mu\in\mathbb{R}\ \text{s.t. } x_0+\bar\mu v\in\overline{\Omega},\ \forall\bar\mu\in[0,\mu],$$

then $w=0$ on an open set of Ω and, by classical unique continuation for eigenfunctions (Holmgren's theorem), $w\equiv 0$.

Now, we give some examples where one of the above situations holds and therefore the approximate controllability property is satisfied.

Examples.

1. *Stationary control.* Assume that the control is located on a time-independent k-dimensional manifold γ. In this case, property (3.12) holds trivially. The set P coincides with γ, and D_x is empty for all $x\in P$. Therefore, the unique continuation property (3.16) is satisfied if and only if γ is not included in the nodal set of any of the eigenfunctions of (3.13).
2. *Time-periodic* $n-1$*-dimensional manifolds.* Assume that $\{\gamma(t)\}_{0\le t\le T}$ is a nonconstant periodic (in time) analytic (in time) family of $n-1$-dimensional manifolds. Then, $\{\gamma(t)\}_{0\le t\le T}$ satisfies property (3.12), and the set P is the range of $\gamma(t)$ over a period $[0,p]$, i.e.,

$$\bigcup_{t\in[0,p]}\gamma(t)=P\subset\Omega.$$

We observe that the dimension of P is n generically. However, this is not always the case. For example, if $n=2$ we may consider $\Omega=(-1,1)\times(-1,1)$ and $\gamma(t)$ the subinterval $[0,\frac{1}{2}+\frac{1}{4}\sin(t)]$ in the axis $y=0$.

When the dimension of P is n, it contains an open set of Ω. In this case, the unique continuation property (3.16) is a consequence of the unique continuation property for the eigenfunctions of (3.13).

We observe that, in general, the dimension of the set P is larger for periodic moving controls than for stationary ones. Therefore, the uniqueness property is much more likely to hold for moving controls.

3. *Pointwise control in* 2*-d.* We assume that $n = 2$, for instance, and $k = 0$. Here the control is localized on a continuous curve $\gamma(t) \subset \Omega \times [0, T]$. Consider the spiral curve around any point $x_0 \in \Omega$:

$$\gamma(t) = x_0 + \frac{\beta}{T + 1 - t}(\cos t, \sin t), \quad \beta > 0.$$

Here the parameter β must be chosen small enough to guarantee that $\gamma(t) \in \Omega$ for all $t \in (-\infty, T]$. In this case, $\gamma(t)$ is analytic and satisfies (3.12), and we have $P = \{x_0\}$ and $D_{x_0} = S^1$. Therefore, the unique continuation property (3.16) holds.

3.3. Rapidly oscillating controllers. Finally, we consider the case of rapidly oscillating controllers, i.e., where the control is located on a k-dimensional manifold $\{\gamma(t/\varepsilon)\}_{0 \le t \le T}$ with $\varepsilon > 0$ a small parameter and $\{\gamma(t)\}_{0 \le t}$ a time-dependent periodic and analytic family of k-dimensional manifolds. To simplify the presentation, we restrict ourselves to the case $k = n - 1$.

We assume without loss of generality that $\gamma(t)$ is periodic of period 2π.

We also assume that the range of $\gamma(t)$ over a period $[0, p]$, i.e., $\bigcup_{t \in [0,p]} \gamma(t) = P \subset \Omega$ has dimension n. As we pointed out in example 2 above (section 3.2) this is not always the case. However, for the purposes of this section, it is natural to consider the case where the supports of the controls scan a larger area.

In this case, system (3.1) reads as follows:

$$\begin{cases} u_t - \Delta u = f(x,t)\delta_{\gamma(t/\epsilon)}(x) & \text{in } Q, \\ u = 0 & \text{on } \Sigma, \\ u(x,0) = u^0(x) & \text{in } \Omega. \end{cases} \tag{3.17}$$

We observe that, by hypothesis, the dimension of P (the range of $\gamma(t)$ over a period) is n and it contains an open set of Ω. Therefore, the unique continuation property (3.16) is a consequence of the unique continuation property for the eigenfunctions of the Laplace operator, and the approximate controllability of system (2.4) holds.

Let us introduce the limit problem

$$\begin{cases} u_t - \Delta u = f(x,t)m_\gamma(x) & \text{in } Q, \\ u = 0 & \text{on } \Sigma, \\ u(x,0) = u^0(x) & \text{in } \Omega, \end{cases} \tag{3.18}$$

where $m_\gamma(x)$ is the limit of $\delta_{\gamma(t/\varepsilon)}(x)$ in the $L^\infty(0, T; H_1')$ weak-* topology as $\varepsilon \to 0$. As we will see, this limit measure $m_\gamma(x)$ is supported in the nonempty open set $\omega \subset \Omega$ defined as the interior set of P.

The following theorem holds.

THEOREM 3.3. *Let us assume that $\{\gamma(t)\}_{t \le T}$ is a nonconstant periodic analytic family of $(n-1)$-dimensional manifolds and $\varepsilon > 0$ is a small parameter.*

Given $T > 0$, $u^0, u^1 \in H_0'$, and $\alpha > 0$, there exists a sequence of approximate controls $f_\varepsilon \in L^2(0, T; L^2(\gamma(t/\varepsilon)))$ (resp., $f_\varepsilon \in L^2(0, T)$ if $n = 1$ and $k = 0$) of system (3.17)*, satisfying* (3.10)*, which is uniformly bounded in $L^2(0, T; L^2(\gamma(t/\varepsilon)))$ (resp., $L^2(0, T)$).*

Moreover, the controls f_ε can be chosen such that they strongly converge in the following sense:

$$f_\varepsilon(x,t)\delta_{\gamma(t/\varepsilon)}(x) \to f(x,t)m_\gamma(x) \text{ in } L^2(0, T; H_1') \quad \text{as } \varepsilon \to 0, \tag{3.19}$$

where f is an approximate control for the limit system (3.18) *so that* (3.10) *holds.*

The measure $m_\gamma(x)$ is the limit of $\delta_{\gamma(t/\varepsilon)}$ in the $L^\infty(0,T;H_1')$ weak- topology as $\varepsilon \to 0$. This limit measure m_γ is supported in the nonempty subset $\omega \subset \Omega$, the interior set of P, and it is characterized by*

$$\int_\omega \varphi(x) m_\gamma(x)\, d\sigma = \frac{1}{2\pi}\int_0^{2\pi}\int_{\gamma(s)} \varphi(x)\, d\sigma\, ds \quad \forall \varphi \in H_1 \text{ if } k \geq 1, \text{ i.e., } n \geq 2$$

$$\int_\omega \varphi(x) m_\gamma(x)\, d\sigma = \frac{1}{2\pi}\int_0^{2\pi} \varphi(\gamma(s))\, ds \quad \forall \varphi \in H_1 \text{ if } k = 0. \tag{3.20}$$

Here σ denotes the surface measure of ω. In the case $k = n-1$, ω is an open subset of Ω and $d\sigma = dx$.

On the other hand, with the above controls, the solutions u_ε of (2.4) *converge strongly in $C([0,T];H_0')$ as $\varepsilon \to 0$ to the solution u of the limit problem* (3.18). *This solution u satisfies* (3.10).

Remark 3.4. 1. As a consequence of the statement in Theorem 3.3, system (3.18) is approximately controllable. In fact, Theorem 3.3 guarantees that the control of (3.18) may be achieved as the limit (when $\varepsilon \to 0$) of the controls of (2.4) in the sense of (3.19).

Note, however, that one could prove directly the approximate controllability of (3.18) since the limit control f acts on an open set of Ω.

2. It is worth noting that for the limit system (3.18), where the control is located on an open subset of Ω, a stronger controllability property is known to hold: the so-called null-controllability property. This means that if $u^1 = 0$, then the control can be chosen to be exact, i.e., the solution of system (3.18) will satisfy $u(x,T) \equiv 0$. Whether this null-controllability property holds for the approximate systems (2.4) and the possible convergence of the controls as $\varepsilon \to 0$ are open problems.

3. When $k < n-1$, the set P will not contain an open subset of Ω but, generically, a $k+1$-dimensional manifold. Therefore, the approximate controllability property of both the approximate systems and the limit system will depend on whether the set ω, the interior of P with respect to the relative topology, belongs to the nodal set of any of the eigenfunctions of the Laplace operator.

If this approximate controllability property holds, then the analogue to Theorem 3.3 still holds. The limit density m_γ is then supported on the manifold ω of dimension $k+1 < n$ and takes the form

$$m_\gamma(x) = g(x)\delta_\omega(x),$$

where $g(x)$ is a density function defined on ω by (3.20).

4. In [B] the numerical approximation of the 1-d case $\Omega = (0,1)$ and $\gamma(t) = x_0 + \delta\cos(t)$ is addressed. Using the standard formal asymptotic expansion method in homogenization and the limit in the sense of distributions of the measure $\delta_{\gamma_\varepsilon(t)}$, the author obtains the first order approximation of a limit control when both ϵ and δ tend to zero ($\epsilon \ll \delta$). This first order approximation consists of two steady distributed controls, concentrated at the points $x = x_0 \pm \delta$, which depend only on the time variable t.

Here we prove that, when $\varepsilon \to 0$, $\delta > 0$ being fixed, the sequence of controls can be chosen to converge, in the sense stated in the above theorem, to a nonsteady distributed control (see (3.22) in example 1 below) acting on the range of $\gamma(t)$ over a period. Namely, the control may vary in x and t. The control obtained in [B]

corresponds to the first order approximation (as $\delta \to 0$) of the limit control, given in (3.22) below, obtained as the limit when $\varepsilon \to 0$ (see [B] for details).

The result above proves that, to some extent, the approximate controls depend continuously on ε. As $\varepsilon \to 0$ the transition is made from controls supported on a manifold of dimension $n-1$ to controls with support in an open subset of Ω.

We finish this section by showing several examples of limit densities $m_\gamma(x)$ in some particular cases.

Examples.

1. *Pointwise control, case* 1. We assume that $\gamma(t)$ is reduced to a unique point at each time $t \in [0,T]$. The trajectory of the control $x = x(t)$ is included in a simple curve $\omega \subset \Omega$, and the control moves forward and backward scanning the curve ω (see Figure 3(a)).

In this case, the integral in (3.20) can be simplified, studying separately the closed intervals where $\gamma(s)$ is one-to-one $\{I_h\}_{h=1}^H$. Note that the whole interval $[0,2\pi]$ is divided into the subintervals I_h. Indeed, if there is a subinterval $I \subset (0,2\pi)$ such that I is not included in $\bigcup_{h=1}^H I_h$, then $\gamma(s)$ must be constant on I and then constant everywhere because of the analyticity of γ.

Note also that the number H of subintervals $I_h \subset [0,2\pi]$ must be finite because of the analyticity of γ. Indeed, at the extremes of I_h, the control has a returning point where its trajectory changes direction and γ' vanishes. If there are infinitely many intervals I_h, there are infinitely many points in a period $[0,2\pi]$ where γ' vanishes. Thus, $\gamma(t)$ must be constant, and this is in contradiction with the hypotheses on $\gamma(t)$.

Let $\omega_h = \gamma(I_h) \subset \Omega$ and let $\gamma_h^{-1} : \omega_h \to I_h$ be the inverse function of ψ in each one of the sets ω_h. Then,

$$\begin{aligned} \frac{1}{2\pi}\int_0^{2\pi} \varphi(\gamma(s))\, ds &= \frac{1}{2\pi}\sum_{h=1}^{H}\int_{I_h} \varphi(\gamma(s))\, \frac{1}{|\gamma'(s)|}\,|\gamma'(s)|\, ds \\ &= \frac{1}{2\pi}\sum_{h=1}^{H}\int_{\omega_h} \varphi(x)\frac{1}{\left|\gamma'(\gamma_h^{-1}(x))\right|} d\omega_h. \end{aligned} \tag{3.21}$$

Then,

$$m_\gamma(x) = \begin{cases} \frac{1}{2\pi}\sum_{h=1}^{H} \frac{1}{|\gamma'(\gamma_h^{-1}(x))|}\delta_{\omega_h} & \text{if } n \geq 2, \\ \frac{1}{2\pi}\sum_{h=1}^{H} \frac{1}{|\gamma'(\gamma_h^{-1}(x))|}\chi_{\omega_h} & \text{if } n = 1. \end{cases}$$

Note that m_γ is defined over the whole curve ω since $\cup_h \omega_h = \omega$.

The function $m_\gamma(x)$ is singular at the extremes of the intervals I_h since $\gamma'(s) = 0$ for some points $s \in [0,2\pi]$. For example, in the 1-d case studied in [B], $\Omega = (0,1)$, $\gamma(t) = x_0 + \delta\cos(t)$, and

$$m_\gamma(x) = \begin{cases} \frac{1}{\pi\sqrt{\delta^2 - (x-x_0)^2}} & \text{if } |x - x_0| < \delta, \\ 0 & \text{otherwise,} \end{cases} \tag{3.22}$$

which is singular at $x = x_0 \pm \delta$. Observe, however, that $m_\gamma(x) \in L^1(\Omega)$ since the integral in (3.21) is well defined for all $\varphi \in H_1$ (see Figure 3).

2. *Pointwise control, case* 2. Now we assume that the trajectory of the control follows a simple closed curve without any returning point. This is only possible if $n \geq 2$ (see Figure 4(a)). In this case, $\gamma : [0,2\pi) \to \Omega$ is one-to-one and $\gamma'(s) \neq 0$ for any $s \in [0,2\pi)$. Then, we have

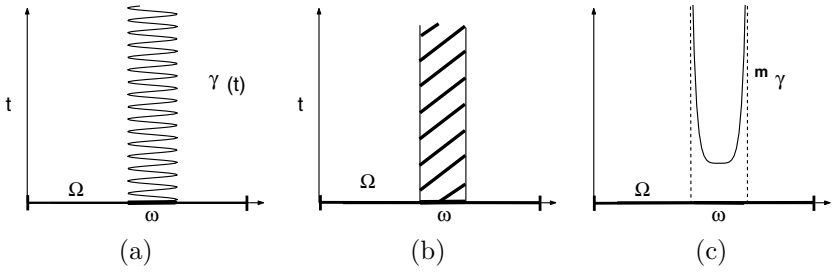


FIG. 3. *The sequence of controls acting on curves $\gamma(t/\varepsilon)$, with $\gamma(s)$ periodic (see* (a)*), converge as $\varepsilon \to 0$ to a control acting in the whole interval ω (see* (b)*) with a density $m_\gamma(x)$ which is singular at the extremes of ω (see* (c)*).*

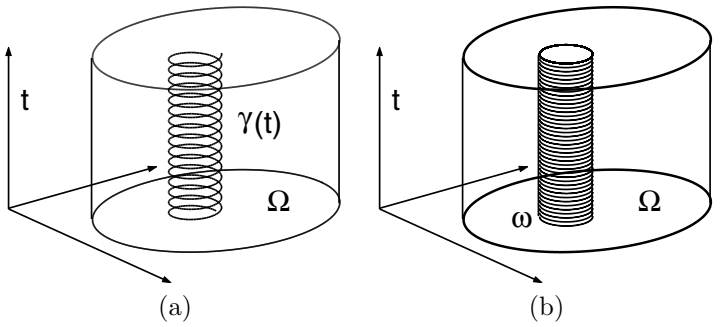


FIG. 4. *The pointwise control located on a periodic circular trajectory in time (see* (a)*) converges, as the period goes to zero, to a control acting on the lateral boundary of the space-time cylinder $\omega \times [0,T]$ (see* (b)*).*

$$m_\gamma(x) = \frac{1}{2\pi} \frac{1}{|\gamma'(\gamma^{-1}(x))|} \delta_\omega(x).$$

In particular, if γ follows a circular trajectory of radius R around $(x_0, y_0) \in \Omega \subset \mathbb{R}^2$, $\gamma(t) = (x_0, y_0) + R(\cos(t), \sin(t))$, then $|\gamma'(\gamma^{-1}(x))| = R$ and m_γ is given by $m_\gamma(x) = R\delta_\omega(x)$ (see Figure 4).

3. *Control on a curve in dimension $n = 2$.* Let $\gamma(t)$ be a curve for any $t \in [0,T]$. To simplify the presentation, we assume that $\{\gamma(t)\}_{0\le t\le T}$ can be described by a unique chart $\psi : V \times [0,T] \to \{\gamma(t)\}_{0\le t\le T}$, where V is an open and bounded subinterval of $\mathbb{R}$.

The measure m_γ satisfies (3.20). It can be computed studying separately the closed time intervals $\{I_h\}_{h=1}^H$, where $\psi(y,s) : V \times I_h \to \Omega$ is one-to-one. Let $\omega_h = \psi(V, I_h) \subset \Omega$ and $\psi_h^{-1} : \omega_h \to V \times I_h$ be the inverse function of ψ. Then,

$$\int_0^{2\pi} \int_{\gamma(s)} \varphi(x)\, d\sigma\, ds = \sum_{h=1}^H \int_{I_h} \int_V \varphi(\psi(y,s)) \frac{\left|\frac{\partial\psi}{\partial y}\right|}{\left|\frac{\partial\psi}{\partial y} \times \frac{\partial\psi}{\partial s}\right|} \left|\frac{\partial\psi}{\partial y} \times \frac{\partial\psi}{\partial s}\right| dy\, ds$$

$$= \sum_{h=1}^H \int_{\omega_h} \varphi(x) \frac{\left|\frac{\partial\psi}{\partial y}(\psi_h^{-1}(x))\right|}{\left|\frac{\partial\psi}{\partial y}(\psi_h^{-1}(x)) \times \frac{\partial\psi}{\partial s}(\psi_h^{-1}(x))\right|} dx. \tag{3.23}$$

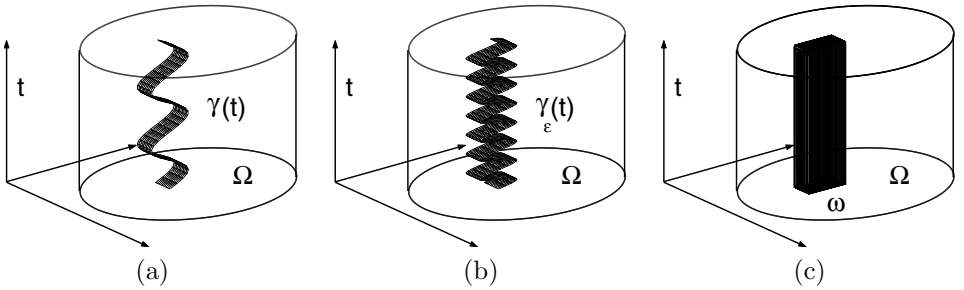


FIG. 5. *The control located on a curve with periodic trajectory in time (see* (a)*) converges, as the period goes to zero (see* (b)*), to a control acting on a* 2*-d domain* ω *for all time (see* (c)*).*

Then,

$$m_\gamma(x) = \frac{1}{2\pi} \sum_{h=1}^{H} \frac{\left|\frac{\partial \psi}{\partial y}(\psi_h^{-1}(x))\right|}{\left|\frac{\partial \psi}{\partial y}(\psi_h^{-1}(x)) \times \frac{\partial \psi}{\partial s}(\psi_h^{-1}(x))\right|} \chi_{\omega_h}.$$

Note that m_γ is defined over the whole ω since $\cup_h \omega_h = \omega$, which is a 2-d set (see Figure 5).

4. Unique continuation for the adjoint system. It is well known that the approximate controllability of system (2.4) is a consequence of the following unique continuation property for the adjoint system: *If* $\varphi \in C([0,T]; H_0)$ *solves*

$$\begin{cases} -\varphi_t - \Delta\varphi = 0 & \text{in } Q, \\ \varphi = 0 & \text{on } \Sigma, \\ \varphi(x,T) = \varphi^0(x) & \text{in } \Omega, \end{cases} \tag{4.1}$$

can we guarantee that

$$\varphi(x,t) = 0 \quad \forall t \in [0,T] \text{ and } \forall x \in \gamma(t) \quad \Rightarrow \quad \varphi \equiv 0? \tag{4.2}$$

In fact, as we will see later on, when this unique continuation property is satisfied, the control can be built as a minimizer of a quadratic, convex, continuous, and coercive functional in a Hilbert space, associated with the adjoint system (4.1).

This section is devoted to analyzing this uniqueness problem. We divide it into two subsections where we study separately the 1-d case and the higher dimensional one.

4.1. The one-dimensional case.

LEMMA 4.1. *Assume that* $\Omega = (0, L)$ *and* $\gamma : [0,T] \to \Omega$ *is a continuous curve which satisfies the hypotheses in the statement of Theorem* 3.1*. Then the unique continuation property* (4.2) *holds for the solutions of the adjoint problem* (4.1)*.*

As we stated in Remark 3.2, the conditions in Theorem 3.1 are sharp in different senses. This is so because these conditions are sharp in the context of Lemma 4.1. We present now three examples that illustrate this fact. After these examples we give the proof of Lemma 4.1.

Example 1. This example shows that condition 2 in Theorem 3.1 is sharp in the sense that if $\gamma : (t_1, T] \to \Omega$ cannot be extended analytically for $t < t_1$, then the

unique continuation property (4.2) for the solutions of the adjoint problem (4.1) may fail.

We consider $\Omega = (0,1)$ and the solution of the backward system (4.1) given by

$$\varphi(x,t) = e^{\lambda_2(t-T)}\sin(\sqrt{\lambda_2}x) + e^{\lambda_4(t-T)}\sin(\sqrt{\lambda_4}x) = e^{4\pi^2(t-T)}\sin(2\pi x) + e^{16\pi^2(t-T)}\sin(4\pi x).$$

For any $t \in [0,T]$ the points $x = x(t) \in \Omega$ for which $\varphi(x,t) = 0$ are characterized by

$$0 = \sin(2\pi x) + e^{12\pi^2(t-T)}\sin(4\pi x) = \sin(2\pi x)(1 + e^{12\pi^2(t-T)}2\cos(2\pi x)).$$

Thus, $\varphi(x(t),t) = 0$ if and only if $x(t) = 1/2 \in \Omega$ (for any $t \in [0,T]$) or $x(t) \in \Omega$ is any of the two roots of the equation

$$\cos(2\pi x(t)) = -\frac{1}{2}e^{-12\pi^2(t-T)} \quad \text{for } t \in (t_1, T], \tag{4.3}$$

where t_1 is defined as $t_1 = T - (\log 2)/(12\pi^2)$. Note that for $t < t_1$ the right-hand side term in (4.3) is smaller than -1 and therefore (4.3) has no roots for $t < t_1$.

We define $\gamma(t)$ as follows:

$$\gamma(t) = \begin{cases} 1/2 & \text{if } t \in [0,t_1], \\ x(t) = \frac{1}{2\pi}\arccos(-\frac{1}{2}e^{-12\pi^2(t-T)}) \text{ with } x(t) \in (0,\frac{1}{2}] & \text{if } t \in [t_1,T]. \end{cases}$$

Clearly $\gamma(t)$ is a continuous curve for which $\varphi(\gamma(t),t) = 0$ for all $t \in [0,T]$. Moreover, $\gamma(t)$ is analytic in $t \in [0,t_1) \cup (t_1,T]$ and it is easy to see (differentiating in (4.3)) that

$$\lim_{\substack{t \to t_1 \\ t > t_1}} \gamma'(t) = \infty.$$

This means that $\gamma : (t_1,T] \to \Omega$ cannot be analytically extended for $t < t_1$.

Thus, we have obtained a curve $\gamma(t)$ for which the unique continuation property (4.2) fails and that shows that condition 2 in Theorem 3.1 is sharp in the sense described above.

Example 2. This example shows that condition 3 is sharp in the sense that if γ cannot be extended analytically to a curve $\bar{\gamma} : (-\infty,T] \to \Omega$, then the unique continuation property (4.2) may fail.

We consider $\Omega = (0,1)$ and the solution of the backwards system (4.1) given by

$$\varphi(x,t) = c_1 e^{\lambda_1(t-T)}\sin(\sqrt{\lambda_1}x) + e^{\lambda_2(t-T)}\sin(\sqrt{\lambda_2}x) = c_1 e^{\pi^2(t-T)}\sin(\pi x) + e^{4\pi^2(t-T)}\sin(2\pi x),$$

with c_1 a constant to be chosen later.

For any $t \in [0,T]$ the points $x = x(t) \in \Omega$ for which $\varphi(x,t) = 0$ are characterized by

$$0 = c_1\sin(\pi x) + e^{3\pi^2(t-T)}\sin(2\pi x) = \sin(\pi x)(c_1 + e^{3\pi^2(t-T)}2\cos(\pi x)).$$

Thus, $\varphi(x(t),t) = 0$ if and only if $x(t) \in \Omega$ is a root of

$$\cos(\pi x(t)) = -\frac{c_1}{2}e^{-3\pi^2(t-T)}. \tag{4.4}$$

We choose $c_1 \in (0,2)$ in such a way that (4.4) has a unique root $x(t) \in \Omega$ for $t \in [0,T]$; i.e., the right-hand side of (4.4) is greater than -1.

Clearly $\gamma(t)$ is a continuous curve for which $\varphi(\gamma(t),t) = 0$ for all $t \in [0,T]$. Moreover, $\gamma(t)$ is analytic in $t \in [0,T]$ but cannot be extended to an analytic curve $\bar{\gamma} : (-\infty, T] \to \Omega$. In fact, its analytic extension is defined implicitly by (4.4), which has no roots if t is very (negatively) large.

Thus, we have obtained a curve $\gamma(t)$ for which the unique continuation property (4.2) fails and that shows that condition 3 in Theorem 3.1 is sharp in the sense described above.

Example 3. This example shows that condition 3 is also sharp in the sense that it is possible to construct solutions of the adjoint heat equation (4.1) that vanish at curves $\gamma(t)$ that do not satisfy (3.11). We consider $\Omega = (0,1)$ and the solution of the backward system (4.1) given by

$$\varphi(x,t) = e^{\lambda_2(t-T)}\sin(\sqrt{\lambda_2}x) - e^{\lambda_3(t-T)}\sin(\sqrt{\lambda_3}x) = e^{4\pi^2(t-T)}\sin(2\pi x) - e^{9\pi^2(t-T)}\sin(3\pi x).$$

It is easy to check that φ satisfies the following properties:

$$\begin{aligned} &\varphi(1/2,t) > 0 \quad \text{if } t \le T,\\ &\varphi(2/3,t) < 0 \quad \text{if } t \le T,\\ &\frac{\partial\varphi}{\partial x}(x,t) < 0 \quad \text{if } x \in [1/2,2/3] \text{ and } t \le T. \end{aligned}$$

Therefore, for any $t \le T$, $\varphi(x,t)$ has a unique zero in $x \in (1/2,2/3)$ at the point $\gamma(t)$, i.e.,

$$\varphi(\gamma(t),t) = 0 \text{ for all } t \le T.$$

On the other hand, by the analytic version of the implicit function theorem (see [H, p. 43], for example), the function $t \to \gamma(t)$ is analytic in $t \in (-\infty, T]$.

Moreover, as $\varphi(\gamma(t),t) = 0$ for all $t \in (-\infty, T]$,

$$\sin(2\pi\gamma(t)) = e^{5\pi^2(t-T)}\sin(3\pi\gamma(t)) \to 0 \text{ as } t \to -\infty.$$

Therefore $\gamma(t) \to 1/2$ as $t \to -\infty$ and, by the Taylor expansion of $\sin(2\pi x)$ and $\sin(3\pi x)$ near $x = 1/2$, we easily deduce that

$$\gamma(t) \sim \frac{1}{2} - \frac{1}{2\pi}e^{5\pi^2(t-T)} \quad \text{as } t \to -\infty.$$

In this way we have constructed an example of a nontrivial solution of (4.1) that vanishes at a nonconstant curve $\gamma(t)$ that does not satisfy (3.11).

Proof of Lemma 4.1. We divide the proof into different parts depending on the hypothesis we make on γ.

Part 1. Assume that γ satisfies hypothesis 1 in the statement of Theorem 3.1. Then there exists an open set $\mathcal{U} \subset (0,T)$ such that $\gamma : \mathcal{U} \to \Omega$ is not analytic at any point $t \in \mathcal{U}$. In this case, we prove the unique continuation property (4.2) by contradiction.

Assume that there exists a nontrivial solution $\varphi \in C([0,T]; H_0)$ of the heat equation (4.1) such that $\varphi(\gamma(t),t) = 0$ for all $t \in \mathcal{U}$. As the Laplace operator generates an

analytic semigroup (see, for example, [P, p. 211]), the solution $\varphi : \Omega \times (-\infty, T) \to \mathbb{R}$ of system (4.1) is analytic. By the analytic version of the implicit function theorem either $\frac{\partial \varphi}{\partial x}(\gamma(t_0), t_0) = 0$ or there exists a unique solution $x = \hat{\gamma}(t)$ of $\varphi(x,t) = 0$ near $(x,t) = (\gamma(t_0), t_0)$ with $\hat{\gamma}$ analytic. The latter is in contradiction with the hypothesis on $x = \gamma(t)$. Therefore, $\frac{\partial \varphi}{\partial x}(\gamma(t_0), t_0) = 0$ and this is true for all $t_0 \in \mathcal{U}$.

As $\gamma(t)$ is a noncharacteristic curve of the heat equation, Holmgren's uniqueness theorem asserts that the only solution φ of the heat equation with Cauchy data $\varphi(\gamma(t_0), t_0) = \frac{\partial \varphi}{\partial x}(\gamma(t_0), t_0) = 0$ on γ is $\varphi \equiv 0$. This is in contradiction with the hypothesis.

Part 2. Assume that γ satisfies hypothesis 2 in the statement of Theorem 3.1. Then there exists $t_1 \in (0,T)$ where $\gamma(t)$ is not analytic and a subinterval $(t_1, t_2) \subset (0,T)$ where γ is analytic and can be extended analytically to a subinterval (t_0, t_2) with $t_1 \in (t_0, t_2)$.

We continue $\gamma(t)$ analytically to $\bar{\gamma} : (t_0, t_2) \to \mathbb{R}$. Observe that the composition $\varphi(\bar{\gamma}(t), t)$ is analytic and therefore, in view of the fact that $\varphi(\bar{\gamma}(t), t) = \varphi(\gamma(t), t) = 0$ for $t \in (t_1, t_2)$,

$$\varphi(\bar{\gamma}(t), t) = 0 \qquad \forall t \in (t_0, t_2). \tag{4.5}$$

Let ω be the open region of $(x,t) \in \mathbb{R} \times (t_0, t_1)$ limited by the curves $x = \gamma(t)$, $x = \bar{\gamma}(t)$ and the horizontal line $t = t_0$. Note that φ vanishes at those parts of the boundary of ω constituted by γ and $\bar{\gamma}$.

Multiplying the adjoint heat equation (4.1) by the solution φ and integrating we easily obtain the following:

$$\begin{aligned} 0 = \int_{\bar{\gamma}(t)}^{\gamma(t)} (-\varphi_t - \varphi_{xx})\varphi \, dx &= -\frac{1}{2}\frac{d}{dt}\int_{\bar{\gamma}(t)}^{\gamma(t)} |\varphi|^2 dx + \int_{\bar{\gamma}(t)}^{\gamma(t)} |\varphi_x|^2 dx \\ &\geq -\frac{1}{2}\frac{d}{dt}\int_{\bar{\gamma}(t)}^{\gamma(t)} |\varphi|^2 dx. \end{aligned} \tag{4.6}$$

Thus, the function $\Phi(t) = \int_{\bar{\gamma}(t)}^{\gamma(t)} |\varphi|^2 dx$ is increasing and positive and vanishes at $t = t_1$ because γ and $\bar{\gamma}$ coincide for $t = t_1$. Therefore,

$$\int_{\bar{\gamma}(t)}^{\gamma(t)} |\varphi|^2 dx = 0 \quad \forall t \in [t_0, t_1],$$

and we deduce that φ must be zero in the open nonempty set ω. By Holmgren's uniqueness theorem we deduce that $\varphi \equiv 0$.

Part 3. Assume that γ satisfies hypothesis 3(a) in the statement of Theorem 3.1. Then, γ is analytic and can be extended analytically to a curve $\bar{\gamma}(t) : (-\infty, T] \to \mathbb{R}$ such that $\bar{\gamma}(t)$ meets the boundary of Ω, i.e., there exists $t_0 \in (-\infty, T]$ such that $\bar{\gamma}(t_0) \in \partial\Omega$.

The solution φ of (4.1) can be extended to all $x \in \mathbb{R}$ by odd extension and periodicity. The resulting φ satisfies the adjoint system (4.1) on $\mathbb{R} \times (-\infty, T)$. Observe that the composition $\varphi(\bar{\gamma}(t), t)$ is analytic and therefore

$$\varphi(\bar{\gamma}(t), t) = 0 \qquad \forall t \in (-\infty, T]. \tag{4.7}$$

By hypothesis, $\bar{\gamma}$ meets the boundary of Ω at $t_0 \in (-\infty, T]$ and $\gamma(t_0)$ must be one of the extremes of Ω, say, $x = 0$. Let ω be the open region of $(x,t) \in \mathbb{R} \times (t_0 - \alpha, t_0)$ with

$\alpha > 0$, bounded by $\bar{\gamma}$ and the axis $x = 0$. Note that ω cannot be an empty set because, in this case, $\bar{\gamma}$ and the axis $x = 0$ would coincide over the interval $(t_0 - \alpha, t_0)$ and, by the analyticity of $\bar{\gamma}$, they would coincide everywhere. This situation is excluded in our hypothesis on $\bar{\gamma}$.

Note also that φ vanishes on the subset of the boundary of ω constituted by $x = 0$ and $\bar{\gamma}$. Then we can argue as in Part 2 above to prove that $\varphi \equiv 0$ on ω and, therefore, $\varphi \equiv 0$.

Part 4. Assume that γ satisfies either hypothesis 3(b) or 3(c) in the statement of Theorem 3.1.

Let $\varphi \in C([0,T]; H_0)$ be a solution of (4.1) with $\varphi(\gamma(t), t) = 0$ for all $t \in [0,T]$. Obviously, this solution can be extended naturally to all $t \leq T$, and the extension is analytic.

On the other hand, by hypothesis, γ can be also extended to an analytic curve $\bar{\gamma} : (-\infty, T] \to \Omega$. Therefore, the composition $\varphi(\bar{\gamma}(t), t)$ is still analytic in $t \in (-\infty, T]$.

Now observe that $\bar{\gamma}$ and γ coincide in $t \in [0,T]$. Therefore, $\varphi(\bar{\gamma}(t), t) = \varphi(\gamma(t), t) = 0$ for any $t \in [0,T]$ and by the unique continuation of analytic functions we deduce that

$$\varphi(\bar{\gamma}(t), t) = 0 \qquad \forall t \in (-\infty, T]. \tag{4.8}$$

Let us introduce the Fourier representation of φ

$$\varphi(x,t) = \sum_{j=1}^{\infty} c_j e^{-\lambda_j (T-t)} w_j(x),$$

where

$$0 < \lambda_1 < \lambda_2 < \cdots < \lambda_j < \cdots$$

are the eigenvalues of (3.13), $w_j(x)$ is the eigenfunction associated to λ_j, and c_j is the Fourier coefficients. We choose $\{w_j(x)\}_{j \geq 1}$ to be orthonormal in $H_0^1(\Omega)$. Observe that $\varphi(T) = \varphi^0 \in H_0 = H_0^1(\Omega)$ and this implies that

$$\sum_{j \geq 1} |c_j|^2 < \infty.$$

From (4.8) we have

$$0 = \varphi(\bar{\gamma}(t), t) = \sum_{j=1}^{\infty} c_j e^{-\lambda_j (T-t)} w_j(\bar{\gamma}(t)) \quad \forall t \in (-\infty, T]. \tag{4.9}$$

We have to prove that this implies that $c_j = 0$ for all $j \geq 1$. We proceed by induction in j. We consider the case $j = 1$. Multiplying the series in (4.9) by $e^{\lambda_1 (T-t)}$ we obtain

$$c_1 w_1(\bar{\gamma}(t)) + \sum_{j=2}^{\infty} c_j e^{(\lambda_1 - \lambda_j)(T-t)} w_j(\bar{\gamma}(t)) = 0 \quad \forall t \in (-\infty, T]. \tag{4.10}$$

The second term on the left-hand side converges to zero as $t \to -\infty$. Indeed,

$$\begin{aligned}\left|\sum_{j=2}^{\infty} c_j e^{(\lambda_1-\lambda_j)(T-t)} w_j(\bar\gamma(t))\right|^2 &\le \left\|\sum_{j=2}^{\infty} c_j e^{(\lambda_1-\lambda_j)(T-t)} w_j\right\|^2_{L^\infty(\Omega)} \\ &\le C\left\|\sum_{j=2}^{\infty} c_j e^{(\lambda_1-\lambda_j)(T-t)} w_j\right\|^2_{H_0^1(\Omega)} = C\sum_{j=2}^{\infty} e^{2(\lambda_1-\lambda_j)(T-t)}|c_j|^2,\end{aligned} \tag{4.11}$$

which converges to zero as $t \to -\infty$. Consequently, the first term in (4.10) converges to zero, as $t \to -\infty$, too.

Assume that γ satisfies hypothesis 3(b) or 3(c) in the statement of Theorem 3.1. Let $x_0 \in \Omega$ be an accumulation point of $\bar\gamma(t)$ as $t \to -\infty$, and consider $t_n \to -\infty$ such that

$$\bar\gamma(t_n) \to x_0.$$

Passing to the limit as $(x_n, t_n) \to (x_0, -\infty)$ in (4.10) we obtain

$$c_1 w_1(x_0) = 0. \tag{4.12}$$

As the zeros of the first eigenfunction w_1 lie on the boundary of Ω we deduce that $c_1 = 0$.

Now we complete the induction argument in j. Assume that $c_j = 0$ for all $j < J$, $J \ge 2$, and let us prove that $c_J = 0$. First of all, we observe that (4.9) reads as follows:

$$0 = \varphi(\bar\gamma(t), t) = \sum_{j=J}^{\infty} c_j e^{-\lambda_j(T-t)} w_j(\bar\gamma(t)) \quad \forall t \in (-\infty, T]. \tag{4.13}$$

Multiplying now the series in (4.13) by $e^{\lambda_J(T-t)}$ we obtain

$$c_J w_J(\bar\gamma(t)) + \sum_{j=J+1}^{\infty} c_j e^{(\lambda_J-\lambda_j)(T-t)} w_j(\bar\gamma(t)) = 0 \quad \forall t \in (-\infty, T]. \tag{4.14}$$

Once again, it is easy to see that the second term on the left-hand side converges to zero as $t \to -\infty$.

Let $x_0 \in \Omega$ be an accumulation point of $\bar\gamma(t)$, as $t \to -\infty$, and consider $t_n \to -\infty$ such that

$$\bar\gamma(t_n) \to x_0.$$

Passing to the limit as $(x_n, t_n) \to (x_0, -\infty)$ in (4.10) we obtain

$$c_J w_J(x_0) = 0. \tag{4.15}$$

Now we distinguish the cases where γ satisfies hypothesis 3(b) or 3(c) in the statement of Theorem 3.1.

Assume that γ satisfies hypothesis 3(b). Then we can choose x_0 in such a way that x_0/L is irrational. As the zeros of the eigenfunctions w_j lie on points $x \in \Omega$ such that x/L is rational, we deduce that $c_J = 0$.

On the other hand, if γ satisfies hypothesis 3(c) in the statement of Theorem 3.1, then (4.15) is not enough to guarantee that $c_J = 0$ because the eigenfunction w_J may vanish at x_0.

Multiplying (4.14) by $(\bar\gamma(t) - x_0)^{-1}$ we obtain

$$0 = \frac{c_J w_J(\bar\gamma(t_n))}{\bar\gamma(t_n) - x_0} + \sum_{j=J+1}^{\infty} \frac{e^{(\lambda_J - \lambda_j)(T-t_n)}}{\bar\gamma(t_n) - x_0} c_j w_j(\bar\gamma(t_n)). \tag{4.16}$$

Due to (4.15), the first term on the right-hand side of (4.16) converges to $c_J w_J'(x_0)$ as $t_n \to -\infty$.

On the other hand, following the argument in (4.11) we easily find the following bound for the second term in (4.16):

$$\begin{aligned}\left| \sum_{j=J+1}^{\infty} \frac{e^{(\lambda_J - \lambda_j)(T-t_n)}}{\bar\gamma(t_n) - x_0} c_j w_j(\bar\gamma(t_n)) \right|^2 &\le \sum_{j=J+1}^{\infty} \frac{e^{-2(T-t_n)(\lambda_{J+1} - \lambda_J)}}{|\bar\gamma(t_n) - x_0|^2} |c_j|^2 \\ &\le \frac{e^{-2(T-t_n)(\lambda_3 - \lambda_2)}}{|\bar\gamma(t_n) - x_0|^2} \sum_{j=J+1}^{\infty} |c_j|^2,\end{aligned}$$

which converges to zero as $t_n \to -\infty$ by the hypothesis (3.11).

Therefore, passing to the limit as $t_n \to -\infty$ in (4.16) we obtain

$$c_J w_J'(x_0) = 0. \tag{4.17}$$

From (4.15), (4.17), and the fact that the eigenfunction w_J satisfies a linear second order ordinary differential equation we deduce that $c_J = 0$. □

4.2. The higher-dimensional case. The following lemma reduces the unique continuation problem (4.2) to a certain unique continuation property for the eigenfunctions of (3.13).

LEMMA 4.2. *Assume that $\{\gamma(t)\}_{0\le t\le T}$ is an analytic family of k-dimensional manifolds on Ω which satisfies the hypothesis* (3.12). *Let us consider the set of accumulation points P, and for each $x \in P$, the set of accumulation directions D_x, as defined in Definition* 3.2. *If the spectral unique continuation property* (3.16) *holds, then the unique continuation property* (4.2) *for the solutions of the adjoint problem* (4.1) *holds as well.*

Proof. Let $\varphi \in C([0,T]; H_0)$ be a solution of (4.1) with $\varphi(x,t) = 0$ for all $(x,t) \in \{\gamma(t)\}_{0\le t\le T}$. Obviously, this solution can be extended naturally to all $t \le T$ by solving (4.1). As the Laplace operator generates an analytic semigroup (see, for example, [P, p. 211]), the solution $\varphi : \Omega \times (-\infty, T) \to \mathbb{R}$ of system (4.1) is analytic.

On the other hand, the family $\{\gamma(t)\}_{0\le t\le T}$ can be extended to an analytic family of manifolds $\{\bar\gamma(t)\}_{-\infty<t\le T}$. Let $\{\psi_\alpha(y,t)\}_{\alpha=1}^{A}$ be a family of charts for $\{\bar\gamma(t)\}_{-\infty<t\le T}$, as in Definition 3.1. Then, for any $y \in V_\alpha$, $\psi_\alpha(y,t)$ is analytic in t and therefore, the composition $\varphi(\psi_\alpha(y,t),t)$ is still analytic in t.

Thus, the fact that $\varphi(x,t) = 0$ for all $(x,t) \in \{\gamma(t)\}_{0\le t\le T}$ implies that $\varphi(\psi_\alpha(y,t),t)$ vanishes for $t \in [0,T]$ and $y \in V_\alpha$ with $\alpha \in A$. Therefore, by the analyticity of $\varphi(\psi_\alpha(y,t),t)$ we have that

$$\varphi(\psi_\alpha(y,t),t) = 0 \quad \forall t \in (-\infty, T] \quad \text{and } \forall y \in V_\alpha \text{ with } \alpha = 1, \dots, A, \tag{4.18}$$

i.e.,

$$\varphi(x,t) = 0 \quad \forall (x,t) \in \{\bar{\gamma}(t)\}_{-\infty<t\leq T}. \tag{4.19}$$

Let us introduce the Fourier representation of φ:

$$\varphi(x,t) = \sum_{j=1}^{\infty} e^{-\lambda_j (T-t)} \sum_{k=1}^{l(j)} c_{j,k} w_{j,k}(x),$$

where

$$0 < \lambda_1 < \lambda_2 < \cdots < \lambda_j < \cdots$$

are the eigenvalues of (3.13), $\{l(j)\}_{j\geq 1}$ is their multiplicity, and $\{w_{j,k}(x)\}_{k=1,\ldots,l(j)}$ is a system of linear independent eigenfunctions associated to λ_j. We choose $\{w_{j,k}(x)\}_{j,k\geq 1}$ to be an orthonormal basis of H_0. Observe that $\varphi(T) = \varphi^0 \in H_0$ and this implies that

$$\sum_{j,k} |c_{j,k}|^2 < \infty.$$

From (4.19) we have

$$0 = \varphi(x,t) = \sum_{j=1}^{\infty} e^{-\lambda_j (T-t)} \sum_{k=1}^{l(j)} c_{j,k} w_{j,k}(x) \quad \forall (x,t) \in \{\gamma(t)\}_{-\infty<0\leq T}. \tag{4.20}$$

We have to prove that this implies that $c_{j,k} = 0$ for all $j \geq 1$ and $k = 1, \ldots, l(j)$. We proceed by induction in j.

First, we consider the case $j = 1$. Observe that the first eigenvalue of the Laplace operator λ_1 is simple. Then $l(1) = 1$ and we have to prove only that $c_{1,1} = 0$. Observe also that, from the spectral unique continuation property (3.16), it suffices to prove that

$$\begin{gathered} c_{1,1} w_{1,1}(x_0 + \mu v) = 0 \quad \forall x_0 \in P,\ \forall v \in D_{x_0}, \\ \text{and } \forall \mu \in \mathbb{R} \text{ s.t. } x_0 + \bar{\mu} v \in \Omega\ \forall \bar{\mu} \in [0,\mu]. \end{gathered} \tag{4.21}$$

Moreover, by the analyticity of the eigenfunction $c_{1,1} w_{1,1}$, this is equivalent to proving that the function $G_{x_0,v,1}(\mu) = c_{1,1} w_{1,1}(x_0 + \mu v)$ satisfies

$$G_{x_0,v,1}(0) = \frac{d^r G_{x_0,v,1}}{d\mu^r}(0) = 0 \quad \forall r \geq 1,\ \forall x_0 \in P, \text{ and } \forall v \in D_{x_0}. \tag{4.22}$$

We use an induction argument in r. We start by proving that $G_{x_0,v,1}(0) = c_{1,1} w_{1,1}(x_0) = 0$.

Multiplying the series in (4.20) by $e^{\lambda_1 (T-t)}$ and taking into account that λ_1 is simple we obtain

$$c_{1,1} w_{1,1}(x) + \sum_{j=2}^{\infty} e^{(\lambda_1 - \lambda_j)(T-t)} \sum_{k=1}^{l(j)} c_{j,k} w_{j,k}(x) = 0 \quad \forall (x,t) \in \{\gamma(t)\}_{-\infty<t\leq T}. \tag{4.23}$$

The second term on the left-hand side converges to zero as $t \to -\infty$ uniformly in $x \in \Omega$. Indeed,

$$\left|\sum_{j=2}^{\infty} e^{(\lambda_1-\lambda_j)(T-t)} \sum_{k=1}^{l(k)} c_{j,k} w_{j,k}(x)\right|^2 \le \left\|\sum_{j=2}^{\infty} e^{(\lambda_1-\lambda_j)(T-t)} \sum_{k=1}^{l(k)} c_{j,k} w_{j,k}\right\|^2_{L^\infty(\Omega)}$$
$$\le C \left\|\sum_{j=2}^{\infty} e^{(\lambda_1-\lambda_j)(T-t)} \sum_{k=1}^{l(k)} c_{j,k} w_{j,k}\right\|^2_{H_1} = C \sum_{j=2}^{\infty} e^{2(\lambda_1-\lambda_j)(T-t)} \sum_{k=1}^{l(k)} |c_{j,k}|^2,$$

which converges to zero as $t \to -\infty$.

Consider now $x_0 \in P$ and $v \in D_{x_0}$. There exists a sequence $(x_n, t_n) \in \{\bar\gamma(t)\}_{-\infty<t\le T}$ such that

$$x_n \to x_0, \quad \frac{x_n - x_0}{\|x_n - x_0\|} \to v, \quad \text{and} \quad t_n \to -\infty. \tag{4.24}$$

Passing to the limit as $(x_n, t_n) \to (x_0, -\infty)$ in (4.23) we obtain

$$G_{x_0,v,1}(0) = c_{1,1} w_{1,1}(x_0) = 0. \tag{4.25}$$

Now we complete the induction argument on r to prove (4.22). We assume that $\frac{d^r G_{x_0,v,1}}{d\mu^r}(0) = 0$ for $r \le R-1$. Then, from the Taylor expansion of $G_{x_0,v,1}$ at $\mu = 0$ we have

$$\frac{d^R G_{x_0,v,1}}{d\mu^R}(0) = \lim_{\mu\to 0} \frac{R!\, G_{x_0,v,1}(\mu)}{\mu^R} = \lim_{\mu_n\to 0} \frac{R!\, G_{x_0,v_n,1}(\mu_n)}{\mu_n^R}, \tag{4.26}$$

where we can choose x_n and v_n as in (4.24) and $\mu_n = \|x_n - x_0\|$.

On the other hand, we have

(4.27)

$$\left|\frac{G_{x_0,v_n,1}(\mu_n)}{\mu_n^R}\right|^2 = \left|c_{1,1}\frac{w_{1,1}(x_n)}{\|x_n - x_0\|^R}\right|^2 = \left|\sum_{j=2}^{\infty} \frac{e^{(\lambda_1-\lambda_j)(T-t_n)}}{\|x_n - x_0\|^R} \sum_{k=1}^{l(j)} c_{j,k} w_{j,k}(x_n)\right|^2$$
$$\le \left\|\sum_{j=2}^{\infty} \frac{e^{(\lambda_1-\lambda_j)(T-t_n)}}{\|x_n - x_0\|^R} \sum_{k=1}^{l(j)} c_{j,k} w_{j,k}\right\|^2_{L^\infty} \le C \left\|\sum_{j=2}^{\infty} \frac{e^{(\lambda_1-\lambda_j)(T-t_n)}}{\|x_n - x_0\|^R} \sum_{k=1}^{l(j)} c_{j,k} w_{j,k}\right\|^2_{H_1}$$
$$= C \sum_{j=2}^{\infty} \frac{e^{2(\lambda_1-\lambda_j)(T-t_n)}}{\|x_n - x_0\|^{2R}} \sum_{k=1}^{l(j)} |c_{j,k}|^2 \le C \frac{e^{-2(\lambda_2-\lambda_1)(T-t_n)}}{\|x_n - x_0\|^{2R}} \sum_{j=2}^{\infty} \sum_{k=1}^{l(j)} |c_{j,k}|^2,$$

which converges to zero as $t_n \to -\infty$ since the factor

$$\frac{e^{-2(\lambda_2-\lambda_1)(T-t_n)}}{\|x_n - x_0\|^{2R}} = \left(\frac{e^{-\frac{\lambda_2-\lambda_1}{R}(T-t_n)}}{\|x_n - x_0\|}\right)^{2R}$$

converges to zero as $t_n \to -\infty$, i.e., as $\mu \to 0$, due to the hypothesis on the accumulation directions given in (3.15).

Here we see why we include only "nonexponential directions" in the definition of D_{x_0} in (3.15).

This concludes the proof of (4.22). From (4.22) we deduce (4.21) and then

$$c_{1,1} = 0.$$

Following an induction argument in j we easily obtain

$$\frac{d^r G_{x_0,v,j}}{d\mu^r}(0) = 0 \quad \forall j \text{ and } \forall x_0 \in P,\ \forall v \in D_{x_0}, \tag{4.28}$$

where $G_{x_0,v,j}(\mu) = \sum_{k=1}^{l(j)} c_{j,k} w_{j,k}(x_0 + \mu v)$.

From (4.28) and the analyticity of the eigenfunctions $w_{j,k}$ we have

$$\sum_{k=1}^{l(j)} c_{j,k} w_{j,k}(x_0 + \mu v) = 0 \quad \forall j,\ \forall x_0 \in P,\ \forall v \in D_{x_0},\ \text{and } \forall \mu \in \mathbb{R} \text{ s.t. } [x_0, x_0 + \mu v] \in \Omega. \tag{4.29}$$

On the other hand, taking into account the fact that $\sum_{k=1}^{l(j)} c_{j,k} w_{j,k}$ is an eigenfunction and by the unique continuation hypothesis (3.16) for the eigenfunctions we obtain

$$\sum_{k=1}^{l(j)} c_{j,k} w_{j,k} \equiv 0 \quad \forall j \geq 1.$$

Therefore $c_{j,k} = 0$ for all $k = 1, \ldots, l(j)$ because of the linear independence of $w_{j,k}$. This concludes the proof of the lemma. □

5. Averaging of rapidly oscillating controls. In this section we prove Theorem 3.3. We introduce the variational approach of the HUM method to characterize the controls f_ε of (3.17) of minimal norm (see [L1] and [L2]). Then we pass to the limit, as $\varepsilon \to 0$, in these equations to obtain the limit control. The main ingredient to establish the convergence is Lemma 5.1 below. The proof of Theorem 3.3 will be given after the proof of Lemma 5.1.

LEMMA 5.1. *Let $\{\gamma(t)\}_{0 \leq t \leq T}$ be an analytic 2π-periodic family of k-dimensional manifolds. Consider a sequence $u_\varepsilon^0 \rightharpoonup u^0$ that weakly converges in H_0. Let u_ε, u be the solutions of the homogeneous system* (3.1) *with $f = 0$, and initial data u_ε^0, u^0 respectively. Let $\gamma_\varepsilon(t) = \gamma(t/\varepsilon)$. Then*

$$\begin{aligned} &\int_0^T \int_{\gamma_\varepsilon(t)} |u_\varepsilon(x,t)|^2 \, d\sigma \, dt \to \int_0^T \langle u(x,t) m_\gamma(x), u(x,t) \rangle_1 dt \quad \text{if } k \geq 1, \\ &\int_0^T |u_\varepsilon(\gamma_\varepsilon(t), t)|^2 \, dt \to \int_0^T \langle u(x,t) m_\gamma(x), u(x,t) \rangle_1 \quad \text{if } k = 0, \end{aligned} \tag{5.1}$$

where $m_\gamma(x)$ is the weak- limit of $\delta_{\gamma(t/\varepsilon)}$ in $L^\infty(0,T;H_1')$, and $\langle \cdot, \cdot \rangle_1$ denotes the duality pairing between H_1 and its dual.*

The density $m_\gamma(x)$ does not depend on the time variable t; it is supported on the nonempty open set ω, defined as the interior set, with respect to the relative topology, of the range of $\{\gamma(t)\}_{0 \leq t \leq 2\pi}$, and it is characterized by (3.20).

Moreover, if $\varphi \in C_0^\infty(\Omega \times (0,T))$, *then*

$$\int_0^T \int_{\gamma_\varepsilon(t)} u_\varepsilon(x,t)\varphi(x,t)\ d\sigma\ dt \to \int_0^T \langle u(x,t)m_\gamma(x), \varphi(x,t)\rangle_1 dt \quad if\ k \geq 1,$$

$$\int_0^T u_\varepsilon(\gamma_\varepsilon(t),t)\varphi(x,t)dt \to \int_0^T \langle u(x,t)m_\gamma(x), \varphi(x,t)\rangle_1 dt \quad if\ k = 0. \tag{5.2}$$

Proof of Lemma 5.1. To simplify the presentation we consider the case $k \geq 1$. The case $k = 0$ is analogous.

The sequence $u_\varepsilon(x,t)$ of solutions of the homogeneous system (3.1) with $f = 0$ and initial data u_ε^0 can be written in the Fourier representation

$$u_\varepsilon(x,t) = \sum_{j=1}^{\infty} e^{-\lambda_j t} \sum_{k=1}^{l(j)} c_{j,k}^\varepsilon w_{j,k}(x).$$

We assume that $(w_{j,k})_{j,k\geq 1}$ constitute an orthonormal basis in H_0. Analogously, the solution $u(x,t)$ of the homogeneous system (2.4) with $f = 0$ and initial data u^0, is

$$u(x,t) = \sum_{j=1}^{\infty} e^{-\lambda_j t} \sum_{k=1}^{l(j)} c_{j,k} w_{j,k}(x).$$

Due to the weak convergence of the initial data $u_\varepsilon^0 \rightharpoonup u^0$ in H_0 we have

$$\sum_{j,k\geq 1} |c_{j,k}^\varepsilon|^2 \leq C, \qquad \sum_{j,k\geq 1} |c_{j,k}|^2 \leq C, \tag{5.3}$$

with C independent of ε. Moreover,

$$c_{j,k}^\varepsilon \to c_{j,k} \qquad \text{as } \varepsilon \to 0 \qquad \forall j,k \geq 1.$$

Let us prove the convergence result stated in (5.1). To avoid the singularity of the solution u_ε at $t = 0$ we divide the left-hand side integral in (5.1) into two parts:

$$\begin{aligned} \int_0^T \int_{\gamma_\varepsilon(t)} |u_\varepsilon(x,t)|^2\ d\sigma\ dt &= \int_0^\delta \int_{\gamma_\varepsilon(t)} |u_\varepsilon(x,t)|^2\ d\sigma\ dt \\ &\quad + \int_\delta^T \int_{\gamma_\varepsilon(t)} |u_\varepsilon(x,t)|^2\ d\sigma\ dt \end{aligned} \tag{5.4}$$

with $\delta > 0$ to be chosen later.

We first estimate the first integral in (5.4). By classical estimates on the heat kernel (see [CH, p. 44]) we know that

$$\|u_\varepsilon(\cdot,t)\|_{L^\infty(\Omega)} \leq Ct^{-\frac{n}{2q}} \left\|u_\varepsilon^0\right\|_{L^q(\Omega)}$$

with C a constant that does not depend on ε. Therefore the first integral in (5.4) can be estimated by

$$\begin{aligned} \int_0^\delta \int_{\gamma_\varepsilon(t)} |u_\varepsilon(x,t)|^2\ d\sigma\ dt &\leq C \left\|u_\varepsilon^0\right\|_{L^q(\Omega)}^2 \int_0^\delta t^{-\frac{n}{q}} \int_{\gamma_\varepsilon(t)} d\sigma\ dt \\ &\leq C \left\|u_\varepsilon^0\right\|_{L^q(\Omega)}^2 \max_{t\in[0,\delta]} (\text{meas } (\gamma(t/\varepsilon))) \int_0^\delta t^{-\frac{n}{q}} dt \\ &= C \left\|u_\varepsilon^0\right\|_{L^q(\Omega)}^2 \frac{\delta^{1-\frac{n}{q}}}{1-\frac{n}{q}} \max_{t\in[0,2\pi]} \text{meas } (\gamma(t)). \end{aligned}$$

We set $q = 2$ if $n = 1$ and $q = 4$ if $n = 2, 3$. Then, by the continuous Sobolev embeddings $H_0^1(\Omega) \to L^\infty(\Omega)$ if $n = 1$ and $H_0^1(\Omega) \to L^4(\Omega)$ for $n = 2, 3$ we easily deduce the estimate

$$\int_0^\delta \int_{\gamma_\varepsilon(t)} |u_\varepsilon(x,t)|^2 \, d\sigma \, dt \le C(\delta, n, \gamma) \left\| u_\varepsilon^0 \right\|_{H_0}^2 \tag{5.5}$$

with $C(\delta, n, \gamma)$ independent of ε and such that

$$C(\delta, n, \gamma) \to 0 \quad \text{as } \delta \to 0.$$

Note that the estimate above holds since the manifold γ is time-periodic and consequently

$$\max_{t\in[0,\delta]} (\text{meas } \gamma(t/\varepsilon)) = \max_{t\in[0,2\pi]} (\text{meas } \gamma(t/\varepsilon)) < \infty$$

as $\varepsilon \to 0$.

Thus, to prove the convergence result stated in (5.1), it suffices to show that the second integral in (5.4), for $\delta > 0$ fixed, tends to

$$\int_\delta^T \langle u(x,t) m_\gamma(x,t), u(x,t) \rangle_1 dt$$

as $\varepsilon \to 0$. Indeed, once this is proved, (5.1) is obtained passing to the limit, as $\delta \to 0$, in (5.4).

We have

$$\begin{aligned}\int_\delta^T \int_{\gamma_\varepsilon(t)} |u_\varepsilon(x,t)|^2 \, d\sigma \, dt &= \int_\delta^T \int_{\gamma_\varepsilon(t)} \sum_{j,i=1}^{\infty} \sum_{k=1}^{l(j)} \sum_{m=1}^{l(i)} e^{-(\lambda_i+\lambda_j)t} c_{j,k}^\varepsilon c_{i,m}^\varepsilon w_{j,k}(x) w_{i,m}(x) d\sigma \, dt \\ &= \sum_{j,i=1}^{\infty} \sum_{k=1}^{l(j)} \sum_{m=1}^{l(i)} \int_\delta^T \int_{\gamma_\varepsilon(t)} e^{-(\lambda_i+\lambda_j)t} c_{j,k}^\varepsilon c_{i,m}^\varepsilon w_{j,k}(x) w_{i,m}(x) d\sigma \, dt.\end{aligned}$$

Now we take the limit as $\varepsilon \to 0$,

$$\begin{aligned}&\lim_{\varepsilon\to 0} \int_\delta^T \int_{\gamma_\varepsilon(t)} |u_\varepsilon(x,t)|^2 \, d\sigma \, dt \\ &\quad = \lim_{\varepsilon\to 0} \sum_{j,i=1}^{\infty} \sum_{k=1}^{l(j)} \sum_{m=1}^{l(i)} \int_\delta^T \int_{\gamma_\varepsilon(t)} e^{-(\lambda_i+\lambda_j)t} c_{j,k}^\varepsilon c_{i,m}^\varepsilon w_{j,k}(x) w_{i,m}(x) \, d\sigma \, dt \\ &\quad = \sum_{j,i=1}^{\infty} \sum_{k=1}^{l(j)} \sum_{m=1}^{l(i)} c_{j,k} c_{i,m} \lim_{\varepsilon\to 0} \int_\delta^T \int_{\gamma_\varepsilon(t)} e^{-(\lambda_i+\lambda_j)t} w_{j,k}(x) w_{i,m}(x) \, d\sigma \, dt. \qquad (5.6)\end{aligned}$$

Interchanging the sum and the limit is justified because of the dominated convergence

theorem. Indeed, each term of the series can be bounded above as follows:

$$\begin{aligned}
&\left| c^{\varepsilon}_{j,k} c^{\varepsilon}_{i,m} \int_{\delta}^{T} \int_{\gamma_\varepsilon(t)} e^{-(\lambda_i+\lambda_j)t} w_{j,k}(x) w_{i,m}(x)\, d\sigma\, dt \right| \\
&\quad \le |c^{\varepsilon}_{j,k} c^{\varepsilon}_{i,m}| \max_{t\in[\delta,T]} \left| \int_{\gamma_\varepsilon(t)} w_{j,k}(x) w_{i,m}(x)\, d\sigma \right| \int_{\delta}^{T} e^{-(\lambda_i+\lambda_j)t} dt \\
&\quad \le \left(\sum_{i=1}^{\infty} \sum_{m=1}^{l(i)} |c^{\varepsilon}_{i,m}|^2 \right) \max_{t\in[0,2\pi]} \left| \int_{\gamma(t)} w_{j,k}(x) w_{i,m}(x)\, d\sigma \right| \frac{e^{-(\lambda_i+\lambda_j)\delta} - e^{-(\lambda_i+\lambda_j)T}}{\lambda_i+\lambda_j} \\
&\quad \le \|u^{\varepsilon}_0\|^2_{H_0} \max_{t\in[0,2\pi]} \left| \int_{\gamma(t)} w_{j,k}(x) w_{i,m}(x)\, d\sigma \right| \frac{e^{-(\lambda_i+\lambda_j)\delta}}{\lambda_i+\lambda_j} \\
&\quad \le C \max_{t\in[0,2\pi]} \left| \int_{\gamma(t)} w_{j,k}(x) w_{i,m}(x)\, d\sigma \right| \frac{e^{-(\lambda_i+\lambda_j)\delta}}{\lambda_i+\lambda_j}, \qquad (5.7)
\end{aligned}$$

since, by hypothesis, the sequence of initial data u^0_ε is uniformly bounded in H_0 (see (5.3)). Now, by the Hölder inequality and the trace theorem,

$$\begin{aligned}
\left| \int_{\gamma(t)} w_{j,k}(x) w_{i,m}(x)\, d\sigma \right| &\le \|w_{j,k}\|_{L^2(\gamma(t))} \|w_{i,m}\|_{L^2(\gamma(t))} \\
&\le C(\gamma(t)) \|w_{j,k}\|_{H_1} \|w_{i,m}\|_{H_1}, \qquad (5.8)
\end{aligned}$$

where $C(\gamma(t))$ is a constant that depends only on the measure of $\gamma(t)$. Note that, in view of the hypothesis on $\gamma(t)$, this constant can be chosen independently of time; i.e., there exists $C(\gamma)$ independent of t such that $C(\gamma(t)) \le C(\gamma)$ for all $t \in [0, 2\pi]$. Moreover, taking into account the normalization of the eigenfunctions in H_0, we have

$$\|w_{j,k}\|_{H_1} \le C\sqrt{\lambda_j}\, \|w_{j,k}\|_{H_0} = C\sqrt{\lambda_j},$$

and the terms in (5.7) can be bounded above by

$$C(\gamma) \frac{\sqrt{\lambda_j}\sqrt{\lambda_i}}{\lambda_j+\lambda_i} e^{-(\lambda_j+\lambda_i)\delta} \le C(\gamma) e^{-(\lambda_j+\lambda_i)\delta} \qquad (5.9)$$

with $C(\gamma)$ a constant that does not depend on ε, i, j, k, m. The sum of all these terms in i, j, k, m is finite due to the well-known asymptotic behavior of the eigenvalues of the Laplace operator. Indeed,

$$\sum_{i,j\ge1} \sum_{m=1}^{l(i)} \sum_{k=1}^{l(j)} e^{-(\lambda_i+\lambda_j)\delta} = \left(\sum_{i\ge1} \sum_{m=1}^{l(i)} e^{-\lambda_i\delta} \right)^2.$$

The last sum is finite. Recall that the number of eigenvalues less than a constant λ, including multiplicity, is asymptotically equal to $\lambda|\Omega|/4\pi$ if $n = 2$, and $\lambda^{3/2}|\Omega|/6\pi^2$ if $n = 3$ (see [CoH, p. 442]). Then, for example, in the case $n = 3$ we have

$$\sum_{i\ge1} \sum_{m=1}^{l(i)} e^{-\lambda_i\delta} = \sum_{k=1}^{\infty} \sum_{k-1\le\lambda_i\le k} l(i) e^{-\lambda_i\delta} \le C \sum_{k=1}^{\infty} k^{3/2} e^{-(k-1)\delta} < \infty.$$

Once we have checked (5.6), we pass to the limit in each one of the terms in the right-hand side of (5.6). Then,

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\int_\delta^T\int_{\gamma_\varepsilon(t)} e^{-(\lambda_i+\lambda_j)t}w_{j,k}(x)w_{i,m}(x)\ d\sigma\ dt\\
&\quad=\lim_{\varepsilon\to 0}\int_\delta^T \langle \delta_{\gamma_\varepsilon(t)}, e^{-(\lambda_i+\lambda_j)t}w_{j,k}(x)w_{i,m}(x)\rangle_1\\
&\quad=\int_\delta^T \langle m_\gamma(x,t), e^{-(\lambda_j+\lambda_i)t}w_{j,k}(x)w_{i,m}(x)\rangle_1\ dt\\
&\quad=\int_\delta^T \langle e^{-\lambda_i t}w_{i,m}(x)m_\gamma(x,t), e^{-\lambda_j t}w_{j,k}(x)\rangle_1\ dt,
\end{aligned}$$

where $m_\gamma(x,t)$ is the weak-* limit of $\delta_{\gamma_\varepsilon(t)}$ in $L^\infty(0,T;H_1)$ and $\langle\cdot,\cdot\rangle_1$ is the duality pairing between H_1 and its dual space.

Thus,

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\int_\delta^T\int_{\gamma_\varepsilon(t)} |u_\varepsilon(x,t)|^2 d\sigma\ dt\\
&\quad=\sum_{j,i=1}^{\infty}\sum_{k=1}^{l(j)}\sum_{m=1}^{l(i)} c_{j,k}c_{i,m}\int_\delta^T \langle e^{-\lambda_i t}w_{i,m}(x)m_\gamma(x,t), e^{-\lambda_j t}w_{j,k}(x)\rangle_1\ dt\\
&\quad=\int_\delta^T \langle u(x,t)m_\gamma(x,t), u(x,t)\rangle_1 dt
\end{aligned}$$

for all $\delta>0$.

It remains to prove that the limit density $m_\gamma(x,t)$ does not depend on the time variable t and that it is supported on the region ω, constituted by the range of $\gamma(t)$ over the period $t\in[0,2\pi]$. In fact, the limit density m_γ is characterized by

$$\int_0^T \langle m_\gamma,\varphi\rangle_1 dt=\lim_{\varepsilon\to 0}\int_0^T\int_{\gamma(t/\varepsilon)}\varphi(x,t)d\sigma\ dt\quad\forall\varphi(x,t)\in L^1(0,T;H_1).$$

Taking into account that $C_0^\infty(\Omega)\times C_0^\infty(0,T)$ is sequentially dense in $C_0^\infty(\Omega\times(0,T))$, which is dense in $L^1(0,T;H_1)$, m_γ is also characterized by

$$\int_0^T \langle m_\gamma,\varphi\rangle_1 dt=\lim_{\varepsilon\to 0}\int_0^T\int_{\gamma(t/\varepsilon)}\varphi(x)\psi(t)d\sigma\ dt\quad\forall\varphi(x)\in H_1,\quad \psi(t)\in L^1(0,T).$$

Note that $F(s)=\int_{\gamma(t/\varepsilon)}\varphi(x)d\sigma$ is a 2π-periodic function and then $F(s/\varepsilon)$ converges weakly to its average in L^2_{loc} as $\varepsilon\to 0$. Therefore,

$$\lim_{\varepsilon\to 0}\int_0^T\int_{\gamma_\varepsilon(t)}\varphi(x)\psi(t)\ d\sigma\ dt=\frac{1}{2\pi}\int_0^{2\pi}\int_{\gamma(s)}\varphi(x)\ d\sigma\ ds\int_0^T\psi(t)\ dt. \tag{5.10}$$

This last integral can be written as an integral over the region $\omega\times(0,T)$, where ω is the interior set of the range of $\{\gamma(t)\}_{t\in[0,2\pi]}$. The control is weighted by a suitable density function m_γ, supported on ω, and independent of the time variable t. Moreover, this limit $m_\gamma(x)$ is characterized by

$$\int_\omega \varphi(x)m_\gamma(x)\ d\sigma=\frac{1}{2\pi}\int_0^{2\pi}\int_{\gamma(s)}\varphi(x)\ d\sigma\ ds\quad\forall\varphi\in H_1. \tag{5.11}$$

Note that m_γ depends, roughly, on the time derivative of $\gamma(t)$ over a period $[0, 2\pi]$.

The proof of (5.2) is similar. We have to take into account only that $C_0^\infty(\Omega) \times C_0^\infty(0,T)$ is sequentially dense in $C_0^\infty(\Omega \times (0,T))$ and then it suffices to check (5.2) for test functions in separated variables. This concludes the proof of the lemma. ∎

Proof of Theorem 3.3. We first restrict ourselves, without loss of generality, to the case where $u^0 = 0$ and $\|u^1\|_{H_0} \geq \alpha$.

Given $\varepsilon > 0$, system (3.1) is approximately controllable. Indeed according to Lemma 4.2 and, in view of the assumptions on the curve $\gamma(t)$, the unique continuation property (4.2) holds for $\gamma_\varepsilon(t) = \gamma(t/\varepsilon)$ for all $\varepsilon > 0$. Then the control that makes (3.10) hold is given by $f_\varepsilon = \overline{\varphi}_\varepsilon(\gamma_\varepsilon(t), t)$, where $\overline{\varphi}_\varepsilon$ solves (4.1), the initial data $\overline{\varphi}_\varepsilon^0$ being the minimizer of the functional

$$J_\varepsilon(\varphi^0) = \frac{1}{2}\int_0^T \int_{\gamma_\varepsilon(t)} |\varphi(x,t)|^2 \, d\sigma dt + \alpha \left\|\varphi^0\right\|_{H_0} - \langle u^1, \varphi^0 \rangle_{H_0', H_0} \tag{5.12}$$

over H_0. Note, in particular, that the coercivity of this functional is guaranteed by the unique continuation property (4.2) (see [FPZ]).

The adjoint system associated to the limit system (3.18) is also given by (4.1) and the corresponding functional associated to (3.18) is given by

$$J(\varphi^0) = \frac{1}{2}\int_0^T \int_\omega |\varphi(x,t)|^2 m_\gamma(x) d\sigma dt + \alpha \left\|\varphi^0\right\|_{H_0} - \langle u^1, \varphi^0 \rangle_{H_0', H_0}, \tag{5.13}$$

where φ is the solution of (4.1) with final data φ^0. Recall that Theorem 3.3 is stated for the particular case $k = n-1$ to simplify the presentation. In the general case $(0 \leq k \leq n-1)$, the first term in (5.13) would be

$$\frac{1}{2}\int_0^T \langle \varphi(x,t) m_\gamma(x), \varphi(x,t) \rangle_1 \, dt, \tag{5.14}$$

and the rest of this proof could be adapted in a straightforward manner. Note that when $k = n-1$, the weak limit of $\delta_{\gamma_\varepsilon(t)}$ is supported in an open subset $\omega \subset \Omega$ and then we can write (5.14) as in (5.13).

We set

$$M_\varepsilon = \inf_{\varphi^0 \in H_0} J_\varepsilon(\varphi^0). \tag{5.15}$$

For each $\varepsilon > 0$ the functional J_ε achieves its minimum M_ε in H_0. This is a consequence of the unique continuation property (4.2) which allows us to prove the coercivity of J_ε for each $\varepsilon > 0$. This unique continuation property is obtained applying the result of Lemma 4.2 to the curve $\gamma_\varepsilon(t) = \gamma(t/\varepsilon)$, which satisfies the hypotheses of Lemma 4.2.

Lemma 5.2 below establishes that the coercivity of J_ε is in fact uniform in ε. Moreover, if $f(t) = \overline{\varphi}_\varepsilon(\gamma(t/\varepsilon), t)$, where $\overline{\varphi}_\varepsilon$ solves (4.1) with data $\overline{\varphi}_\varepsilon^0$, the solution of (2.4) satisfies (3.10).

LEMMA 5.2. *We have*

$$\liminf_{\substack{\|\varphi^0\|_{H_0} \to \infty \\ \varepsilon \to 0}} \frac{J_\varepsilon(\varphi^0)}{\|\varphi^0\|_{H_0}} \geq \alpha. \tag{5.16}$$

Furthermore, the minimizers $\{\overline{\varphi_\varepsilon^0}\}_{\varepsilon \geq 0}$ *are uniformly bounded in* H_0.

Proof of Lemma 5.2. Let us consider sequences $\varepsilon_j \to 0$ and $\varphi^0_{\varepsilon_j} \in H_0$ such that $\|\varphi^0_{\varepsilon_j}\|_{H_0} \to \infty$ as $j \to \infty$.

Let us introduce the normalized data

$$\eta^0_{\varepsilon_j} = \frac{\varphi^0_{\varepsilon_j}}{\|\varphi^0_{\varepsilon_j}\|_{H_0}}$$

and the corresponding solutions of (4.1):

$$\eta_{\varepsilon_j} = \frac{\varphi_{\varepsilon_j}}{\|\varphi^0_{\varepsilon_j}\|_{H_0}}.$$

We have

$$I_j = \frac{J_{\varepsilon_j}(\varphi^0_{\varepsilon_j})}{\|\varphi^0_{\varepsilon_j}\|_{H_0}} = \frac{1}{2}\|\varphi^0_{\varepsilon_j}\|_{H_0} \int_0^T \int_{\gamma_{\epsilon_j}(t)} |\eta_{\varepsilon_j}(x,t)|^2 d\sigma dt + \alpha - \langle u^1, \psi^0_{\varepsilon_j}\rangle_{H_0',H_0}.$$

We distinguish the following two cases:

Case 1. $\liminf_{j\to\infty} \int_0^T \int_{\gamma_{\epsilon_j}(t)} |\eta_{\varepsilon_j}(x,t)|^2 d\gamma_\varepsilon dt > 0$. In this case, we have clearly $\liminf_{j\to\infty} I_j = \infty$.

Case 2. $\liminf_{j\to\infty} \int_0^T \int_{\gamma_{\epsilon_j}(t)} |\eta_{\varepsilon_j}(x,t)|^2 d\gamma_\varepsilon dt = 0$. In this case we argue by contradiction. Assume that there exists a subsequence, still denoted by the index j, such that

$$\int_0^T \int_{\gamma_{\epsilon_j}(t)} |\eta_{\varepsilon_j}(x,t)|^2 d\sigma dt \to 0 \tag{5.17}$$

and

$$\liminf_{j\to\infty} I_j < \alpha. \tag{5.18}$$

By extracting a subsequence, still denoted by the index j, we have

$$\eta^0_{\varepsilon_j} \rightharpoonup \eta^0 \text{ weakly in } H_0$$

and therefore

$$\eta_{\varepsilon_j} \rightharpoonup \eta \text{ weakly-* in } L^\infty(0,T;H_0),$$

where η is the solution of (4.1) with initial data η^0. By Lemma 5.1 we have

$$\eta = 0 \text{ in } \gamma_\varepsilon(t) \times (0,T).$$

Now, recall that, by hypothesis, γ_ε is a strategic curve and then Lemma 4.2 establishes that $\eta^0 = 0$. Thus

$$\eta^0_{\varepsilon_j} \rightharpoonup 0 \text{ weakly in } H_0$$

and therefore

$$\liminf_{j\to\infty} I_j \geq \liminf_{j\to\infty} (\alpha - \langle u^1, \eta^0_{\varepsilon_j}\rangle_{H_0',H_0}) = \alpha$$

since u_j^1 converges strongly in H_0. This is in contradiction with (5.18) and concludes the proof of (5.16).

On the other hand, it is obvious that $I_\varepsilon \leq 0$ for all $\varepsilon > 0$. Thus, (5.16) implies the uniform boundedness of the minimizers in H_0.

Concerning the convergence of the minimizers we have the following lemma.

LEMMA 5.3. *The minimizers $\overline{\varphi}_\varepsilon^0$ of J_ε converge strongly in H_0 as $\varepsilon \to 0$ to the minimizer $\overline{\varphi}^0$ of J in* (5.13) *and M_ε converges to*

$$M = \inf_{\overline{\varphi}^0 \in H_0} J(\overline{\varphi}^0). \tag{5.19}$$

Moreover, the corresponding solutions $\overline{\varphi}_\varepsilon$ of (4.1) *converge in $C([0,T];H_0)$ to the solution $\overline{\varphi}$ as $\varepsilon \to 0$.*

Proof of Lemma 5.3. We adapt a classical argument in Γ-convergence (see [DM]). By extracting a subsequence, which we still denote by ε, we have

$$\overline{\varphi}_\varepsilon^0 \rightharpoonup \eta^0 \text{ weakly in } H_0$$

as $\varepsilon \to 0$. It is sufficient to check that $\overline{\varphi}^0 = \eta^0$ or, equivalently, that η^0 is the minimizer of J, i.e.,

$$J(\eta^0) \leq J(\varphi^0) \quad \forall \varphi^0 \in H_0. \tag{5.20}$$

We know that

$$\overline{\varphi}_\varepsilon \rightharpoonup \eta \text{ weakly-* in } L^\infty(0,T;H_0),$$

where η is the solution of (4.1) with initial data ψ^0. By Lemma 5.1 we deduce that

$$J(\eta^0) = \lim_{\varepsilon \to 0} J_\varepsilon(\overline{\varphi}_\varepsilon^0). \tag{5.21}$$

On the other hand, for each $\varphi^0 \in H_0$ we have

$$\lim_{\varepsilon \to 0} J_\varepsilon(\overline{\varphi}_\varepsilon^0) \leq \lim_{\varepsilon \to 0} J_\varepsilon(\varphi^0). \tag{5.22}$$

Observe also that for $\varphi^0 \in H_0$ fixed, Lemma 5.1 ensures that

$$\lim_{\varepsilon \to 0} J_\varepsilon(\varphi^0) = J(\varphi^0). \tag{5.23}$$

Combining (5.21)–(5.23) it is easy to see that (5.20) holds.

This concludes the proof of the weak convergence of the minimizers and it also shows that

$$\liminf_{\varepsilon \to 0} M_\varepsilon \geq M = J(\bar{\varphi}^0) = \limsup_{\varepsilon \to 0} J_\varepsilon(\overline{\varphi}^0) \geq \limsup_{\varepsilon \to 0} J_\varepsilon(\overline{\varphi}_\varepsilon^0) = \limsup_{\varepsilon \to 0} M_\varepsilon. \tag{5.24}$$

Therefore we deduce that $M_\varepsilon \to M$.

Observe that (5.19), combined with the weak convergence of $\overline{\varphi}_\varepsilon^0$ in H_0, implies that

$$\lim_{\varepsilon \to 0} \left(\frac{1}{2} \int_0^T \int_{\gamma_\varepsilon(t)} |\overline{\varphi}_\varepsilon(x,t)|^2 d\sigma \, dt + \alpha \left\| \overline{\varphi}_\varepsilon^0 \right\|_{H_0} \right) = \frac{1}{2} \int_0^T \int_\omega |\overline{\varphi}|^2 m_\gamma(x) d\omega dt + \alpha \left\| \overline{\varphi}^0 \right\|_{H_0},$$

since the last term in $J_\varepsilon(\bar{\varphi}_\varepsilon^0)$, which is linear in $\bar{\varphi}_\varepsilon^0$, passes trivially to the limit.

This identity, combined with the weak convergence of $\overline{\varphi}_\varepsilon^0$ to $\overline{\varphi}^0$ in H_0 and Lemma 5.1, implies that

$$\overline{\varphi}_\varepsilon^0 \to \overline{\varphi}^0 \text{ strongly in } H_0. \tag{5.25}$$

Therefore, we have

$$\overline{\varphi}_\varepsilon \to \overline{\varphi} \text{ strongly in } C([0,T];H_0).$$

This concludes the proof of Theorem 3.3 when $u^0 = 0$ and $\|u^1\|_{L^2(\Omega)} \geq \alpha$.

Let us consider now the case where u^0 is nonzero. We set $v^1 = v(T)$ where v is the solution of (2.4) with $f = 0$. Now observe that the solution u of (2.4) can be written as $u = v + w$ where w is the solution of (2.4) with zero initial data that satisfies $w(T) = u(T) - v^1$. In this way, the controllability problem for u can be reduced to a controllability problem for w with zero initial data $w^0 = 0$. This is the problem we solved. The proof is now complete. □

Remark 5.1. The proof guarantees that the coercivity property (5.16) is also true for the limit functional J. This fact could also be proved arguing directly on J. In this case we would use the corresponding unique continuation property for the solutions of the adjoint limit system.

6. Further results and open problems. Let us describe briefly some generalizations and open problems related with the results in this paper.

1. In this paper we have restricted ourselves to space dimensions $n \leq 3$. However, only the results for space dimension $n = 1$, which was treated separately, are based on arguments that cannot be generalized to higher dimensions. The proofs given for the cases $n = 2, 3$ do not depend on the dimension and can be easily generalized to any space dimension n, for it is sufficient to make an appropriate choice of the Sobolev spaces (3.3) where system (3.1) is well-posed.

2. As we mentioned in the introduction, when the control acts on an open nonempty subset of the domain Ω for all $0 \leq t \leq T$ the heat equation is null-controllable. The problem of null-controllability for a system with a singular control concentrated on the curve γ is completely open. It is well known that the null-controllability of system (3.1) is equivalent to the following observability inequality for the solutions of the adjoint system (4.1):

$$\int_\Omega |\varphi(x,0)|^2 dx \leq C \int_0^T \int_{\gamma(s)} |\varphi(x,t)|^2 d\sigma\, ds \qquad \forall \varphi \text{ solution of (4.1).} \tag{6.1}$$

This observability inequality is much stronger than the unique continuation property (4.2). The arguments and techniques developed in this paper do not allow us to obtain (6.1). In recent years, Carleman estimates have been used systematically as the most efficient tool to obtain observability inequalities (see [FI] and [FZ]), but they do not seem to be sufficient to obtain quantitative results as (6.1) when $\gamma(t)$ is of dimension $k \leq n - 1$.

3. The results of this paper (Theorem 3.3) show that increasing the time-oscillations of the control region improves the controllability properties of the heat equation. It would be interesting to further analyze this property in the context of the cost of controlling the system. In other words, the initial and final data u^0, u^1 being fixed, as well as $\alpha > 0$, it would be interesting to analyze the size of the minimal control allowing us to achieve (3.10) and its behavior as $\varepsilon \to 0$. This was done in [FZ]

in the case of controls acting on open subsets of Ω independent of the time. But the techniques in [FZ], based on the use of Carleman inequalities, do not seem to apply in the present situation.

4. The results in Theorem 3.3 are also valid when considering controls that simultaneously guarantee the approximate control condition (2.2) and the exact control of a finite-dimensional projection (see [Z2] for details). In this case, the proof is very similar. We only have to replace the term $\alpha \left\|\varphi^0\right\|_{H_0}$ by $\alpha \left\|(I-\pi_E)\varphi^0\right\|_{H_0}$ in both the functionals J_ε and J, given in (5.12) and (5.13), respectively. Here, I represents the identity and π_E the orthogonal projection of H_0 on a finite-dimensional subspace E. It is not difficult to see that the arguments in the proof of Theorem 3.3 can be easily adapted to this situation.

5. The main ingredients to prove the approximate controllability results of this paper, and in particular Theorem 3.2, are the Fourier decomposition of solutions and the time-analyticity of the underlying semigroup. Therefore, we can easily extend the results to more general equations where these two properties hold. This is the case, for instance, in equations of the form

$$\rho(x)u_t - \ \text{div}\ (A(x)\nabla u) = 0,$$

with Dirichlet, Neuman, or mixed boundary conditions.

The situation is more complex when considering time-dependent coefficients. Then, in general, it is not possible to write the solutions in Fourier series. However, there are some particular cases where a certain decomposition is still possible. For example, consider the heat equation

$$\rho(t)u_t - \Delta u = 0$$

with $0 < \rho_m \leq \rho(t) \leq \rho_M < \infty$ and some boundary conditions. With the change of variable

$$s(t) = \int_0^t \rho^{-1}(r)dr,$$

the equation is transformed into the constant coefficient heat equation

$$u_s - \Delta u = 0,$$

for which a Fourier decomposition is known. Coming back to the original variable t we obtain a decomposition of the solutions that allows us to adapt the results of this paper.

But the problem is completely open for general linear equations with coefficients depending on both space and time.

6. The techniques used in this paper cannot be adapted to the semilinear case. Indeed, the approximate controllability of the semilinear heat equation is usually derived from the approximate controllability of the linearized equation with a potential. As this potential depends on both the space and time variables, our techniques do not apply, as pointed out above.

7. The techniques used in this paper cannot be adapted to wave or plate equations since the time-analyticity of the solutions fails. For the wave equation there are some partial controllability results, for the 1-d case, when the control acts on a point that follows particular time-dependent trajectories (see [K2]).

REFERENCES

[B] M. Berggrem, *Optimal Control of Time Evolution Systems: Controllability Investigations and Numerical Algorithms*, Ph.D. thesis, Rice University, Houston, 1995.

[CH] T. Cazenave and A. Haraux, *An Introduction to SemiLinear Evolution Equations*, The Clarendon Press, Oxford University Press, New York, 1998.

[CoH] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 1, Wiley, New York, 1989.

[DM] G. Dal Maso, *An Introduction to* Γ*-convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser Boston, Boston, 1992.

[DN] P. Donato and A. Nabil, *Approximate controllability of linear parabolic equations in perforated domains*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 21–38.

[DF] H. Donnelly and C. Fefferman, *Nodal sets of eigenfunctions on Riemmanian manifolds*, Invent. Math., 93 (1988), pp. 161–183.

[FPZ] C. Fabre, J.-P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[FZ] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Advances Differential Equations, 5 (2000), pp. 465–514.

[FI] A. V. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Series 34, Research Institute of Mathematics, Global Analysis Research Center, Seoul National University, Seoul, Korea, 1996.

[H] M. Hervé, *Les fonctions analytiques*, Presses Universitaires de France, Paris, 1982.

[JL] D. Jerison and G. Lebeau, *Nodal sets of sums of eigenfunctions, in Harmonic Analysis and Partial Differential Equations* (Chicago, IL, 1996), Chicago Lectures in Math., University of Chicago Press, Chicago, 1999, pp. 223–239.

[K] A. Khapalov, *Controllability of the wave equation with a moving point control*, Appl. Math. Optim., 31 (1995), pp. 155–175.

[K2] A. Khapalov, *Mobile point controls versus locally distributed ones for the controllability of the semilinear parabolic equation*, SIAM J. Control Optim., 40 (2001), pp. 231–252.

[LR] G. Lebeau and L. Robiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.

[L] F. H. Lin, *Nodal sets of solutions of elliptic and parabolic equations*, Comm. Pure Appl. Math. 44 (1991), pp. 287–308.

[L1] J.-L. Lions, *Contrôlabilité exacte, stabilisation et perturbations de systèmes distribués*. Tomes 1 and 2, Rech. Math. Appl. 8 and 9, Masson, Paris, 1988.

[L2] J.-L. Lions, *Some Methods in the Mathematical Analysis of Systems and Their Control*, Gordon and Breach, New York, 1981.

[L3] J.-L. Lions, *Pointwise control for distributed systems*, in Control and Estimation in Distributed Parameter Systems, H.T. Banks, ed., Frontiers Appl. Math. 11, SIAM, Philadelphia, 1992, pp. 1–39.

[LM] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Vol. III, Grundlehren Math. Wiss. 183, Springer-Verlag, New York, Heidelberg, 1973.

[N] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[P] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1993.

[Pu] C. C. Pugh, *Real Mathematical Analysis*, Springer-Verlag, New York, 2002.

[Z1] E. Zuazua, *Approximate controllability for linear parabolic equations with rapidly oscillating coefficients*, Control Cybernet., 23 (1994), pp. 793–801.

[Z2] E. Zuazua, *Some problems and results on the controllability of partial differential equations*, in European Congress of Mathematics, Vol. II (Budapest, 1996), Progr. Math. 169, Birkhäuser, Basel, Switzerland, 1998, pp. 276–311.

[Z3] E. Zuazua, *Controllability of partial differential equations*, Discrete Contin. Dynam. Systems, 8 (2002), pp. 469–513.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1435–1466

SPECTRAL FACTORIZATION BY SYMMETRIC EXTRACTION FOR DISTRIBUTED PARAMETER SYSTEMS*

J. J. WINKIN†, F. M. CALLIER†, B. JACOB‡, AND J. R. PARTINGTON§

Abstract. The spectral factorization problem of a scalar coercive spectral density is considered in the framework of the Callier–Desoer algebra of distributed parameter system transfer functions. Criteria are given for the infinite product representation of a meromorphic coercive spectral density of finite order and for the convergence of infinite product representations of spectral factors, i.e., for the convergence of the symmetric extraction method for solving the spectral factorization problem of such spectral density. These convergence criteria are applied to the solution of the linear-quadratic optimal control problem by spectral factorization for a specific class of semigroup Hilbert state-space systems with a Riesz-spectral generator. The speed of convergence of the symmetric extraction method is also considered. As an example a damped vibrating string model is handled.

Key words. distributed parameter systems, spectral factorization, coercivity, meromorphic function, entire function, finite order, infinite product, symmetric extraction, convergence analysis

AMS subject classifications. Primary, 47A68, 47A70, 49R20; Secondary, 93B52, 93C05

DOI. 10.1137/S0363012902416456

1. Introduction.

The spectral factorization problem plays a central role in the framework of the fractional representation approach (which is also known in the literature as the "factorization approach") for feedback control system design; see, e.g., [8], [35]. In addition, spectral factorization constitutes an essential step in the solution of the linear-quadratic (LQ) optimal control problem for infinite-dimensional state-space systems; see, e.g., [9], [10], [18], [33], [36] and the references therein. The spectral factorization problem is also used as a main tool for solving linear operator inequalities (Lur'e equations); see, e.g., [16], [19]. As far as the LQ-optimal control problem is concerned, it is shown in [9] and [10] that the latter is solvable by spectral factorization for C_0-semigroup Hilbert state-space systems with bounded observation and control operators and with finite-dimensional output and input spaces. The philosophy developed in those papers has been extended, e.g., in [33] and [36] to C_0-semigroup Hilbert state-space systems with unbounded observation and control operators. In those references, the authors analyze spectral factorization problems of operator-valued Popov functions, giving an H^∞ spectral factor and showing notably the existence of solutions to related operator Riccati equations. However, they do not develop any method to perform the spectral factorization iteratively, which is done here similarly as on the heat diffusion model dealt with in [10].

Fundamental questions concerning the spectral factorization problem have been studied in the literature: in particular the existence and multiplicity of spectral factors and the continuity of the spectral factorization mapping have been analyzed; see, e.g., [12], [23], [22] and the references therein. As far as computational questions are

*Received by the editors October 22, 2002; accepted for publication (in revised form) April 7, 2004; published electronically January 27, 2005.
http://www.siam.org/journals/sicon/43-4/41645.html
†Department of Mathematics, University of Namur (FUNDP), Rempart de la Vierge 8, B-5000 Namur, Belgium (joseph.winkin@fundp.ac.be, frank.callier@fundp.ac.be).
‡Fachbereich Mathematik, Universität Dortmund, D-44221 Dortmund, Germany (birgit.jacob@math.uni-dortmund.de).
§School of Mathematics, University of Leeds, Leeds LS2 9JT, United Kingdom (J.R.Partington@leeds.ac.uk).

concerned, several methods have been developed and analyzed for the approximate computation of spectral factors or solutions of related operator Riccati equations for distributed parameter systems; see, e.g., the references cited in [9], [10], and [36]. The LQ-control based normalized coprime fraction spectral factorization problem considered in [10] was solved by an ad hoc iterative symmetric extraction method for a one-dimensional heat diffusion model, involving only elementary rational factors with real poles and zeros (first order case). The symmetric extraction method of spectral factorization was also studied in detail for multivariable finite-dimensional state-space systems in [4].

The major aim of this paper is to extend this method to a class of single-variable distributed parameter (i.e., in particular infinite-dimensional state-space) systems, including standard models like heat diffusion or wave propagation, and thereby involving the symmetric extraction of elementary rational factors with complex conjugate poles and zeros (second order case) as well. More specifically, this paper is devoted to the description and the convergence analysis of the symmetric extraction method for the spectral factorization of a scalar coercive spectral density, which is assumed to be a meromorphic function of finite order (see, e.g., [27]), in the framework of the Callier–Desoer algebra of distributed parameter system transfer functions (see, e.g., [5], [6], [11], [17]). Criteria for the infinite product representation of a meromorphic coercive spectral density of finite order and for the convergence of the symmetric extraction method of spectral factorization are developed, which extend some previous partial results (see [8], [10]). These criteria are mainly based on the knowledge of the comparative asymptotic behavior of the spectral density poles and zeros. Some comments concerning the speed of convergence of such a method are also given. Moreover, the symmetric extraction method is shown to work for the spectral factorization of a coprime fraction (coercive) spectral density. The analysis is performed in the framework of C_0-semigroup Hilbert state-space systems, whose infinitesimal generator is a Riesz-spectral operator, with eigenvalues satisfying some asymptotic conditions, and with transfer function in the Callier–Desoer algebra. The results are illustrated by an example, namely, the LQ-control based normalized coprime fraction spectral factorization problem for a vibrating string model. Some of these results were reported in [13], [14].

The methodology for the convergence analysis of the symmetric extraction method of spectral factorization which is used here is based on entire function theory; see, e.g., [37], [3], [26]. In particular, Hadamard's theorem on the infinite product representation of entire functions of finite order plays a central role here. Basically the analysis developed in this paper uses the material which is contained in [37] and which is paramount for the proof of a related important result, viz., Akhiezer's theorem, concerning the spectral factorization of entire functions of exponential type; see, e.g., [25, p. 567], [3, Theorem 7.5.1, p. 125].

Related results dealing with matrix-valued functions can also be found, e.g., in [29] and [30, Theorem 2.1]. These contributions do not deal with the symmetric case, whereas the present paper does. In addition, e.g., in [29], the starting point of the analysis is the function to be factorized (requiring a realization step in the analysis). Here the starting point of the analysis is basically a system transfer function, the main objective being to apply the methodology in a system theoretic framework.

The paper is organized as follows. Some preliminaries concerning the frequency domain framework, and properties of coercive spectral densities and invertible spectral factors are given in section 2. Fundamental results concerning the representation of a meromorphic coercive spectral density of finite order as an infinite product of

elementary rational spectral densities are developed in section 3. These results are used in section 4 in order to establish spectral criteria for the convergence of the symmetric extraction method for the computation of a spectral factor. Section 5 is devoted to the implementation of this method for solving the LQ-optimal control based spectral factorization problem for a vibrating string model. Finally, section 6 contains some concluding remarks and perspectives.

2. Preliminary concepts and results. The analysis of the symmetric extraction method of spectral factorization is performed in the framework of the Callier–Desoer transfer function algebra (see, e.g., [5], [6], [11], [17, section 7.1]). The latter is briefly described below.

Let $\sigma \leq 0$. An impulse response f is said to be in $\mathcal{A}(\sigma)$ if for $t < 0, f(t) = 0$, and for $t \geq 0, f(t) = f_a(t) + f_{sa}(t)$, where the regular functional part $f_a \in L^1_\sigma$, i.e., $\exp(-\sigma\cdot)\ f_a(\cdot)$ is in $L^1(0,\infty)$, and the singular atomic part $f_{sa} := \sum_{i=0}^{\infty}\ f_i\delta(.-t_i)$, where $t_0 = 0, t_i > 0$ for $i = 1, 2, \ldots$, and $f_i \in \mathbb{C}$ for $i = 0, 1, \ldots$ with $\sum_{i=0}^{\infty}\ |f_i| \exp(-\sigma t_i) < \infty$. The norm of a distribution f in $\mathcal{A}(\sigma)$ is defined by

$$||f||_{\mathcal{A}(\sigma)} := \int_0^\infty\ |f_a(t)|\ e^{-\sigma t}\ dt + \sum_{i=0}^{\infty}\ |f_i|\ e^{-\sigma t_i}\ .$$

The Laplace transform of a distribution f is denoted by $\hat{f}$, and the class of Laplace transforms of elements in $\mathcal{A}(\sigma)$ is denoted by $\hat{\mathcal{A}}(\sigma)$. The norm of $\hat{f}$ in $\hat{\mathcal{A}}(\sigma)$ is defined by

$$||\hat{f}||_{\hat{\mathcal{A}}(\sigma)} := ||f||_{\mathcal{A}(\sigma)}\ .$$

An impulse response f is said to be in $\mathcal{A}_-(\sigma)$ if $f \in \mathcal{A}(\sigma_1)$ for some $\sigma_1 < \sigma$. We write $\mathcal{A}_-$ for $\mathcal{A}_-(0)$. $\mathcal{A}(\sigma)$ and $\mathcal{A}_-$ are convolution algebras. By $\hat{\mathcal{A}}_-(\sigma)$ and $\hat{\mathcal{A}}_-$ we denote the classes of Laplace transforms of elements in $\mathcal{A}_-(\sigma)$ and $\mathcal{A}_-$, respectively. Then $\hat{\mathcal{A}}_-$ is our selected class of distributed proper-stable transfer functions. It contains the multiplicative subset $\hat{\mathcal{A}}_-^\infty$ of transfer functions that are bounded away from zero at infinity in $\mathbb{C}_+$, i.e., that are biproper-stable. The Callier–Desoer algebra $\hat{\mathcal{B}}$ of possibly unstable transfer functions consists of those $\hat{f}$ such that $\hat{f} = \hat{n}\hat{d}^{-1}$ with $\hat{n} \in \hat{\mathcal{A}}_-$ and $\hat{d} \in \hat{\mathcal{A}}_-^\infty$. A transfer function is in $\hat{\mathcal{B}}$ if and only if it is the sum of a completely unstable strictly proper rational function and a stable transfer function in $\hat{\mathcal{A}}_-$; hence $\hat{d}$ above can always be chosen biproper-stable *rational*; see [11], [17].

DEFINITION 2.1. *A complex–valued function f is said to be* (1) parahermitian *if* $f(s) \equiv f_*(s) := \overline{f(-\overline{s})}$, (2) real *if* $\overline{f(s)} \equiv f(\overline{s})$, *and* (3) real parahermitian *if* $\overline{f(s)} \equiv f(\overline{s})$ *and* $f(s) \equiv f_*(s)$.

A function $\hat{F}$ is said to be a (real) spectral density *if $\hat{F}$ is (real) parahermitian such that $\hat{F} = \hat{F}_* = \hat{G}_* + \hat{G}$, where $\hat{G}$ is in $\hat{\mathcal{A}}_-$, and $\hat{F}$ is nonnegative on the imaginary axis, i.e., $\hat{F}(j\omega) \geq 0$ for all $\omega \in \mathbb{R}$. A spectral density $\hat{F}$ is said to be* coercive *if there exists $\eta > 0$ such that $\hat{F}(j\omega) \geq \eta$ for all $\omega \in \mathbb{R}$. A transfer function $\hat{R}$ in $\hat{\mathcal{A}}_-$ is said to be a* spectral factor *of a spectral density $\hat{F}$ if $\hat{F}(j\omega) = \hat{R}_*(j\omega)\hat{R}(j\omega)$ for all $\omega \in \mathbb{R}$. A spectral factor $\hat{R}$ is said to be* invertible *if $\hat{R}^{-1}$ is in $\hat{\mathcal{A}}_-$.*

A spectral density is also called a *Popov function* in the literature; see, e.g., [36] and the references therein. It is known that a spectral density has an invertible spectral factor if and only if it is coercive, and that spectral factors are unique up to multiplication by a constant of modulus one (see, e.g., [8], [9], [12]); furthermore,

to coercive real spectral densities correspond real spectral factors unique up to the $\pm$ sign. Moreover, a coercive spectral density $\hat{F}$ such that $\hat{F}(\infty) = 1$, i.e., $\hat{F} = \hat{G}_* + \hat{G}$ with $\hat{G} = G_0 + \hat{G}_a \in \hat{\mathcal{A}}_-$ and $\operatorname{Re} G_0 = 2^{-1}$, has a unique invertible *standard* spectral factor $\hat{R} = 1 + \hat{R}_a \in \hat{\mathcal{A}}_-$, i.e., such that $\hat{R}(\infty) = 1$. The following properties will be needed for the analysis of the following sections, especially in the proof of Theorem 3.4. The proof of the following lemma can be found in [8].

LEMMA 2.2 (algebraic properties of coercive spectral densities).

(a) *If $\hat{F}$ is a coercive (real) spectral density, then so is its inverse $\hat{F}^{-1}$.*

(b) *If $\hat{F}$ and $\hat{G}$ are coercive (real) spectral densities, then so is their product $\hat{F}\cdot\hat{G}$.*

Remark 2.1. All impulse responses f considered below have no delayed impulses (delays); i.e., their singular atomic part is of the form $f_{sa} = f_0\delta(\cdot)$.

In this paper we are essentially interested in meromorphic spectral densities. Recall that a function f is said to be *meromorphic* (in $\mathbb{C}$) if there exists a countable set $S \subset \mathbb{C}$ such that S has no limit point, f is holomorphic in S^c, and f has a pole at each point of S; see, e.g., [31, p. 241]. In particular a meromorphic coercive real spectral density $\hat{F}$ with a meromorphic inverse has a countable set of zeros $\mathcal{Z} \subset \mathbb{C}$ and a countable set of poles $\mathcal{P} \subset \mathbb{C}$ such that $\mathcal{Z}$ and $\mathcal{P}$ have no limit points, there exists a vertical strip S_δ, $\delta > 0$, such that $\mathcal{Z} \cap S_\delta = \emptyset$ and $\mathcal{P} \cap S_\delta = \emptyset$, and for any $z \in \mathcal{Z}$ and any $p \in \mathcal{P}$, $-z$, $\overline{z}$, $-\overline{z}$ are in $\mathcal{Z}$ and $-p$, $\overline{p}$, $-\overline{p}$ are in $\mathcal{P}$.

In addition, the order of a meromorphic function is defined as follows (see, e.g., [27, Chapters VI and VIII]): For a meromorphic function f, with no poles at $s = 0$, define the counting function $N(r, f)$ and the proximity function $m(r, f)$, where $r \geq 0$, respectively, by

$$N(r, f) \;:=\; \int_0^r \frac{n(t, f)}{t}\, dt,$$

where $n(t, f)$ denotes the number of poles of f (counting their multiplicities) in the closed disk $\{s \in \mathbb{C} : |s| \leq t\}$, and

$$m(r, f) \;:=\; \frac{1}{2\pi}\int_0^{2\pi} \log^+ |f(re^{j\theta})|\, d\theta,$$

where $\log^+ x := \max\{0, \log x\}$ and f is assumed to have no poles on the circle $\{s \in \mathbb{C} : |s| = r\}$. The function T which is given by

$$T(r) := T(r, f) := m(r, f) + N(r, f)$$

is called the *characteristic function* of f. Observe that T is positive and monotonically increasing for $r > 0$. The *order* of f is defined to be the order of its characteristic function T, viz.,

$$\rho \;:=\; \limsup_{r\to\infty} \frac{\log\, T(r)}{\log\, r}.$$

In particular, for an entire function f, let $M(r)$ be its maximum modulus defined by

$$M(r) \;:=\; \max\{|f(s)| : |s| = r\};$$

then the functions T and $\log\, M$ are of the same order, viz.,

$$\rho \;=\; \limsup_{r\to\infty} \frac{\log\, \log\, M(r)}{\log\, r}$$

(see, e.g., [27, p. 216] and [37]). Thus an entire function f is of finite order if and only if there exists a constant $k > 0$ such that $\max\{|f(s)| : |s| = r\} \leq \exp(r^k)$ for r large (see, e.g., [37]). Moreover, the sum, the product, and the quotient of two meromorphic functions of finite order are meromorphic functions of finite order as well; see, e.g., [27, p. 216]. The following concept will be very useful for the analysis of meromorphic spectral densities of finite order.

DEFINITION 2.3. *A* paraconjugate symmetric (ps-)family $\mathbf{M}$ with countably many defining parameters $\mu_n \in \mathbf{M}$ *is a countable family of complex numbers* (ρ_l), *containing* (μ_n) *as a subfamily, such that* (a) $\mathbf{M}$ *is paraconjugate symmetric, i.e., the* ρ_l*'s are either real such that* $\rho_l = \mu_n$ *and* $-\mu_n \in \mathbb{R}$, *with* $\mu_n < 0$, *or complex nonreal such that* $\rho_l = \mu_n, \overline{\mu_n}, -\mu_n$ *and* $-\overline{\mu_n} \in \mathbb{C}$, *with* $\operatorname{Re}\mu_n < 0$, $\operatorname{Im}\mu_n > 0$, $n \in \mathbb{N}$, (b) *the defining parameters may be finitely repeated in* $\mathbf{M}$, *i.e.,* (μ_n) *(or, equivalently,* (ρ_l)*) does not contain any constant subfamily, and* (c) *all the points* ρ_l *of* $\mathbf{M}$ *are located outside a vertical strip containing the imaginary axis in its interior, i.e., there exists some* $\kappa > 0$ *such that, for all* n, $|\operatorname{Re}\mu_n| \geq \kappa$.

It turns out that there is a strong connection between meromorphic real spectral densities and real parahermitian entire functions; see Lemma 2.5 below. This result is based on the following additional lemma concerning the infinite product representation of real parahermitian entire functions, which also will be needed in the proof of Theorem 3.4.

LEMMA 2.4 (infinite product representation of real parahermitian entire functions).

(1) *Consider a ps-family* $\mathbf{M}$ *with defining parameters* μ_n. *Assume that*

$$\sum_{n=1}^{\infty} \frac{1}{|\mu_n|^2} < \infty. \tag{2.1}$$

Then there exists an entire function P *of finite order such that* P *has a zero at each point of* $\mathbf{M}$ *and no other zero in* $\mathbb{C}$, *and such that* (a) P *has a product factorization of the form*

$$P(s) = \prod_{n=1}^{\infty} P_n(s), \tag{2.2}$$

where

$$P_n(s) = 1 - \left(\frac{s}{\mu_n}\right)^2, \quad \text{with } \mu_n \in \mathbb{R} \text{ such that } \mu_n < 0, \tag{2.3}$$

and

$$P_n(s) = \left(1 - \left(\frac{s}{\mu_n}\right)^2\right)\left(1 - \left(\frac{s}{\overline{\mu_n}}\right)^2\right), \quad \text{with } \mu_n \in \mathbb{C}\backslash\mathbb{R} \text{ such that } \operatorname{Re}\mu_n < 0, \tag{2.4}$$

whence, for all $\omega \in \mathbb{R}$, $P(j\omega) \geq 0$. *Moreover, the convergence of the infinite product in* (2.2) *is uniform and absolute on closed discs* $D(r) := \{s \in \mathbb{C} : |s| \leq r\}$, *and* (b) P *is real parahermitian, whence for all* $s \in \mathbb{C}$, $P(s) = P(-s)$, *and there exists* $\delta > 0$ *such that, for all* $s \in S_\delta$, $P(s) \neq 0$.

(2) *Let f be a real parahermitian entire function having zeros in $\mathbb{C}$ as described in part* (1)*; then $f(s)$ has the product representation*

$$f(s) = e^{g(s)}\, P(s), \tag{2.5}$$

where $P(s)$ is as in part (1) *and $g(s)$ is a real parahermitian entire function. If in addition f is of finite order ρ, then $g(s)$ is a polynomial of degree $\delta[g] \le \rho$.*

Proof. See, e.g., [37, pp. 54–57] for more detail.

Assume that $\mathbf{M} =: (\rho_l)_{l=1}^{\infty}$. Let p be the least nonnegative integer such that $\sum_{l=1}^{\infty} \frac{1}{|\rho_l|^{p+1}} < \infty$. It follows from (2.1) that p is 0 or 1. Then

$$P(s) := \prod_{l=1}^{\infty} E\left(\frac{s}{\rho_l}, p\right), \tag{2.6}$$

where

$$E\left(\frac{s}{\rho_l}, p\right) = \left(1 - \frac{s}{\rho_l}\right) e^{\frac{ps}{\rho_l}} \quad \text{for } p = 0, 1$$

is the *canonical product of genus* p associated with the sequence $(\rho_l)_{l=1}^{\infty}$. By the reasoning of [37, pp. 55–56] it is an entire function, which has a zero at each point ρ_l and no other zeros in $\mathbb{C}$. Moreover, the convergence of its infinite product is uniform and absolute on closed discs $D(r)$, whence it can be reordered arbitrarily. Now as the factors $E(\frac{s}{\rho_l}, p)$ can be grouped to form factors of the form (2.3) or (2.4) that have real coefficients and are invariant when s is exchanged for $-s$, there holds that in either case (i.e., $p = 0$ or $p = 1$), $P(s)$ is real parahermitian and can be rewritten as

$$P(s) = \prod_{n=1}^{\infty} P_n(s) = P(-s), \tag{2.7}$$

where the P_n's are the polynomial functions given by (2.3)–(2.4). In addition, the entire function P is of *finite order* $\rho \le 2$ (i.e., $|P(s)| \le e^{|s|^{\rho+\epsilon}}$ for $|s|$ large [37, pp. 63–64], for any $\epsilon > 0$). Indeed by [37, pp. 65–66], *the exponent of convergence of the zeros* ρ_l of the canonical product in (2.2)–(2.4) is the greatest lower bound of the nonnegative numbers α such that $\sum_{l=1}^{\infty} |\rho_l|^{-\alpha} < \infty$. By (2.1) it is less than or equal to 2. Then, by [37, Theorem 6, p. 69], the entire function P is of order $\rho \le 2$. Finally, the function $P(s)$ satisfies all conclusions of part (1) of the lemma.

Part (2) follows by the reasoning around Weierstrass's factorization theorem [37, pp. 55–57] and the reasoning of the proof of part (1). The last statement concerning an entire function f of finite order follows by Hadamard's theorem; see, e.g., [37, Theorem 9, p. 74]. □

Remark 2.2. Let f be a real parahermitian entire function of finite order $\rho < 2$ with zeros as in part (1) of Lemma 2.4, whence for some $\delta > 0$, for all $s \in S_\delta$, $f(s) \neq 0$, and for all $\omega \in \mathbb{R}$, $f(j\omega) \ge 0$. Then f has a product factorization of the form

$$f(s) = kP(s), \tag{2.8}$$

where k is a positive constant and $P(s)$ is of the form (2.2)–(2.4).

Indeed, f has the product factorization (2.5), where the entire function g is a polynomial of degree $\delta[g] \le \rho < 2$. Since g is a real coefficient polynomial function of $(-s^2)$, it should be a constant c, such that (2.8) holds with $k = e^c$.

It follows from the lemma above that a coercive real spectral density that is a meromorphic function of finite order can be written as a ratio of two real parahermitian entire functions of finite order, provided that its poles tend to infinity sufficiently fast: see the following lemma.

LEMMA 2.5. *Consider a coercive real spectral density $\hat{F}$ given by $\hat{F} = \hat{F}_* = \hat{G}_* + \hat{G}$, where $\hat{G} \in \hat{\mathcal{A}}_-$ is such that $G_{sa} = G_0\delta(\cdot)$ for some $G_0 \in \mathbb{C}$, whence $\hat{F}$ is holomorphic in some open vertical strip containing the imaginary axis $S_\delta := \{s \in \mathbb{C} : \mathrm{Re}\ s \in (-\delta, \delta)\}$, where $\delta > 0$. Assume that $\hat{F}$ is a meromorphic function of finite order. Let the poles of $\hat{F}$ form a ps-family $\mathbf{P}$ with defining parameters p_n and let*

$$\sum_{n=1}^{\infty} \frac{1}{|p_n|^2} < \infty.$$

Then $\hat{F}$ can be written as a fraction

$$\hat{F}(s) = \frac{N(s)}{D(s)},$$

where the denominator D and numerator N are real parahermitian entire functions of finite order, such that $D(s) = D(-s)$, $N(s) = N(-s)$, and the zeros and poles of $\hat{F}$ are those of N and D, respectively.

Proof. Let the ps-family $\mathbf{P}$ be given by $\mathbf{P} =: (\pi_l)$. By Lemma 2.4 part (1), there exists a real parahermitian entire function D of finite order such that D has a zero at each pole π_l of $\hat{F}$ and no other zeros in $\mathbb{C}$. Now, as in the proof of [31, Theorem 15.12, p. 327], consider the function $N := \hat{F}\ D$, which is obviously real parahermitian and such that $\hat{F} = N\ D^{-1}$. Moreover, the singularities of N at the points π_l are removable, whence N can be extended such that it is holomorphic in $\mathbb{C}$, i.e., entire, and N and D have no common zeros in $\mathbb{C}$. Finally, $N = \hat{F}\ D$ is of finite order, since so are $\hat{F}$ and D. □

Remark 2.3. (α) The converse of Lemma 2.5 obviously holds: any coercive real spectral density whose poles satisfy the condition above and which can be written as a fraction as in the statement of Lemma 2.5 is a meromorphic function of finite order.

(β) The proof above contains the essential arguments of the proofs of [31, Theorems 15.11 and 15.12, pp. 326–327]). Lemma 2.5 will be used in the proof of Lemma 4.7.

(γ) Related results concerning the canonical representation of any meromorphic function of finite order are given in [27, pp. 218–221].

Later on, after having transformed the spectral density under study, we shall also need the following auxiliary technical result.

LEMMA 2.6 (entire coercive real spectral density of finite order without zeros). *Let $\hat{F}$ be a coercive real spectral density given by $\hat{F} = \hat{F}_* = \hat{G}_* + \hat{G}$, where $\hat{G} \in \hat{\mathcal{A}}_-$ is such that $G_{sa} = G_0\delta(\cdot)$ for some $G_0 \in \mathbb{C}$. Assume that*

(2.9) *$\hat{F}$ is an entire function of finite order without zeros.*

In addition, assume that the limit of $\hat{F}$ at infinity exists on the imaginary axis such that

$$\hat{F}(\pm j\infty) := \lim_{|\omega| \to \infty} \hat{F}(j\omega) = 1, \tag{2.10}$$

(or, equivalently, Re $G_0 = 2^{-1}$*). Then* $\hat{F}$ *is a constant function, i.e.,* $\hat{F}(s) = 1$*, for all* $s \in \mathbb{C}$.

Proof. By Hadamard's theorem [37, Theorem 9, p. 74], $\hat{F}$ has the form

$$\hat{F}(s) = e^{g(s)}, \tag{2.11}$$

where $g(s)$ is a polynomial, i.e., $g(s) = \sum_{k=0}^{n} g_k s^k$. One has $\hat{F}(0) = e^{g(0)} = e^{g_0}$ is real and positive. Hence $g_0 = \log(\hat{F}(0))$ is real. Moreover, as $\hat{F}$ is real, i.e., $\hat{F}(\overline{s}) = \overline{\hat{F}(s)}$, there holds by (2.11) and the continuity of $g(s)$ that there exists a unique integer l such that $g(\overline{s}) = \overline{g(s)} + jl2\pi$, which for $s = 0$ reads $\mathrm{Im}(g_0) = l\pi$, whence $l = 0$ as g_0 is real. Thus $g(\overline{s}) = \overline{g(s)}$, i.e., g is a real polynomial or, equivalently, g has real coefficients.

In addition there holds that $\hat{F}$ is real parahermitian, whence $\hat{F}(s) = \hat{F}(-s)$. A similar reasoning using (2.11) shows then that g is a real polynomial in s^2, whence it can be rewritten as a real polynomial h in $-s^2$, i.e.,

$$g(s) = \sum_{k=0}^{m} g_{2k} s^{2k} = \sum_{k=0}^{m} h_{2k}(-s^2)^k =: h(-s^2) \in \mathbb{R}[-s^2] \quad \text{with} \quad h_{2k} = (-1)^k g_{2k}.$$

Observe now that, by the structure of $\hat{F}$, its coercivity, and (2.10), $\hat{F}(j\omega)$ is a real positive uniformly continuous function on $\omega \in \overline{\mathbb{R}}$ (i.e., the extended real line) and bounded as well as bounded away from zero on $\overline{\mathbb{R}}$. Hence $\log(\hat{F}(j\omega))$ is a real uniformly continuous function on $\omega \in \overline{\mathbb{R}}$ and bounded above and below on $\overline{\mathbb{R}}$; moreover, by (2.10), $\lim_{|\omega|\to\infty} \log(\hat{F}(j\omega)) = 0$. Furthermore, there holds that $\log(\hat{F}(j\omega)) = h(\omega^2)$ for all $\omega \in \mathbb{R}$, whence $\lim_{|\omega|\to\infty} h(\omega^2) = 0$ with $h(\omega^2) \in \mathbb{R}[\omega^2]$. As a consequence, the polynomial $h(\omega^2)$ is identical to $\log(\hat{F}(j\omega))$ on $\omega \in \overline{\mathbb{R}}$, bounded there above and below, and zero at infinity. Hence $h(\omega^2)$ must reduce to a constant polynomial which is identically zero, i.e., $g(s) \equiv 0$. Thus $\hat{F}(s) \equiv 1$. ∎

Remark 2.4. (α) Often (2.11) reads $\hat{F}(s) = ke^{g(s)}$, where k is a positive constant. Then with $k = e^c$, where $c \in \mathbb{R}$, one gets $\hat{F}(s) = e^{g(s)+c}$, where $g(s)+c$ is a polynomial. Hence (2.11) holds without loss of generality.

(β) Assumption (2.9) is realized for a meromorphic coercive real spectral density of finite order for which, "after the removal of the poles and zeros," there remains an entire function of finite order without zeros, or, equivalently, "after the removal of the poles" there remains an entire function of finite order (see Lemmas 2.5 and 2.4 (part 2)).

(γ) In Lemma 2.6, the assumption that the order of the spectral density $\hat{F}$ (as an entire function) is finite cannot be omitted. This fact is illustrated by the following simple example. Consider the function $\hat{F}$ given by

$$\hat{F}(s) := \exp\left(2 \cdot \frac{\sinh\ s}{s}\right).$$

Observe that $\hat{F}$ is a coercive real spectral density of the form $\hat{F} = \hat{G}_* + \hat{G}$, where $\hat{G} = G_0 + \hat{G}_a \in \hat{\mathcal{A}}_-$ and Re $G_0 = 2^{-1}$. Indeed, let $\hat{R}$ be the function defined by

$$\hat{R}(s) := \exp\left(\hat{g}(s)\right), \quad \text{where} \quad \hat{g}(s) := \frac{1 - e^{-s}}{s}.$$

Then $\hat{g}$ belongs to $\hat{\mathcal{A}}_-$ as the Laplace transform of the function of finite support $g := \chi_{[0,1]}$; i.e., $g(t) = 1$ if $0 \le t \le 1$ and $g(t) = 0$ elsewhere. Moreover, $\hat{g}$ is strictly

proper, i.e., $\hat{g}(\infty) = 0$, where $\hat{g}(\infty)$ should be interpreted as the limit of $\hat{g}(s)$ as $|s| \to \infty$ in any right half-plane strictly containing the closed right half-plane. Hence $\hat{R}$ is in $\hat{\mathcal{A}}_-$ together with its inverse $\hat{R}^{-1} = \exp(-\hat{g}(s))$, since they are exponentials of elements of a Banach algebra, viz., $\hat{\mathcal{A}}(\sigma) \subset \hat{\mathcal{A}}_-$, $\sigma < 0$; moreover, $\hat{R}(\pm j\infty) = 1$. Since $\hat{F}(s) = \hat{R}(-s)\cdot\hat{R}(s)$, it follows that $\hat{F}$ is a coercive real spectral density with invertible standard real spectral factor $\hat{R}$. In addition, $\hat{g}$ is an entire function, whence $\hat{R}$ is an entire function without zeros, and so is $\hat{F}$. However, $\hat{F}$ is not a constant function. Observe that there is no contradiction with Lemma 2.6, since $\hat{F}$ is of infinite order. Indeed, the function $2\cdot\frac{\sinh s}{s}$ is an entire function which is not a polynomial. Hence, in view of Hadamard's theorem (see, e.g., [37, p. 74]), $\hat{F}$ cannot be of finite order.

3. Meromorphic spectral densities of finite order. The main objective of this section is to show that, under certain conditions, a coercive real spectral density that is a meromorphic function of finite order can be written as an infinite product of coercive real rational spectral densities. First it is shown that, under certain technical conditions, such an infinite product is necessarily a coercive real spectral density.

3.1. Product of rational spectral densities.

THEOREM 3.1 (infinite product of coercive rational spectral densities). *Consider a function $\hat{F}$ given, for all s in some vertical strip symmetric with respect to the imaginary axis, by an infinite product of pole-zero pairs of the form*

$$\hat{F}(s) = \prod_{n=1}^{\infty} \hat{F}_n(s), \tag{3.1}$$

where the elementary factors $\hat{F}_n$ are coercive real rational spectral densities, which are given either by

$$\hat{F}_n(s) = \frac{z_n^2 - s^2}{p_n^2 - s^2}, \tag{3.2}$$

where z_n and $p_n \in \mathbb{R}$, with z_n and $p_n < 0$, or by

$$\hat{F}_n(s) = \frac{(z_n^2 - s^2)(\overline{z}_n^2 - s^2)}{(p_n^2 - s^2)(\overline{p}_n^2 - s^2)}, \tag{3.3}$$

where z_n and $p_n \in \mathbb{C}\backslash\mathbb{R}$, with Re z_n *and* Re $p_n < 0$. *Consider the standard invertible (real) spectral factors $\hat{R}_n$ of the spectral densities $\hat{F}_n$, which are such that $\hat{R}_n(\infty) = 1$ and which are given by*

$$\hat{R}_n(s) = \frac{z_n - s}{p_n - s} \tag{3.4}$$

(first order factor) when $\hat{F}_n$ is defined by (3.2) *and by*

$$\hat{R}_n(s) = \frac{(z_n - s)(\overline{z}_n - s)}{(p_n - s)(\overline{p}_n - s)} \tag{3.5}$$

(second order factor) when $\hat{F}_n$ is defined by (3.3)*, respectively. Assume that there exists a constant $\sigma < 0$ such that $\hat{R}_n$ and $\hat{R}_n^{-1}$ are in $\hat{\mathcal{A}}(\sigma)$, for all n, with*

$$\sum_{n=1}^{\infty} ||(R_n)_a||_{\mathcal{A}(\sigma)} < \infty \tag{3.6}$$

and

$$\sum_{i=1}^{\infty} ||(R_n^{-1})_a||_{\mathcal{A}(\sigma)} < \infty. \tag{3.7}$$

Then the following assertions hold:

(a) *The infinite product in* (3.1) *converges to* $\hat{F}$ *in the Banach algebra*

$$\widehat{L\Delta}(\sigma) := \{\hat{f} = \hat{f}_- + \hat{f}_+ \ : \ (f_-)_* \text{ and } f_+ \ \in \mathcal{A}(\sigma)\}$$

equipped with the norm

$$||\hat{f}||_\sigma := ||f||_\sigma := ||(f_-)_*||_{\mathcal{A}(\sigma)} + ||f_+||_{\mathcal{A}(\sigma)};$$

(b) *The function* $\hat{F}$ *is a coercive real spectral density such that* $\hat{F} = \hat{F}_* = \hat{G}_* + \hat{G}$, *where* $\hat{G}$ *is in* $\hat{\mathcal{A}}(\sigma) \subset \hat{\mathcal{A}}_-$.

Proof. (a) The proof goes along the lines of [10, proof of Theorem 5].

(b) It follows from assertion (a) and from the fact that every elementary factor $\hat{F}_n$ is real parahermitian and positive semidefinite on the imaginary axis, that $\hat{F}$ is a real spectral density of the form $\hat{F} = \hat{G}_* + \hat{G}$ for some $\hat{G}$ in $\hat{\mathcal{A}}(\sigma)$. Finally observe that, by the fact that there exists a constant $\sigma < 0$ such that $\hat{R}_n$ and $\hat{R}_n^{-1}$ are in $\hat{\mathcal{A}}(\sigma)$ for all n, every elementary factor spectral density $\hat{F}_n$ is coercive such that, for some $\eta > 0$, $\prod_{n=1}^{N} \hat{F}_n(j\omega) \geq \eta$ for all $\omega \in \mathbb{R}$ and for all $N \geq 1$. Hence the spectral density $\hat{F}$ is coercive. ∎

Remark 3.1. Convergence in the $\widehat{L\Delta}(\sigma)$-norm implies convergence in the sup-norm on a vertical strip without singularities containing the $j\omega$-axis in its interior.

This result leads to a criterion for the convergence of an infinite product of coercive rational spectral densities, which is based on the knowledge of the spectrum, i.e., more precisely, on the knowledge of the comparative asymptotic behavior of the spectral density poles and zeros, p_n and z_n, as n tends to infinity; see Corollary 3.3 below. The proof of this spectral criterion is based on the following preliminary result.

LEMMA 3.2 (estimates of $||(R_n)_a||_{\mathcal{A}(\sigma)}$ and $||(R_n^{-1})_a||_{\mathcal{A}(\sigma)}$). *Consider the rational elementary factors* $\hat{R}_n$, $n = 1, 2, \ldots$, *given by* (3.4)–(3.5), *and assume that there exists a constant* $\sigma < 0$ *such that* $2 \cdot |\sigma| \leq \min(|\text{Re } p_n|, |\text{Re } z_n|)$ *for all* n. *Then* $\hat{R}_n$ *and* $\hat{R}_n^{-1}$ *are in* $\hat{\mathcal{A}}(\sigma)$ *for all* n *and the following inequalities hold for all* n: *when* $\hat{R}_n$ *is given by* (3.4), *then*

$$||(R_n)_a||_{\mathcal{A}(\sigma)} \leq 2 \, \frac{|z_n - p_n|}{|p_n|},$$

and

$$||(R_n^{-1})_a||_{\mathcal{A}(\sigma)} \leq 2 \, \frac{|z_n - p_n|}{|z_n|},$$

and when $\hat{R}_n$ *is given by* (3.5), *then*

$$||(R_n)_a||_{\mathcal{A}(\sigma)} \leq 4 \, \frac{|z_n - p_n|}{|\text{Re } p_n|} \left(1 + \frac{|z_n - p_n|}{|\text{Re } p_n|}\right),$$

and

$$||(R_n^{-1})_a||_{\mathcal{A}(\sigma)} \leq 4 \, \frac{|z_n - p_n|}{|\text{Re } z_n|} \left(1 + \frac{|z_n - p_n|}{|\text{Re } z_n|}\right).$$

Proof. For the case of a first order factor, see [10, Fact 1]. Concerning the case of a second order factor, we derive only the inequality for $(R_n)_a$. The other one can be proved similarly. Now, for all $t \geq 0$,

$$(R_n)_a(t) = (p_n - z_n) \cdot e^{p_n t} + (\overline{p_n - z_n}) \cdot e^{\overline{p_n} t} + \frac{|p_n - z_n|^2}{(\overline{p_n} - p_n)} \cdot (e^{\overline{p_n} t} - e^{p_n t}),$$

whence, upon noting that $|\overline{p_n} - p_n| = 2|\text{Im } p_n|$ and $|e^{\overline{p_n} t} - e^{p_n t}| = 2 \cdot e^{\text{Re } p_n t} \cdot |\sin(\text{Im } p_n t)|$, one obtains

$$|(R_n)_a(t)| \leq 2\ |z_n - p_n| \cdot e^{\text{Re } p_n t} + |z_n - p_n|^2 \cdot t \cdot e^{\text{Re } p_n t}.$$

By using the assumption that $|\text{Re } p_n| \geq 2 \cdot |\sigma|$, one gets

$$||e^{\text{Re } p_n t}||_{\mathcal{A}(\sigma)} = \int_0^\infty e^{(\text{Re } p_n - \sigma)t} dt = |\text{Re } p_n - \sigma|^{-1} \leq \frac{2}{|\text{Re } p_n|},$$

and

$$||t \cdot e^{\text{Re } p_n t}||_{\mathcal{A}(\sigma)} = \int_0^\infty t \cdot e^{(\text{Re } p_n - \sigma)t} dt = |\text{Re } p_n - \sigma|^{-2} \leq \frac{4}{|\text{Re } p_n|^2}.$$

As a result one gets

$$||(R_n)_a||_{\mathcal{A}(\sigma)} \leq 4\ \frac{|z_n - p_n|}{|\text{Re } p_n|} \left(1 + \frac{|z_n - p_n|}{|\text{Re } p_n|}\right). \qquad \square$$

COROLLARY 3.3 (spectral criterion for the convergence of an infinite product of coercive rational spectral densities). *Consider a function $\hat{F}$ given by* (3.1)–(3.3) *for all s in some vertical strip symmetric with respect to the imaginary axis. Let $\hat{R}_n$, $n = 1, 2, \ldots$, be the rational elementary factors defined by* (3.4)–(3.5), *with $2 \cdot |\sigma| \leq \min(|\text{Re } p_n|, |\text{Re } z_n|)$, for all n, for some $\sigma < 0$. Now assume that*

$$\sum_{n=1}^{\infty} \frac{|z_n - p_n|}{|\text{Re } p_n|} < \infty, \tag{3.8}$$

and

$$\sum_{n=1}^{\infty} \frac{|z_n - p_n|}{|\text{Re } z_n|} < \infty. \tag{3.9}$$

Then the conclusions of Theorem 3.1 *hold.*

Proof. Consider the series

$$\sum_{n=1}^{\infty} \frac{|z_n - p_n|}{|\text{Re } p_n|} \left(1 + \frac{|z_n - p_n|}{|\text{Re } p_n|}\right).$$

By (3.8) the sequence

$$\left(\frac{|z_n - p_n|}{|\text{Re } p_n|}\right)_{n=1}^{\infty}$$

is in l^1, hence also in l^∞, and thus also in l^2. Therefore the series above converges, and in view of Lemma 3.2, (3.6) holds. Moreover, by similar arguments (3.7) holds. Hence the conclusion follows by Theorem 3.1. □

Remark 3.2. The conclusions of Corollary 3.3 still hold if conditions (3.8) and (3.9) are replaced, respectively, by

$$\frac{|z_n - p_n|}{|\mathrm{Re}\ p_n|} = O\left(\frac{1}{n^\alpha}\right) \tag{3.10}$$

and

$$\frac{|z_n - p_n|}{|\mathrm{Re}\ z_n|} = O\left(\frac{1}{n^\alpha}\right) \tag{3.11}$$

for some exponent $\alpha > 1$.

3.2. Product representation of meromorphic spectral densities. The fact that a coercive real spectral density $\hat{F}$ has an infinite product representation of the form (3.1)–(3.3) is not automatically satisfied in applications. Typically one should check this by using the Weierstrass factorization theorem for entire functions and related results; see, e.g., [37, Chapter 2, part 1], [26, section 7.1, p. 343], and the references therein. This was done in [10, p. 765] for a heat diffusion example. In the following theorem, conditions on the poles and zeros of a meromorphic spectral density of finite order are derived, under which this methodology can be used.

THEOREM 3.4 (criterion for the infinite product representation of a meromorphic coercive spectral density of finite order). *Let $\hat{F}$ be a coercive real spectral density given by $\hat{F} = \hat{F}_* = \hat{G}_* + \hat{G}$, where $\hat{G} \in \hat{\mathcal{A}}_-$ is such that $G_{sa} = G_0\delta(\cdot)$ for some $G_0 \in \mathbb{C}$; whence $\hat{F}$ is holomorphic in some open vertical strip containing the imaginary axis, namely, S_δ, where $\delta > 0$. Assume that the limit of $\hat{F}$ at infinity exists in this vertical strip such that*

$$\hat{F}(\infty) := \lim_{|s|\to\infty;\ s\in S_\delta} \hat{F}(s) = \lim_{|\omega|\to\infty;\ -\delta<\sigma<\delta} \hat{F}(\sigma + j\omega) = 1, \tag{3.12}$$

(or, equivalently, $\mathrm{Re}\ G_0 = 2^{-1}$*). In addition, assume that $\hat{F}$ is a meromorphic function of finite order such that*

(1) *the poles of $\hat{F}$ form a ps-family* $\mathbf{P}$ *with defining parameters p_n,*
(2) *the zeros of $\hat{F}$ form a ps-family* $\mathbf{Z}$ *with defining parameters z_n, and*
(3)

$$\sum_{n=1}^{\infty} \frac{1}{|p_n|^2} < \infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \frac{1}{|z_n|^2} < \infty. \tag{3.13}$$

Assume that the set of zeros (poles, respectively) of $\hat{F}$ consists of countably many real zeros (poles, respectively) and countably many complex zeros (poles, respectively) such that its zeros and poles can be associated by a one-to-one relationship, leading to elementary factors of the form (3.17)–(3.18). *Finally, assume that conditions* (3.8)–(3.9) *hold.*

Then (a) *$\hat{F}$ can be written as a fraction*

$$\hat{F}(s) = \frac{N(s)}{D(s)}, \tag{3.14}$$

where the denominator D and numerator N are real parahermitian entire functions of finite order such that $D(s) = D(-s)$ and $N(s) = N(-s)$ and the zeros and poles of $\hat{F}$ are those of N and D, respectively;

(b) *the spectral density $\hat{F}$ admits an infinite product representation of pole-zero pairs that is of the form*

$$\hat{F}(s) = \prod_{n=1}^{\infty} \hat{F}_n(s), \tag{3.15}$$

and the inverse spectral density $\hat{F}^{-1}$ admits the infinite product representation

$$\hat{F}(s)^{-1} = \prod_{n=1}^{\infty} \hat{F}_n(s)^{-1}, \tag{3.16}$$

where the elementary factors $\hat{F}_n$ are coercive real rational spectral densities that are given either by

$$\hat{F}_n(s) = \frac{z_n^2 - s^2}{p_n^2 - s^2} \tag{3.17}$$

if z_n and $p_n \in \mathbb{R}$, with z_n and $p_n < 0$, or by

$$\hat{F}_n(s) = \frac{(z_n^2 - s^2)(\overline{z}_n^2 - s^2)}{(p_n^2 - s^2)(\overline{p}_n^2 - s^2)} \tag{3.18}$$

if z_n and $p_n \in \mathbb{C}\backslash\mathbb{R}$, with Re z_n *and* Re $p_n < 0$.

Remark 3.3. (α) Conditions (3.13) are not the most general ones encountered in the theory of infinite product representation of entire functions; see, e.g., [26, p. 358] and [37, Theorem 1, pp. 55–56]. However, according to [26], these conditions are applicable to many problems. In addition these assumptions together with the fact that a spectral density is parahermitian lead to a simpler structure for the corresponding infinite product elementary factors.

(β) Concerning the definitions of the elementary factors (3.17)–(3.18), it is natural and usual to take the upper half-plane zeros and poles in increasing order of real part. This is done in this way in the application dealt with in this paper; see section 5.

Proof of Theorem 3.4. (a) follows from Lemma 2.5.

(b) *Step* 1. D and N have the infinite product representations

$$D(s) = e^{G_D(s)} \cdot \prod_{n=1}^{\infty} D_n(s), \tag{3.19}$$

and

$$N(s) = e^{G_N(s)} \cdot \prod_{n=1}^{\infty} N_n(s), \tag{3.20}$$

respectively, where G_D and G_N are polynomial functions and where D_n and N_n are the polynomial functions given by

$$D_n(s) = \begin{cases} 1 - (\frac{s}{p_n})^2 & \text{if } p_n \in \mathbb{R} \text{ with } p_n < 0, \\ (1 - (\frac{s}{p_n})^2) \cdot (1 - (\frac{s}{\overline{p}_n})^2) & \text{if } p_n \in \mathbb{C} \text{ with Re } p_n < 0 \end{cases} \tag{3.21}$$

and by

$$N_n(s) = \begin{cases} 1 - (\frac{s}{z_n})^2 & \text{if } z_n \in \mathbb{R} \text{ with } z_n < 0, \\ (1 - (\frac{s}{z_n})^2) \cdot (1 - (\frac{s}{\overline{z_n}})^2) & \text{if } z_n \in \mathbb{C} \text{ with } \operatorname{Re} z_n < 0, \end{cases} \tag{3.22}$$

respectively. Moreover, the convergence of the infinite products is uniform and absolute on any closed disc $D(r)$.

Indeed this follows from the data concerning D and N, where in particular (3.13) holds, and from Lemma 2.4, part (2).

Step 2. On closed discs $D(r)$ where small neighborhoods of the poles are omitted, the spectral density $\hat{F}$ has the infinite product representations

$$\hat{F}(s) \;=\; e^{H(s)} \cdot \prod_{n=1}^{\infty} \frac{N_n(s)}{D_n(s)} = \; e^{H(s)} \cdot \prod_{n=1}^{\infty} \phi_n \cdot \prod_{n=1}^{\infty} \hat{F}_n(s), \tag{3.23}$$

where $H := G_N - G_D$ is a polynomial function, the $\hat{F}_n(s)$ are given by (3.17)–(3.18), and where the constants ϕ_n are given by

$$\phi_n = \begin{cases} \left|\frac{p_n}{z_n}\right|^2 & \text{if } p_n,\ z_n \in \mathbb{R} \text{ with } p_n \text{ and } z_n < 0, \\ \left|\frac{p_n}{z_n}\right|^4 & \text{if } p_n,\ z_n \in \mathbb{C}\backslash\mathbb{R} \text{ with } \operatorname{Re} p_n \text{ and } \operatorname{Re} z_n < 0. \end{cases} \tag{3.24}$$

Moreover, the convergence of the s-dependent products is uniform and absolute on the aforementioned punctured discs.

Indeed this follows from identity (3.14), from (3.19)–(3.22), and from assumptions (3.8)–(3.9). First it can be shown that for n sufficiently large, for the case that $p_n,\ z_n \in \mathbb{R}$ with p_n and $z_n < 0$,

$$\left|\hat{F}_n(s) - 1\right| = \; O\left\{\left|1 - \left(\frac{z_n}{p_n}\right)^2\right|\right\},$$

and for the case that $p_n,\ z_n \in \mathbb{C}$ with $\operatorname{Re} p_n$ and $\operatorname{Re} z_n < 0$,

$$\left|\hat{F}_n(s) - 1\right| = \; O\left\{\left|1 - \left|\frac{z_n}{p_n}\right|^4\right| + 2\left|1 - \left(\frac{z_n}{p_n}\right)^2\right|\right\}.$$

In addition, observe that $\hat{F}(\infty) = 2 \cdot \operatorname{Re} G_0$; whence, by assumption (3.12), $\operatorname{Re} G_0 = 2^{-1}$. Moreover, it follows from (3.13) that $|p_n| \to \infty$ and $|z_n| \to \infty$ as $n \to \infty$. Finally by assumptions (3.8)–(3.9) and the inequalities

$$\left|1 - \left(\frac{z_n}{p_n}\right)\right| \leq \frac{|z_n - p_n|}{|\operatorname{Re} p_n|}$$

and

$$\left|1 - \left(\frac{p_n}{z_n}\right)\right| \leq \frac{|z_n - p_n|}{|\operatorname{Re} z_n|},$$

there holds that the spectral density poles and zeros will be asymptotically close (as $n \to \infty$), i.e.,

$$\lim_{n\to\infty} \left(\frac{z_n}{p_n}\right) = 1 \quad \text{and} \quad \lim_{n\to\infty} \left(\frac{p_n}{z_n}\right) = 1.$$

It then follows easily that

$$\sum_{n=1}^{\infty} \left| 1 - \left(\frac{z_n}{p_n} \right)^2 \right| < \infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \left| 1 - \left(\frac{p_n}{z_n} \right)^2 \right| < \infty. \tag{3.25}$$

Hence by (3.25), the infinite product $\prod_{n=1}^{\infty} \hat{F}_n(s)$ converges uniformly and absolutely in any closed disk $D(r)$ where small neighborhoods of the poles are omitted. Next by (3.25) it can also be shown that the infinite product $\prod_{n=1}^{\infty} \phi_n$ is well defined. Now, on closed discs $D(r)$ where small neighborhoods of the poles are omitted, there holds for any $M \in \mathbb{N}$

$$\frac{\prod_{n=1}^{M} N_n}{\prod_{n=1}^{M} D_n} = \prod_{n=1}^{M} \phi_n \cdot \prod_{n=1}^{M} F_n.$$

Hence, as the limits exist for $M \to \infty$ uniformly and absolutely on such discs, Step 2 follows.

Step 3. There holds that

$$e^{H(s)} \cdot \prod_{n=1}^{\infty} \phi_n \equiv 1, \tag{3.26}$$

whence, in view of (3.23), (3.15) and (3.17)–(3.18) hold.

Indeed observe that the infinite product $\prod_{n=1}^{\infty} \hat{F}_n(s)$ converges uniformly and absolutely in any closed disk not containing any of its poles and that $\prod_{n=1}^{\infty} \hat{F}_n(\infty) = 1$, such that $\prod_{n=1}^{\infty} \hat{F}_n(s)$ exists at infinity and converges uniformly in the vertical strip S_δ by Corollary 3.3 and (3.8)–(3.9). Thus

$$\prod_{n=1}^{\infty} \hat{F}_n(s) \to 1 \quad \text{as} \quad |s| \to \infty \quad \text{in} \quad S_\delta. \tag{3.27}$$

Since, by assumption, $\hat{F}(\infty) = 1$, it follows from (3.23) and (3.27) that

$$e^{H(\infty)} \cdot \prod_{n=1}^{\infty} \phi_n := \lim_{|s| \to \infty;\ s \in S_\delta} e^{H(s)} \cdot \prod_{n=1}^{\infty} \phi_n = 1. \tag{3.28}$$

Observe that, by Corollary 3.3, the function $\prod_{n=1}^{\infty} \hat{F}_n$ is a coercive real spectral density; whence, in view of Lemma 2.2, so is the function

$$e^{H(s)} \cdot \prod_{n=1}^{\infty} \phi_n = \hat{F}(s) \cdot \left(\prod_{n=1}^{\infty} \hat{F}_n(s) \right)^{-1},$$

which in addition, as H is a polynomial, is an entire function of finite order without zeros. Hence, by Lemma 2.6 (see also Remark 2.4 (α)), it follows from (3.28) that Step 3 holds.

Step 4. Observe that $\hat{F}^{-1}$ is a meromorphic function of finite order and mutatis mutandis satisfies the same conditions as $\hat{F}$. Hence, by reasoning similar to that above, (3.16) and (3.17)–(3.18) hold. □

4. Spectral factorization by symmetric extraction.

4.1. Main results. By the proof of [10, Theorem 5], the following result holds.

THEOREM 4.1 (criterion for infinite product representation of invertible spectral factors). *Consider a coercive real spectral density $\hat{F}$ given by* (3.1)–(3.3) *for all s in some vertical strip symmetric with respect to the imaginary axis. Let $\hat{R}_n$, $n = 1, 2, \dots$, be the rational elementary factors defined by* (3.4)–(3.5). *Assume that there exists a constant $\sigma < 0$ such that $\hat{R}_n$ and $\hat{R}_n^{-1}$ are in $\hat{\mathcal{A}}(\sigma)$, for all n, with*

$$\sum_{n=1}^{\infty} ||(R_n)_a||_{\mathcal{A}(\sigma)} < \infty \tag{4.1}$$

and

$$\sum_{i=1}^{\infty} ||(R_n^{-1})_a||_{\mathcal{A}(\sigma)} < \infty. \tag{4.2}$$

Then the invertible standard spectral factor $\hat{R}$ in $\hat{\mathcal{A}}_-$ of $\hat{F}$ is given by the infinite product representation

$$\hat{R}(s) = \prod_{n=1}^{\infty} \hat{R}_n(s) = \lim_{N\to\infty} \prod_{n=1}^{N} \hat{R}_n(s), \tag{4.3}$$

where the limit is taken in the framework of the topology induced by the norm $||\cdot||_{\hat{\mathcal{A}}(\sigma)}$ on the Banach algebra $\hat{\mathcal{A}}(\sigma)$.

We are now ready to get a spectral criterion for the convergence of the symmetric extraction method of spectral factorization: it is based on the knowledge of the comparative asymptotic behavior of the spectral density poles and zeros.

THEOREM 4.2 (spectral criterion for the convergence of the symmetric extraction method). *Consider a coercive real spectral density $\hat{F}$ given by* (3.1)–(3.3) *for all s in some vertical strip symmetric with respect to the imaginary axis. Let $\hat{R}_n$, $n = 1, 2, \dots$, be the rational elementary factors defined by* (3.4)–(3.5), *with $2 \cdot |\sigma| \leq \min(|\mathrm{Re}\ p_n|, |\mathrm{Re}\ z_n|)$, for all n, for some $\sigma < 0$. Now assume that $\hat{R}_n$ is a first order factor given by* (3.4) *or a second order factor given by* (3.5) *satisfying* (3.8) *and* (3.9). *Then* (a) *the conclusions of Theorem* 4.1 *hold. In particular, the sequence*

$$\left(\prod_{n=1}^{N} \hat{R}_n\right)_{N\geq 1} \tag{4.4}$$

of invertible approximate (rational) spectral factors converges to the invertible standard spectral factor $\hat{R} \in \hat{\mathcal{A}}_-$ of $\hat{F}$ in the $\hat{\mathcal{A}}(\sigma)$ norm, and the sequence

$$\left(\prod_{n=1}^{N} \hat{R}_n^{-1}\right)_{N\geq 1} \tag{4.5}$$

converges to the corresponding inverse spectral factor $\hat{R}^{-1} \in \hat{\mathcal{A}}_-$;

(b) *with $\hat{W}_N \in \widehat{L\Delta}^+(\sigma) \subset \hat{\mathcal{A}}_-$ denoting the approximate spectral factor defined by*

$$W_N := \prod_{i=1}^{N} R_i,$$

the spectral factor, inverse spectral factor, and spectral factorization relative errors *can be estimated for all* $N = 1, 2, \ldots$, *respectively, by the following inequalities:*

$$\|(W_N - R) * R^{-1}\|_\sigma \;\le\; \exp S_{N+1} - 1 \;\le\; 2S_{N+1}, \tag{4.6}$$

where the last inequality holds if

$$S_{N+1} \;:=\; 4 \sum_{i=N+1}^{\infty} \frac{|z_n - p_n|}{|\mathrm{Re}\ z_n|} \left(1 + \frac{|z_n - p_n|}{|\mathrm{Re}\ z_n|}\right) \;\le\; 1;$$

$$\|(W_N^{-1} - R^{-1}) * R\|_\sigma \;\le\; \exp T_{N+1} - 1 \;\le\; 2T_{N+1}, \tag{4.7}$$

where the last inequality holds if

$$T_{N+1} \;:=\; 4 \sum_{i=N+1}^{\infty} \frac{|z_n - p_n|}{|\mathrm{Re}\ p_n|} \left(1 + \frac{|z_n - p_n|}{|\mathrm{Re}\ p_n|}\right) \;\le 1;$$

and finally

$$\|((W_N)_* \cdot W_N - F) * F^{-1}\|_\sigma \;\le\; \exp(2S_{N+1}) - 1 \;\le\; 4S_{N+1}, \tag{4.8}$$

where the last inequality holds if $S_{N+1} \;\le\; 2^{-1}$.

Proof. (a) The conclusion follows directly from Theorem 4.1, by the proof of Corollary 3.3.

(b) The relative error estimates (4.6)–(4.8) can be derived by following the lines of the proof of [10, Theorem 5, pp. 766–767] and by using Lemma 3.2. □

Remark 4.1. (α) The conclusions of Theorem 4.2 still hold if conditions (3.8) and (3.9) are replaced by conditions (3.10) and (3.11). Note that this remark is also applicable to any other result that holds here under these conditions, e.g., Theorem 3.4.

(β) Typically, in applications, e.g., LQ-optimal control or spectral factorization of a normalized coprime fraction spectral density (see, e.g., [9], [10]) for an infinite-dimensional (without loss of generality) stable system, the p_n's and $\overline{p}_n$'s are the poles of the open-loop transfer function, and the z_n's and $\overline{z}_n$'s are the poles of the closed-loop transfer function.

(γ) The symmetric extraction method works very well for the heat diffusion equation; see [10], [7]. Indeed, in that case, the spectral density zeros and poles are all real, and the corresponding relative spectral errors $|z_n - p_n| \cdot |p_n|^{-1}$ and $|z_n - p_n| \cdot |z_n|^{-1}$ tend to zero exponentially fast as n tends to infinity, whence (3.10) and (3.11) obviously hold for $\alpha = \infty$, i.e., for any $\alpha > 1$.

(δ) The speed of convergence of the sequences (4.4) and (4.5) towards an invertible spectral factor $\hat{R}$ and its inverse $\hat{R}^{-1}$, respectively, is dictated by the magnitude of the parameter α of conditions (3.10) and (3.11). The larger it is, the better is the speed of convergence of the symmetric extraction method. However, this speed of convergence might not be as good as in the heat equation example mentioned above; see Example 4.1.

(ϵ) It is possible to compute absolute and relative error estimates, in the $\hat{\mathcal{A}}(\sigma)$-norm, for the spectral factor as well as for its inverse, especially when $||(R_n)_a||_{\mathcal{A}(\sigma)}$ and $||(R_n^{-1})_a||_{\mathcal{A}(\sigma)}$ are of the order of the general term of a converging power series; see [10].

Example 4.1. Consider the following coercive spectral density $\hat{F}(s)$ given by (3.1) with a countable number of elementary factors of the form (3.3) with complex conjugate poles and zeros such that

(1) for some $\sigma < 0$

$$2 \cdot |\sigma| \leq \min(|\text{Re } p_n|, |\text{Re } z_n|), \quad n = 1, 2, \dots ,$$

(2) for some constants $a > 0$ and $b > 0$

$$p_n = -a + j \cdot b \cdot n, \quad n = 1, 2, \dots ,$$

(3) for n sufficiently large,

$$|z_n - p_n| = O(\frac{1}{n^2}).$$

Then by Theorem 4.2 any invertible spectral factor of the spectral density $\hat{F}$ can be approximated arbitrarily precisely in the $\hat{\mathcal{A}}(\sigma)$-norm, by an invertible approximate (rational) spectral factor of the form (4.4). Here convergence is achieved but at a much slower speed ($\alpha = 2$) than in the heat equation example mentioned above ($\alpha = \infty$). Note that, in section 5, a physical example (vibrating string) is chosen to illustrate the possibility of slow convergence.

COROLLARY 4.3. *Let $\hat{F}$ be a coercive real spectral density satisfying the assumptions of Theorem* 3.4. *Then $\hat{F}$ admits an infinite product representation of pole-zero pairs that is of the form* (3.1)–(3.3) *for all s in some vertical strip symmetric with respect to the imaginary axis. Let $\hat{R}_n$, $n = 1, 2, \dots$, be the rational elementary factors defined by* (3.4)–(3.5), *with $2 \cdot |\sigma| \leq \min(|\text{Re } p_n|, |\text{Re } z_n|)$, for all n, for some $\sigma < 0$.*

Then the sequence

$$\left(\prod_{n=1}^{N} \hat{R}_n \right)_{N \geq 1} \tag{4.9}$$

of invertible approximate (rational) spectral factors converges to the invertible standard spectral factor $\hat{R} \in \hat{\mathcal{A}}_-$ of $\hat{F}$ in the $\hat{\mathcal{A}}(\sigma)$-norm, and the sequence

$$\left(\prod_{n=1}^{N} \hat{R}_n^{-1} \right)_{N \geq 1} \tag{4.10}$$

converges to the corresponding inverse spectral factor $\hat{R}^{-1} \in \hat{\mathcal{A}}_-$.

Proof. The conclusion follows from Theorems 3.4 and 4.2. □

We conclude this subsection by yet another spectral criterion for the convergence of the symmetric extraction spectral factorization method. Its sufficient conditions are stronger than those established in the previous results. However, they fit specific classes of systems quite well, as shown in the next subsection.

THEOREM 4.4 (spectral criterion for the convergence of the symmetric extraction method). *Let $\hat{F}$ be a coercive real spectral density given by $\hat{F} = \hat{F}_* = \hat{G}_* + \hat{G}$, where $\hat{G} \in \hat{\mathcal{A}}_-$ is such that $G_{sa} = G_0 \delta(\cdot)$ for some $G_0 \in \mathbb{C}$; whence $\hat{F}$ is holomorphic in some open vertical strip containing the imaginary axis, namely, $S_\delta := \{s \in \mathbb{C} : \text{Re } s \in (-\delta, \delta)\}$, where $\delta > 0$. Assume that the limit of $\hat{F}$ at infinity exists in this vertical strip such that*

$$\hat{F}(\infty) := \lim_{|s| \to \infty;\ s \in S_\delta} \hat{F}(s) = \lim_{|\omega| \to \infty;\ -\delta < \sigma < \delta} \hat{F}(\sigma + j\omega) = 1, \tag{4.11}$$

(or, equivalently, $\operatorname{Re} G_0 = 2^{-1}$*). In addition, assume that* $\hat{F}$ *is a meromorphic function of finite order and is given by the fraction*

$$\hat{F}(s) = \frac{N(s)}{D(s)}, \tag{4.12}$$

where the denominator D *and numerator* N *are real parahermitian entire functions of finite order such that* $D(s) = D(-s)$ *and* $N(s) = N(-s)$.

Moreover, (1) *the zeros of* D *form a ps-family* $\mathbf{P}$ *with defining parameters* p_n,

(2) *the zeros of* N *form a ps-family* $\mathbf{Z}$ *with defining parameters* z_n,

(3) *one assumes that*

$$\sum_{n=1}^{\infty} \frac{1}{|p_n|^2} < \infty, \tag{4.13}$$

and

(4) *the pole-zero absolute error sequence is absolutely summable; i.e.,*

$$\sum_{n=1}^{\infty} |z_n - p_n| < \infty. \tag{4.14}$$

Then (a) $\hat{F}$ *admits the infinite product representation of pole-zero pairs* (3.1)–(3.3).

(b) *In addition, let* $\hat{R}_n$, $n = 1, 2, \ldots$, *be the rational elementary factors defined by* (3.4)–(3.5), *with* $2 \cdot |\sigma| \le \min(|\operatorname{Re} p_n|, |\operatorname{Re} z_n|)$, *for all* n, *for some* $\sigma < 0$. *Then the sequence* $(\prod_{n=1}^{N} \hat{R}_n)_{N \ge 1}$ *of invertible approximate (rational) spectral factors converges to the exact invertible standard spectral factor* $\hat{R} \in \hat{\mathcal{A}}_-$ *of* $\hat{F}$ *in the* $\hat{\mathcal{A}}(\sigma)$*-norm, and the sequence* $(\prod_{n=1}^{N} \hat{R}_n^{-1})_{N \ge 1}$ *converges to the corresponding inverse spectral factor* $\hat{R}^{-1} \in \hat{\mathcal{A}}_-$.

Proof. In view of Corollary 4.3, it suffices to check that the assumptions of Theorem 3.4 hold. Since the spectral density $\hat{F}$ is coercive, there exists some $\gamma > 0$ such that, for all $n \ge 1$,

$$|\operatorname{Re} z_n| \ge \gamma.$$

Therefore by condition (4.14) there holds

$$\sum_{n=1}^{\infty} \frac{|z_n - p_n|}{|\operatorname{Re} z_n|} < \infty;$$

i.e., assumption (3.9) holds.

Now observe that

$$\sum_{n=1}^{\infty} \left| \frac{1}{z_n} - \frac{1}{p_n} \right| < \infty. \tag{4.15}$$

Indeed, for all $n \ge 1$,

$$\left| \frac{1}{z_n} - \frac{1}{p_n} \right| \le \gamma^{-1} \cdot |z_n - p_n| \cdot |p_n|^{-1},$$

where, by (4.13) and (4.14), the sequences $(|z_n - p_n|)$ and $(|p_n|^{-1})$ are, respectively, in l^1 and in l^∞. Whence (4.15) holds. It follows by (4.15) and (4.13) that (3.13)

holds. Finally observe that, by the holomorphicity of F in the strip S_δ, one has for all $n \geq 1$

$$|\mathrm{Re}\ p_n| \geq \delta.$$

This together with (4.14) implies

$$\sum_{n=1}^{\infty} \frac{|z_n - p_n|}{|\mathrm{Re}\ p_n|} < \infty,$$

i.e., assumption (3.8) holds. Hence all assumptions of Theorem 3.4 are valid and we are done. □

Remark 4.2. It can be shown that Theorem 4.4 can be applied to the heat diffusion model studied in [10], [7] (see Remark 4.1 (γ)). Actually this result is applicable to an important class of semigroup state-space systems in the framework of the LQ-optimal control problem for such systems. This question is addressed in the following subsection.

4.2. Semigroup state-space systems. Consider a single-input C_0-semigroup state-space system with bounded control and observation operators (see, e.g., [17], [28]), viz.,

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad y(t) = Cx(t), \quad t \geq 0, \tag{4.16}$$

where $x(t) \in \mathrm{H}$, with H a separable Hilbert state-space with inner product $\langle \cdot\ ,\ \cdot \rangle$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}^p$, and

(1) $A : D(A) \subset \mathrm{H} \to \mathrm{H}$ is the generator of a C_0-semigroup $(e^{At})_{t \geq 0} \subset \mathbf{L}(\mathrm{H})$,

(2) $B \in \mathbf{L}(\mathbb{R}, \mathrm{H})$ is a bounded linear control operator given by

$$Bu = bu \quad \text{for all } u \in \mathbb{R}\ , \quad b \in \mathrm{H},$$

(3) $C \in \mathbf{L}(\mathrm{H}, \mathbb{R}^p)$ is a bounded linear observation operator.

Furthermore assume that A is a Riesz-spectral operator [17] with discrete spectrum

$$\sigma(A) = \sigma_p(A) = \{\lambda_n\ :\ n \in \mathbb{N}\} \subset \mathbb{C} \tag{4.17}$$

consisting of simple eigenvalues such that

$$\delta := \inf\ \{\ |\lambda_n - \lambda_m|\ :\ n, m \in \mathbb{N}\ ,\ n \neq m\} > 0 \tag{4.18}$$

and

$$\mu := \sup \left\{ \sum_{\substack{l=1 \\ l \neq n}}^{\infty} \frac{1}{|\lambda_l - \lambda_n|^2}\ :\ n \in \mathbb{N} \right\} < \infty. \tag{4.19}$$

Remark 4.3. Since the operator A is the (infinitesimal) generator of a C_0-semigroup of bounded linear operators $(e^{At})_{t \geq 0}$ on H, it holds (see [17]) that $\sup\ \{\mathrm{Re}\ \lambda_n\ :\ n \in \mathbb{N}\} < \infty$.

Finally assume that

(4.20) (A, B) is exponentially stabilizable and (C, A) is exponentially detectable.

Observe that, by (4.20), there exists some $\sigma < 0$ such that the spectrum of A can be decomposed according to

$$\sigma(A) \;=\; (\sigma(A) \cap \overset{o}{\mathbb{C}}_{\sigma-}) \;\dot\cup\; (\sigma(A) \;\cap\; \mathbb{C}_+), \tag{4.21}$$

where $\overset{o}{\mathbb{C}}_{\sigma-}$ denotes the open left half-plane $\{s \in \mathbb{C} \;:\; \mathrm{Re}\; s < \sigma\}$ and $\mathbb{C}_+$ denotes the closed right half-plane $\{s \in \mathbb{C} \;:\; \mathrm{Re}\; s \geq 0\}$. The two sets on the right-hand side of identity (4.21) are disjoint and the unstable spectrum $\sigma(A) \;\cap\; \mathbb{C}_+$ is a finite point set. Moreover, the following holds.

LEMMA 4.5. *The* nonzero *eigenvalues* $\lambda_n, n \in \mathbb{N}$, *of* A *satisfy*

$$\sum_{n=1,\lambda_n \neq 0}^{\infty} \;\frac{1}{|\lambda_n|^2} \;<\; \infty. \tag{4.22}$$

Proof. This follows immediately from (4.19). □

Now for system (4.16)–(4.20) consider the LQ-optimal control problem: for any initial state $x_0 \in \mathrm{H}$, find a square-integrable control $u_0 \in L^2(0,\infty;\mathbb{R})$ which minimizes the cost functional

$$J(x_0,u) := \int_0^\infty \;(\|Cx(t)\|^2 \;+\; \|u(t)\|^2)\;\, dt.$$

It is well known (see, e.g., [17] and the references cited therein) that the optimal control $u_0(t)$ is given by

$$u_0(t) = K_0\; x_0(t), \qquad x_0(t) = e^{(A+BK_0)t}x_0,$$

where the optimal feedback operator $K_0 \in \mathcal{L}(\mathrm{H},\mathbb{R})$ is given by

$$K_0 = -B^*\; Q_0,$$

where the operator $Q_0 \in \mathbf{L}(H)$ is the unique nonnegative self-adjoint solution of the operator Riccati equation on the domain of the operator A:

$$A^*Q_0 + Q_0A + C^*C - Q_0BB^*Q_0 = 0 \quad \text{on}\;\; D(A),$$

where $Q_0(D(A)) \subset D(A^*)$. Moreover, the optimal feedback K_0 is stabilizing; i.e., the feedback semigroup $(e^{(A+BK_0)t})_{t\geq 0}$ is exponentially stable, and $K_0 \in \mathbf{L}(\mathrm{H},\mathbb{R})$ is also given by

$$K_0\; x \;=\; \langle k_0\;,\; x\rangle \quad \text{for all } x \in \mathrm{H}, \quad k_0 \in \mathrm{H}. \tag{4.23}$$

We have then the following.

LEMMA 4.6. *Consider the* C_0*-semigroup state-space system given by* (4.16)–(4.20). *Let* $(\phi_n)_{n\in\mathbb{N}}$ *be a Riesz basis of eigenvectors of* A *(corresponding to the eigenvalues* λ_n*) and let* $(\psi_n)_{n\in\mathbb{N}}$ *be the corresponding biorthogonal dual Riesz basis of eigenvectors of the adjoint operator* A^*. *Consider the LQ-optimal feedback operator* $K_0 \in \mathcal{L}(H,\mathbb{R})$ *given by* (4.23). *Then the feedback semigroup generator*

$$A_c := A + BK_0 = A + b\;\langle k_0\;,\cdot\rangle \tag{4.24}$$

has a discrete spectrum of eigenvalues $\lambda_{cn}, n \in \mathbb{N}$, *with*

$$\sigma(A_c) = \sigma_p(A_c) = \{\lambda_{cn} \ : \ n \in \mathbb{N}\}, \tag{4.25}$$

and corresponding eigenvectors forming a Riesz basis of H. Moreover,

$$|\lambda_{cn} - \lambda_n| \ = \ O\left(\ |\langle k_0, \phi_n\rangle \cdot \langle b, \psi_n\rangle| \ \right) \quad \textit{for} \ \ n \ \textit{sufficiently large}, \tag{4.26}$$

whence

$$\sum_{n=1}^{\infty} |\lambda_{cn} - \lambda_n| < \infty. \tag{4.27}$$

Proof. Results (4.25)–(4.26), where $\{\lambda_{cn} : n \in \mathbb{N}\}$ is replaced by its closure, follow by (4.17)–(4.19), [34, Theorem 2.1], and [15, Appendix B, p. 66]. Now in (4.26) the sequences $(\langle k_0, \phi_n\rangle)$ and $(\langle b, \psi_n\rangle)$ are in l^2, whence the product sequence $(\langle k_0, \phi_n\rangle \cdot \langle b, \psi_n\rangle)$ is in l^1. Hence (4.27) holds. Furthermore, by Lemma 4.5, (4.22) holds.

Now as $(e^{(A+BK_0)t})_{t\geq 0}$ is exponentially stable and $\sigma(A)$ is as in (4.21), there exists some $\eta > 0$ and $M \in \mathbb{N}$ such that for all $n \geq M$, $\min(|\mathrm{Re}\lambda_n|, |\mathrm{Re}\lambda_{cn}|) \geq \eta$. This together with (4.27) and (4.22) and arguments similar to those in the proof of Theorem 4.4, gives

$$\sum_{n=M}^{\infty} \ \left| \frac{1}{\lambda_n} - \frac{1}{\lambda_{cn}} \right| < \infty.$$

Consequently

$$\sum_{n=1}^{\infty} \ \frac{1}{|\lambda_{cn}|^2} \ < \ \infty.$$

Hence $\sigma_p(A_c)$ cannot have limit points in $\mathbb{C}$, and the last equality of (4.25) holds. □

Remark 4.4. In view of [24, Corollary 4.6], the feedback semigroup generator A_c given by (4.24) is a *Riesz-spectral operator* [17, p. 41], whenever its eigenvalues $\lambda_{cn}, n \in \mathbb{N}$ are simple. *This will be tacitly assumed in what follows.*

It is also known (see [9], [10]) that the LQ-optimal control problem can be solved by the spectral factorization of a specific spectral density, whose spectral factor gives the state-feedback operator K_o via a Diophantine equation. In the sequel we shall concentrate upon that spectral factorization problem. More precisely, assuming that the operator pair (A, B) is exponentially stabilizable, there exists a stabilizing feedback operator $K \in \mathcal{L}(\mathrm{H}, \mathbb{R})$, i.e., such that the C_0-semigroup $(e^{(A+BK)t})_{t\geq 0}$ is exponentially stable. Moreover, with the spectrum of A as in (4.21), the feedback $K = [0 \ K_2]$ (K_2 = vector) can be chosen such that

$$\sigma(A + BK) \ = \ (\sigma(A) \cap \ \overset{o}{\mathbb{C}}_{\sigma-}) \ \dot{\cup} \ \Sigma, \tag{4.28}$$

where $\Sigma \subset \overset{o}{\mathbb{C}}_{\sigma-}$ is a finite set having the same number of elements as $\sigma(A) \ \cap \ \mathbb{C}_+$. Under these conditions the pair $(\hat{\mathcal{N}}, \hat{\mathcal{D}}) \in \hat{\mathcal{A}}_-^{p\times 1} \times \hat{\mathcal{A}}_-$ defined by

$$(\hat{\mathcal{N}}(s), \hat{\mathcal{D}}(s)) := (C(sI - A - BK)^{-1}B, \ 1 + K(sI - A - BK)^{-1}B) \tag{4.29}$$

generates a right fraction of the semigroup state-space system transfer function $\hat{G}(s) = C(sI - A)^{-1}B \in \hat{\mathcal{B}}^{p\times 1}$ with no common zeros in $\mathbb{C}_+$, where $\hat{\mathcal{D}}$ is a biproper stable rational function whose zeros are in $\sigma(A) \cap \mathbb{C}_+$ and whose poles are in Σ. Moreover, $\hat{\mathcal{D}}$ equals 1 at infinity.

Now consider the function $\hat{F}(s)$ defined by

$$\hat{F} := \hat{\mathcal{N}}_* \, \hat{\mathcal{N}} + \hat{\mathcal{D}}_* \, \hat{\mathcal{D}} = \hat{\mathcal{D}}_* \, (1 + \hat{G}_* \, \hat{G}) \, \hat{\mathcal{D}}. \tag{4.30}$$

It is shown in [9, Theorem 3, pp. 70–71] and [10, Theorem 3, pp. 761–762] that $\hat{F}$ is a coercive spectral density whose spectral factorization is the main step towards the solution of the LQ-optimal control problem, i.e., for the computation of the feedback operator K_0. Moreover for the specific case at hand one has the following.

LEMMA 4.7. *Under the assumptions of Lemma* 4.6, *let* $\hat{G}(s) = C(sI - A)^{-1}B$ *be the transfer function of the* C_0*-semigroup state-space system* (4.16)–(4.20). *Consider the real function* $\hat{F}$ *given by* (4.30), *where* $(\hat{\mathcal{N}}, \hat{\mathcal{D}})$ *is the right fraction* (4.29) *with no common zeros in* $\mathbb{C}_+$ *of* $\hat{G}(s)$, *where* $K \in \mathbf{L}(H, \mathbb{R})$ *is a stabilizing feedback operator such that* (4.28) *holds.*

Then $\hat{F}$ *is a coercive real spectral density such that* $\hat{F}$ *is holomorphic in a vertical strip* S_δ *for some* $\delta > 0$ *and such that* $\hat{F}(\infty) = 1$, *i.e.,* (4.11) *holds. Moreover,* $\hat{F}$ *is a meromorphic function of finite order* $\rho \leq 2$ *and can be described as a fraction of real parahermitian entire functions, i.e.,*

$$\hat{F}(s) = \frac{N(s)}{D(s)}, \tag{4.31}$$

where the functions $D = D_*$ *and* $N = N_*$ *are entire functions with countable zero sets* $\mathcal{Z}[D]$ *and* $\mathcal{Z}[N]$, *respectively, such that, with* $\mathcal{P}[\hat{F}]$ *denoting the set of poles of* $\hat{F}$ *and* $\sigma < 0$ *chosen such that* (4.21) *holds,*

$$\mathcal{Z}[D] = \mathcal{P}[\hat{F}] \subset \{p, -\overline{p} \, : \, p \in (\sigma(A) \cap \overset{o}{\mathbb{C}}_{\sigma-}) \, \dot{\cup} \, \Sigma\}, \tag{4.32}$$

and

$$\mathcal{Z}[N] = \mathcal{Z}[\hat{F}] \subset \{z, -\overline{z} \, : \, z \in \sigma(A_c) = \sigma(A + BK_0)\}. \tag{4.33}$$

Remark 4.5. (α) When the C_0-semigroup $(e^{At})_{t\geq 0}$ is exponentially stable, one can choose the feedback K to be zero. In this case, the right fraction $(\hat{\mathcal{N}}, \hat{\mathcal{D}})$ defined by (4.29) is given by $(\hat{\mathcal{N}}, \hat{\mathcal{D}}) = (\hat{G} \, , \, 1)$, and the spectral density reads

$$\hat{F} = 1 + \hat{G}_* \, \hat{G}. \tag{4.34}$$

Furthermore, the denominator entire function D in (4.31) is such that

$$\mathcal{Z}[D] = \mathcal{P}[\hat{F}] \subset \{p, -\overline{p} \, : \, p \in \sigma(A)\}. \tag{4.35}$$

(β) In general the inclusions in (4.32) and (4.33) are not equalities. This is due to the fact that the system is not necessarily approximately controllable and/or observable, whence the numerator and denominator can be simplified by common zero cancellations.

In standard examples, like the heat diffusion (see [10]) and the vibrating string (see below), the spectral density above is obtained (by applying the Laplace transform

to the PDE describing the system) as a fraction of entire functions where an infinite number of common zero cancellations may occur.

Proof of Lemma 4.7. Property (4.11) follows directly from (4.30) and (4.29), which ensure that $\hat{G}(s)$ is zero at infinity in $\mathbb{C}_{\sigma+}$ and that $\hat{\mathcal{D}}$ equals 1 at infinity. By [17, Lemma 4.3.10, p. 183], the transfer function $\hat{G}(s)$ is given by

$$\hat{G}(s) \;=\; \sum_{n=1}^{\infty} \; (C\phi_n) \;\; \overline{\langle b, \psi_n \rangle} \;\; (s-\lambda_n)^{-1}, \tag{4.36}$$

which is holomorphic in the resolvent set $\rho(A)$, whose complement $\sigma(A)$ is a pure point spectrum of isolated points by (4.18). It follows, using [31, Definition 10.41, p. 241]), that $\hat{G}(s)$ is meromorphic in $\mathbb{C}$ with poles contained in $\sigma(A)$. Thus, upon noting that in (4.30) $\hat{\mathcal{D}}$ is a biproper rational function, there holds that the spectral density $\hat{F}$ given by (4.30) is real parahermitian meromorphic in $\mathbb{C}$, with poles given by the inclusion in (4.32). Consequently, by Lemma 2.5, $\hat{F} = N/D$, where N and D are real parahermitian entire functions with no common zeros in $\mathbb{C}$; i.e., (4.31) holds with $\mathcal{Z}[D] = \mathcal{P}[\hat{F}]$ and $\mathcal{Z}[N] = \mathcal{Z}[\hat{F}]$. Hence (4.32) follows and we have to show that the inclusion of (4.33) holds. Now, by [9, Theorem 2, p. 67], the inverse spectral density $\hat{F}^{-1}$ can be written as

$$\hat{F}^{-1} \;=\; \hat{W}\, \hat{W}_*,$$

where

$$\hat{W}(s) \;=\; \hat{\mathcal{D}}(s)^{-1}\, [1 + K_0(sI - A - BK_0)^{-1}B] \;\in\; \hat{\mathcal{A}}_-.$$

Therefore the zero set of $\hat{F}$, i.e., the pole set of $\hat{F}^{-1}$, satisfies

$$\mathcal{Z}[\hat{F}] = \mathcal{P}[\hat{F}^{-1}] \;\subset\; \{z, -\overline{z} \;:\; z \in \sigma(A + BK_0)\},$$

and the inclusion in (4.33) holds. Finally $\hat{F}$ is holomorphic in a vertical strip S_δ for some $\delta > 0$ by its pole structure and because it is coercive.

It remains to be proved that the transfer function $\hat{G}(s) = C(sI - A)^{-1}B$ is a meromorphic function *of finite order*, whence so will be the spectral density $\hat{F}$ given by (4.30). Since C and B are bounded linear operators of finite-rank, and (ϕ_n) and (ψ_n) are Riesz bases, the sequences $((C\phi_n))$ and $(\overline{\langle b, \psi_n \rangle})$ are in l^2. It follows by (4.36) that

$$\hat{G}(s) = \sum_{n=1}^{\infty} \; d_n (s-\lambda_n)^{-1},$$

where the product sequence $(d_n) := ((C\phi_n)) \cdot \overline{\langle b, \psi_n \rangle})$ is in l^1, i.e., absolutely summable, and convergence is pointwise. It follows that

$$|\hat{G}(s)| \leq \frac{K}{d(s, \sigma(A))}, \tag{4.37}$$

for some constant K, where $d(s, \sigma(A))$ denotes the distance between s and $\sigma(A)$, which is given by

$$d(s, \sigma(A)) \;:=\; \inf \{|s - \lambda| \;:\; \lambda \in \sigma(A)\}.$$

In view of Lemma 4.5, $|\lambda_n| \longrightarrow \infty$ as $n \to \infty$. Hence, for any positive real number R, there exists a nonnegative integer $n(R)$ such that

$$|\lambda_n| > R \ \text{ for all } n > n(R).$$

Because of assumption (4.18), the maximum number of poles of the function $\hat{G}$ in a disc $\{s \in \mathbb{C} : |s| \leq R\}$ of arbitrarily large radius R, namely, $n(R)$, is such that

$$n(R) = O(R^2). \tag{4.38}$$

Moreover, in the annulus defined by the circles centered at the origin with radii R and $R+1$, respectively, the number of poles of $\hat{G}$ is $O(R)$, i.e., is equal to κR for some constant κ. These poles can be arranged in increasing order of modulus, say,

$$R_0 := R \leq R_1 \leq \cdots \leq R_{\kappa R} \leq R_{\kappa R+1} := R+1.$$

This set is formed of at most $\kappa R + 2$ numbers which are contained in an interval of length one. So the maximum gap between two consecutive numbers among them is at least $(\kappa R + 1)^{-1}$. Now if a point $s \in \mathbb{C}$ is such that $|s|$ lies in the middle of this gap, then the distance from s to the nearest pole of $\hat{G}$ is at least $(2(\kappa R + 1))^{-1}$, i.e., $O(R^{-1})$. Hence

$$d(s, \sigma(A)) = O\left(\frac{1}{R}\right) \text{ on a circle } C(0, r_R) := \{s \in \mathbb{C} : |s| = r_R\}, \text{ where } R < r_R < R+1.$$

Thus, by inequality (4.37),

$$|\hat{G}(s)| = O(R) \ \text{ on the circle } C(0, r_R). \tag{4.39}$$

It follows from (4.38) and (4.39) that the counting and proximity functions of $\hat{G}$ satisfy, respectively,

$$N(r_n, \hat{G}) = O(r_n^2) \ \text{ and } \ m(r_n, \hat{G}) = O(\log r_n) \ \text{ as } n \to \infty,$$

where the sequence of points (r_n) is such that, for all n, $n < r_n < n+1$; whence

$$T(r_n) = O\left(r_n^2\right) \quad \text{as } n \to \infty.$$

Since the characteristic function $T(r) := T(r, \hat{G})$ of $\hat{G}$ is a monotonically increasing function of $r > 0$, it follows that

$$T(r) = O(r^2) \ \text{ as } r \to \infty. \tag{4.40}$$

Observe that the order ρ of the meromorphic function $\hat{G}$ is the lower bound of all positive numbers k such that $T(r) = O(r^k)$ as $r \to \infty$. Hence, in view of (4.40), the function $\hat{G}$ is of finite order $\rho \leq 2$. □

We are now ready to show that the symmetric extraction method of spectral factorization works for such systems.

THEOREM 4.8. *Let the assumptions of Lemma* 4.6 *hold. Consider the coercive real spectral density* $\hat{F}$ *given by* (4.30), *where* $(\hat{\mathcal{N}}, \hat{\mathcal{D}})$ *is the right-coprime fraction of the transfer function* $\hat{G}(s) = C(sI - A)^{-1}B$, *which is given by* (4.29) *for some stabilizing feedback operator* $K \in \mathbf{L}(H, \mathbb{R})$ *such that* (4.28) *holds. Then the symmetric*

extraction method of spectral factorization of the spectral density $\hat{F}$ is convergent, i.e., the conclusions (a) *and* (b) *of Theorem* 4.4 *hold.*

Proof. From the initial stabilization procedure described above, it is clear that one may assume without loss of generality that the open-loop C_0-semigroup $(e^{At})_{t\geq 0}$ is exponentially stable. Hence without loss of generality, poles and zeros of $\hat{F}$ are a subset of, respectively, the λ_n and the λ_{cn} mentioned above. The conclusions then follow directly from Lemmas 4.5–4.7, by using Theorem 4.4. □

5. Example: Vibrating string with low damping. The main result of the last section, viz., Theorem 4.8, is now used in order to apply the symmetric extraction method to a lowly damped vibrating string model, with the purpose of illustrating a case of slow convergence.

In what follows $z(t,x)$ denotes the vertical position of a damped vibrating string at the place $x \in [0,1]$ and time $t \geq 0$ that is described by the PDE

$$z_{tt}(t,x) = z_{xx}(t,x) - 2\beta z_t(t,x) + b(x)u(t), \tag{5.1}$$

where the damping parameter $\beta \in (0,\pi)$ (low damping) and for all $t \geq 0$, $z(t,0) = z(t,1) = 0$; moreover, $u(t) \in \mathbb{R}$ is a scalar input, and $b(x)$ is a window function given for $\nu_i > 0$ small and $[x_i - \nu_i, x_i + \nu_i] \subset [0,1]$ by

$$b(x) := (2\nu_i)^{-1}\chi_{[x_i-\nu_i,x_i+\nu_i]}(x), \quad x \in [0,1]. \tag{5.2}$$

The scalar output $y(t) \in \mathbb{R}$ is given by

$$y(t) := \int_0^1 c(x)z(t,x)dx, \tag{5.3}$$

where $c(x)$ is a window function which for $\nu_o > 0$ small and $[x_o - \nu_o, x_o + \nu_o] \subset [0,1]$ reads

$$c(x) := (2\nu_o)^{-1}\chi_{[x_o-\nu_o,x_o+\nu_o]}(x), \quad x \in [0,1]. \tag{5.4}$$

It is moreover assumed that

$$x_i - \nu_i > 0, \quad x_o - \nu_o > x_i + \nu_i, \quad \text{and} \quad x_o + \nu_o < 1. \tag{5.5}$$

In order to show that the theory of subsection 4.2 applies to this example, we first derive a semigroup state-space model of the form (4.16) for this system. The reader is referred to [17, Examples 2.2.5 and 2.3.8] and [1, Example 3.5.3] for more detail. Consider the Hilbert space $\mathrm{H} = \mathrm{L}^2(0,1)$ with standard scalar product $\langle\cdot,\cdot\rangle_2$, which is antilinear in its second argument. Let $A:\ (D(A) \subset \mathrm{H}) \to \mathrm{H}$ be the generator of a C_0-semigroup $(e^{At})_{t\geq 0}$ on H given by

$$Az = z'', \qquad D(A) = \{z \in \mathrm{H}^2(0,1) :\ z(1) = 0,\ z(0) = 0\} = \mathrm{H}^2(0,1) \cap \mathrm{H}_0^1(0,1). \tag{5.6}$$

Since $A = A^* < 0$, A generates on H an analytic semigroup that is exponentially stable. Moreover, $D[(-A)^{\frac{1}{2}}]$ equipped with the graph norm of $(-A)^{\frac{1}{2}}$ is a Hilbert space that can be identified with $\mathrm{H}_0^1(0,1)$ equipped with the norm $||z'||_2$ for any $z \in \mathrm{H}_0^1(0,1)$ [1, Example 3.5.3]. In this sense it is possible to consider the Hilbert space

$$\mathcal{H} := D[(-A)^{\frac{1}{2}}] \oplus \mathrm{H} = \mathrm{H}_0^1(0,1) \oplus \mathrm{H}, \tag{5.7}$$

with scalar product

$$\langle \zeta, \eta \rangle_{\mathcal{H}} := \langle \zeta_1', \eta_1' \rangle_2 + \langle \zeta_2, \eta_2 \rangle_2 \qquad \forall\, \zeta = \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix},\ \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \in \mathcal{H}. \tag{5.8}$$

Recall now $b(\cdot)$ and $c(\cdot)$ given by (5.2) and (5.4), and define $B \in \mathbf{L}(\mathbb{R}, \mathrm{H})$ and $C \in \mathbf{L}(\mathrm{H}, \mathbb{R})$ by, respectively,

$$Bu := b(\cdot)u \quad \forall\, u \in \mathbb{R} \quad \text{and} \quad Cz := \langle c(\cdot), z(\cdot) \rangle_2 = \int_0^1 c(x) z(x) dx \quad \forall\, z \in \mathrm{H}. \tag{5.9}$$

Consider now $\mathcal{A}:\ (D(\mathcal{A}) \subset \mathcal{H}) \to \mathcal{H}$ given by

$$\mathcal{A} := \begin{bmatrix} 0 & I \\ A & -2\beta I \end{bmatrix}, \qquad D(\mathcal{A}) = D(A) \oplus \mathrm{H}_0^1(0,1), \tag{5.10}$$

and $\mathcal{B} \in \mathbf{L}(\mathbb{R}, \mathcal{H})$ and $\mathcal{C} \in \mathbf{L}(\mathcal{H}, \mathbb{R})$ defined by

$$\mathcal{B}u := \begin{bmatrix} 0 \\ Bu \end{bmatrix} \quad \forall\, u \in \mathbb{R} \quad \text{and} \quad \mathcal{C}\zeta := \begin{bmatrix} C & 0 \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \quad \forall\, \zeta \in \mathcal{H}. \tag{5.11}$$

Observe that $\mathcal{C} \in \mathbf{L}(\mathcal{H}, \mathbb{R})$ because $\langle c(\cdot), z(\cdot) \rangle_2 = \langle -\int_0^{\cdot} c(\xi) d\xi, z'(\cdot) \rangle_2$ for any $z \in \mathrm{H}_0^1(0,1)$. It turns out that $\mathcal{A}$ is the generator of an exponentially stable C_0-semigroup $(e^{\mathcal{A}t})_{t \geq 0}$ of contraction on $\mathcal{H}$ [17, Example 2.2.5] and $\mathcal{A}$ is a Riesz-spectral operator [17, Definition 2.3.4, Example 2.3.8]. More precisely, $\mathcal{A}$ is a Riesz-spectral operator that has for $k \in \mathbb{Z}_0$ countably many complex eigenvalues λ_k given by

$$\lambda_k = -\beta + j \operatorname{sign}(k) \sqrt{(k\pi)^2 - \beta^2}, \tag{5.12}$$

with primal Riesz basis of eigenvectors

$$\phi_k(x) = \begin{bmatrix} \phi_{k1}(x) \\ \phi_{k2}(x) \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda_k \end{bmatrix} \frac{\operatorname{sign}(k) \sin(k\pi x)}{\lambda_k} \quad x \in [0,1] \quad \forall k \in \mathbb{Z}_0, \tag{5.13}$$

and dual Riesz basis of eigenvectors

$$\psi_k(x) = \begin{bmatrix} \psi_{k1}(x) \\ \psi_{k2}(x) \end{bmatrix} = \begin{bmatrix} 1 \\ -\overline{\lambda_k} \end{bmatrix} \frac{\operatorname{sign}(k) \sin(k\pi x)}{j \operatorname{Im}(\lambda_k)} \quad x \in [0,1] \quad \forall k \in \mathbb{Z}_0, \tag{5.14}$$

such that

$$||\phi_k||_{\mathcal{H}}^2 = 1 \quad \text{and} \quad ||\psi_k||_{\mathcal{H}}^2 = \frac{(k\pi)^2}{(k\pi)^2 - \beta^2}.$$

Hence by [17, Theorem 2.3.5] the spectrum of $\mathcal{A}$ satisfies

$$\sigma(\mathcal{A}) = \overline{\{\lambda_k : k \in \mathbb{Z}_0\}},$$

where the eigenvalues λ_k are given by (5.12) and the growth constant of the semigroup $(e^{\mathcal{A}t})_{t \geq 0}$ generated by $\mathcal{A}$ is given by

$$\omega_0 := \inf_{t>0} \left(\frac{1}{t} \log ||e^{\mathcal{A}t}|| \right) = \sup_{k \in \mathbb{Z}_0} \operatorname{Re}\, \lambda_k = -\beta.$$

Thus, for $\sigma \in (-\beta, 0]$, $||e^{\mathcal{A}t}|| \leq M\ \exp(\sigma t)$ for all $t \geq 0$, and for such σ, $(e^{\mathcal{A}t})_{t\geq 0}$ is σ-exponentially stable [17, Definition 5.1.1].

Upon identifying $\zeta_1(t)(\cdot) := z(t,\cdot)$ and $\zeta_2(t)(\cdot) := z_t(t,\cdot)$, the PDE model described by (5.1)–(5.5) can be given an infinite-dimensional state-space description of the form (4.16) on the state-space $\zeta = [\zeta_1, \zeta_2]^T \in \mathcal{H}$ given by

$$\dot{\zeta} = \mathcal{A}\zeta + \mathcal{B}u(t) \quad \text{and} \quad y(t) = \mathcal{C}\zeta(t), \tag{5.15}$$

where one uses the mild solution of the state differential equation, viz.,

$$\zeta(t) = e^{\mathcal{A}t}\zeta(0) + \int_0^t e^{\mathcal{A}(t-\tau)}\mathcal{B}u(\tau)d\tau, \quad t \geq 0, \quad \zeta(0) \in \mathcal{H}, \tag{5.16}$$

and where $\mathcal{A}$ is a Riesz-spectral operator satisfying condition (4.17).

In addition, it follows from the fact that the semigroup $(e^{\mathcal{A}t})_{t\geq 0}$ is exponentially stable that (4.20) holds. Thus it remains to be proved that (4.18) and (4.19) hold. Now observe that, for all $k, l \in \mathbb{Z}_0$ such that $k \neq l$, there holds

$$|\lambda_k - \lambda_l| \ \geq \sqrt{\pi^2 - \beta^2};$$

whence (4.18) holds with

$$\delta = \inf\ \{\ |\lambda_k - \lambda_l|\ :\ k, l \in \mathbb{Z}_0\ ,\ k \neq l\}\ \geq \sqrt{\pi^2 - \beta^2}\ \ >\ 0.$$

Finally, for all k, l in $\mathbb{Z}_0$ such that $k \neq l$, there holds

$$|\lambda_l - \lambda_k|^2\ >\ (l\pi)^2 - \beta^2\ >\ 0,$$

when $\text{sign}(k) \neq \text{sign}(l)$, and

$$|\lambda_l - \lambda_k|^2\ =\ \left(\frac{\pi^2\ (k^2 - l^2)}{\sqrt{(k\pi)^2 - \beta^2} + \sqrt{(l\pi)^2 - \beta^2}}\right)^2\ >\ \pi^2\ (k-l)^2\ >\ 0,$$

when $\text{sign}(k) = \text{sign}(l)$. It follows that, for all k in $\mathbb{Z}_0$,

$$\sum_{\substack{l\ \in\ \mathbb{Z}_0 \\ l\neq k}} \frac{1}{|\lambda_l - \lambda_k|^2}\ \leq\ \sum_{l\ \in\ \mathbb{N}} \frac{1}{(l\pi)^2 - \beta^2} + \frac{1}{\pi^2} \cdot \sum_{\substack{l\ \in\ \mathbb{Z} \\ l\neq k}} \frac{1}{(l-k)^2}.$$

Hence condition (4.19) is satisfied with

$$\mu\ =\ \sup\ \left\{\ \sum_{\substack{l\ \in\ \mathbb{Z}_0 \\ l\neq k}} \frac{1}{|\lambda_l - \lambda_k|^2}\ :\ k \in \mathbb{Z}_0\right\}\ \leq\ \sum_{l\ \in\ \mathbb{N}} \frac{1}{(l\pi)^2 - \beta^2} + \frac{1}{\pi^2} \cdot \sum_{l\ \in\ \mathbb{Z}_0} \frac{1}{l^2} <\ \infty.$$

In view of Remark 4.5 (α), it follows from the exponential stability of the semigroup $(e^{\mathcal{A}t})_{t\geq 0}$ that the corresponding LQ-optimal control based (coercive) spectral density $\hat{F}$ to be factorized can without loss of generality be chosen to be

$$\hat{F} = 1 + \hat{g}_*\ \hat{g}, \tag{5.17}$$

where $\hat{g} \in \hat{\mathcal{A}}_-$ is the vibrating string model transfer function, which is given by

$$\hat{g}(s) = \frac{\sin(\sqrt{\rho_s}(1-x_o))}{\sin(\sqrt{\rho_s})} \cdot \frac{\sin(\sqrt{\rho_s}\nu_o)}{\sqrt{\rho_s}\nu_o} \cdot \frac{\sin(\sqrt{\rho_s}x_i)}{\sqrt{\rho_s}} \cdot \frac{\sin(\sqrt{\rho_s}\nu_i)}{\sqrt{\rho_s}\nu_i} \tag{5.18}$$

or, equivalently,

$$\hat{g}(s) = \frac{\sinh(\sqrt{r_s}(1-x_o))}{\sinh(\sqrt{r_s})} \cdot \frac{\sinh(\sqrt{r_s}\nu_o)}{\sqrt{r_s}\nu_o} \cdot \frac{\sinh(\sqrt{r_s}x_i)}{\sqrt{r_s}} \cdot \frac{\sinh(\sqrt{r_s}\nu_i)}{\sqrt{r_s}\nu_i}, \tag{5.19}$$

where $\rho_s := -s(2\beta + s)$ and $r_s := -\rho_s$.

Remark 5.1. As the semigroup $(e^{\mathcal{A}t})_{t\geq 0}$ generated by $\mathcal{A}$ is σ-exponentially stable for $\sigma \in (-\beta, 0]$, and $\mathcal{B} \in \mathbf{L}(\mathbb{R}, \mathcal{H})$ and $\mathcal{C} \in \mathbf{L}(\mathcal{H}, \mathbb{R})$, there holds by [17, Lemma 7.3.1] that the transfer function given above belongs to the class $\hat{\mathcal{A}}_-(\sigma)$ and to the class $\hat{\mathcal{A}}(\sigma)$ for $\sigma \in (-\beta, 0]$. As the corresponding impulse response $g(t)$ has no impulses, one has $\exp(-\sigma\cdot)g(\cdot) \in \mathrm{L}^1(0,\infty)$, $\sup_{\mathrm{Re}\ s\geq\sigma} |\hat{g}(s)| < \infty$, and $\hat{g}(s)$ is zero at infinity in $\mathbb{C}_{\sigma+} := \{s \in \mathbb{C} : \mathrm{Re}\ s \geq \sigma\}$.

Now observe that the spectral density $\hat{F}$ can be written as

$$\hat{F} \ = \ \frac{N}{D}, \tag{5.20}$$

where N and D are the real parahermitian entire functions given, respectively, by

$$N(s) \ := \ n(-s)\cdot n(s) + d(-s)\cdot d(s) \qquad \text{and} \qquad D(s) \ := \ d(-s)\cdot d(s), \tag{5.21}$$

where d and n are, respectively, the denominator and numerator of the transfer function $\hat{g}$, which are the entire functions given, respectively, by

$$d(s) \ := \ \frac{\sin(\sqrt{\rho_s})}{\sqrt{\rho_s}} \ = \ \frac{\sinh(\sqrt{r_s})}{\sqrt{r_s}}, \tag{5.22}$$

and

$$n(s) := \hat{g}(s)\ d(s), \tag{5.23}$$

where the zeros of d are exactly the open-loop eigenvalues λ_k, given by (5.12). Observe that the numerator N and denominator D of the spectral density $\hat{F}$ above may have infinitely many common zeros (see Remark 4.5 (β)). In addition observe that, in view of (5.21)–(5.22), the entire functions N and D are of finite order (see, e.g., [37, Example 1, p. 76]).

It follows from the analysis above that, by Theorem 4.8, the spectral factorization by symmetric extraction of the spectral density $\hat{F}$ given by (5.20)–(5.22) is convergent; i.e., the conclusions (a) and (b) of Theorem 4.4 hold.

Numerical results are presented in Table 5.1. These results were obtained for the following parameter values: $\beta = 2$, $x_i = 0.02$, $x_o = 1 - x_i = 0.98$, and $\nu_i = \nu_o = 0.01$. It is found that the closed-loop eigenvalues λ_{cn} have numerically a constant real part equal to -2 and hence are vertically distant from the open-loop ones by $|\lambda_n - \lambda_{cn}|$. One can observe that the convergence is slow. Moreover, the absolute and relative errors are overall decreasing in an oscillatory manner. Further numerical evidence leads us to conjecture that $|\lambda_n - \lambda_{cn}|$ is of order $n^{-\alpha}$, where α is slightly larger than one.

TABLE 5.1
Eigenvalues, eigenvalue errors $\delta_n := |\lambda_n - \lambda_{cn}|$ *(versus the sequences* $(1/n)$ *and* $(1/n^2)$*).*

n	λ_n	δ_n	$n \cdot \delta_n$	$n^2 \cdot \delta_n$
1	-2+2.42j	1.55e-7	1.55e-7	1.55e-7
2	-2+5.96j	3.26e-7	6.52e-7	1.30e-6
3	-2+9.21j	6.05e-7	1.81e-6	5.44e-6
4	-2+12.41j	9.79e-7	3.92e-6	1.57e-5
5	-2+15.58j	1.43e-6	7.16e-6	3.58e-5
16	-2+50.23j	6.28e-6	1.01e-4	1.62e-3
17	-2+53.37j	6.31e-6	1.07e-4	1.82e-3
18	-2+56.51j	6.25e-6	1.12e-4	2.02e-3
19	-2+59.66j	6.10e-6	1.16e-4	2.22e-3
20	-2+62.80j	5.86e-6	1.17e-4	2.34e-3
21	-2+65.94j	5.56e-6	1.17e-4	2.46e-3
22	-2+69.09j	5.21e-6	1.15e-4	2.53e-3
23	-2+72.23j	4.81e-6	1.11e-4	2.55e-3
24	-2+75.37j	4.38e-6	1.05e-4	2.52e-3
25	-2+78.51j	3.93e-6	9.83e-5	2.46e-3
36	-2+113.08j	4.19e-7	1.51e-5	5.44e-4
37	-2+116.22j	3.01e-7	1.12e-5	4.14e-4
38	-2+119.36j	2.11e-7	8.01e-6	3.04e-4
39	-2+122.51j	1.42e-7	5.56e-6	2.17e-4
40	-2+125.65j	9.26e-8	3.71e-6	1.48e-4

This is theoretically confirmed by the facts that (1) by Lemma 4.6, (5.11), and (5.14), $|\lambda_n - \lambda_{cn}| = O(\frac{x_n}{n})$, where (x_n) is a square-summable sequence, whence $\alpha > 1$, and (2) the linearized, i.e., Newton–Raphson, estimate of $|\lambda_n - \lambda_{cn}|$ is $O(\frac{1}{n^2})$. Thus as nonlinear perturbations do not improve the speed of convergence, one has $\alpha \in (1, 2]$: the situation is comparable with that of Example 4.1.

Notice further that the tail-sums used in (4.6)–(4.8) are here of order $\frac{1}{n^{\alpha-1}}$. Hence the error analysis of Theorem 4.2 reveals that approximate spectral factorization will be achieved very slowly, the main reason being the asymptotically linear distribution of the spectra along a vertical line in the open left half-plane. A better situation is to be expected when this is not the case, i.e., acceleration by the fact that the real parts of the closed-loop eigenvalues Re λ_{cn} tend to $-\infty$.

6. Conclusion. As we have seen, the symmetric extraction method may be applied to a wide class of distributed parameter systems, for which its convergence has been established theoretically.

Another example for which the symmetric extraction method is appropriate is the beam equation with structural damping (see, e.g., [32, pp. 131–133]), and in this case one would expect the convergence to be faster, since the real parts of the eigenvalues tend to $-\infty$. More generally, one would expect the convergence to be faster when the semigroup is analytic, since in that case the spectrum lies in a sector contained in some left half-plane; see, e.g., [2].

Other possible techniques for approaching the spectral factorization problem for distributed parameter systems include a direct approximation of the spectral density function, but this needs to be treated with caution, since the mapping from spectral density to spectral factor is discontinuous in the uniform norm (see, e.g., [22]). It would also be of interest to extend the present methods to multivariable systems (the finite-dimensional case was analyzed in [4]), but this introduces additional function-theoretic difficulties.

As a referee has observed, there may be connections between the factorization

approach taken here and the invariant subspace approach. This could be an interesting topic for further research, in particular for the special class of Riesz-spectral systems; see [24].

REFERENCES

[1] S. P. Banks, *State-Space and Frequency-Domain Methods in the Control of Distributed Parameter Systems*, IEEE Topics in Control Ser. 3, Peter Peregrinus, London, 1983.

[2] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Vol. I, Systems Control Found. Appl., Birkhaüser Boston, Boston, 1992.

[3] R. P. Boas, *Entire Functions*, Academic Press, New York, 1954.

[4] F. M. Callier, *On polynomial matrix spectral factorization by symmetric extraction*, IEEE Trans. Automat. Control, 30 (1985), pp. 453–464.

[5] F. M. Callier and C. A. Desoer, *An algebra of transfer functions of distributed linear time–invariant systems*, IEEE Trans. Circuits Systems, 25 (1978), pp. 651–662.

[6] F. M. Callier and C. A. Desoer, *Correction to "An algebra of transfer functions of distributed linear time–invariant systems,"* IEEE Trans. Circuits Systems, 26 (1979), p. 360.

[7] F. M. Callier and L. Dumortier, *Partially stabilizing LQ-optimal control for stabilizable semigroup systems*, Integral Equations Operator Theory, 32 (1998), pp. 119–151.

[8] F. M. Callier and J. Winkin, *The spectral factorization problem for SISO distributed systems*, in Modelling, Robustness and Sensitivity Reduction in Control Systems, NATO Adv. Sci. Inst. Ser. F Comput. Systems Sci. 34, R. F. Curtain, ed., Springer-Verlag, Berlin, 1987, pp. 463–489.

[9] F. M. Callier and J. Winkin, *Spectral factorization and LQ-optimal regulation for multivariable distributed systems*, Internat. J. Control, 52 (1990), pp. 55–75.

[10] F. M. Callier and J. Winkin, *LQ-optimal control of infinite-dimensional systems by spectral factorization*, Automatica J. IFAC, 28 (1992), pp. 757–770.

[11] F. M. Callier and J. Winkin, *Infinite dimensional system transfer functions*, in Analysis and Optimization of Systems: State and Frequency Domain Approaches to Infinite-Dimensional Systems, Lecture Notes in Control and Inform. Sci. 185, R. F. Curtain, A. Bensoussan, and J. L. Lions, eds., Springer-Verlag, Berlin, New York, 1993, pp. 72–101.

[12] F. M. Callier and J. Winkin, *The spectral factorization problem for multivariable distributed parameter systems*, Integral Equations Operator Theory, 34 (1999), pp. 270–292.

[13] F. M. Callier and J. Winkin, *On spectral factorization by symmetric extraction for distributed parameter systems*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 1112–1117.

[14] F. M. Callier and J. Winkin, *Spectral factorization by symmetric extraction for distributed parameter systems*, in Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems, Perpignan, France, 2000; available on CD-ROM, SI21B_2.

[15] R. F. Curtain, *Pole assignment for distributed systems by finite-dimensional control*, Automatica J. IFAC, 21 (1985), pp. 57–67.

[16] R. F. Curtain, *Linear operator inequalities for strongly stable weakly regular linear systems*, Math. Control Signals Systems, 14 (2001), pp. 299–337.

[17] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, Heidelberg, Germany, 1995.

[18] P. Grabowski, *The LQ-controller problem: An example*, IMA J. Math. Contr. Inform., 11 (1994), pp. 355–368.

[19] P. Grabowski and F. M. Callier, *On the circle criterion for boundary control systems in factor form: Lyapunov stability and Lur'e equations*, Internal report 2002/05, University of Namur, Namur, Belgium, 2002.

[20] G. H. Hardy, *A Course of Pure Mathematics*, Cambridge Math. Lib. Ser., Cambridge University Press, Cambridge, UK, 1992.

[21] E. Hille, *Analytic Function Theory*, Vol. I, Ginn, Boston, 1959.

[22] B. Jacob and J. R. Partington, *On the boundedness and continuity of the spectral factorization mapping*, SIAM J. Control Optim., 40 (2001), pp. 88–106.

[23] B. Jacob, J. Winkin, and H. Zwart, *Continuity of the spectral factorization on a vertical strip*, Systems Control Lett., 37 (1999), pp. 183–192.

[24] C. R. Kuiper and H. J. Zwart, *Connections between the algebraic Riccati equation and the Hamiltonian for Riesz-spectral systems*, J. Math. Systems, Estim. Control, 6 (1996), pp. 1–48.

[25] B. Ja. Levin, *Distribution of zeros of entire functions*, AMS Translations of Mathematical Monographs 5, AMS, Providence, RI, 1980.
[26] J. E. Marsden, *Basic Complex Analysis*, W. H. Freeman, New York, 1973.
[27] R. Nevanlinna, *Analytic Functions*, Springer-Verlag, Berlin Heidelberg, 1970.
[28] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.
[29] C. L. Prather and A. C. M. Ran, *A Hadamard factorization theorem for entire matrix valued functions*, Oper. Theory Adv. Appl., 19 (1986), pp. 359–372.
[30] C. L. Prather and A. C. M. Ran, *Factorization of a class of meromorphic matrix valued functions*, J. Math. Anal. Appl., 127 (1987), pp. 413–422.
[31] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, New York, 1974.
[32] D. L. Russell, *On mathematical models for the elastic beam with frequency-proportional damping*, in Control and Estimation in Distributed Parameter Systems, SIAM Frontiers Appl. Math. 11, H. T. Banks, ed., SIAM, Philadelphia, 1992, pp. 125–169.
[33] O. J. Staffans, *Quadratic optimal control through coprime and spectral factorizations*, Abo Akademi Reports on Computer Science and Mathematics, 29 (1996), pp. 131–138.
[34] S.-H. Sun, *On spectrum distribution of completely controllable linear systems*, SIAM J. Control Optim., 19 (1981), pp. 730–743.
[35] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
[36] M. Weiss and G. Weiss, *Optimal control of stable weakly regular linear systems*, Math. Control Signals Systems, 10 (1997), pp. 287–330.
[37] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1467–1492

ERGODIC CONTROL FOR CONSTRAINED DIFFUSIONS: CHARACTERIZATION USING HJB EQUATIONS*

VIVEK BORKAR† AND AMARJIT BUDHIRAJA‡

Abstract. Recently in [A. Budhiraja, *SIAM J. Control Optim.*, 42 (2003), pp. 532–558] an ergodic control problem for a class of diffusion processes, constrained to take values in a polyhedral cone, was considered. The main result of that paper was that under appropriate conditions on the model, there is a Markov control for which the infimum of the cost function is attained. In the current work we characterize the value of the ergodic control problem via a suitable Hamilton–Jacobi–Bellman (HJB) equation. The theory of existence and uniqueness of classical solutions, for PDEs in domains with corners and reflection fields which are oblique, discontinuous, and multivalued on corners, is not available. We show that the natural HJB equation for the ergodic control problem admits a unique continuous viscosity solution which enables us to characterize the value function of the control problem. The existence of a solution to this HJB equation is established via the classical vanishing discount argument. The key step is proving the precompactness of the family of suitably renormalized discounted value functions. In this regard we use a recent technique, introduced in [V. S. Borkar, *Stochastic Process Appl.*, 103 (2003), pp. 293–310], of using the Athreya–Ney–Nummelin pseudoatom construction for obtaining a coupling of a pair of embedded, discrete time, controlled Markov chains.

Key words. ergodic control, optimal Markov control, controlled reflected diffusions, constrained processes, HJB equation, viscosity solutions, domains with corners, oblique Neumann problem, pseudoatom, coupling

AMS subject classifications. 93E20, 60H30, 60J60

DOI. 10.1137/S0363012902417619

1. Introduction.

In a recent work [10] an ergodic control problem for a class of constrained diffusion processes, in polyhedral cones, was studied. Such controlled constrained diffusion processes arise in the heavy traffic analysis of single class open queuing networks with state dependent arrival and service rates with control in the marginal service rates (cf. [24]). The study of optimal control of queuing networks in heavy traffic via the analysis of the control problem for a suitable limit controlled diffusion process is currently an active area of research (cf. [23], [21], [22], [20], [26], [13], [12], [24], [1]). The control problems considered in the above works (excepting the last two papers) are somewhat different from that in [10] (and the current work) in that they correspond to the control of sequencing and routing of jobs in the network. In the diffusion limit such control problems lead to rather nontrivial singular control problems with state constraints. In contrast, the current paper (and also [24], [1]) considers the problem of drift control for a diffusion which is constrained to take values in a polyhedral domain via the action of a suitable Skorohod map. The domain $G \subset \mathbb{R}^k$, which is the state space of the controlled Markov process, is given as an intersection of N half spaces G_i; $i = 1, \ldots, N$. Associated with each G_i is a vector d_i which defines the "direction of constraint" in the relative interior of ∂G_i. At a

*Received by the editors November 12, 2002; accepted for publication (in revised form) April 16, 2004; published electronically January 27, 2005. Research supported in part by a grant for "Nonlinear Studies" from the Indian Space Research Organization and the Defense Research and Development Organization, Government of India, administered through the Indian Institute of Science.

http://www.siam.org/journals/sicon/43-4/41761.html

†School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India (borkar@tifr.res.in).

‡Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260 (budhiraj@email.unc.edu).

point $x \in \partial G$ where several faces meet, there is more than one possible direction of constraint; in fact, the set of permissible directions is a cone denoted by $d(x)$. Roughly speaking, the constrained version of a given unrestricted trajectory in $\mathbb{R}^k$ is obtained by pushing back the trajectory, whenever it is about to exit the domain, in one of the permissible directions of constraint using the minimal force required to keep the trajectory inside the domain. Precise definitions will be given in section 2. The constraining mechanism is described via the notion of a Skorohod problem. Under appropriate conditions on $(d_i)_{i=1}^N$ it follows from the results in [16] that one can define the "Skorohod map," denoted as $\Gamma(\cdot)$, which takes an unrestricted trajectory $\psi(\cdot)$ and maps it to a trajectory $\phi(\cdot) \doteq \Gamma(\psi)(\cdot)$ such that $\phi(t) \in G$ for all $t \in (0, \infty)$.

The controlled constrained diffusion process that we will study is obtained as a solution to the equation

$$X(t) = \Gamma\left(X(0) + \int_0^{\cdot} b(X(s), u(s))ds + \int_0^{\cdot} \sigma(X(s))dW(s)\right)(t), \quad t \in [0, \infty), \tag{1.1}$$

where $W(\cdot)$ is a standard Wiener process, $b : G \times U \to \mathbb{R}^k$; $\sigma : G \to \mathbb{R}^{k\times k}$ are suitable coefficients, U is a given control set, and $u(\cdot)$ is a U valued "admissible" control process. The cost of interest is the ergodic cost criterion

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T k(X(s), u(s))ds, \tag{1.2}$$

where the limit above is taken almost surely (a.s.) and $k : G \times U \to \mathbb{R}$ is a suitable map.

In control theory, one of the most desirable features of a good control is that it should depend only on the current value of the state and not on the whole history of the state and/or the control process. Namely, one is interested in obtaining controls $u(\cdot)$ such that there exists some measurable map $v : G \to U$ satisfying $u(t) = v(X(t))$ a.s. for all $t \in [0, \infty)$. Under such a control the solution to (1.1) becomes a Markov process and for this reason the map $v(\cdot)$ is referred to as a "Markov control." The main result of [10] is that, under appropriate conditions on the model (Conditions 2.2, 2.4, 2.5, and 2.8 in section 2), there is a Markov control for which the infimum of the cost in (1.2) is attained.

The other important goal in stochastic control theory is the characterization of the value function of the control problem via a suitable Hamilton–Jacobi–Bellman (HJB) equation. For unconstrained diffusions this problem has been extensively studied and we refer the reader to [7] for a detailed account. For the controlled Markov processes in the present work, the problem is quite challenging since the domain in which the process is constrained to lie is not smooth (because of the corners where the faces meet) and the reflection field is oblique, discontinuous, and multivalued at the boundary points which lie on more than one face. The theory of existence and uniqueness of classical solutions for PDEs in such domains is not available. However, using the fundamental ideas of Crandall and Lions [15] and Lions [25], Dupuis and Ishii [17] have developed an existence and uniqueness theory of viscosity solutions for fully nonlinear second order elliptic PDEs on such domains. In this work we will show that the value of the ergodic control problem introduced above can be characterized via the unique viscosity solution of an appropriate HJB equation. The usual approach to the HJB equation for the ergodic control is via the "vanishing discount method" (cf. [14], [9], [7], [4], [27]). In this approach one first studies the value function $V_\alpha(x)$

of the discounted control problem

$$V_\alpha(x) = \inf_u \mathbb{E}\left(\int_0^\infty e^{-\alpha s} k(X^x(s), u(s))ds\right), \tag{1.3}$$

where $\alpha \in (0,\infty)$, the infimum is taken over all admissible controls u, and $X^x(\cdot)$ is the solution of (1.1) with $X(0) \equiv x$. For $f \in C_b^2(G)$ let $Lf : G \times U \to \mathbb{R}$ be defined as

$$(Lf)(x,u) \doteq \frac{1}{2}\sum_{i,j=1}^k a_{i,j}(x)\frac{\partial^2 f}{\partial x_i \partial x_j}(x) + \sum_{i=1}^k b_i(x,u)\frac{\partial f}{\partial x_i}(x), \quad (x,u) \in G \times U, \tag{1.4}$$

where $a_{ij}(x) \doteq \sigma(x)\sigma^T(x)$. Using results of [17] we will show that the value function $V_\alpha(x)$ is the unique viscosity solution (see Definition 3.3) of the following HJB equation:

$$\begin{aligned} \inf_{u\in U}\,(L\psi(x,u) + k(x,u) - \alpha\psi(x)) = 0, \quad x \in G, \\ \langle \nabla\psi(x), d_i\rangle = 0, \quad x \in \partial G,\ i \in \mathrm{In}(x), \end{aligned} \tag{1.5}$$

where $\mathrm{In}(x) \doteq \{i \in \{1,2,\ldots,N\} : x \in \partial G_i\}$. We remark that the work [17] considers the case where the domain is bounded; however, by a slight modification the techniques there can be used to cover the case in the present work.

In order to study the HJB equation of the ergodic control problem, we need to take the limit as $\alpha \to 0$. The key step in this program is to show that the family

$$\{\overline{V}_\alpha(\cdot) \doteq V_\alpha(\cdot) - V_\alpha(0),\ \alpha \in (0,\infty)\} \tag{1.6}$$

is precompact in $C(G)$. The classical derivation (see Theorem VI.3.1 of [7]) of such a result makes use of certain gradient estimates on $V_\alpha(x)$, uniform in α, which we are unable to prove for the model considered in the present work. Another approach based on viscosity solutions, taken in [4], proves the above precompactness by making some strong stability assumptions on the model (a restoring force toward bounded sets that grows without bound as $|x| \to \infty$) which are not natural for the constrained diffusion models that arise from the heavy traffic analysis of queuing networks. In the present work we prove the precompactness of the family in (1.6) by using the Athreya–Ney–Nummelin pseudoatom construction which was recently introduced in the context of partially observed ergodic control problems in [5]. The importance of pseudoatom construction ideas in ergodic control problems has also been pointed out in [28]. Using this construction, the precompactness of the family of (renormalized) discounted value functions for a partially observed control problem was proved in [8]. One of the key requirements for the coupling methods used in the above cited works to work, is the existence of a suitable Lyapunov function for the underlying controlled Markov processes. For the processes considered in the present work, the existence of such a Lyapunov function was proved in [2]. Using this Lyapunov function one can show that a Foster-type drift criterion is satisfied for an appropriate embedded discrete time controlled Markov chain. This, along with the pseudoatom construction, enables us to show that the coupling time, for two embedded controlled Markov chains driven by the same Markov control and independent noise processes but with two different initial conditions, has finite moments. The above step is the main ingredient to the proof of the precompactness of (1.6).

Once the precompactness is proved, one can take the limit of $(\alpha V_\alpha(0), \overline{V}_\alpha(\cdot))$, along a subsequence, as $\alpha \to 0$. Then by stability (under perturbations) properties of

viscosity solutions it follows that the limit, denoted as $(\rho, V(\cdot))$, is a viscosity solution of the HJB equation for the ergodic control problem (see (5.2)). The rest of the work involves showing that this equation admits a unique solution and that ρ is the infimum, over all admissible controls, of the cost function in (1.2).

The control problem considered in this work is motivated by average cost per unit time control problems for open queuing networks in heavy traffic. Consider a stochastic processing network consisting of k service stations each having input streams of jobs, possibly from outside and from other service stations in the network. The routing of the jobs in the network is probabilistic but uncontrolled and fixed (see [24]). More precisely, on completion of service at the ith station a customer is routed to station j with probability p_{ij}. The interarrival times of external streams of customers, in general, need not be independent, and their probability distribution could be state dependent. Here the state of the system is given by the vector of queue lengths at the various stations in the network. The system is "single class" in that the service rates at a given station do not depend on the customer type; however, the service rates could be state dependent. Furthermore, a system manager may exercise control to adjust the marginal service rates. This is the only mode of control in the system. The goal of the system manager is to control the marginal service rates in order to minimize a long term average cost per unit time where the cost function could depend on the queue lengths and the control process. Such a control problem was considered in [24] under a further restriction that the buffer lengths at all the service stations are finite. The authors showed that under appropriate assumptions on the arrival, service processes, routing probability matrix, and suitable heavy traffic conditions, such a control problem can be approximated by an ergodic control problem for certain controlled diffusions in compact polyhedral domains with control appearing only in the drift term. The compactness of the domain is a consequence of assumption of finite buffers at every station. More precisely, the authors show that the value function for the suitably scaled queuing network problem converges to the value function of the diffusion control problem as the scaling parameter approaches its limit. Furthermore, it is shown that an almost optimal solution for the limiting diffusion control problem can be used to obtain a near optimal solution to the network control problem when the network is close to heavy traffic. Thus it becomes of key importance to obtain methods for computing or numerically approximating controls for the limit diffusion control problem. In this work we consider the infinite buffer analogue of the control problem studied in [24]. Since in this setting one does not have the compactness of the state space which is critically exploited in the analysis of [24], the precise connection between the ergodic cost problem for the network and the corresponding control problem for the diffusion is still an open issue. Our goal in the current work is to take a first step in this direction by studying the properties of the formal diffusion control problem that arises in the heavy traffic analysis of the above described control problem. The main result of this work, namely, the characterization of the value function of the limit control problem as a unique solution of a suitable HJB equation, is the first step in numerically solving for an almost optimal control for the diffusion control problem. The eventual goal, of course, is to establish the convergence of the value function of the network control problem to the value function of the diffusion control problem and then obtain near optimal controls for the underlying queuing network. This will be studied in our future work.

The paper is organized as follows. In section 2 we present some preliminary definitions and known results that will be used in this work. Section 3 is devoted to showing that the value function of the discounted cost problem is the unique solution

of the HJB equation in (1.5). In section 4 we present the pseudoatom construction and use it to show the precompactness of the family in (1.6). In section 5, by taking the limit as $\alpha \to 0$ we obtain a viscosity solution of the HJB equation for the ergodic control problem. Finally, we characterize the value function by showing that the equation admits a unique viscosity solution.

2. Preliminaries and background results. Let $G \subset \mathbb{R}^k$ be a polyhedral cone with the vertex at the origin given as the intersection of half spaces G_i, $i = 1, \ldots, N$. Each half space G_i is associated with a unit vector n_i via the relation $G_i = \{x \in \mathbb{R}^k : \langle x, n_i \rangle \geq 0\}$, where $\langle \cdot, \cdot \rangle$ denotes the usual inner product in $\mathbb{R}^k$. Denote the boundary of a set $B \subset \mathbb{R}^k$ by ∂B. We will denote the set $\{x \in \partial G : \langle x, n_i \rangle = 0\}$ by F_i. For $x \in \partial G$, define the set, $n(x)$, of inward normals to G at x by $n(x) \doteq \{r : |r| = 1,\ \langle r, x - y \rangle \leq 0,\ \forall y \in G\}$. With each face F_i we associate a unit vector d_i such that $\langle d_i, n_i \rangle > 0$. This vector defines the *direction of constraint* associated with the face F_i. For $x \in \partial G$ define $d(x) \doteq \{d \in \mathbb{R}^k : d = \sum_{i \in \mathrm{In}(x)} \alpha_i d_i;\ \alpha_i \geq 0;\ \|d\| = 1\}$. We will denote the collection of all subsets of $\{1, \ldots, N\}$ by Λ.

Let $D([0, \infty) : \mathbb{R}^k)$ denote the set of functions mapping $[0, \infty)$ to $\mathbb{R}^k$ that are right continuous and have limits from the left. We endow $D([0, \infty) : \mathbb{R}^k)$ with the usual Skorohod topology. Let $D_G([0, \infty) : \mathbb{R}^k) \doteq \{\psi \in D([0, \infty) : \mathbb{R}^k) : \psi(0) \in G\}$. For $\eta \in D([0, \infty) : \mathbb{R}^k)$ let $|\eta|(T)$ denote the total variation of η on $[0, T]$ with respect to the Euclidean norm on $\mathbb{R}^k$.

DEFINITION 2.1. *Let $\psi \in D_G([0, \infty) : \mathbb{R}^k)$ be given. Then $(\phi, \eta) \in D([0, \infty) : \mathbb{R}^k) \times D([0, \infty) : \mathbb{R}^k)$ solves the Skorohod problem (SP) for ψ with respect to G and d if and only if $\phi(0) = \psi(0)$, and for all $t \in [0, \infty)$ (1) $\phi(t) = \psi(t) + \eta(t)$; (2) $\phi(t) \in G$; (3) $|\eta|(t) < \infty$; (4) $|\eta|(t) = \int_{[0,t]} I_{\{\phi(s) \in \partial G\}} d|\eta|(s)$; (5) There exists (Borel) measurable $\gamma : [0, \infty) \to \mathbb{R}^k$ such that $\gamma(t) \in d(\phi(t))$ ($d|\eta|$ almost everywhere (a.e.)) and $\eta(t) = \int_{[0,t]} \gamma(s) d|\eta|(s)$.*

On the domain $D \subset D_G([0, \infty) : \mathbb{R}^k)$ on which there is a unique solution to the SP we define the Skorohod map (SM) Γ as $\Gamma(\psi) \doteq \phi$, if $(\phi, \psi - \phi)$ is the unique solution of the SP posed by ψ. The following is the key assumption made in [10] on the data defining the SP.

Condition 2.2. (a) There exists a compact, convex set $B \in \mathbb{R}^k$ with $0 \in B^0$ such that if $v(z)$ denotes the set of inward normals to B at $z \in \partial B$, then for $i = 1, 2, \ldots, N$, $z \in \partial B$ and $|\langle z, n_i \rangle| < 1$ implies that $\langle v, d_i \rangle = 0$ for all $v \in v(z)$. (b) There exists a map $\pi : \mathbb{R}^k \to G$ such that if $y \in G$, then $\pi(y) = y$, and if $y \notin G$, then $\pi(y) \in \partial G$, and $y - \pi(y) = \alpha\gamma$ for some $\alpha \leq 0$ and $\gamma \in d(\pi(y))$. (c) For every $x \in \partial G$, there is $n \in n(x)$ such that $\langle d, n \rangle > 0$ for all $d \in d(x)$.

An important consequence of the above assumption is the regularity of the SM in the following sense.

THEOREM 2.3 (Dupuis and Ishii [16]). *Under Condition* 2.2 *the SM is well defined on all of $D_G([0, \infty) : \mathbb{R}^k)$, i.e., $D = D_G([0, \infty) : \mathbb{R}^k)$, and the SM is Lipschitz continuous in the following sense. There exists a $K < \infty$ such that for all $\phi_1, \phi_2 \in D_G([0, \infty) : \mathbb{R}^k)$*

$$\sup_{0 \leq t < \infty} |\Gamma(\phi_1)(t) - \Gamma(\phi_2)(t)| < K \sup_{0 \leq t < \infty} |\phi_1(t) - \phi_2(t)|. \tag{2.1}$$

In the rest of the paper Condition 2.2 will always be taken to hold. We refer the reader to [18] for sufficient conditions and examples for which the above condition holds. We will also assume without loss of generality that $K \geq 1$.

We now introduce the controlled constrained diffusion processes that will be studied in this paper. Throughout this paper we will assume the relaxed control framework; i.e., there is a compact metric space S such that the control set is $U \doteq \mathcal{P}(S)$ (the space of all probability measures on S endowed with the weak convergence topology). All topological spaces in this paper will be endowed with their natural Borel σ-field. For a topological space $\mathcal{K}$, we will denote its Borel σ-field by $\mathcal{B}(\mathcal{K})$. The space of all real, measurable and bounded functions defined on $\mathcal{K}$ will be denoted as $BM(\mathcal{K})$, the subset of $BM(\mathcal{K})$ consisting of continuous functions will be denoted by $C_b(\mathcal{K})$, and the space of all probability measures on $(\mathcal{K}, \mathcal{B}(\mathcal{K}))$ will be denoted by $\mathcal{P}(\mathcal{K})$. The space $\mathcal{P}(\mathcal{K})$ will be endowed with the weak convergence topology. For $A \in \mathcal{B}(\mathcal{K})$, $\mathcal{I}_A(\cdot)$ will denote the indicator function of the set A. Also, we will denote by $C_b^2(G)$ the space of real valued, bounded and twice continuously differentiable functions on G. By a filtered probability space $(\Omega, \mathcal{F}, P, (\mathcal{F}_t))$ we will mean a probability space $(\Omega, \mathcal{F}, P)$ endowed by a filtration $(\mathcal{F}_t)_{t\geq 0}$ satisfying the usual hypothesis. A pair of stochastic processes $(u(\cdot), W(\cdot))$ defined on some filtered probability space $(\Omega, \mathcal{F}, P, (\mathcal{F}_t))$ is said to be an admissible pair if $W(\cdot)$ is an $\mathcal{F}_t$-standard Wiener process and $u(\cdot)$ is a U valued, measurable, $\{\mathcal{F}_t\}$ adapted process.

We will consider controlled constrained diffusion processes of the form defined in (1.1), where for $(x, u) \in G \times U$, $b(x, u) \doteq \int_S \bar{b}(x, \alpha)u(d\alpha)$ and the coefficients $\sigma : G \to \mathbb{R}^{k\times k}$ and $\bar{b} : G \times S \to \mathbb{R}^k$ satisfy the following conditions.

Condition 2.4. There exists $r \in (0, \infty)$ such that

(i) $\bar{b}$ is a continuous map and for all $x, y \in G$ and $\alpha \in S$

$$\|\bar{b}(x, \alpha) - \bar{b}(y, \alpha)\| + \|\sigma(x) - \sigma(y)\| \leq r\|x - y\|.$$

(ii) For all $x \in G$ and $\alpha \in S$

$$\|\bar{b}(x, \alpha)\| + \|\sigma(x)\| \leq r.$$

We will also assume the following nondegeneracy assumption on σ.

Condition 2.5. There exists $c_0 \in (0, \infty)$ such that for all $x \in G$ and $\alpha \in \mathbb{R}^k$ $\alpha'(\sigma(x)\sigma'(x))\alpha \geq c_0\alpha'\alpha$.

In the rest of the paper, in addition to Condition 2.2, Conditions 2.4 and 2.5 will also be assumed to hold. Under these conditions, it follows, via the Lipschitz property of the SM and the usual fixed point arguments, that (1.1) admits a unique strong solution (cf. Theorem 2.6 of [10]). If $X(\cdot)$ solves (1.1), then (cf. Theorem 3.5.1 of [23]) there exist continuous, increasing $\mathcal{F}_t$ adapted processes $\{Y_i(\cdot);\ 1 \leq i \leq N\}$ such that

$$X(t) = X(0) + \int_0^t b(X(s), u(s))ds + \int_0^t \sigma(X(s))dW(s) + \sum_{i=1}^N d_i Y_i(t) \tag{2.2}$$

for all t, a.s. Furthermore, $Y_i(0) = 0$ and for all $t > 0$ $\int_0^t \mathcal{I}_{F_i}(X(s))dY_i(s) = Y_i(t)$, a.s., $i = 1, \ldots, N$. The following lemma essentially says that in considering admissible controls, we can without loss of generality restrict ourselves to controls that are adapted with respect to the filtration generated by $(X(\cdot), Y(\cdot))$. The proof is similar to Theorem 1.2.2, p. 18, of [7] and is therefore omitted.

LEMMA 2.6. *Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ be a filtered probability space on which is given an admissible pair $(u(\cdot), W(\cdot))$. Let $X(\cdot)$ be a solution to* (1.1) *with the corresponding boundary processes $\{Y_i(\cdot)\}_{i=1}^N$ as in* (2.2). *Then there exists an enlargement*

$(\overline{\Omega}, \overline{\mathcal{F}}, \{\overline{\mathcal{F}}_t\}, \overline{P})$ of the above probability space on which is given an $\{\overline{\mathcal{F}}_t\}$ Wiener process $\tilde{W}(\cdot)$ and $\mathcal{P}(S)$ valued measurable stochastic process $\tilde{u}(\cdot)$ such that for a.e. $t \in [0,\infty)$, $\tilde{u}(t)$ is $\mathcal{F}_t^{X,Y}$ measurable ($\tilde{u}$ is called a feedback control), where $\mathcal{F}_t^{X,Y}$ denotes the P completion of $\sigma\{X(s);\ \{Y_i(s)\}_{i=1}^N;\ 0 \le s \le t\}$, and $X(\cdot)$ solves

$$X(t) = X(0) + \int_0^t b(X(s), \tilde{u}(s))ds + \int_0^t \sigma(X(s))d\tilde{W}(s) + \sum_{i=1}^N d_i Y_i(t).$$

Henceforth, without loss of generality, we will assume that all controls are feedback controls. Next we introduce Markov controls. We begin with the following definition.

DEFINITION 2.7. *Let $v : G \to U$ be a measurable map. We say that the equation*

$$X(t) = \Gamma\left(X(0) + \int_0^{\cdot} b(X(s), v(X(s)))ds + \int_0^{\cdot} \sigma(X(s))dW(s)\right)(t), \quad X(0) \sim \mu, \tag{2.3}$$

admits a weak solution if there exists a filtered probability space $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$ on which is given an $\{\mathcal{F}_t\}$ Wiener process $W(\cdot)$ and an $\mathcal{F}_t$ adapted process $X(\cdot)$ with continuous paths such that $X(0)$ has the probability law μ and for all t the equality in (2.3) *holds a.s. We say that* (2.3) *admits a unique weak solution if whenever there are two sets of such spaces and processes denoted as $(\Omega^i, \mathcal{F}^i, P^i, \mathcal{F}_t^i)$, $(W^i(\cdot), X^i(\cdot))$, $i = 1, 2$, then the probability law of $X^1(\cdot)$ is the same as that of $X^2(\cdot)$.*

With an abuse of terminology, the map v will be referred to as a "Markov control." Under the standing assumptions of this paper there is a unique weak solution for (2.3), and denoting the law of the solution process $X(\cdot)$, when $X(0) = x$ a.s., by P_x^v it can be shown that $\{P_x^v\}_{x\in G}$ is a strongly Feller Markov family (cf. Theorem 2.9 of [10]).

We will call a Markov control v a stable Markov control (SMC) if the corresponding Markov family $\{P_x^v\}_{x\in G}$ is positive recurrent and has a unique invariant measure. We will now present the blanket stability condition, introduced in [10] under which all Markov controls are stable.

Define

$$\mathcal{C} \doteq \left\{ -\sum_{i=1}^N \alpha_i d_i : \alpha_i \ge 0;\ i \in \{1, \dots, N\} \right\}. \tag{2.4}$$

The cone $\mathcal{C}$ was used to characterize stability for a certain class of constrained diffusion processes in [11, 3].

Let $\delta \in (0,\infty)$ be fixed. Define the set

$$\mathcal{C}(\delta) \doteq \{v \in \mathcal{C} : \operatorname{dist}(v, \partial C) \ge \delta\}. \tag{2.5}$$

The blanket stability condition below, which will be assumed throughout this paper, stipulates the permissible drifts in the underlying diffusion.

Condition 2.8. There exists a $\delta \in (0,\infty)$ such that for all $(x,u) \in G \times U$, $b(x,u) \in \mathcal{C}(\delta)$.

Under the assumptions made above the results of [3] show that all Markov controls are SMC, namely, that the following theorem holds.

THEOREM 2.9. *The Markov family $\{P_x^v\}_{x\in G}$ defined above is positive recurrent and admits a unique invariant measure, denoted as η_v.*

In this work we are interested in a control problem with an ergodic cost criterion. Namely, we are interested in minimizing, over the class of all admissible controls, the cost

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t k(X(s),u(s))ds, \tag{2.6}$$

where $X(\cdot)$ is given as a solution of (2.2) on some filtered probability space with an admissible pair $(u(\cdot),W(\cdot))$, the limit above is taken a.s. on the corresponding probability space, and $k : G\times U \to \mathbb{R}$ is a map defined as follows. For $(x,u)\in G\times U$, $k(x,u) \doteq \int_S \overline{k}(x,\alpha)u(d\alpha)$, where $\overline{k}$ is in $C_b(G\times S)$.

Under the assumptions made above, the following result on the existence of an optimal Markov control was proved in [10].

THEOREM 2.10. *There exists a Markov control $\overline{v}(\cdot)$ such that if for some $\mu \in \mathcal{P}(G)$, $\overline{X}(\cdot)$ is the corresponding process solving* (2.3) *(with v there replaced by $\overline{v}$), on some filtered probability space, with the probability law of $\overline{X}(0)$ being μ, then*

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T k(\overline{X}(s),\overline{v}(\overline{X}(s)))ds = \inf \operatorname{ess\,inf} \limsup_{T\to\infty} \frac{1}{T}\int_0^T k(X(s),u(s))ds \tag{2.7}$$

a.s., where the outside infimum on the right-hand side above is taken over all controlled processes $X(\cdot)$ with an arbitrary initial distribution and solving (1.1) *over some filtered probability space with some admissible pair $(W(\cdot),u(\cdot))$.*

3. The discounted cost problem. One of the important goals in optimal control theory is to derive the HJB equation for the value function and characterize the value function as the unique solution (in an appropriate class) of the PDE.

The classical approach to the HJB equation for the ergodic control is by the "vanishing discount method" (cf. [14], [9], [7], [4]). In this approach the first step is to study the value function $V_\alpha(x)$ of the discounted control problem defined via (1.3). In this section we will characterize the value function $V_\alpha(x)$ via a suitable HJB equation.

With an abuse of notation, for $\alpha \in S$ we will write $(Lf)(x,\delta_{\{\alpha\}})$ merely as $(Lf)(x,\alpha)$, where $\delta_{\{\alpha\}}$ denotes the probability measure concentrated at the point α. Thus with this notation, for $(x,u)\in G\times U$, $(Lf)(x,u) = \int_S (Lf)(x,\alpha)u(d\alpha)$. For $i=1,2,\dots,N$ and $f\in C_b^2(G)$ let $D_if : G\to\mathbb{R}$ be defined as $(D_if)(x) \doteq \langle d_i, \nabla f(x)\rangle$, $x\in G$. The natural HJB equation associated with the control problem (1.3) is the one given in (1.5). A theory of classical solutions for such a PDE is not available, and therefore we will consider solutions in the viscosity sense [15], [25], [17]. We begin with the following proposition. The proof is identical to Theorem III.2.1 of [7] and, therefore, is omitted.

PROPOSITION 3.1. *Let $\alpha\in(0,\infty)$ and let V_α be defined by* (1.3). *Then $V_\alpha \in C_b(G)$.*

Now let $(X^x(\cdot),Y^x(\cdot))$ be given as a solution of (2.2) with $X(0)\equiv x$. Let η be stopping time with respect to the natural filtration of $(X^x(\cdot),Y^x(\cdot))$. Then one can prove the following dynamic programming principle, exactly along the lines of Theorem III.1.3 of [7] (cf. comments above equation (III.1.8) of [7]).

$$V_\alpha(x) = \inf \mathbb{E}\left(\int_0^\eta e^{-\alpha t}k(X^x(t),u(t))dt + e^{-\alpha\eta}V_\alpha(X^x(\eta))\right), \tag{3.1}$$

where the infimum is taken over all feedback controls u. The following lemma will be useful in controlling the reflection term $(Y(\cdot))$ in (2.2).

LEMMA 3.2. *Let $(X^x(\cdot),Y^x(\cdot))$ be given as a solution of* (2.2) *with $X(0)\equiv x$. Then for every $m\in\mathbb{N}$, there exists a $C_m\in(0,\infty)$, which is independent of the initial condition x, the control process u, and $t\in(0,\infty)$ such that*

$$\sup_{i\in\{1,2,\dots,N\}} \mathbb{E}(Y_i(t))^m \le C_m t^{\frac{m}{2}}.$$

Proof. From Lemma 4.2 of [10], there exists a $g \in C_b^2(G)$ such that

$$\langle \nabla g(x), d_i \rangle \geq 1, \quad \forall x \in F_i,\ i \in \{1, \ldots, N\}. \tag{3.2}$$

An application of Itô's formula yields that for all $i \in \{1, 2, \ldots, N\}$

$$Y_i(t) \leq \left| \int_0^t (Lg)(X(s), v(X(s)))ds \right| + \left| \int_0^t \langle \nabla g(X(s)), \sigma(X(s))dW(s) \rangle \right| + |g(X_t) - g(X_0)|.$$

The result now follows on recalling that g is in $C_b^2(G)$ and using the boundedness of b, σ, Burkholder–Gundy inequalities, the Lipschitz property of the SM, and Gronwall's inequality. □

Next we say what we mean by the viscosity solution of (1.5). Denote by S^k the space of $k \times k$ real symmetric matrices. Let $F_\alpha : G \times \mathbb{R} \times \mathbb{R}^k \times S^k \to \mathbb{R}$ be defined as

$$F_\alpha(x, r, p, M) \doteq -\frac{1}{2} \operatorname{Tr}(aM) + \alpha r - \inf_{u \in U} \{\langle b(x, u), p \rangle + k(x, u)\}. \tag{3.3}$$

Also, define $F_{\alpha,*}$ and F_α^* as maps from $G \times \mathbb{R} \times \mathbb{R}^k \times S^k$ to $\mathbb{R}$, as follows:

$$F_{\alpha,*}(x, r, p, M) = \begin{cases} F_\alpha(x, r, p, M) & \text{if } x \in G^o, \\ F_\alpha(x, r, p, M) \wedge \min\{-\langle d_i, p \rangle;\ i \in \mathrm{In}(x)\} & \text{if } x \in \partial G, \end{cases} \tag{3.4}$$

$$F_\alpha^*(x, r, p, M) = \begin{cases} F_\alpha(x, r, p, M) & \text{if } x \in G^o, \\ F_\alpha(x, r, p, M) \vee \max\{-\langle d_i, p \rangle;\ i \in \mathrm{In}(x)\} & \text{if } x \in \partial G. \end{cases} \tag{3.5}$$

DEFINITION 3.3. *We say that $\phi \in C_b(G)$ is a viscosity solution of* (1.5) *if and only if for all $x_0 \in G$ the following hold.*

1. *For all $\psi \in C^2(G)$, such that x_0 is a strict maximum point of $\phi - \psi$,*

$$F_{\alpha,*}(x_0, \phi(x_0), (D\psi)(x_0), (D^2\psi)(x_0)) \leq 0. \tag{3.6}$$

2. *For all $\psi \in C^2(G)$, such that x_0 is a strict minimum point of $\phi - \psi$,*

$$F_\alpha^*(x_0, \phi(x_0), (D\psi)(x_0), (D^2\psi)(x_0)) \geq 0. \tag{3.7}$$

We now have the following result.

THEOREM 3.4. *The value function V_α defined via* (1.3) *is a viscosity solution of* (1.5).

Proof. Let $x_0 \in G$ and $\psi \in C^2(G)$ be such that x_0 is a strict maximum point of $V_\alpha - \psi$. We will show that (3.6) holds with ϕ there replaced by V_α. Let $\kappa \doteq \psi(x_0) - V_\alpha(x_0)$ and fix $u \in U$. Then

$$\begin{aligned} \psi(x_0) &= V_\alpha(x_0) + \kappa \\ &\leq \mathbb{E}\left(\int_0^h e^{-\alpha s} k(X^{x_0}(s), u)ds + e^{-\alpha h} V_\alpha(X^{x_0}(h)) \right) + \kappa \\ &\leq \mathbb{E}\left(\int_0^h e^{-\alpha s} k(X^{x_0}(s), u)ds + (e^{-\alpha h} - 1) V_\alpha(X^{x_0}(h)) + \psi(X^{x_0}(h)) \right), \end{aligned}$$

where the first inequality follows from the dynamic programming principle (3.1) and the second inequality uses the minimality of κ. The above inequality yields

$$\begin{aligned} 0 \leq \frac{1}{h} \mathbb{E}\left(\int_0^h e^{-\alpha s} k(X^{x_0}(s), u)ds \right) + \frac{1}{h} \mathbb{E}(\psi(X^{x_0}(h)) - \psi(x_0)) \\ + \frac{(e^{-\alpha h} - 1)}{h} \mathbb{E}(V_\alpha(X^{x_0}(h))). \end{aligned}$$

Taking the limit as $h \to 0$ in the above inequality and using Proposition 3.1 and Itô's formula we have that

$$\begin{aligned} 0 &\le k(x_0,u) + (L\psi)(x_0,u) - \alpha V_\alpha(x_0) \\ &\quad + \limsup_{h\to 0} \sum_{i=1}^N \frac{1}{h}\mathbb{E}\bigg(\int_0^h \langle d_i, D\psi(X^{x_0}(s))\rangle dY_i(s)\bigg) \\ &= -F_\alpha(x_0, V_\alpha(x_0), D\psi(x_0), D^2\psi(x_0)) \\ &\quad + \limsup_{h\to 0} \sum_{i=1}^N \frac{1}{h}\mathbb{E}\bigg(\int_0^h \langle d_i, D\psi(X^{x_0}(s))\rangle dY_i(s)\bigg). \end{aligned} \tag{3.8}$$

We will now show that for every $i \in \{1,\dots,N\}$

$$\langle d_i, D\psi(x_0)\rangle < 0 \;\Rightarrow\; \limsup_{h\to 0} \frac{1}{h}\mathbb{E}\bigg(\int_0^h \langle d_i, D\psi(X^{x_0}(s))\rangle dY_i(s)\bigg) \le 0. \tag{3.9}$$

Clearly (3.8) and (3.9) will prove that (3.6) holds (with ϕ replaced by V_α). Now suppose that $i \in \{1,\dots,N\}$ is such that $\langle d_i, D\psi(x_0)\rangle < 0$. Then there exists $\epsilon > 0$ such that

$$\langle d_i, D\psi(y)\rangle < 0 \quad \forall y \in G \text{ satisfying } |x_0 - y| \le \epsilon. \tag{3.10}$$

Now

$$\begin{aligned} &\limsup_{h\to 0} \frac{1}{h}\mathbb{E}\bigg(\int_0^h \langle d_i, D\psi(X^{x_0}(s))\rangle dY_i(s)\bigg) \\ &\quad \le |D\psi|_\infty \limsup_{h\to 0} \frac{1}{h}\mathbb{E}(1_{\{\sup_{0\le s\le h}|X^{x_0}(s)-x_0|\ge\epsilon\}} Y_i(h)) \\ &\quad \le |D\psi|_\infty \limsup_{h\to 0} \sqrt{\mathbb{E}(Y_i^2(h))}\, \frac{\sqrt{\mathbb{E}(\sup_{0\le s\le h}|X^{x_0}(s)-x_0|^5)}}{h\epsilon^{5/2}} \\ &\quad \le C|D\psi|_\infty \sqrt{\mathbb{E}(Y_i^2(1))} \limsup_{h\to 0} \frac{h^{5/4}}{h\epsilon^{5/2}} \\ &\quad = 0 \end{aligned}$$

for a suitable constant C, where the second inequality follows on applying the Cauchy–Schwarz inequality and Chebyshev's inequality; the final inequality follows on using the Lipschitz property of the SM, boundedness of drift and diffusion coefficients, and Burkholder–Gundy inequalities. This proves (3.9) and hence part 1 of Definition 3.3.

Next let $\psi \in C^2(G)$ be such that x_0 is a strict minimum point of $V_\alpha - \psi$. To complete the proof we need to show that (3.7) holds with ϕ there replaced by V_α. From (3.1) we have that

$$V_\alpha(x_0) = \inf \mathbb{E}\left(\int_0^\epsilon e^{-\alpha t} k(X^{x_0}(t), u(t))dt + e^{-\alpha\epsilon} V(X^{x_0}(\epsilon))\right).$$

Let $u^\epsilon(\cdot)$ be a feedback control such that

$$V_\alpha(x_0) + \epsilon^2 \ge \mathbb{E}\left(\int_0^\epsilon e^{-\alpha t} k(X^{x_0,\epsilon}(t), u^\epsilon(t))dt + e^{-\alpha\epsilon} V(X^{x_0,\epsilon}(\epsilon))\right),$$

where $X^{x_0,\epsilon}(\cdot)$ solves (1.1) with $u(\cdot)$ there replaced by $u^\epsilon(\cdot)$ and $X(0) \equiv x$. Let κ be as before. Then

$$\psi(x_0) = V_\alpha(x_0) + \kappa \geq \mathbb{E}\left(\int_0^\epsilon e^{-\alpha t} k(X^{x_0,\epsilon}(t), u^\epsilon(t))dt \right.$$
$$\left. + (e^{-\alpha\epsilon} - 1)V(X^{x_0,\epsilon}(\epsilon)) + \psi(X^{x_0,\epsilon}(\epsilon))\right) - \epsilon^2.$$

Dividing by ϵ and taking the limit as $\epsilon \to 0$, we have

$$0 \geq \liminf_{\epsilon\to 0} \frac{1}{\epsilon}\mathbb{E}\left(\int_0^\epsilon e^{-\alpha t} k(X^{x_0,\epsilon}(t), u^\epsilon(t))dt \right.$$
$$\left. + (e^{-\alpha\epsilon} - 1)V(X^{x_0,\epsilon}(\epsilon)) + \psi(X^{x_0,\epsilon}(\epsilon)) - \psi(x_0)\right)$$
$$= \liminf_{\epsilon\to 0} \frac{1}{\epsilon}\mathbb{E}\left(\int_0^\epsilon e^{-\alpha t} k(x_0, u^\epsilon(t))dt \right.$$
$$\left. + \int_0^\epsilon (L\psi)(x_0, u^\epsilon(t))dt + \sum_{i=1}^N \int_0^\epsilon (D_i\psi)(X^{x_0,\epsilon}(t))dY_i(t)\right) - \alpha V_\alpha(x_0).$$

The last step follows on using the continuity and/or Lipschitz properties of k, $L\psi$, σ, and b and observing that

$$\limsup_{\epsilon\to 0} \mathbb{E}\left(\sup_{0\leq s\leq\epsilon} |X^{x_0,\epsilon}(s) - x_0|\right)^p \leq \limsup_{\epsilon\to 0} \sup \mathbb{E}\left(\sup_{0\leq s\leq\epsilon} |X^{x_0}(s) - x_0|\right)^p = 0,$$

where the supremum on the right-hand side is taken over all admissible controls. Thus we have that

$$0 \geq \liminf_{\epsilon\to 0} \frac{1}{\epsilon}\mathbb{E}\left(\int_0^\epsilon k(x_0, u^\epsilon(t))dt + \int_0^\epsilon (L\psi)(x_0, u^\epsilon(t))dt \right.$$
$$\left. + \sum_{i=1}^N \int_0^\epsilon (D_i\psi)(X^{x_0,\epsilon}(t))dY_i(t)\right) - \alpha V_\alpha(x_0)$$
$$\geq \inf_u (k(x_0, u) + (L\psi)(x_0, u)) - \alpha V_\alpha(x_0) + \liminf_{\epsilon\to 0} \sum_{i=1}^N \frac{1}{\epsilon}\mathbb{E}\left(\int_0^\epsilon (D_i\psi)(X^{x_0,\epsilon}(t))dY_i(t)\right)$$
$$= -F_\alpha(x_0, V_\alpha(x_0), D\psi(x_0), D^2\psi(x_0)) + \liminf_{\epsilon\to 0} \sum_{i=1}^N \frac{1}{\epsilon}\mathbb{E}\left(\int_0^\epsilon (D_i\psi)(X^{x_0,\epsilon}(t))dY_i(t)\right).$$

From the above inequality one can prove part 2 of Definition 3.3 exactly in the way part 1 was proved from (3.8). This proves the theorem. □

Next, we will show that under the standing assumptions of this paper, there is a unique viscosity solution of (1.5). For $n \in \mathbb{N}$, let $B_n \doteq \{x \in G \mid |x| < n\}$. Let $\psi \in C_b(G)$ be given. We begin by considering the following equation:

$$\begin{aligned} \inf_{u\in U} (L\phi(x, u) + k(x, u) - \alpha\phi(x)) &= 0, && x \in G \cap B_n, \\ \langle \nabla\phi(x), d_i\rangle &= 0, && x \in \partial G \cap B_n,\ i \in \mathrm{In}(x), \\ \phi(x) &= \psi(x), && x \in \partial B_n. \end{aligned} \tag{3.11}$$

DEFINITION 3.5. *We say that $\phi \in C_b(G)$ is a viscosity solution of* (3.11) *if* 1. *and* 2. *in Definition* 3.3 *hold for all $x_0 \in G \cap B_n$ and $\phi(x) = \psi(x)$ for all $x \in \partial B_n$.*

For $x \in \overline{B}_n$, let $X^x(\cdot)$ be given as a solution of (1.1) with $X(0) \equiv x$ and some admissible pair $(u(\cdot), W(\cdot))$. Let

$$\tau_n \equiv \tau_n(x) \doteq \inf\{t : X^x(t) \in B_n^c\} \tag{3.12}$$

and define

$$V^n(x) \doteq \inf \mathbb{E}\left(\int_0^{\tau_n} e^{-\alpha s} k(X^x(s), u(s))ds + e^{-\alpha\tau_n}\psi(X^x(\tau_n))\right), \tag{3.13}$$

where the infimum above is taken over all admissible controls.

The existence part of the following result is proved exactly as for Theorem 3.4. The proof of uniqueness, essentially, follows from results in [17]. A sketch of the argument is provided in the appendix for the reader's convenience.

THEOREM 3.6. *Let $\alpha \geq 0$ and let $V^n(\cdot)$ be defined via* (3.13). *Then $V^n(\cdot)$ is the unique viscosity solution of* (3.11).

An immediate consequence of the above theorem is the following result.

THEOREM 3.7. *Let $\alpha \in (0, \infty)$. Then $V_\alpha(\cdot)$ defined via* (1.3) *is the unique viscosity solution of* (1.5).

Proof. From Theorem 3.4 we know that $V_\alpha(\cdot)$ is a viscosity solution of (1.5). Now let $\tilde{V}$ be another viscosity solution of (1.5). Let $\tau_n(x)$ be defined via (3.12). Define

$$\phi(x) \doteq \inf \mathbb{E}\left(\int_0^{\tau_n(x)} e^{-\alpha s} k(X^x(s), u(s))ds + e^{-\alpha\tau_n(x)}\tilde{V}(X^x(\tau_n(x)))\right),$$

where the infimum is taken over all admissible controls. From Theorem 3.6, ϕ is the unique viscosity solution of (3.11), with ψ there replaced by $\tilde{V}$. However, since $\tilde{V}$ solves (1.5), clearly it is also a solution of (3.11) (once more with ψ there replaced by $\tilde{V}$). Thus we have that $\phi = \tilde{V}$ and so

$$\tilde{V}(x) = \inf \mathbb{E}\left(\int_0^{\tau_n(x)} e^{-\alpha t} k(X^x(t), u(t))dt + e^{-\alpha\tau_n(x)}\tilde{V}(X^x(\tau_n(x)))\right).$$

Also from (3.1) we have that the above equality holds with $\tilde{V}$ replaced by V_α. Thus we have that for $x \in G$ and n large enough so that $x \in B_n$,

$$\begin{aligned}|\tilde{V}(x) - V_\alpha(x)| &\leq \sup|\mathbb{E}(e^{-\alpha\tau_n(x)}\tilde{V}(X^x(\tau_n(x))) - e^{-\alpha\tau_n(x)}V_\alpha(X^x(\tau_n(x))))| \\ &\leq (|\tilde{V}|_\infty + |V_\alpha|_\infty)\sup\left(\mathbb{E}\big(e^{-\alpha\tau_n(x)}\big)\right),\end{aligned}$$

where the supremum in the above display is taken over all admissible controls. Using the boundedness of the drift and diffusion coefficients and the Lipschitz property of the Skorohod map, it follows that $\sup(\mathbb{E}(e^{-\alpha\tau_n(x)})) \to 0$ as $n \to \infty$. This shows that $\tilde{V}(x) = V_\alpha(x)$ for all $x \in G$. □

4. The vanishing discount limit. In this section we will show that if V_α is given via (1.3) and $\overline{V}_\alpha(x) \doteq V_\alpha(x) - V_\alpha(0)$, $x \in G$, then the family $\{\overline{V}_\alpha;\ \alpha \in (0, \infty)\}$ is precompact in $C(G)$. We begin with the following result, which says that the infimum of the cost in (1.3) over all admissible controls coincides with the infimum taken over all Markov controls. For the proof of this result for unconstrained diffusions, see Theorem II.4.2 of [7]. The result in the setting of constrained diffusions, of the form considered

in this work, is proved in a similar manner and therefore the proof is omitted. We refer the reader to [10] where the equivalence of infimum over all admissible controls and infimum over all Markov controls, under an ergodic cost criterion, is proved for constrained diffusions of the form considered in this work.

THEOREM 4.1. *For $\alpha \in (0,\infty)$ let $V_\alpha(\cdot)$ be defined via* (1.3). *Then for all $x \in G$*

$$V_\alpha(x) = \inf_v \mathbb{E}\left(\int_0^\infty e^{-\alpha s} k(X^x(s), v(X^x(s)))ds\right), \tag{4.1}$$

where $X^x(\cdot)$ is given as the unique weak solution of (2.3) *with $X(0) \equiv x$ and the infimum above is taken over all Markov controls v.*

We will also need the following "finite time horizon" equicontinuity result. For $M_0 \in (0,\infty)$ let

$$E_{M_0} \doteq \{x \in G : |x| \le M_0\}.$$

THEOREM 4.2. *Let $M, M_0, \epsilon \in (0,\infty)$. Define*

$$\begin{aligned}\Lambda(\epsilon, M, M_0) \doteq \sup_{x_1,x_2 \in E_{M_0},\, |x_1-x_2|\le\epsilon} \sup_v \int_0^M |&\mathbb{E}k(X^{x_1}(s), v(X^{x_1}(s))) \\ &-\mathbb{E}k(X^{x_2}(s), v(X^{x_2}(s)))|ds,\end{aligned} \tag{4.2}$$

where for $i = 1, 2$, $X^{x_i}(\cdot)$ is given as the unique weak solution of (2.3) *with $X(0) \equiv x_i$ and the inside supremum is taken over all Markov controls v. Then, for all $M, M_0 \in (0,\infty)$, $\Lambda(\epsilon, M, M_0) \to 0$ as $\epsilon \to 0$.*

The proof of the above theorem follows from results in [6]. We provide a sketch in the appendix.

Observe that, from Theorem 4.1, for all $x_1, x_2 \in G$ and $\alpha \in (0,\infty)$

$$\begin{aligned}|V_\alpha(x_1) - V_\alpha(x_2)| \le \sup_v \Big|&\mathbb{E}\int_0^\infty e^{-\alpha s} k(X^{x_1}(s), v(X^{x_1}(s)))ds \\ &- \mathbb{E}\int_0^\infty e^{-\alpha s} k(X^{x_2}(s), v(X^{x_2}(s)))ds\Big|,\end{aligned}$$

where for $i = 1, 2$, $X^{x_i}(\cdot)$ is given as the unique weak solution of (2.3) with $X(0) \equiv x_i$ and the inside supremum is taken over all Markov controls v. Now fix $\epsilon, M_0 \in (0,\infty)$ and let $M \in \mathbb{N}_0$. Suppose that $x_1, x_2 \in E_{M_0}$ and $|x_1 - x_2| \le \epsilon$. Then the expression on the right-hand side above can be bounded by

$$\begin{aligned}\sup_{x_1,x_2\in E_{M_0}} \sup_v \Big|&\mathbb{E}\int_M^\infty e^{-\alpha s} k(X^{x_1}(s), v(X^{x_1}(s)))ds \\ &- \mathbb{E}\int_M^\infty e^{-\alpha s} k(X^{x_2}(s), v(X^{x_2}(s)))ds\Big| + \Lambda(\epsilon, M, M_0).\end{aligned} \tag{4.3}$$

The main step in the proof of precompactness of $\{\overline{V}_\alpha;\ \alpha \in (0,\infty)\}$ is the following.

PROPOSITION 4.3. *For $M_0 \in (0,\infty)$ and $M \in \mathbb{N}$, let*

$$\begin{aligned}\Psi(M, M_0) \doteq \sup_{\alpha\in(0,\infty)} \sup_{x_1,x_2\in E_{M_0}} \sup_v \Big|&\mathbb{E}\int_M^\infty e^{-\alpha s} k(X^{x_1}(s), v(X^{x_1}(s)))ds \\ &- \mathbb{E}\int_M^\infty e^{-\alpha s} k(X^{x_2}(s), v(X^{x_2}(s)))ds\Big|.\end{aligned}$$

Then, for all $M_0 \in (0,\infty)$, $\Psi(M, M_0) \to 0$ as $M \to \infty$.

In order to prove the above proposition, we will need some stability properties of the underlying constrained diffusion process. Following [5], we begin by an embedding of the continuous time control problem in a discrete time control problem. Define

$$\mathcal{U} \doteq \{\theta : [0,1] \to U : \theta \text{ is a measurable map}\}.$$

We endow $\mathcal{U}$ with the coarsest topology under which, for every $e \in L^2[0,1]$ and $f \in C_b(S)$, the map $\Psi : \mathcal{U} \to \mathbb{R}$ defined as $\Psi(u) \doteq \int_0^1 e(t) \int_S f(\theta) u_t(d\theta)dt$ is continuous. Let $\hat{\Phi} \subset \mathcal{P}(C([0,1] : \mathbb{R}^k) \times \mathcal{U})$ be the class of all probability measures which correspond to the probability law of some admissible pair $(u(t), W(t))_{0 \le t \le 1}$. It follows from Chapter 1 of [7] that $\hat{\Phi}$ is a compact metric space.

Let $\phi \in \hat{\Phi}$ and let $(u(t), W(t))_{0 \le t \le 1}$ be the corresponding admissible pair on a filtered probability space $(\Omega, \mathcal{F}, P, (\mathcal{F}_t))$. Define

$$\hat{k}_\alpha(x, \phi) \doteq \mathbb{E}\left(\int_0^1 e^{-\alpha s} k(X^x(s), u(s))ds\right), \tag{4.4}$$

where $X^x(\cdot)$ is given as a solution of (1.1) with $X(0) \equiv x$. Let v be a Markov control. Define $\varrho_v : G \to \hat{\Phi}$ as follows. For $x \in G$, $\varrho_v(x)$ is defined as the probability law of $\{v(X^x(t)), W(t)\}_{0 \le t \le 1}$, where $W(\cdot)$ is a Wiener process and $X^x(\cdot)$ is given as the unique weak solution of (2.3) with $X(0) \equiv x$.

Setting $\hat{\alpha} \doteq e^{-\alpha}$ we have that for $M \in \mathbb{N}$

$$\begin{aligned} &\mathbb{E}\left(\int_M^\infty e^{-\alpha t} k(X^x(t), v(X^x(t)))dt\right) \\ &\quad = \sum_{n=M}^\infty \mathbb{E}\left(\mathbb{E}\left(\int_n^{n+1} e^{-\alpha t} k(X^x(t), v(X^x(t)))dt \mid \mathcal{F}_n^X\right)\right) \\ &\quad = \sum_{n=M}^\infty \hat{\alpha}^n \mathbb{E}(\hat{k}_\alpha(X_n^x, \varrho_v(X_n^x))), \end{aligned} \tag{4.5}$$

where $\mathcal{F}_n^X \doteq \sigma\{X(s) : 0 \le s \le n\}$ and $X_n^x \doteq X^x(n)$.

Note that $\{X_n^x\}$ is a controlled Markov chain with control set $\hat{\Phi}$, Markov control ϱ_v, and (controlled) transition probability kernel $\hat{p}(x_1, \phi, dy_1)$ given as follows:

$$\int_G f(y)\hat{p}(x, \phi, dy) \doteq \mathbb{E}(f(\xi)), \quad x \in G,\ f \in BM(G),\ \phi \in \hat{\Phi}, \tag{4.6}$$

where $\xi \doteq X(1)$ and $X(\cdot)$ is given via (1.1) with $X(0) \equiv x$ and the control pair $(u(t), W(t))_{0 \le t \le 1}$ having the probability law ϕ.

We now introduce a Lyapunov function for the controlled Markov chain $\{X_n^x\}$. This Lyapunov function was constructed in [2].

THEOREM 4.4 (see [2]). *There exists a function $F : G \to \mathbb{R}$ such that it is twice continuously differentiable on $G \setminus \{0\}$ and such that the following hold.*

(a) *There exist $c_1, c_2 \in (0, \infty)$ such that $c_1|x| \le F(x) \le c_2|x|$ for all $x \in G$.*

(b) *For all $\epsilon > 0$ there exists $M \in (0, \infty)$ such that $(x \in G,\ |x| \ge M)$ implies $\|D^2F(x)\| \le \epsilon$.*

(c) *There exists $c \in (0, \infty)$ such that $Df(x) \cdot r \le -c$, for $r \in \mathcal{C}(\delta)$ and $x \in G \setminus \{0\}$, and $Df(x) \cdot d \le -c$, for $d \in d(x)$ and $x \in \partial G \setminus \{0\}$.*

(d) *There exists* $L \in (0,\infty)$ *such that* $\sup_{x\in G}|Df(x)| \le L$.

The following theorem is an immediate consequence of the above theorem. For the sake of completeness we include a sketch in the appendix.

THEOREM 4.5. *There exist* $c_0, \ell_0, M_0 \in (0,\infty)$ *such that for any admissible pair* $(u(\cdot), W(\cdot))$ *on some filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$, $x \in G$, *and* $X^x(\cdot)$ *given by* (1.1), *we have that*

$$\mathbb{E}(F(X^x_{n+1}) \mid \mathcal{F}_n) - F(X^x_n) \le -c_0 1_{X^x_n \in B^c} + M_0 1_{X^x_n \in B}, \tag{4.7}$$

where $B \doteq \{x \in G \mid |x| \le \ell_0\}$ *and* $X^x_n \doteq X^x(n)$.

We now introduce a controlled probability transition kernel on $H \doteq G \times G$, with control set $\hat{\Phi} \times \hat{\Phi}$ defined as follows. For $\overline{x} \equiv (x_1, x_2) \in H$ and $\overline{\phi} \equiv (\phi_1, \phi_2) \in \hat{\Phi} \times \hat{\Phi}$, let $\overline{p}(\overline{x}, \overline{\phi}, d\overline{y}) \in \mathcal{P}(H)$ be defined as

$$\overline{p}(\overline{x}, \overline{\phi}, d\overline{y}) \doteq \hat{p}(x_1, \phi_1, dy_1)\hat{p}(x_2, \phi_2, dy_2). \tag{4.8}$$

Let v be a Markov control and for $x_1, x_2 \in G$, let $X^{x_i}(\cdot)$ be given via (2.3) with $X^{x_i}(0) \equiv x_i$ and driving Wiener process $W^{(i)}$, $i = 1, 2$. The Wiener processes $W^{(1)}$ and $W^{(2)}$ are taken to be independent of each other. Denote $X^{x_i}(n)$ by $X^{x_i}_n$. It is easy to see that if $\overline{x} \doteq (x_1, x_2)$, then $\{\overline{X}^{\overline{x}}_n\} \doteq \{(X^{x_1}_n, X^{x_2}_n)\}$ is an H valued controlled Markov chain, starting at $\overline{x}$, with the controlled probability transition kernel $\overline{p}(\overline{x}, \phi, d\overline{y})$ and the Markov control $\overline{\varrho}_v(\overline{x}) \equiv (\varrho_v(x_1), \varrho_v(x_2))$. Also, as noted earlier, for $i = 1, 2$, $X^{x_i}_n$ is a G valued controlled Markov chain, starting at x_i, with the controlled probability transition kernel $\hat{p}(x, \phi, dy)$ and the Markov control ϱ_v.

The pseudoatom construction. We will now proceed, as in [5], to adapt the Athreya–Ney–Nummelin construction of a pseudoatom to the current problem. Let $H \doteq G \times G$ and B be as in the statement of Theorem 4.5. Define $B^* \doteq B \times B$ and let $H^* \doteq H \times \{0, 1\} = G \times G \times \{0, 1\}$.

Let λ denote the Lebesgue measure on G. Define $\nu \in \mathcal{P}(H)$ as

$$\nu(A) \doteq \frac{(\lambda \times \lambda)(A \cap B^*)}{(\lambda(B))^2}. \tag{4.9}$$

Using the uniform nondegeneracy of the diffusion coefficient in (1.1), it follows that there exists $0 < \delta^* < 1$ such that

$$\overline{p}(x, \phi, A) \ge \delta^* 1_{B^*}(x)\nu(A), \quad \forall x \in H,\ A \in \mathcal{B}(H),\ \phi \in \hat{\Phi} \times \hat{\Phi}. \tag{4.10}$$

For a set $A \in \mathcal{B}(H)$ we let $A_0 \doteq A \times \{0\}$ and $A_1 \doteq A \times \{1\}$. For every $\mu \in \mathcal{P}(G \times G)$ we define a $\mu^* \in \mathcal{P}(H^*)$ as follows. For $A \in \mathcal{B}(H)$

$$\begin{aligned} \mu^*(A_0) &\doteq (1 - \delta^*)\mu(AB^*) + \mu(A(B^*)^c), \\ \mu^*(A_1) &\doteq \delta^* \mu(AB^*). \end{aligned} \tag{4.11}$$

Clearly, $\mu^*(A_0) + \mu^*(A_1) = \mu(A)$ and if $A \subset (B^*)^c$, then $\mu^*(A_0) = \mu(A)$.

Define $\varrho^*_v : H^* \to \hat{\Phi} \times \hat{\Phi}$ as follows. For $(x_1, x_2, i) \in H^*$, $\varrho^*_v(x_1, x_2, i) \doteq \overline{\varrho}_v(x_1, x_2)$. On a suitable probability space $(\Omega^*, \mathcal{F}^*, P^*)$, define an H^* valued controlled Markov chain $Z_n \equiv (X^*_n, i^*_n)$, where $X^*_n \equiv (X^{1,*}_n, X^{2,*}_n)$, with the control set $\hat{\Phi} \times \hat{\Phi}$ and the Markov control ϱ^*_v such that:

(1) The controlled transition kernel of Z_n is given as follows. For $\overline{z} \equiv (z,i) \in H^*$ and $\overline{\phi} \in \hat{\Phi} \times \hat{\Phi}$

$$q(\overline{z},\overline{\phi},d\overline{y}) = \begin{cases} \overline{p}^*(z,\overline{\phi},dy) & \text{if } \overline{z} \in H_0 \setminus B_0^*, \\ \dfrac{1}{1-\delta^*}(\overline{p}^*(z,\overline{\phi},dy) - \delta^*\nu^*(d\overline{y})) & \text{if } \overline{z} \in B_0^*, \\ \nu^*(d\overline{y}) & \text{if } \overline{z} \in H_1, \end{cases} \tag{4.12}$$

where $\overline{y} \equiv (y,j) \in H^*$.

(2) The initial distributions are given as follows. For $A \in \mathcal{B}(H)$

$$P^*(Z_0 \in A_0) \doteq ((1-\delta^*)1_{AB^*}(\overline{x}) + 1_{A(B^*)^c}(\overline{x})),$$
$$P^*(Z_0 \in A_1) \doteq \delta^* 1_{AB^*}(\overline{x}).$$

The above construction assures that the probability laws of $\{X_n^*, \varrho_v^*(X_n^*)\}_{n\in\mathbb{N}_0}$ and $\{\overline{X_n^{\overline{x}}}, \overline{\varrho}_v(\overline{X_n^{\overline{x}}})\}_{n\in\mathbb{N}_0}$ are the same. Using the Lyapunov function in (4.4) one can now show, in a similar manner as in Lemma 3.3 of [5], that the hitting time of B_1^* by the controlled Markov chain $\{Z_n\}$ has finite moments of all orders. Stability properties of the underlying (controlled) Markov processes using the Foster-type drift criterion in (4.7) are quite well studied in the literature. For example, [19] obtained functional central limit theorems for Markov processes and gives conditions under which the invariant distribution is a continuous function of its transition kernel. In [30], using Foster–Lyapunov inequalities the authors obtained criteria for Harris recurrence, ergodicity, and geometric ergodicity for general Right processes in terms of the underlying extended generator. See also [31] for more on regularity properties of Markov processes that follow from such a drift criterion. The following theorem is another standard consequence of the drift inequality (4.7) and thus we omit the proof.

THEOREM 4.6. *Let*

$$\tau(x_1,x_2) \doteq \inf\{n \in \mathbb{N}_0 : Z_n \in B_1^*\}.$$

Then there exists an $r \in (1,\infty)$ *such that for every* $M_0 \in (0,\infty)$,

$$\sup_{x_i\in G;\, |x_i|\le M_0;\, i=1,2} \sup \mathbb{E}^*(r^{\tau(x_1,x_2)}) < \infty,$$

where the inner supremum is taken over all Markov controls v.

The main idea in the proof is to show that one can find a small enough $\delta_0 \in (0,\infty)$ such that $\mathcal{V}(y) \doteq e^{\delta_0 F(y)}$, $y \in G$ satisfies the geometric drift condition (V4) of section 15.2.2 of [29]. (See Theorem 16.3.1 of [29].) Then define $\overline{\mathcal{V}} : H^* \to \mathbb{R}_+$ as $\overline{\mathcal{V}}(x_1,x_2,i) \doteq \mathcal{V}(x_1) + \mathcal{V}(x_2)$, where $(x_1,x_2,i) \in H^*$. And, for $n \in \mathbb{N}_0$, setting $\Gamma_n^* \doteq \sigma(X_m^*, i_m^*;\ m \le n)$, one can check that

$$\mathbb{E}^*(\overline{\mathcal{V}}(Z_{n+1}) \mid \Gamma_n^*) - \overline{\mathcal{V}}(Z_n) \le -\beta\overline{\mathcal{V}}(Z_n) + 2b$$

for some $\beta \in (0,1)$ and $b \in (0,\infty)$. The result now follows from Theorem 15.2.5 of [29].

We now prove Proposition 4.3.

Proof of Proposition 4.3. Let v be a Markov control and let $x_1, x_2 \in E_{M_0}$. Using the discrete time embedding introduced above, one has that

$$\left|\mathbb{E}\int_M^\infty e^{-\alpha s}k(X^{x_1}(s), v(X^{x_1}(s)))ds - \mathbb{E}\int_M^\infty e^{-\alpha s}k(X^{x_2}(s), v(X^{x_2}(s)))ds\right|$$

can be rewritten as

$$\left|\sum_{n=M}^{\infty}\hat{\alpha}^n\mathbb{E}(\hat{k}_\alpha(X_n^{x_1},\varrho_v(X_n^{x_1}))-\hat{k}_\alpha(X_n^{x_2},\varrho_v(X_n^{x_2})))\right|.$$

From the pseudoatom construction we have that $\{(X_n^{x_1},X_n^{x_2})\}_{n\in\mathbb{N}_0}$ has the same probability law as $\{(X_n^{1,*},X_n^{2,*})\}_{n\in\mathbb{N}_0}$. Thus the expression in the above display is the same as

$$\left|\sum_{n=M}^{\infty}\hat{\alpha}^n\mathbb{E}^*(\hat{k}_\alpha(X_n^{1,*},\varrho_v(X_n^{1,*}))-\hat{k}_\alpha(X_n^{2,*},\varrho_v(X_n^{2,*})))\right|.$$

Let $\tau\equiv\tau(x_1,x_2)$ be as in Theorem 4.6. Then the above expression can be written as

$$\begin{aligned}&\left|\sum_{n=M}^{\infty}\hat{\alpha}^n\mathbb{E}^*1_{\tau<n}(\hat{k}_\alpha(X_n^{1,*},\varrho_v(X_n^{1,*}))-\hat{k}_\alpha(X_n^{2,*},\varrho_v(X_n^{2,*})))\right|\\&\quad+\left|\sum_{n=M}^{\infty}\hat{\alpha}^n\mathbb{E}^*1_{\tau\geq n}(\hat{k}_\alpha(X_n^{1,*},\varrho_v(X_n^{1,*}))-\hat{k}_\alpha(X_n^{2,*},\varrho_v(X_n^{2,*})))\right|\\&\equiv T_1+T_2.\end{aligned}$$

Observing that for $m\in\mathbb{N}$, $X_{\tau+m}^{1,*}$ and $X_{\tau+m}^{2,*}$ have the same conditional law given Γ_τ^*, we have that $T_1=0$. Next, using the boundedness of $\hat{k}_\alpha$ we have that $T_2\leq\kappa\mathbb{E}^*(\tau-M)^+$ for a suitable constant κ. Note that $\kappa\mathbb{E}^*(\tau-M)^+$ can be bounded above by $\kappa\sup\frac{\mathbb{E}^*(\tau^2)}{M}$, where the supremum is taken over all $x_1,x_2\in E_{M_0}$ and Markov controls v. Combining the above observations we have that $\Psi(M,M_0)$ is bounded above by $\kappa\sup\frac{\mathbb{E}^*(\tau^2)}{M}$. The result now follows from Theorem 4.6. □

As an immediate consequence of the above proposition we have the following result.

THEOREM 4.7. *The family* $\{\overline{V}_\alpha(\cdot):\ \alpha\in(0,1)\}$ *is precompact in* $C(G)$.

Proof. Fix $M_0\in(0,\infty)$ and let $\delta>0$ be arbitrary. From Proposition 4.3 we can find $M\in\mathbb{N}$ large enough so that $\Psi(M,M_0)<\frac{\delta}{2}$. Next, from Theorem 4.2 find ϵ small enough so that $\Lambda(\epsilon,M,M_0)<\frac{\delta}{2}$. Using these bounds in (4.3) we have that for all $x_1,x_2\in E_{M_0}$ with $|x_1-x_2|\leq\epsilon$ and $\alpha\in(0,\infty)$, $|V_\alpha(x_1)-V_\alpha(x_2)|\leq\delta$. Since δ and M_0 are arbitrary, the result follows. □

Remark 4.8. By imitating the proof of Lemma 3.4 of [5] one can show that there is a $\varsigma\in(0,\infty)$ such that for $x\in G$

$$\begin{aligned}|\overline{V}_\alpha(x)|&\leq\varsigma F(x)\\&\leq\varsigma c_2|x|,\end{aligned}$$

where the last inequality follows from Theorem 4.4. This shows that if $V(\cdot)$ is any limit point of $\overline{V}_\alpha(\cdot)$, then $|V(x)|\leq\varsigma c_2|x|$ for all $x\in G$.

5. The HJB equation for the ergodic control problem. Let V_α and $\overline{V}_\alpha$ be as in the previous section. From (1.3) and the boundedness of k it follows that $\sup_{\alpha\in(0,1)}\alpha V_\alpha(0)<\infty$. This, along with Theorem 4.7, implies that there exists a sequence $\alpha_n\to0$ as $n\to\infty$ and $\rho\in(0,\infty)$, $V\in C(G)$ such that

$$\lim_{n\to\infty}\alpha_nV_{\alpha_n}(0)=\rho,\quad\lim_{n\to\infty}\overline{V}_{\alpha_n}=V,\tag{5.1}$$

where the second limit is taken uniformly on compact sets of G.

For $\rho^* \in [0,\infty)$, consider the following equation:

$$\begin{aligned} \inf_{u\in U}(L\psi(x,u)+k(x,u)-\rho^*) &= 0, \quad x\in G, \\ \langle \nabla\psi(x), d_i\rangle &= 0, \quad x\in \partial G,\ i\in \mathrm{In}(x). \end{aligned} \tag{5.2}$$

A viscosity solution to the above equation is defined in a similar manner as that defined for (1.5) in Definition 3.3. We begin with the following result.

THEOREM 5.1. *Let (V,ρ) be given via* (5.1). *Then V is a viscosity solution to* (5.2) *with $\rho^* = \rho$.*

Proof. The proof is a slight variation of the arguments in [25], [4]. Let $F : G\times\mathbb{R}\times\mathbb{R}^k\times S^k \to \mathbb{R}$ be defined as

$$F(x,r,p,M) \doteq -\frac{1}{2}\,\mathrm{Tr}(aM) + \rho - \inf_{u\in U}\{\langle b(x,u),p\rangle + k(x,u)\}. \tag{5.3}$$

Define F_* and F^* as maps from $G\times\mathbb{R}\times\mathbb{R}^k\times S^k$ to $\mathbb{R}$ via (3.4) and (3.5), with F_α there replaced by F. Fix $x_0\in G$ and let $\psi\in C^2(G)$ be such that x_0 is a strict maximum point of $V-\psi$. We would like to show that

$$F_*(x_0, V(x_0), (D\psi)(x_0), (D^2\psi)(x_0)) \le 0. \tag{5.4}$$

Define $V_n(\cdot) \doteq V_{\alpha_n}(\cdot)$, $\overline{V}_n(\cdot) \doteq \overline{V}_{\alpha_n}(\cdot)$, and $F_n(\cdot) \doteq F_{\alpha_n}(\cdot)$. Using the fact that $\overline{V}_n$ converges to V uniformly on compacts, we can find an $N_0\in(0,\infty)$ and a sequence $\{x_n\}\subset G$ such that $x_n\to x_0$ and x_n is a local maximum of $V_n(\cdot)-\psi(\cdot)$ for all $n\ge N_0$. From Theorem 3.4 we then have that

$$F_n(x_n, V_n(x_n), (D\psi)(x_n), (D^2\psi)(x_n)) \wedge \min\{-\langle D\psi(x_n), d_i\rangle;\ i\in \mathrm{In}(x_n)\} \le 0. \tag{5.5}$$

Define $\tilde{F}_n : G\times\mathbb{R}\times\mathbb{R}^k\times S^k\to\mathbb{R}$ as

$$\tilde{F}_n(x,r,p,M) \doteq -\frac{1}{2}\,\mathrm{Tr}(aM) + \alpha_n r - \alpha_n V_n(0) - \inf_{u\in U}\{\langle b(x,u),p\rangle + k(x,u)\}. \tag{5.6}$$

Then from (5.5) we have that

$$\tilde{F}_n(x_n, \overline{V}_n(x_n), (D\psi)(x_n), (D^2\psi)(x_n)) \wedge \min\{-\langle D\psi(x_n), d_i\rangle;\ i\in \mathrm{In}(x_n)\} \le 0. \tag{5.7}$$

Next note that, as $n\to\infty$, $x_n\to x_0$, $\overline{V}_n(x_n)\to V(x_0)$, $(D\psi)(x_n)\to(D\psi)(x_0)$, and $(D^2\psi)(x_n)\to(D\psi)(x_0)$. Furthermore, since as $n\to\infty$, $\tilde{F}_n(\cdot)\to F(\cdot)$ uniformly on compacts, we have that

$$\tilde{F}_n(x_n, \overline{V}_n(x_n), (D\psi)(x_n), (D^2\psi)(x_n)) \to F(x_0, V(x_0), (D\psi)(x_0), (D^2\psi)(x_0)).$$

Also the lower semicontinuity property of $\mathrm{In}(\cdot)$ implies that

$$\min\{-\langle D\psi(x_0), d_i\rangle;\ i\in \mathrm{In}(x_0)\} \le \liminf_{n\to\infty} \min\{-\langle D\psi(x_n), d_i\rangle;\ i\in \mathrm{In}(x_n)\}.$$

Using the above two displays in (5.5) we have (5.4).

In a similar manner one shows that for all $x_0\in G$ and $\psi\in C^2(G)$, such that x_0 is a strict maximum point of $V-\psi$, we have that

$$F^*(x_0, V(x_0), (D\psi)(x_0), (D^2\psi)(x_0)) \ge 0.$$

This proves the result. ∎

We now characterize the value of the ergodic cost problem via the solution of (5.2). We will denote the right-hand side of (2.7) by $\overline{\rho}$.

THEOREM 5.2. *Let $(V(\cdot),\rho)$ be a solution of* (5.2). *Then $\rho=\overline{\rho}$.*

Proof. We begin by noting that, for all $r\in(0,\infty)$, $V(\cdot)$ is a viscosity solution of

$$\begin{aligned}\inf_{u\in U}\left((L-r)\psi(x,u)+k(x,u)-\rho+rV(x)\right)=0,\quad x\in G,\\ \langle\nabla\psi(x),d_i\rangle=0,\quad x\in\partial G,\ i\in \mathrm{In}(x).\end{aligned}\tag{5.8}$$

This, in view of Theorem 3.7, implies that for all $x\in G$

$$V(x)=\inf\mathbb{E}\left(\int_0^\infty e^{-rt}(k(X^x(t),u(t))-\rho+rV(X^x(t)))dt\right),\tag{5.9}$$

where the infimum is taken over all admissible pairs $(u(\cdot),W(\cdot))$.

From the proof of Lemma 4.5 of [3] it follows that for all $M\in(0,\infty)$

$$\sup_{x\in G,\,|x|\le M,\,(u(\cdot),W(\cdot))\text{ admissible}}\ \sup_{t\in[0,\infty)}\mathbb{E}|X^x(t)|<\infty,\tag{5.10}$$

where $X^x(\cdot)$ is given as a solution of (1.1) with $X(0)\equiv x$ and the control pair $(u(\cdot),W(\cdot))$. From this bound it follows via a slight modification of the proof of Lemma 6.4 of [10] that

$$\sup_v\int_G|x|\eta_v(dx)<\infty,$$

where the supremum is taken over all Markov controls.

Next let $\overline{v}:G\to U$ be as in Theorem 2.10 and let $\eta_{\overline{v}}$ be as in Theorem 2.9. Then, in view of the above observation and Remark 4.8, we have that

$$\begin{aligned}r\int_G V(x)\eta_{\overline{v}}(dx)+\rho&\le r\int_G\left(\mathbb{E}\left(\int_0^\infty e^{-rt}(k(X^x(t),\overline{v}(X^x(t)))+rV(X^x(t)))dt\right)\right)\eta_{\overline{v}}(dx)\\&=\int_G k(x,\overline{v}(x))\eta_{\overline{v}}(dx)+r\int_G V(x)\eta_{\overline{v}}(dx).\end{aligned}$$

Thus we have that

$$\rho\le\int_G k(x,\overline{v}(x))\eta_{\overline{v}(dx)}=\overline{\rho},\tag{5.11}$$

where the last equality follows from Theorem 2.10.

Now we prove the reverse inequality. Let $(u^r(\cdot),W^r(\cdot))$ be an admissible control which is optimal for the cost function in (5.9). The existence of such a control follows via the usual compactness arguments (cf. Chapter 2 of [7]). We denote by $X^r(\cdot)$ the process defined via (1.1) with $X_0\equiv x$ and $(u(\cdot),W(\cdot))$ there replaced by $(u^r(\cdot),W^r(\cdot))$. Then for all $T\ge0$

$$V(x)=\mathbb{E}\left(\int_0^T e^{-rs}(k(X^r(t),u^r(t))-\rho+rV(X^r(t)))dt+e^{-rT}V(X^r(T))\right).\tag{5.12}$$

By picking a subsequence if necessary, we can assume, without loss of generality, that as $r\to0$, $(u^r(\cdot),W^r(\cdot),X^r(\cdot))$ converge weakly to some $(u^*(\cdot),W^*(\cdot),X^*(\cdot))$, where $u^*(\cdot)$, $W^*(\cdot)$ is an admissible pair and X^* is the corresponding controlled diffusion

starting at x. Using the continuity and boundedness of V and k, we then have by taking the limit as $r \to 0$ in (5.12) that

$$V(x) = \mathbb{E}\bigg(\int_0^T (k(X^*(t), u^*(t)) - \rho)dt + V(X^*(T)) \bigg).$$

Dividing by T and taking the limit as $T \to \infty$, we have, on recalling that $V(z) \le \varsigma c_2 |z|$ for all $z \in G$ and using (5.10), that

$$\rho = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\bigg(\int_0^T k(X^*(t), u^*(t))dt \bigg) \ge \overline{\rho},$$

where the last inequality once more follows from Theorem 2.10. Combining the above display with (5.11), we have the result. □

As an immediate consequence of the above, we have the following result which is proved exactly in the same way as Theorem 4.4 of [4].

THEOREM 5.3. *(V, ρ) given via (5.1) is the unique viscosity solution of (5.2) satisfying $V(0) = 0$ and the linear growth condition $|V(x)| \le c|x|$ for some $c \in (0, \infty)$.*

Appendix. In this section we sketch the proofs of Theorems 3.6, 4.2, and 4.5. We begin with Theorem 3.6. The proof is essentially taken from [17]; however, we reproduce it here for the reader's convenience. We begin with the following lemma, which is Lemma 3.1 of [17].

LEMMA A.1. *Let $u \in USC(G)$ and let $v \in LSC(G)$. Define $w \in USC(G \times G)$ by $w(x, y) = u(x) - v(y)$. For $x \in G$ and a real valued function f on G, define*

$$D^+ f(x) \doteq \{p \in \mathbb{R}^k : f(x + h) \le f(x) + \langle p, h\rangle + o(|h|) \text{ for } x + h \in G \text{ and as } h \to 0\},$$
$$D^- f(x) \doteq \{p \in \mathbb{R}^k : f(x + h) \ge f(x) + \langle p, h\rangle + o(|h|) \text{ for } x + h \in G \text{ and as } h \to 0\},$$
$$D^{2,+} f(x) \doteq \bigg\{ (p, A) \in \mathbb{R}^k \times S^k : f(x + h) \le f(x) + \langle p, h\rangle + \frac{1}{2}\langle Ah, h\rangle + o(|h|^2) \text{ for } x + h \in G \text{ and as } h \to 0 \bigg\}$$

and

$$D^{2,-} f(x) \doteq \bigg\{ (p, A) \in \mathbb{R}^k \times S^k : f(x + h) \ge f(x) + \langle p, h\rangle + \frac{1}{2}\langle Ah, h\rangle + o(|h|^2) \text{ for } x + h \in G \text{ and as } h \to 0 \bigg\}.$$

Also define $\overline{D}^{2,+} f(x)$ as the set of points $(r, p, A) \in \mathbb{R} \times \mathbb{R}^k \times S^k$ for which there is a sequence $\{(x_n, p_n, A_n)\} \subset G \times \mathbb{R}^k \times S^k$ such that $(p_n, A_n) \in D^{2,+} f(x_n)$ and such that $x_n \to x$, $u(x_n) \to r$, $p_n \to p$, and $A_n \to A$ as $n \to \infty$. Similarly define $\overline{D}^{2,-} f(x)$. Let $\alpha, \beta > 0$, $p, q \in \mathbb{R}^n$, and $x, y \in G$. Assume that

$$\bigg(p, q, \alpha \begin{pmatrix} I & -I \\ -I & I \end{pmatrix} + \beta \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \bigg) \in D^{2,+} w(x, y).$$

Then there are $X, Y \in S^k$ for which

$$-C\alpha \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \le \begin{pmatrix} X - \beta I & 0 \\ 0 & Y - \beta I \end{pmatrix} \le C\alpha \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}$$

and

$$(u(x), p, X) \in \overline{D}^{2,+}u(x) \quad \textit{and} \quad (v(y), -q, -Y) \in \overline{D}^{2,-}v(y),$$

where $C \in (1, \infty)$ *is an absolute constant.*

Proof of Theorem 3.6. As stated above in Theorem 3.6, the proof that $V^n(\cdot)$ defined via (3.13) is a viscosity solution of (3.5) follows exactly as the proof of Theorem 3.4. Now let $V_1(\cdot)$ and $V_2(\cdot)$ be two viscosity solutions of (3.5). Let g be as in Lemma 4.2 of [10] (cf. (3.2)). For $\gamma, \beta \in (0, \infty)$ define

$$V_{\gamma\beta}(x) \doteq V_1(x) - \gamma g(x) + \beta \quad \text{and} \quad U_{\gamma\beta}(x) \doteq V_2(x) - \gamma g(x) - \beta, \quad x \in G.$$

Define $\tilde{F}_{\alpha,*}$ and $\tilde{F}^*_\alpha$ as maps from $G \times \mathbb{R} \times \mathbb{R}^k \times S^k$ to $\mathbb{R}$, via (3.4) and (3.5), respectively, by replacing the set $\{-\langle d_i, p\rangle;\ i \in \mathrm{In}(x)\}$ in (3.4) by the set $\{-\langle d_i, p\rangle + \gamma;\ i \in \mathrm{In}(x)\}$ and in (3.5) by the set $\{-\langle d_i, p\rangle - \gamma;\ i \in \mathrm{In}(x)\}$. Then as in Theorem 2.1 of [17], to every $\beta \in (0, \infty)$ there exists a $\gamma \equiv \gamma(\beta) \le \beta$ in $(0, \infty)$ such that

$$\tilde{F}_{\alpha,*}(x, U_{\gamma,\beta}(x), p, A) \le 0 \quad \text{for } x \in G \text{ and } (p, A) \in D^{2,+}U_{\gamma,\beta}(x), \tag{A.1}$$

$$\tilde{F}^*_\alpha(x, V_{\gamma,\beta}(x), p, A) \ge 0 \quad \text{for } x \in G \text{ and } (p, A) \in D^{2,-}V_{\gamma,\beta}(x). \tag{A.2}$$

Now fix a β and the corresponding γ in $(0, \infty)$. Suppose that

$$\kappa_0 \doteq \max_{x \in \overline{G} \cap B_n} (U_{\gamma\beta}(x) - V_{\gamma\beta}(x)) > 0. \tag{A.3}$$

Then noting that $U_{\gamma\beta}(x) - V_{\gamma\beta}(x) \le 0$ for all $x \in \partial B_n$, it follows through standard maximum principle arguments that

$$\kappa_0 = (U_{\gamma\beta}(z) - V_{\gamma\beta}(z)) \quad \text{for some } z \in \partial G \cap B_n. \tag{A.4}$$

We now show that (A.4) leads to a contradiction.

Let F be as in (3.3). Then by using the boundedness and Lipschitz property of the coefficients one can show that the following hold.

- There is a function $m_1 \in C([0, \infty))$ satisfying $m_1(0) = 0$ such that for all $\theta_0 \ge 1$, $x, y \in G \cap \overline{B}_n$, $r \in \mathbb{R}$, $P \in \mathbb{R}^k$, and $X, Y \in S^k$,

$$F_\alpha(y, r, p, -Y) - F_\alpha(x, r, p, X) \le m_1(|x - y|(|p| + 1) + \theta_0|x - y|^2)$$

whenever

$$-\theta_0 \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \le \begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix} \le \theta_0 \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}.$$

- There is a neighborhood U of ∂G in $G \cap \overline{B}_n$ and a function $m_2 \in C([0, \infty))$ satisfying $m_2(0) = 0$ for which

$$|F(x, r, p, X) - F(x, r, q, Y)| \le m_2(|p - q| + \|X - Y\|) \tag{A.5}$$

for $x \in U$, $r \in \mathbb{R}$, $p, q \in \mathbb{R}^k$, and $X, Y \in S^k$.

Clearly, we can find an open neighborhood V of z (which is as small as we want) such that

$$\mathrm{In}(x) \subset \mathrm{In}(z) \quad \text{for } x \in V \cap \partial G,\ V \cap G \subset U, \tag{A.6}$$

and

$$\langle y-x, n_i\rangle \le \theta|x-y| \quad \text{for } i\in \mathrm{In}(z),\ x\in V\cap G_i, \text{ and } y\in V\cap \partial G_i. \tag{A.7}$$

Also, from Theorem 4.1 of [17] we can find a family $\{w_\epsilon\}_{\epsilon>0}$ of continuous functions on $\overline{V}\times\overline{V}$ and a positive constant θ having the following property: For any $\epsilon>0$ and $x,y\in V$, there are $p\equiv p(\epsilon,x,y), q\equiv q(\epsilon,x,y)\in\mathbb{R}^k$ such that for all $i\in \mathrm{In}(z)$, the following hold (cf. equations (3.15)–(3.19) of [17]):

$$w_\epsilon(x,x)=0, \qquad w_\epsilon(x,y)\ge \theta\frac{|x-y|^2}{\epsilon}, \tag{A.8}$$

$$\langle d_i, p\rangle \ge -\frac{|x-y|^2}{\epsilon} \quad \text{if } \langle y-x,n_i\rangle \ge -\theta|x-y|, \tag{A.9}$$

$$\langle d_i, q\rangle \ge -\frac{|x-y|^2}{\epsilon} \quad \text{if } \langle y-x,n_i\rangle \le \theta|x-y|, \tag{A.10}$$

$$|p+q|\le \frac{|x-y|^2}{\epsilon}, \quad |q|\le \frac{|x-y|}{\epsilon}, \tag{A.11}$$

and

$$\left(p, q, \frac{1}{\epsilon}\begin{pmatrix} I & -I\\ -I & I\end{pmatrix} + \frac{|x-y|^2}{\epsilon}\begin{pmatrix} I & 0\\ 0 & I\end{pmatrix}\right)\in D^{2,+}w_\epsilon(x,y). \tag{A.12}$$

Henceforth, we will write $U_{\gamma,\beta}$ as u and $V_{\gamma,\beta}$ as v. Now fix $\delta>0$ and define $\tilde{u}\in USC(G)$ by

$$\tilde{u}(x)\doteq u(x)-\frac{\delta}{2}|x-z|^2. \tag{A.13}$$

Clearly, z is the unique maximum point of $\tilde{u}-v$. Fix $\epsilon>0$ and define $\phi\in USC([\overline{V}\cap G]\times[\overline{V}\cap G])$ by

$$\phi(x,y)\doteq \tilde{u}(x)-v(y)-w_\epsilon(x,y).$$

Let $(\overline{x},\overline{y})\equiv(\overline{x}(\epsilon),\overline{y}(\epsilon))\in[\overline{V}\cap G]\times[\overline{V}\cap G]$ be a maximum point of ϕ. Then it follows exactly as in [17] (cf. equation (3.23) of that paper) that as $\epsilon\to 0$,

$$\frac{|\overline{x}-\overline{y}|^2}{\epsilon}\to 0, \quad \overline{x},\,\overline{y}\to z, \quad \tilde{u}(\overline{x})\to\tilde{u}(z), \quad \text{and} \quad v(\overline{y})\to v(z). \tag{A.14}$$

Now let ϵ be small enough so that $\overline{x},\overline{y}\in V$. Define $w(x,y)\doteq\tilde{u}(x)-v(y)$, $x,y\in G$. Let $p\doteq p(\epsilon,\overline{x},\overline{y})$, $q\doteq q(\epsilon,\overline{x},\overline{y})$. From (A.12) it follows that

$$\left(p, q, \frac{1}{\epsilon}\begin{pmatrix} I & -I\\ -I & I\end{pmatrix} + \frac{|\overline{x}-\overline{y}|^2}{\epsilon}\begin{pmatrix} I & 0\\ 0 & I\end{pmatrix}\right)\in D^{2,+}w_\epsilon(\overline{x},\overline{y})\subset D^{2,+}w(\overline{x},\overline{y}),$$

where the last inclusion follows on noting that $(\overline{x},\overline{y})$ is the maximum point of ϕ. Hereafter, denote $\frac{|\overline{x}-\overline{y}|^2}{\epsilon}$ by κ. From Lemma A.1 one can find matrices $X,Y\in S^k$ such that

$$-\frac{C}{\epsilon}\begin{pmatrix} I & 0\\ 0 & I\end{pmatrix}\le\begin{pmatrix} X & 0\\ 0 & Y\end{pmatrix}\le\frac{C}{\epsilon}\begin{pmatrix} I & -I\\ -I & I\end{pmatrix}+C\kappa\begin{pmatrix} I & 0\\ 0 & I\end{pmatrix}$$

and

$$(\tilde{u}(\overline{x}), p, X) \in \overline{D}^{2,+}\tilde{u}(\overline{x}) \quad \text{and} \quad (v(\overline{y}), -q, -Y) \in \overline{D}^{2,-}v(\overline{y}), \tag{A.15}$$

where C is as in Lemma A.1.

Using (A.13) one has that

$$(u(\overline{x}), p + \delta(\overline{x} - z), X + \delta I) \in \overline{D}^{2,+}u(\overline{x}). \tag{A.16}$$

Also, using (A.14) we have that for ϵ small enough and $i \in \mathrm{In}(z)$

$$\langle d_i, p + \delta(\overline{x} - z)\rangle + \gamma \geq \langle d_i, p\rangle + \frac{\gamma}{2}. \tag{A.17}$$

Using (A.9), (A.10), and (A.14) we have that for $i \in \mathrm{In}(z)$ and small enough ϵ

$$\langle d_i, p\rangle + \frac{\gamma}{2} > 0 \quad \text{if } \overline{x} \in \partial G_i, \quad \text{and} \quad -\langle d_i, q\rangle - \frac{\gamma}{2} < 0 \quad \text{if } \overline{y} \in \partial G_i. \tag{A.18}$$

From (A.6) we have that

$$V \cap \partial G \subset \bigcup_{i \in \mathrm{In}(z)} \partial G_i. \tag{A.19}$$

Combining (A.19), (A.18), and (A.17) we have that, for small enough ϵ,

$$\begin{aligned} &\langle d_i, p + \delta(\overline{x} - z)\rangle + \gamma > 0 \quad \text{if } \overline{x} \in \partial G \text{ and } i \in \mathrm{In}(\overline{x}) \quad \text{and} \\ &\langle d_i, -q\rangle - \gamma < 0 \quad \text{if } \overline{y} \in \partial G \text{ and } i \in \mathrm{In}(\overline{y}). \end{aligned} \tag{A.20}$$

This, along with (A.1), (A.2), (A.16), (A.15), shows that we can find $\epsilon_\delta \in (0, \infty)$ such that for all $0 \leq \epsilon \leq \epsilon_\delta$

$$F_\alpha(\overline{x}, u(\overline{x}), p + \delta(\overline{x} - z), X + \delta I) \leq 0 \leq F_\alpha(\overline{y}, v(\overline{y}), -q, -Y).$$

Thus we have that

$$\begin{aligned} 0 &\geq F_\alpha(\overline{x}, u(\overline{x}), p + \delta(\overline{x} - z), X + \delta I) - F_\alpha(\overline{y}, v(\overline{y}), -q, -Y) \\ &\geq F_\alpha(\overline{x}, u(\overline{x}), -q, X - C\kappa I) - F_\alpha(\overline{y}, u(\overline{x}), -q, -Y + C\kappa I) \\ &\quad + \alpha(u(\overline{x}) - v(\overline{y})) - m_2(\kappa + \delta|\overline{x} - z| + \delta + C\kappa) - m_2(C\kappa) \\ &\geq \alpha(u(\overline{x}) - v(\overline{y})) - m_1(|\overline{x} - \overline{y}| + 2C\kappa) - m_2(\kappa + \delta|\overline{x} - z| + \delta + C\kappa) - m_2(C\kappa), \end{aligned}$$

where the second inequality is obtained from (A.15), while the third inequality is obtained from (A.16). Taking the limit as $\epsilon \to 0$ we get that $\alpha(u(z) - v(z)) \leq 0$, which contradicts (A.3) and (A.4). This shows that $(U_{\gamma\beta}(x) - V_{\gamma\beta}(x)) \leq 0$ for all $x \in \overline{G \cap B_n}$. Taking the limit as $\beta \to 0$, we see that $V_2(x) \leq V_1(x)$ for all $x \in \overline{G \cap B_n}$. Reversing the roles of V_1 and V_2, we see that we must have $V_2(x) = V_1(x)$ for all $x \in \overline{G \cap B_n}$. □

Sketch of the proof of Theorem 4.2. Let $\mathrm{SM} \doteq \{v : G \to U \mid v \text{ is measurable}\}$ be the space of all Markov controls. Let $d(\cdot, \cdot)$ be the metric on SM introduced in [6] (see also Theorem II.2.2 of [7]). Then SM is compact under the metric topology of d.

Let $\Lambda(\epsilon, M, M_0)$ be as in the statement of the theorem and arguing via contradiction suppose that $\limsup_{\epsilon \to 0} \Lambda(\epsilon, M, M_0) \neq 0$. Then one can find $x_n, y_n, x \in E_{M_0}$ and $v_n, v \in \mathrm{SM}$ such that $x_n \to x$, $y_n \to x$, $v_n \to v$, and

$$\limsup_{n \to \infty} \int_0^M |\mathbb{E}k(X^{x_n}(s), v_n(X^{x_n}(s))) - \mathbb{E}k(X^{y_n}(s), v_n(X^{y_n}(s)))|\, ds \neq 0, \tag{A.21}$$

where $X^{x_n}(\cdot)$ (and $X^{y_n}(\cdot)$) are defined via (2.3) with $X(0) \equiv x_n$(resp., y_n) and the Markov control v^n. In a manner similar to Theorem II.2.3 of [7] one can show that for all $t \in (0,\infty)$ the probability laws of $X^{x_n}(t)$ and $X^{y_n}(t)$ converge in total variation from the probability law of $X(t)$, where $X(\cdot)$ is defined via (2.3) with $X(0) \equiv x$. From this and recalling that $k(x,v) \doteq \int_S \overline{k}(x,\alpha)v(d\alpha)$, where $\overline{k} \in C_b(G \times S)$, it then follows that for all $s \in (0,\infty)$

$$|\mathbb{E}k(X^{x_n}(s), v_n(X^{x_n}(s))) - \mathbb{E}k(X(s), v_n(X(s)))|$$

and

$$|\mathbb{E}k(X^{y_n}(s), v_n(X^{y_n}(s))) - \mathbb{E}k(X(s), v_n(X(s)))|$$

converge to 0 as $n \to \infty$. Combining the above displays we have that for all $s \in (0,\infty)$

$$|\mathbb{E}k(X^{x_n}(s), v_n(X^{x_n}(s))) - \mathbb{E}k(X^{y_n}(s), v_n(X^{y_n}(s)))|$$

converges to 0 as $s \to \infty$. Finally, an application of the dominated convergence theorem gives that

$$\int_0^M |\mathbb{E}k(X^{x_n}(s), v_n(X^{x_n}(s))) - \mathbb{E}k(X^{y_n}(s), v_n(X^{y_n}(s)))|ds$$

converges to 0 as $n \to \infty$. This contradicts (A.21), and thus the result follows. □

Proof of Theorem 4.5. An application of Itô's formula gives that

$$\begin{aligned} F(X^x_{n+1}) \le F(X^x_n) &+ \int_n^{n+1} \langle DF(X^x(s)), b(X^x(s), u(s))\rangle ds \\ &+ \sum_{i=1}^N \int_n^{n+1} \langle DF(X^x(s)), d_i\rangle dY_i(s) \\ &+ \frac{1}{2}\int_n^{n+1} \mathrm{tr}(\sigma^*(X^x(s))D^2F(X^x(s))\sigma(X^x(s)))ds \\ &+ \int_n^{n+1} \langle DF(X^x(s)), \sigma(X^x(s))dW(s)\rangle. \end{aligned}$$

Conditioning with respect to $\mathcal{F}_n$ and using parts (c) and (d) of Theorem 4.4, we have that

$$\begin{aligned} &\mathbb{E}(F(X^x_{n+1}) \mid \mathcal{F}_n) - F(X^x_n) \\ &\quad \le -c + \frac{1}{2}\mathbb{E}\left(\int_n^{n+1} \mathrm{tr}(\sigma^*(X^x(s))D^2F(X^x(s))\sigma(X^x(s)))ds \mid \mathcal{F}_n\right). \end{aligned} \tag{A.22}$$

Let $\tilde{g}(x) \doteq \frac{1}{2}\mathrm{tr}(\sigma^*(x)D^2F(x)\sigma(x))$. From part (b) of Theorem 4.4 we can find $\ell_0 \in (0,\infty)$ such that $|\tilde{g}(x)| \le \frac{c}{2}$ for all $x \in G$ such that $|x| > \frac{\ell_0}{2}$. This implies that

$$\begin{aligned} &\frac{1}{2}\int_n^{n+1} \mathrm{tr}(\sigma^*(X^x(s))D^2F(X^x(s))\sigma(X^x(s)))ds \\ &\quad \le |\tilde{g}|_\infty 1_{X^x_n \in B} + \left(|\tilde{g}|_\infty 1_{\sup_{n\le s\le n+1}|X^x(s)-X^x_n| > \frac{\ell_0}{2}} + \frac{c}{2}\right) 1_{X^x_n \in B^c}. \end{aligned}$$

From the above display and (A.22) we have that

$$\begin{aligned}
&\mathbb{E}(F(X_{n+1}^x) \mid \mathcal{F}_n) - F(X_n^x) \\
&\quad \le -c + \left(|\tilde{g}|_\infty \mathbb{E}(1_{\sup_{n\le s\le n+1} |X^x(s)-X_n^x| > \frac{\ell_0}{2}} \mid \mathcal{F}_n) + \frac{c}{2} \right) 1_{X_n^x \in B^c} + |\tilde{g}|_\infty 1_{X_n^x \in B} \\
&\quad \le \left(-\frac{c}{2} + |\tilde{g}|_\infty \frac{\mathbb{E}(\sup_{n\le s\le n+1} |X^x(s) - X_n^x| \mid \mathcal{F}_n)}{\ell_0/2} \right) 1_{X_n^x \in B^c} + |\tilde{g}|_\infty 1_{X_n^x \in B}.
\end{aligned}$$

Using the boundedness of the drift and diffusion coefficients we can find $\overline{\ell}$ such that

$$\mathbb{E}\left(\sup_{n\le s\le n+1} |X^x(s) - X_n^x| \mid \mathcal{F}_n \right) \le \overline{\ell},$$

and without loss of generality we can assume that ℓ_0 is large enough so that $\frac{2|\tilde{g}|_\infty \overline{\ell}}{\ell_0} < \frac{c}{4}$. Using these bounds in the above display, we have that

$$\mathbb{E}(F(X_{n+1}^x) \mid \mathcal{F}_n) - F(X_n^x) \le -\frac{c}{4} 1_{X_n^x \in B^c} + |\tilde{g}|_\infty 1_{X_n^x \in B}.$$

The result now follows on setting $c_0 = \frac{c}{4}$ and $M_0 = |\tilde{g}|_\infty$. □

Acknowledgments. We would like to thank the referee and an associate editor for several suggestions that led to improvements in the manuscript.

REFERENCES

[1] B. Ata, J. M. Harrison, and L. A. Shepp, *Drift Rate Control of a Brownian Processing System*, preprint.

[2] R. Atar and A. Budhiraja, *Stability properties of constrained jump-diffusion processes*, Electron. J. Probab., 7 (2002), 31 (electronic).

[3] R. Atar, A. Budhiraja, and P. Dupuis, *On positive recurrence of constrained diffusion processes*, Ann. Probab., 29 (2001), pp. 979–1000.

[4] G. K. Basak, V. S. Borkar, and M. K. Ghosh, *Ergodic control of degenerate diffusions*, Stochastic Anal. Appl., 1 (1997), pp. 1–17.

[5] V. S. Borkar, *Dynamic programming for ergodic control with partial observations*, Stochastic Process. Appl., 103 (2003), pp. 293–310.

[6] V. S. Borkar, *A topology for Markov controls*, Appl. Math. Optim., 20 (1989), pp. 55–62.

[7] V. S. Borkar, *Optimal Control of Diffusion Processes*, Longman Scientific and Technical, Harlow, UK, 1989.

[8] V. S. Borkar and A. Budhiraja, *A further remark on dynamic programming for partially observed Markov processes*, Stochastic Process. Appl., 112 (2004), pp. 79–93.

[9] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions.* II. *Adaptive control*, Appl. Math. Optim., 21 (1990), pp. 191–220.

[10] A. Budhiraja, *An ergodic control problem for constrained diffusion processes: Existence of optimal Markov control*, SIAM J. Control Optim., 42 (2003), pp. 532–558.

[11] A. Budhiraja and P. Dupuis, *Simple necessary and sufficient conditions for the stability of constrained processes*, SIAM J. Appl. Math., 59 (1999), pp. 1686–1700.

[12] A. Budhiraja and A. P. Ghosh, *A large deviations approach to asymptotically optimal control of crisscross network in heavy traffic*, Ann. Appl. Probab., to appear.

[13] M. Chen, C. Pandit, and S. P. Meyn, *In search of sensitivity in network optimization*, Queueing Systems Theory Appl., 44 (2003), pp. 313–363.

[14] R. M. Cox, *Stationary and discounted control of diffusion processes*, Ph.D. thesis, Columbia University, New York, NY, 1984.

[15] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[16] P. Dupuis and H. Ishii, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics Stochastics Rep., 35 (1991), pp. 31–62.

[17] P. Dupuis and H. Ishii, *On oblique derivative problems for fully nonlinear second-order elliptic PDE's on domains with corners*, Hokkaido Math. J., 20 (1991), pp. 135–164.

[18] P. Dupuis and K. Ramanan, *Convex duality and the Skorokhod Problem.* I, II, Probab. Theory Related Fields, 115 (1999), pp. 153–195, 197–236.

[19] P. W. Glynn and S. P. Meyn, *A Liapounov bound for solutions of the Poisson equation*, Ann. Probab., 24 (1996), pp. 916–931.

[20] J. M. Harrison and J. A. Van Miegham, *Dynamic control of Brownian networks: State space collapse and equivalent workload formulation*, Ann. Appl. Probab., 7 (1997), pp. 747–771.

[21] J. M. Harrison and L. Wein, *Scheduling networks of queues: Heavy trafic analysis of a simple open network*, Queueing Systems Theory Appl., 5 (1989), pp. 265–280.

[22] F. P. Kelly and C. N. Laws, *Dynamic routing in open queuing networks: Brownian models, cut constraints and resource pooling*, Queueing Systems Theory Appl., 13 (1993), pp. 47–86.

[23] H. J. Kushner, *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*, Springer-Verlag, New York, May 2001.

[24] H. J. Kushner and L. F. Martins, *Limit theorems for pathwise average cost per unit time problems for controlled queues in heavy traffic*, Stochastics Stochastics Rep., 42 (1993), pp. 25–51.

[25] P.-L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations.* II. *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.

[26] L. F. Martins and H. J. Kushner, *Routing and singular control for queuing networks in heavy traffic*, SIAM J. Control Optim., 28 (1990), pp. 1209–1233.

[27] S. P. Meyn, *The policy iteration algorithm for average reward Markov decision processes with general state space*, IEEE Trans. Automat. Control, 42 (1997), pp. 1663–1680.

[28] S. P. Meyn, *Stability, performance evaluation, and optimization*, Handbook of Markov Decision Processes, Internat. Ser. Oper. Res. Management Sci. 40, Kluwer Academic Publishers, Boston, 2002, pp. 305–346.

[29] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[30] S. Meyn and R. Tweedie, *Stability of Markovian processes,* III. *Foster-Lyapunov criteria for continuous-time processes*, Adv. in Appl. Probab., 25 (1993), pp. 518–548.

[31] S. Meyn and R. Tweedie, *Generalized resolvents and Harris recurrence of Markov processes*, in Doeblin and Modern Probability (Blaubeuren, 1991), Contemp. Math. 149, AMS, Providence, RI, 1991, pp. 227–250.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1493–1520

CAUSAL AND STABLE INPUT/OUTPUT STRUCTURES ON MULTIDIMENSIONAL BEHAVIORS*

J. WOOD†, V. R. SULE‡, AND E. ROGERS†

Abstract. In this work we study multidimensional (nD) linear differential behaviors with a distinguished independent variable, called "time." We define in a natural way causality and stability of input/output structures with respect to this distinguished direction. We make an extension of some results in the theory of partial differential equations, demonstrating that causality is equivalent to a property of the transfer matrix which is essentially hyperbolicity of the P^c operator defining the behavior $(\mathcal{B}^c)_{0,y}$. We also quote results which in effect characterize time-autonomy for the general systems case.

Stability is likewise characterized by a property of the transfer matrix. We prove this result for the 2D case and for the case of a single equation; for the general case it requires solution of an open problem concerning the geometry of a particular set in $\mathbb{C}^n$. In order to characterize input/output stability we also develop new results on inclusions of kernels, freeness of variables, and closure with respect to $\mathcal{S}, \mathcal{S}'$ and associated spaces, which are of independent interest. We also discuss stability of autonomous behaviors, which we believe to be governed by a corresponding condition.

Key words. stability, causality, partial differential equations, multidimensional systems, behavioral approach, hyperbolic systems, input/output structures, time-autonomy.

AMS subject classifications. 13C05, 13C12, 35B37, 93A99, 93B25, 93B55

DOI. 10.1137/S0363012903429979

1. Introduction.

In this paper we are concerned with questions of causality and stability for systems defined by PDEs. We consider these problems in the framework of multidimensional (nD) behaviors (note that this is quite distinct from the infinite-dimensional systems framework of, e.g., [2]). To date, the theory of nD behaviors has almost entirely considered the independent variables on an equal footing. However, in an apparent majority of applications, particularly in the case of systems given by PDEs, one of the independent variables, "time," is distinguished and plays a special role. Recent work by Sasane [26, 24, 25] attempts to develop nD behavioral theory in this less symmetrical and more applicable situation.

This consideration is particularly significant when we discuss a concept such as stability, which is naturally associated with the passage of "time." Stability of course may be divided into two concepts: stability with respect to initial conditions (i.e., stability of an autonomous behavior) and input/output stability. The current work was motivated by consideration of the first concept but has led only to a (partial) characterization of the second!

We therefore begin our main exposition with a discussion of stability of autonomous behaviors in section 3. Our argument is motivated by the rough principle that a behavior should be classified as "unstable" only when it contains trajectories which are unstable (in whatever sense) but for which the corresponding initial conditions (whatever we may mean by these) are, nevertheless, (in some sense) stable. The two-dimensional (2D) discrete definition of stability by Valcher [33] applies this

*Received by the editors June 23, 2003; accepted for publication (in revised form) April 26, 2004; published electronically January 27, 2005.

http://www.siam.org/journals/sicon/43-4/42997.html

†School of Electronics and Computer Science, University of Southampton, Southampton, SO17, 1BJ, UK (etar@ecs.soton.ac.uk).

‡Department of Electrical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai, 400 076, India (vrs@ee.iitb.ac.in).

principle. Unfortunately, for PDEs, "initial conditions" is a much harder concept to understand and work with. Section 3, however, proposes a possible condition for stability in this sense, which we call the characteristic variety (CV) condition. Roughly, the idea behind this condition is to classify trajectories as unstable only if they are both "blowing up" in the time direction and "physically reasonable" in the spatial direction. Section 3.1 brings hyperbolic systems into the discussion, since hyperbolicity is for certain important classes of systems a consequence of the CV condition. One particular property of hyperbolic systems is that they are time-autonomous as defined in [25]; we quote a result from the PDE literature which effectively characterizes this property for the general systems case.

The remainder of the paper is devoted to input/output stability and also input/output causality, on which the former concept rests. Essentially, a system (or rather, an input/output structure on a given behavior) is defined as causal if for any input with zero past, there exists a corresponding output with zero past. Strictly speaking, this definition only makes sense if the "past" contains an initial condition set for the zero-input behavior; we therefore restrict our attention to systems for which this behavior is time-autonomous. Bringing in some important results from the literature on PDEs, we demonstrate in section 4 that causality is a consequence of hyperbolicity of the P operator. More strongly, we show that causality is equivalent to a property of the transfer matrix of the system, which is in turn equivalent to hyperbolicity of the P^c operator in the description of the controllable part.

We then move on to stability, which is defined in terms of trajectories in $\mathcal{S}$ or in $\mathcal{S}'$ having support in the half-space $t \geq 0$. An input/output structure is defined as stable if any causal output response to an input of this type is also necessarily of this type. The growth restrictions on these trajectories in the spatial directions convey the notion that they are "physically reasonable," and the growth restrictions in the temporal direction suggest stability. In section 5, we give some background results from the PDE literature on convolution operators for these special cases; these results effectively characterize input/output stability in this sense for the special case $p(\delta)y = u$. The general case requires some "structure theory" for behaviors over $\mathcal{S}_+$ (signals of $\mathcal{S}$ with support in the half-plane) and its dual space, which we develop in section 6. For completeness, using the same methodology we also develop analogous results for the spaces $\mathcal{S}$ and $\mathcal{S}'$. In particular, for $n \leq 2$ and for the case where the P operator is a single polynomial (covering the single-input single-output case in particular), we characterize inclusion of behaviors over $\mathcal{S}'$ and freeness of variables over $\mathcal{S}$ (these being essentially dual problems). The more general case for $n > 2$ is not proved, since the arguments used depend critically on proving a geometric property "ideal-convexity" of a particular set (the "bad frequency" set) in $\mathbb{C}^n$. To date, we have not been able to prove this property for $n \geq 2$; this is discussed in section 5.3.

Section 7 finally defines input/output stability, and characterizes it, again for the cases $n \leq 2$ or P equal to a single polynomial (e.g., single-input single-output). It is shown that an input/output structure is both causal and stable (in terms of trajectories in $\mathcal{S}$ with zero past) if and only if the least common denominator of the transfer matrix satisfies the CV condition introduced in section 3, i.e., if and only if the system has no controllable unstable poles. Thus it appears that input/output stability may be equivalent to stability of the zero input behavior, as happens in the standard one-dimensional (1D) case. We also give a similar sufficient condition for stability with respect to trajectories in $\mathcal{S}'$ with zero past; this requires no prior assumptions on the system. Extension of all results to the general nD case requires only a proof that the set of unstable frequencies is ideal-convex. We summarize in section 8.

2. Behaviors, classical spaces, and pole structure. We begin by briefly reviewing some concepts and results from the theory of nD behaviors; see, e.g., [22, 34] for the 1D case and, e.g., [17, 20, 35, 40] for general background on the continuous nD case.

2.1. Classical and associated spaces. We consider solutions to behaviors in the classical spaces from the theory of distributions and so begin by recalling these and associated spaces. We denote the classical spaces by $\mathcal{C}^\infty$ (smooth functions), $\mathcal{D}'$ (distributions), $\mathcal{C}_0^\infty$ (compactly smooth functions), $\mathcal{E}'$ (compactly supported distributions), $\mathcal{S}$ (rapidly decreasing functions), and $\mathcal{S}'$ (tempered distributions). Here all functions and functionals are taken to be complex-valued. Recall (e.g., [23]) that rapidly decreasing functions are those functions which decay faster than any polynomial grows; a precise definition is given in section 5.1. The tempered distributions may be thought of as distributions that grow no faster than some polynomial (see, e.g., [8, sec. 7.1] for a detailed treatment of $\mathcal{S}$ and $\mathcal{S}'$). Following [4] we also define, for any of the classical spaces $\mathcal{W}$, the spaces [4],

$$\mathcal{W}_+ := \{w \in \mathcal{W} \mid \operatorname{supp} w \subseteq \mathbb{R}^n_+\}, \tag{1}$$

$$\mathcal{W}_- := \{w \in \mathcal{W} \mid \operatorname{supp} w \subseteq \mathbb{R}^n_-\}, \tag{2}$$

$$\mathcal{W}_\oplus := \mathcal{W}/\mathcal{W}_-, \tag{3}$$

$$\mathcal{W}_\ominus := \mathcal{W}/\mathcal{W}_+. \tag{4}$$

Here $\mathbb{R}_+ := \mathbb{R}^{n-1} \times [0, +\infty)$ and $\mathbb{R}_- := \mathbb{R}^{n-1} \times (-\infty, 0]$.

The spaces $\mathcal{S}_+$ and $\mathcal{S}'_\oplus$ will prove particularly important in what follows. It is clear that any element $f \in \mathcal{S}_\oplus$ may be regarded as an element of the dual of $\mathcal{S}_+$, according to

$$\langle f, \phi\rangle := \langle \overline{f}, \iota(\phi)\rangle,\ \phi \in \mathcal{S}_+, \tag{5}$$

where $\overline{f}$ is any element of $\mathcal{S}'$ which projects to f, and $\iota : \mathcal{S}_+ \to \mathcal{S}$ is the natural inclusion. As noted in [4], $\mathcal{S}'_\oplus$ is in fact equal to the dual space of $\mathcal{S}_+$.

Denote by $\mathbb{C}[s]$ the polynomial ring in n indeterminates $s = s_1, \ldots, s_n$ with complex coefficients. We associate with any polynomial matrix $R = R(s) \in \mathbb{C}^{g\times q}$ the differential operator $R(\delta) := R(\partial/\partial x_1, \ldots, \partial/\partial x_n)$, $x_1, \ldots, x_n$ being independent variables in the space $\mathbb{R}^n$. This operator maps $\mathcal{W}^q$ to $\mathcal{W}^g$ for any of the spaces $\mathcal{W}$ listed above (the action on factors $\mathcal{W}_\oplus$, $\mathcal{W}_\ominus$ being induced in the obvious way). We remark that in the theory of PDEs, it is more usual to consider operators in the form $w \mapsto R(1/\imath)\partial)w$. For this reason, certain results in the theory of PDEs concerning the algebraic structure of operators change form in a straightforward way when translated into the current framework.

Recall also that for any of the classical pairs of dual spaces $\mathcal{W}, \mathcal{W}'$, and any polynomial matrix $R \in \mathbb{C}[s]^{g\times q}$, the *adjoint matrix* $R^*(s) := R^T(-s)$ has the property that

$$\langle R(\partial)f, \phi\rangle = \langle f, R^*(\partial)\phi\rangle \tag{6}$$

for any $f \in \mathcal{W}^q$, $\phi \in (\mathcal{W}')^g$.

2.2. Behaviors, associated varieties, and time-autonomy. For any of the spaces $\mathcal{W}$ discussed in section 2.1, and for a polynomial matrix $R \in \mathbb{C}[s]^{g\times q}$,

denote as usual

$$\ker_{\mathcal{W}} R = \{w \in \mathcal{W}^q \mid R(\delta)w = 0\}, \tag{7}$$

$$\operatorname{im}_{\mathcal{W}} R = \{w \in \mathcal{W}^g \mid \exists\, l \in \mathcal{W}^q \text{ s.t. } w = R(\partial)l\}. \tag{8}$$

In this situation, we say that R is a *kernel representation matrix* of the behavior $\mathcal{B} = \ker_{\mathcal{R}} R$. $\mathcal{W}$ is referred to as the *the signal space;* the signal space of a behavior is taken to be $\mathcal{D}'$ unless otherwise specified.

For the operator $R(\partial)$ or behavior $\ker_{\mathcal{D}'} R$, the associated *system module* or *module of formal quantities* is defined as $\mathcal{M} := \mathbb{C}[s]^{1\times q}/\mathbb{C}[s]^{1\times g}R$. This object is also standard in PDE theory, and some relationships between $\mathcal{M}$ and $\mathcal{B}$ are drawn out in [35]. In particular, the behavior $\mathcal{B}$ (for any signal space $\mathcal{W}$) may be identified with $\operatorname{Hom}_{\mathbb{C}[s]}(\mathcal{M}, \mathcal{W})$ [13, 17].

Given a polynomial matrix $R \in \mathbb{C}[s]^{g\times q}$, recall that a *universal or minimal left annihilator* is a matrix $L \in \mathbb{C}[s]^{h\times g}$ for some h, such that the rows of L generate the set of polynomial vectors v with $vR = 0$. Then the "fundamental principle" of Ehrenpreis–Palamodov states that $\operatorname{im}_{\mathcal{W}} R = \ker L$ for $\mathcal{W} = \mathcal{D}'$ or $\mathcal{W} = \mathcal{C}^\infty$. Equivalently, these two signal spaces (modules) are *injective.* This property is a major component of a very rich relationship between system modules $\mathcal{M}$ and behaviors $\mathcal{B}$, introduced into behavioral theory by Oberst [17]. We will also use standard facts and results concerning the *associated primes* of $\mathcal{M}$; see, e.g., [3] for the background here.

Let $\mathcal{B} = \ker_{\mathcal{D}'} R$ with $R \in \mathbb{C}[s]^{g\times q}$; denote by $\mathcal{V}(\mathcal{B})$ the *characteristic variety*

$$\mathcal{V}(\mathcal{B}) := \{\zeta \in \mathbb{C}^n \mid \operatorname{rank} R(\zeta) < q\} \tag{9}$$

which is well known to depend only on $\mathcal{B}$ and to be equal to the variety of the ideal

$$\operatorname{ann} \mathcal{M} := \{r \in \mathbb{C}[s] \mid rx = 0 \text{ for all } x \in \mathcal{M}\}. \tag{10}$$

The points of $\mathcal{V}(\mathcal{B})$ are precisely the frequencies ζ for which $\mathcal{B}$ admits *polynomial exponential trajectories* $p(x)\exp(\langle\zeta, x\rangle)$, p a polynomial function; see, e.g., [18, 37] for a discussion in the behavioral context. Since the associated primes of $\mathcal{M}$ include the primes minimal with respect to the property of including ann $\mathcal{M}$, $\mathcal{V}(\mathcal{B})$ is equal to the union of the varieties of the associated primes of $\mathcal{M}$.

We may further consider the *projective closure* $\overline{\mathcal{V}}$ of a variety $\mathcal{V} = \mathcal{V}(I)$ for some ideal I, which is the smallest projective variety containing $\mathcal{V}$ according to the inclusion $\mathbb{C}^n \to \mathbb{PC}^n$, $\zeta \mapsto (\zeta, 1)$. $\overline{\mathcal{V}}$ is the set of zeros of the homogeneous ideal equal to the set of all homogeneous polynomials in $\mathbb{C}[s, z]$ mapping I under $p(s, z) \mapsto p(s, 1)$. The *variety at infinity* $\tilde{\mathcal{V}}$ is defined as the intersection of $\tilde{\mathcal{V}}$ with the "hyperplane at infinity" $\mathbb{C}^n \times 0 \subseteq \mathbb{PC}^n$ (but $\tilde{\mathcal{V}}$ is regarded as an affine variety in $\mathbb{C}^n$). Since the hyperplane at infinity is defined by the additional equation $z = 0$, $\tilde{\mathcal{V}}$ may easily be seen to equal the set of zeros of the principal parts $pr(p)$ of all $p \in I$, where $pr(p)$ is the sum of all terms of p with the highest total degree. The variety at infinity may be computed, e.g., using Gröbner bases [3, sec. 15.10.5].

A vector $v \in \mathbb{R}^n\backslash 0$ is said to be a *characteristic direction* for $\mathcal{M}$ or for ann $\mathcal{M}$, or for the associated system of PDEs, if $v \in \tilde{\mathcal{V}}(\text{ann } \mathcal{M})$; otherwise it is said to be *noncharacteristic.* Recall now the definition of time-autonomy due to Sasane, Thomas, and Willems [25].

DEFINITION 2.1. *A behavior $\mathcal{B}$ is called* time-autonomous *if any trajectory is determined by its restriction to the half-space $\{x \in \mathbb{R}^n \mid x_n < 0.\}$ The behavior is autonomous if the characteristic variety is not all of $\mathbb{C}^n$.*

Thus for a behavior $\mathcal{B}$ with signal space $\mathcal{D}'$, time-autonomy is equivalent to the absence of nonzero solutions in $\mathcal{D}'_+$, so means that if a trajectory is zero in the "past" ($\mathbb{R}^n_-$) it must remain zero in the "future" ($\mathbb{R}^n_+$).

Nonzero solutions over $\mathcal{D}'_+$ or $\mathcal{C}^\infty_+$ (or more generally in a specified half-space) are *null solutions.* The following result characterizing their existence was first proved by Hormander [6] in the smooth, single polynomial case, and can be found in [15] for the smooth systems case and [16] for the distributional and smooth systems cases.

THEOREM 2.2. *A behavior $\mathcal{B}$ in $(\mathcal{D}')^q$ or in $(\mathcal{C}^\infty)^q$ has no null solutions, i.e., is time-autonomous, if and only if $(0,\dots,0,1)$ is a noncharacteristic direction for the system.*

Note. Recent work [27] has also given a characterization of time-autonomy for a class of 2D systems. Further consideration of how this relates to the above result is left as a topic for future research.

The *(Willems) closure* of a submodule $\mathcal{N}$ of $\mathbb{C}[s]^{1\times q}$ with respect to a signal space $\mathcal{W}$ is defined [20] as

$$\mathcal{N}^{\perp\perp} := \{v' \in \mathbb{C}[s]^{1\times q} \mid v'(\partial)w = 0 \text{ for all } w \in \mathcal{N}^\perp\}, \tag{11}$$

where

$$\mathcal{N}^\perp := \{w \in \mathcal{W}^q \mid v(\partial)w = 0 \text{ for all } v \in \mathcal{N}\}. \tag{12}$$

Notice that $\mathcal{N} \subseteq \mathcal{N}^{\perp\perp}$ and if $\mathcal{N}_1^\perp \subseteq \mathcal{N}_2^\perp$ are two behaviors, then $\mathcal{N}_2^{\perp\perp} \subseteq \mathcal{N}_1^{\perp\perp}$.

2.3. Pole structure. We now recall some results from [37] concerning the pole structure of nD behaviors. The material on input/output structures etc. may also be found in many other places in the literature.

Recall first that a *(free) input/output structure* (x, y) on a behavior $\mathcal{B}$ with a general signal space $\mathcal{W}$ is a partition of the system variables into m input variables u and p output variables y with the properties that

1. the projection of the behavior onto the u variables equals $\mathcal{W}^m$ (we say the variables u are *free over* $\mathcal{W}$), and
2. the *zero-input behavior*

$$\mathcal{B}_{0,y} := \{(u, y) \in \mathcal{B} \mid u = 0\} \tag{13}$$

is autonomous, i.e., has no free variables.

For a given kernel representation, by writing the system equations in the form

$$P(\partial)y = Q(\partial)u,$$

we equivalently have that P has full column rank and the rank of $(-Q, P)$ is equal to the rank of Q. When these conditions apply, there is a unique rational function matrix G with $PG = Q$, called the *transfer matrix.*

The *controllable part* $\mathcal{B}^c$ of $\mathcal{B}$ defined as the (unique) maximal controllable subbehavior of $\mathcal{B}$, possesses the same input/output structures as $\mathcal{B}$ and admits the same transfer matrix with respect to any such input/output structure. We do not define controllability here but refer the reader to [20]. The zero-input behavior $(\mathcal{B}^c)_{0,y}$ of the controllable part has a special structure.

LEMMA 2.3 (see [37, Thm. 5.3]). *Let $\mathcal{B}^c$ be a controllable behavior with given input/output structure (u, y). Let $\mathcal{M}'$ be the system module associated to the zero-input behavior $(\mathcal{B}^c)_{0,y}$. Then the associated primes of $\mathcal{M}'$ are all principal, and the ideal* ann $\mathcal{M}'$ *is generated by the least common denominator of the transfer matrix.*

For convenience, we will call a finitely generated module with the property that its associated primes are all principal a *principal module*; there is no standard term as far as we know.

The factor space $\mathcal{B}/\mathcal{B}^c$ has the structure of an "abstract behavior" as defined in [21]; it may be realized as any behavior of the form $R^c(\partial)(\mathcal{B})$, where R^c is a kernel representation matrix of $\mathcal{B}^c$. The behavior $\mathcal{B}/\mathcal{B}^c$ is autonomous and is the natural analogue of the autonomous part in 1D behavioral theory. It may be identified with the set of all classes of mutually concatenable trajectories in $\mathcal{B}$ [40], and so this behavior may be called the "obstruction to controllability."

The *pole variety*, *controllable pole variety*, and *uncontrollable pole variety* of $\mathcal{B}$ (with a specified input/output structure) are defined, respectively, as $\mathcal{V}(\mathcal{B}_{0,y})$, $\mathcal{V}((\mathcal{B}^c)_{0,y})$, and $\mathcal{V}(\mathcal{B}/\mathcal{B}^c)$. The points of the uncontrollable variety have an interpretation as input decoupling zeros, as discussed in [39]. The three sets are related as follows.

LEMMA 2.4. *We have*

$$\mathcal{V}(\mathcal{B}_{0,y}) = \mathcal{V}((\mathcal{B}^c)_{0,y}) \cup \mathcal{V}(\mathcal{B}/\mathcal{B}^c), \tag{14}$$

$$\widetilde{\mathcal{V}}(\mathcal{B}_{0,y}) = \widetilde{\mathcal{V}}((\mathcal{B}^c)_{0,y}) \cup \widetilde{\mathcal{V}}(\mathcal{B}/\mathcal{B}^c). \tag{15}$$

Proof. Equation (15) is derived in [37] from a standard general result. Equation (16) must also be a consequence of a standard result, but one that we have not found, so we derive it here. Let $\mathcal{M}, \mathcal{M}'$, and $\mathcal{M}''$ denote the system modules corresponding, respectively, to the behaviors $\mathcal{B}_{0,y}, \mathcal{B}/\mathcal{B}^c$, and $(\mathcal{B}^c)_{0,y}$; then $\mathcal{M}' \subseteq \mathcal{M}$ with factor $\mathcal{M}''$, and it is straightforward from this to see that

$$\operatorname{ann} \mathcal{M} \subseteq \operatorname{ann} \mathcal{M}' \cap \mathcal{M}'' \subseteq \operatorname{rad} \operatorname{ann} \mathcal{M}, \tag{16}$$

where rad denotes the radical of an ideal. Write

$$J := \{pr(d) \mid d \in \operatorname{ann} \mathcal{M}\}$$

and define J' and J'' analogously with respect to $\mathcal{M}'$ and $\mathcal{M}''$, respectively. Clearly, $J \subseteq J' \cap J''$. Moreover, if $r \in J' \cap J''$, say, $r = pr\ d_1 = pr\ d_2$ with $d_1 \in \operatorname{ann} \mathcal{M}'$, $d_2 \in \operatorname{ann} \mathcal{M}''$, then we find that $d_1 d_2 \in \operatorname{ann} \mathcal{M}$. We also see that for any two polynomials $p, q \in \mathbb{C}[s]$, $pr(p) \cdot pr(q) = pr(pq)$, and so $r^2 = pr(d_1 d_2) \in J$, and so $r \in \operatorname{rad} J$. This proves that

$$J \subseteq J' \cap J'' \subseteq \operatorname{rad} J.$$

Consequently, $\mathcal{V}(J) = \mathcal{V}(J') \cup \mathcal{V}(J'')$. However, $\mathcal{V}(J) = \tilde{\mathcal{J}}(\mathcal{B}_{0,y})$, etc., so we have proved (16). ∎

3. Stability of autonomous behaviors. In this section we consider an autonomous behavior $\mathcal{B}$ given by a kernel representation matrix R, which necessarily has full column rank q. Furthermore, we assume that one of the independent variables "time" (t) is distinguished; without loss of generality we will always take this to be the last variable listed in the coordinate system for $\mathbb{R}^n$. Under what conditions should $\mathcal{B}$ be referred to as a "stable" behavior?

Stability in this context should mean that $\mathcal{B}$ contains no physically reasonable trajectories which grow in time at an unacceptably fast rate in some sense (e.g., which are unbounded). We might call this "stability with respect to the initial conditions."

In [20, 30], $\mathcal{B}$ was defined to be *stable with respect to a cone* $C \subseteq \mathbb{R}^n$ if every smooth trajectory of $\mathcal{B}$ tends to 0 along every half-line in C. This is characterized in [20, 30], subject to some minor assumptions, by the condition that no projection onto the real space of $\mathcal{V}(\mathcal{B})$ should lie in the polar cone of C with positive distance from the boundary of this polar.

Let us consider the heat or diffusion equation in one spatial variable

$$\frac{\partial^2 w}{\partial x^2} = \frac{\partial w}{\partial t}. \tag{17}$$

This system was used recently by Sasane in [26, 24] to motivate an alternative signal space to $\mathcal{C}^\infty$, $\mathcal{D}'$; here we will consider it in a similar spirt. We find that the characteristic variety of the system (17) or of its behavior $\mathcal{B}$, is

$$\mathcal{V}(\mathcal{B}) = \{(\eta, \xi) \in \mathbb{C}^2 \mid \eta^2 = \xi\},$$

and the behavior contains trajectories of the form

$$\exp(\Re(\eta)x + \Re(\eta^2)t)\exp(\imath\Im(\eta)x + \imath\Im(\eta^2)t)$$

for all $\eta \in \mathbb{C}$. Hence $\mathcal{B}$ contains trajectories which are unbounded on the positive t-axis, corresponding to the choices $\Re(\eta^2) > 0$, and so is unstable in the sense introduced in [20, 30]. However, note that if $\Re(\eta^2) > 0$, then $\Re(\eta) \neq 0$; i.e., any solution which is unbounded on the $+t$-axis is also unbounded (indeed, exponentially growing) on the x-axis. In other words, the only way to get unbounded temporal behavior in this system is to start with exponentially growing initial spatial data! Indeed, we would prefer to consider the heat equation as "stable"; with no external input of heat, heat should diffuse in time and never blow up. In this paper, we take the view that the initial data and trajectories which are exponentially growing spatially are physically unrealistic. With these considerations in mind we introduce the following concept.

DEFINITION 3.1. *An autonomous behavior* $\mathcal{B}$, *or its characteristic variety* $\mathcal{V} = \mathcal{V}(\mathcal{B})$, *is said to* satisfy the CV condition *if*

$$\mathcal{V} \cap \mathcal{X}^+ = \emptyset, \mid \mathcal{X}^+ := \imath\mathbb{R}^{n-1} \times \overline{\mathbb{C}^+}, \tag{CV}$$

where $\overline{\mathbb{C}^+}$ *denotes the closed right-half plane. We say that* $\mathcal{B}$ *or* $\mathcal{V}$ *satisfies* the weak CV (WCV) condition *if the same holds but for the open right-half plane* $\mathbb{C}^+$ *instead of* $\overline{\mathbb{C}^+}$. *We also say that a polynomial or ideal* satisfies the CV condition *or* satisfies the WCV condition *if the corresponding condition is satisfied by the variety of the polynomial/ideal.*

For later use we also define

$$\mathcal{X}^- := \imath\mathbb{R}^{n-1} \times \overline{\mathbb{C}^-}, \tag{18}$$

where $\overline{\mathbb{C}^-}$ denotes the closed left-half plane.

Recalling the description of the points of the characteristic variety in terms of polynomial exponential trajectories, we note that a behavior satisfies the CV condition if and only if it contains no polynomial exponential trajectories which are bounded at $t = 0$ (corresponding to the spatial frequency components being imaginary) but which do not decay along the positive t-axis (corresponding to the temporal frequency components being in $\overline{\mathbb{C}^+}$). This observation applies equally well to both complex and real-valued trajectories. We therefore think of points of $\mathcal{X}^+$ as *unstable frequencies.*

Similarly, a behavior satisfies the WCV condition if and only if it contains no polynomial exponential trajectories which are bounded at $t = 0$ but grow faster than a polynomial in the $+t$-direction. Note that the behavior defined by the heat equation certainly satisfies the WCV condition, as if η is imaginary, then $\Re(\eta^2) < 0$ gives rise to a trajectory which is exponentially decaying in time.

As a working definition, we consider an autonomous behavior to be stable when it satisfies the CV condition. This attempts to capture the idea that a behavior is unstable when it contains trajectories which are well behaved at $t = 0$ but do not decay to 0 as $t \to +\infty$. The concept of "stability with respect to initial conditions" also lies behind the definition of stability for 2D discrete systems given by Valcher [33] (there is little parallel in the mathematics of the continuous and discrete cases, but the underlying philosophy is the same). As we have seen, in terms of polynomial exponential trajectories Definition 3.1 seems very appropriate, and by the same considerations the CV condition is certainly necessary for stability in the general sense we have been indicating. Moreover, since the polynomial exponential solutions of a system are dense in the smooth solutions, at least in the case of complex coefficients [7, Thm. 7.6.14], it is not unreasonable to conjecture that the CV condition is also sufficient for this type of stability. However, this important problem remains open, and we rely on later sections to fully motivate the CV condition.

Let us cover the other basic classical examples for general n. We have seen that the behavior of the heat equation satisfies the WCV condition, though it does not satisfy the CV condition; this holds in any number of spatial dimensions. This behavior of the wave equation and of the gradient operator also satisfy the WCV condition but not the CV condition. The kernel of the Laplace operator does not satisfy either.

3.1. Hyperbolic systems. It is also interesting to note that the WCV condition implies the *Gårding condition,* which is necessary for hyperbolicity of an autonomous system given by a single polynomial. Here is the condition,

$$\{\Re(\xi) \mid \exists \eta \in \imath\mathbb{R}^{n-1},\ \xi, \eta \in \mathcal{V}\} \subseteq \mathbb{R} \text{ is bounded above.} \tag{19}$$

We now discuss hyperbolicity, giving the definition for the systems case which is more complex than the better known definition for a single polynomial. The following definition is identical to one of the equivalent definitions given by Nacinovich [16], adjusted only in respect of the fact that our systems are defined via $P(\partial)w = 0$, whereas he uses the more standard $P((1/\imath)\partial)w = 0$. Also we have specialized the definition to hyperbolicity in a fixed direction.

DEFINITION 3.2. *A system $P(\partial)w = 0$, operator $P(\partial)$, associated system module $\mathcal{M}$, or behavior $\mathcal{B}$, is called* hyperbolic (in the direction t) *if for every associated prime I of $\mathcal{M}$ we can find a constant $0 < c < 1$ such that (where $\Re(\eta, \xi)$ denotes the real part vector of the complex $n + 1$ tuple (η, ξ))*

$$\Re(\xi) \leq c|\Re(\eta, \xi)| + c^{-1} \text{ for every } (\eta, \xi) \in \mathcal{V}(I). \tag{20}$$

The following result, also adapted from [16], links the definition to the more familiar one for a single polynomial.

THEOREM 3.3. *Let $P \in \mathbb{C}[s]^{g\times q}$, and let $\mathcal{M}$ be the system module, i.e., $\mathcal{M} = \mathbb{C}[s]^{1\times q}/\mathbb{C}[s]^{1\times g}P$. Suppose that $\mathcal{M}$ is principal. Then $P(\partial)$ is hyperbolic if and only if $(0, \ldots, 0, 1)$ is a noncharacteristic direction for the system $P(\partial)y = 0$, and the Gårding condition* (20) *holds for the characteristic variety of the system.*

Proof. We refer to the remark following [16, Prop. 6.1], which states that when the associated primes of $\mathcal{M}$ are all principal, hyperbolicity is equivalent to the requirement

that $(0, \dots, 0, 1)$ be noncharacteristic, together with the Gårding condition for each variety $\mathcal{V}(I)$, I an associated prime of $\mathcal{M}$. Since the number of associated primes is finite, the latter condition is, however, equivalent to the Gårding condition for the characteristic variety itself. □

It is easy to see that the conditions of Theorem 3.3 hold in particular when P is a single polynomial, as expected. Note from Lemma 2.3 that this condition is also met for the module corresponding to the zero-input behavior of any controllable behavior.

We remark that hyperbolicity is equivalent to solvability of the "noncharacteristic" Cauchy problem in many different formulations [9, 10], which is of great importance and deserves investigation in the context of control systems theory. Essentially, hyperbolicity allows the unique continuation of initial data in a large class on $t = 0$ to trajectories on the half-space $t \geq 0$. We will note in the next section its connections to causality.

Note also that hyperbolic behaviors are in particular time-autonomous (in the general case this is a consequence of Theorem 4.2 in the next section). Next, we link the WCV condition to hyperbolicity.

LEMMA 3.4. *Let $P(\partial)$ be a partial differential operator with kernel $\mathcal{B}$ and system module $\mathcal{M}$. Suppose that $\mathcal{M}$ is principal and that $\mathcal{B}$ satisfies the WCV condition and is time-autonomous. Then the system is hyperbolic and therefore admits a solution to the noncharacteristic Cauchy problem.*

In the case where $\mathcal{M}$ is principal, the CV property, together with time-autonomy, is of course a much stronger property than hyperbolicity; for example, in two dimensions the kernel of the operator $(\partial/\partial t - 1)$ is hyperbolic but does not satisfy the WCV condition. The relationship between these two properties will become clearer when we examine stable input/output structures. This, however, will require us to consider causality in the continuous space-time input/output framework.

4. Causal input/output structures. We are interested in this section with the question of when a given input/output structure is causal. Following Zampieri [38] for the discrete case (in which the past and future are defined with respect to a cone), we introduce the following definition of causality.

DEFINITION 4.1. *Suppose that (u, y) is an input/output structure on $\mathcal{B}$ and $\mathcal{B}_{0,y}$ is time-autonomous. Then the input/output structure is said to be* causal (with respect to $\mathcal{C}^\infty$) *if for any smooth input u with support in $\mathbb{R}^n_+$, there exists a smooth output y (necessarily unique) with support in $\mathbb{R}^n_+$, such that $(u, y) \in \mathcal{B}$.*

Causality with respect to $\mathcal{D}'$ may also be defined in the obvious way; where no solution type is specified, we take causality to be meant in the smooth sense.

Thus causality indicates that for any input u with zero past, there is a corresponding output y (intuitively, the output corresponding to zero initial conditions) with a past which is determined by that of u, and is therefore also zero. Notice that this interpretation of Definition 4.1 only makes complete sense if the complementary half-space $\mathbb{R}^n_-$ contains the domain of a complete set of Cauchy data for $\mathcal{B}_{0,y}$, and this is the reason for insisting a priori that $\mathcal{B}_{0,y}$ be time-autonomous. Time-autonomy of $\mathcal{B}_{0,y}$ clearly means that an output trajectory is entirely determined by its value in $\mathbb{R}^n_-$, together with the input trajectory. Physical systems should of course always be causal in the intuitive sense, but not all autonomous behaviors arising in the physical context are time-autonomous. For example, the behavior defined by the heat equation (18) is not time-autonomous, though in a more restricted and perhaps physically better motivated sense, it is (see [24]). A general solution of the heat equation is, however, not uniquely defined by its past!—essentially because an input may be supplied via

a boundary condition. For time-autonomous $\mathcal{B}_{0,y}$, however, causality in the sense of Definition 4.1 captures the intuitive concept.

Hyperbolicity is intimately connected to causality. To demonstrate this, we give some results essentially taken from [16]; these results are very well known in the PDE community for the case of a single polynomial operator (see, e.g., [9, Thm. 12.5.4]), in which case the matrix L below is 0.

THEOREM 4.2. *Let $P \in \mathbb{C}[s]^{g\times q}$ with universal left annihilator matrix $L \in \mathbb{C}[s]^{h\times g}$. The following are equivalent:*

1. *$P(\partial)$ is hyperbolic.*
2. *The system*

$$P(\partial)y = u \tag{21}$$

has a unique solution $y \in (\mathcal{C}_+^\infty)^p$ for all $u \in \ker_{\mathcal{C}_+^\infty} L$.

3. *The system (21) has a unique solution $y \in (\mathcal{D}'_+)^p$ for all $u \in \ker_{\mathcal{D}'_+} L$.*

Proof. In [16, Thm. 5.1], it is stated that $P(\partial)$ is hyperbolic if and only if $\operatorname{Ext}^i_{\mathbb{C}[s]}(\mathcal{M}, \mathcal{C}_+^\infty) = 0$, $i = 0, 1$. For $i = 0$ this means that $P(\partial)$ admits no smooth null solutions; for $i = 1$ it means that the sequence

$$(\mathcal{C}'_+)^p \overset{P(\partial)}{\to} (\mathcal{C}_+^\infty)^g \overset{L(\partial)}{\to} (\mathcal{C}_+^\infty)^h$$

is exact, which together with condition 2 [16, Thm. 5.2] gives the result for distributions, which establishes the equivalence of 1–3. ∎

Notice that Theorem 4.2 in particular gives a restricted form of the "fundamental principle"—over $\mathcal{C}_+^\infty$ or $\mathcal{D}'_+$ when P (or its module) is hyperbolic, the system (21) has a solution y for given u if and only if u satisfies the necessary "compatibility conditions." Indeed, the results of Nacinovich characterize hyperbolicity in terms of the vanishing of $\operatorname{Ext}^i_{\mathbb{C}[s]}$, as indicated in the proof above.

Moreover, at least in the case of a single polynomial [9, Thm. 12.5.4], hyperbolicity guarantees the existence of a fundamental solution to the system (21) having support in $\mathbb{R}^n_+$, and in both smooth and distributional cases the causal input-to-output map is given simply by convolution with the fundamental solution, which in the current context we can also refer to as the impulse response. As discussed in [30], hyperbolic systems are in this sense, therefore, the natural analogue of standard 1D (lumped) systems.

In [30], Shankar considers causality in a slightly different sense, namely to mean the existence of a continuous linear shift-invariant map from the set of smooth inputs with support in some cone contained in $\mathbb{R}^n_+$, to the set of outputs of the same type. For the system (21) with P equal to a single polynomial p, he proves that the existence of such a causal input-to-output map guarantees time-autonomy of the zero-input behavior and thereby (using Theorem 4.2) hyperbolicity of p also.

It is easy to see that hyperbolicity is still sufficient for results of this type when a Q term is added to the equations.

COROLLARY 4.3. *Let $\mathcal{B}$ be a behavior defined by the equations*

$$P(\partial)y = Q(\partial)u$$

forming an input/output structure. If P is a hyperbolic operator, then $\mathcal{B}_{0,y}$ is time-autonomous and the input/output structure is causal.

Proof. Suppose that P is hyperbolic. Time-autonomy of $\mathcal{B}_{0,y}$ is immediate from the uniqueness of solutions y given in Theorem 4.2, and causality follows from this result also. ∎

It is open to discussion as to whether the converse of Corollary 4.3 holds. The next lemma shows that under a simple assumption, all the trajectories of $\mathcal{B}$ with support in $\mathcal{D}'_+$ are contained in the controllable part. We will then use this reasoning to demonstrate that $\mathcal{B}$ and $\mathcal{B}^c$ have the same causal input/output structures.

LEMMA 4.4. *Let $\mathcal{B}$ be a behavior in $(\mathcal{D}')^q$. If $\mathcal{B}/\mathcal{B}^c$ is time-autonomous, then $\mathcal{B} \cap (\mathcal{D}')^q = \mathcal{B}^c \cap (\mathcal{D}'_+)^q$.*

Proof. Suppose that $\mathcal{B}/\mathcal{B}^c$ is time-autonomous. The inclusion $\supseteq$ is trivial. Now suppose $w \in \mathcal{B} \cap (\mathcal{D}'_+)^q$, and let R^c be a kernel representation matrix of $\mathcal{B}^c$. Since $\mathcal{B}/\mathcal{B}^c$ is time-autonomous, $R^c(\partial)w \in R^c(\partial)(\mathcal{B}) \cong \mathcal{B}/\mathcal{B}^c$ must vanish, so $w \in \mathcal{B}^c$. ∎

An analogous argument may be used to generalize Zampieri's Lemma 3.3 and, hence, other results in [38] from the 2D to nD discrete case.

COROLLARY 4.5. *Let $\mathcal{B}$ be a behavior with controllable part $\mathcal{B}^c$ and a given input/output structure (u, y) (which is necessarily an input/output structure on $\mathcal{B}^c$ also). Then the following hold:*

1. *$\mathcal{B}_{0,y}$ is time-autonomous if and only if both $\mathcal{B}/\mathcal{B}^c$ and $(\mathcal{B}^c)_{0,y}$ are.*
2. *Under the equivalent conditions of claim 1, (u, y) is a causal input/output structure on $\mathcal{B}$ if and only if it is a causal input/output structure on $\mathcal{B}^c$.*

Proof.

1. This claim is immediate from Theorem 2.2 together with Lemma 2.4.
2. Suppose that the conditions of claim 1 hold. If (u, y) is causal for $\mathcal{B}^c$, then it is trivial that it is also causal for $\mathcal{B}$, and the converse follows directly from Lemma 4.4. ∎

Corollary 4.3, Lemma 4.4, and Corollary 4.5 apply to causality in both the smooth and distributional senses.

One consequence of Corollary 4.5 is that when causality of an input/output structure is defined (i.e., when the zero-input behavior is time-autonomous), whether or not it holds is determined purely by the controllable part of the behavior and therefore by the transfer matrix. This motivates the following definition.

DEFINITION 4.6. *Call a transfer matrix G* causal *if its least common denominator is hyperbolic,* stable *if this polynomial obeys the* CV *condition, and* weakly stable *if this polynomial obeys the WCV condition.*

Notice that for $n = 1$, stability of G agrees with the classical concept, and causality of G is automatic.

Suppose we are given a behavior $\mathcal{B}$ with input/output structure (u, y) and transfer matrix G. Due to Lemma 2.3, G is causal if and only if $(\mathcal{B}^c)_{0,y}$ is hyperbolic. Similarly, G is stable if and only if $(\mathcal{B}^c)_{0,y}$ obeys the CV condition. The following new result shows that causality of G corresponds to causality of the corresponding input/output structures when the latter are defined.

THEOREM 4.7. *Suppose that $\mathcal{B}$ is a behavior with an input/output structure such that $\mathcal{B}_{0,y}$ is time-autonomous. Then the input/output structure is causal with respect to $\mathcal{C}^\infty$ if and only if the associated transfer matrix is causal. These conditions imply that the input/output structure is causal with respect to $\mathcal{D}'$.*

Proof. Let $\mathcal{B}$ be given with $\mathcal{B}_{0,y}$ time-autonomous and let $P(\partial)y = Q^c(\partial)u$ be a description of the controllable part of $\mathcal{B}^c$. Write $\mathcal{M}$ for the system module of the behavior $\mathcal{B}^c$ and $\mathcal{M}'$ for the system module of the operator $P^c(\partial)$, corresponding to the behavior $(\mathcal{B}^c)_{0,y}$. We have an exact sequence (e.g., [37])

$$0 \to F \to \mathcal{M} \to \mathcal{M}' \to 0$$

for a free submodule F of $\mathcal{M}$, with rank equal to the number of inputs m. From this

we obtain the long exact sequence in Ext (see, e.g., [3]),

$$0 \to \mathrm{Hom}_{\mathbb{C}[s]}(\mathcal{M}', \mathcal{C}_+^\infty) \to \mathrm{Hom}_{\mathbb{C}[s]}(\mathcal{M}, \mathcal{C}_+^\infty) \xrightarrow{\rho} \mathrm{Hom}_{\mathbb{C}[s]}(F, \mathcal{C}_+^\infty)$$

$$\to \mathrm{Ext}^1_{\mathbb{C}[s]}(\mathcal{M}', \mathcal{C}_+^\infty) \to \mathrm{Ext}^1_{\mathbb{C}[s]}(\mathcal{M}, \mathcal{C}_+^\infty) \to 0, \tag{22}$$

the last "0" term occurring since F is free. Recall that $\mathrm{Hom}_{\mathbb{C}[s]}(\mathcal{M}', \mathcal{C}_+^\infty)$ is identified with the $\mathcal{C}_+^\infty$ solutions in $(\mathcal{B}^c)_{0,y}$ and $\mathrm{Hom}_{\mathbb{C}[s]}(\mathcal{M}, \mathcal{C}_+^\infty)$ with the $\mathcal{C}_+^\infty$ solutions in $\mathcal{B}^c$. Also, $\mathrm{Hom}_{\mathbb{C}[s]}(F, \mathcal{C}_+^\infty) = (\mathcal{C}_+^\infty)^m$ and the map ρ is the projection map $(u, y) \in (\mathcal{C}_+^\infty)^{m+p} \mapsto u$. Since $\mathcal{B}^c$ is controllable, $\mathcal{M}$ is torsionfree [20]. Furthermore, we have an exact sequence

$$0 \to \mathcal{C}_+^\infty \to \mathcal{C} \to \mathcal{C}_\ominus^\infty \to 0.$$

Note that $\mathcal{C}_\ominus^\infty$ is the set of restrictions of smooth functions to the set $\mathbb{R}^n_-$; it is therefore the direct limit as $\epsilon \mapsto 0$ of the sets $\mathcal{C}^\infty(H_\epsilon)$ on $H_\epsilon = \{x \in \mathbb{R}^n,\ x_n < \epsilon\}$ for $\epsilon > 0$. Since H_ϵ is convex, $\mathcal{C}^\infty(H_\epsilon)$ is known to be an injective module (e.g., [19, Cor. VII.8.4]), and now $\mathcal{C}_\ominus^\infty$ is injective as the direct limits of injectives (e.g., [11, Thm. (3.46)]). Therefore $\mathcal{C}_+^\infty$ has injective dimension 1, and so, $\mathcal{M}$ being torsionfree, $\mathrm{Ext}^1_{\mathbb{C}[s]}(\mathcal{M}, \mathcal{C}_+^\infty) = 0$ by [36, Thm. 4.8]. From (22), we now see that the cokernel of ρ equals $\mathrm{Ext}^1_{\mathbb{C}[s]}(\mathcal{M}', \mathcal{C}_+^\infty)$.

Suppose now that (x, y) is causal for $\mathcal{B}$. Then by Corollary 4.5, it is also causal for $\mathcal{B}^c$, so the map ρ is surjective, due to which $\mathrm{Ext}^1_{\mathbb{C}[s]}(\mathcal{M}', \mathcal{C}_+^\infty) = 0$. Also, as $\mathcal{B}_{0,y}$ is time-autonomous, $(\mathcal{B})^c_{0,y}$ is time-autonomous by Corollary 4.5, and so $\mathrm{Ext}^0_{\mathbb{C}[s]}(\mathcal{M}', \mathcal{C}_+^\infty) = 0$ also. By Theorem 4.2, $P^c(\partial)$ is hyperbolic, which means that G is causal.

Conversely, suppose that G is causal. As observed preceding the theorem, $P^c(\partial)$ is hyperbolic and so $\mathrm{Ext}^1_{\mathbb{C}[s]}(\mathcal{M}', \mathcal{C}_+^\infty)$ vanishes by Theorem 4.2. Thus the map ρ is surjective, so the variables u are free over $\mathcal{C}_+^\infty$, i.e., (u, y) is causal for $\mathcal{B}^c$ and so for $\mathcal{B}$ by Corollary 4.5. This converse argument also applies to distributional solutions. □

In particular, Theorem 4.7 establishes that causality of a given input/output structure may be tested (when it is defined) merely by looking at the least common denominator d of the transfer matrix G. In fact, since the prior condition of time-autonomy enforces that $(\mathcal{B}^c)_{0,y}$ be time-autonomous (by Corollary 4.5) and therefore that $(0, \ldots, 0, 1)$ be noncharacteristic for d, we have that (u, y) is causal if and only if d satisfies the Gårding condition. Unfortunately, it is not immediately clear how the condition may be tested. One possibly tractable necessary and sufficient condition is given in [16, Prop. 6.1]; $\mathcal{M}$ is hyperbolic if and only if $(0, \ldots, 0, -1)$ does not appear in any of the asymptotic cones (at infinity) of the real parts of the varieties of the associated primes of $\mathcal{M}$; for brevity we omit a precise description.

We remark also that a well-known necessary condition for hyperbolicity of d is that the principal part of $pr(d)$ of d itself be hyperbolic (e.g., [9, Thm. 12.4.2]). The Gårding condition on $pr(d)$ is equivalent to requiring that the roots ζ of $pr(d)(\nu, \zeta)$ all be real for any real $\nu \neq 0$ (e.g., [9, Thm. 12.4.3]). Any further analysis of computational issues is outside the scope of this paper.

5. Convolution operators and ideal convexity. Before we tackle the subject of stable input/output structures, we need to explore two areas of background material. This first is convolution operators and Fourier transforms on the classical and other related spaces, as developed by Gindikin and Volevich [4]. This will lead us to a necessary and sufficient condition for input/output stability of the system

$p(\partial)y = u$. The second area is "ideal-convexity" of a region in complex space, which is a necessary property for the extension of certain results from polynomials to ideals (and thereby general systems).

5.1. Convolution operators on $\mathcal{S}, \mathcal{S}'$. The following material is largely taken from [4, secs. 1.1–1.2]. For $s \in \mathbb{N}$ and $l \in \mathbb{R}$, let $\mathcal{C}^{(s)}_{(l)}$ denote the space of s-times continuously differentiable functions f on $\mathbb{R}^n$ with finite Hölder norm

$$|f|^{(s)}_{(l)} := \sup_{x\in\mathbb{R}^n, \alpha\in\mathbb{N}^n, |\alpha|\le s} (1+x^2)^{l/2} |\partial^\alpha f(x)|. \tag{23}$$

Here $|\alpha|$ denotes the total of the components of α, and δ^α is a shorthand for the operator $p(\partial)$, where $p = (s_1^{\alpha_1} s_2^{\alpha_2}, \ldots, s_n^{\alpha_n})$. The following elementary lemma is not in [4] but will prove useful.

LEMMA 5.1. *If $f, g \in \mathcal{C}^{(s)}_{l/2}$ for some s, l, then $fg \in \mathcal{C}^{(s)}_{l}$.*

Proof. Suppose that f and g are in $\mathcal{C}^{(s)}_{(l/2)}$. We have

$$\sup_{x\in\mathbb{R}^n,\ |\alpha|\le s} (1+x^2)^{l/2} \, |\partial^\alpha (fg)(x)| = \sup_{x\in\mathbb{R}^n,\ |\alpha|\le s} (1+x^2)^{l/2} \left| \sum_{i\le\alpha} \beta_{i,\alpha} (\partial^i f)(x) (\partial^{\alpha-i} g)(x) \right|,$$

where the sum on the right-hand side is taken over all multi-indices i which are, componentwise, less than or equal to α, and $\beta_{i,\alpha}$ are constants depending only on i, α. Hence we obtain

$$\begin{aligned}
\sup_{x\in\mathbb{R}^n,\ |\alpha|\le s} (1+x^2)^{l/2} \, |\partial^\alpha (fg)(x)| &\le \sup_{x\in\mathbb{R}^n,\ |\alpha|\le s} \sum_{i\le\alpha} (1+x^2)^{l/2} |\beta_{i,\alpha}| \, |(\partial^i f)(x)| \, |(\partial^{\alpha-i} g)(x)| \\
&\le \max_{|\alpha|\le s} \sum_{i\le\alpha} |\beta_{i,\alpha}| \left(\sup_{x\in\mathbb{R}^n} (1+x^2)^{l/4} |(\partial^i f)(x)| \right) \\
&\quad \times \left(\sup_{x\in\mathbb{R}^n} (1+x^2)^{l/4} |(\partial^{\alpha-i} g)(x)| \right),
\end{aligned}$$

which is finite since for any α the sum given is a finite linear combination of terms which are finite by supposition. Therefore $fg \in \mathcal{C}^{(s)}_{(l)}$ as required. ∎

Recall that $\mathcal{S}$ is defined as the intersection of all the spaces $\mathcal{C}^{(s)}_{(l)}$. Also of interest is the set

$$\mathcal{L} := \bigcap_s \bigcup_l \mathcal{C}^{(s)}_{(l)} \tag{24}$$

which is the set of all smooth functions, each derivative of which grows no faster than some power of x (which power may depend on the derivative), and

$$\mathcal{O} := \bigcup_l \bigcap_s \mathcal{C}^{(s)}_{(l)} \tag{25}$$

the set of all smooth functions, each derivative of which grows no faster than some power of x which is independent of the derivative. ("$\mathcal{M}$" is used rather than "$\mathcal{L}$" in [4].) $\mathcal{O}$ may be thought of as the set of (at most) slowly growing smooth functions; its dual space is denoted by $\mathcal{O}'$ and may be thought of as the space of rapidly decreasing distributions. Clearly $\mathcal{S} \subseteq \mathcal{O} \subseteq \mathcal{L} \subseteq \mathcal{C}^\infty$. We collect some basic facts and results from [4, sec. 1.1].

LEMMA 5.2.

1. *$\mathcal{L}$ is a ring with respect to multiplication.*
2. *$\mathcal{S}$ is closed under multiplication by elements of $\mathcal{L}$.*
3. *$\mathcal{F}(\mathcal{S}) = \mathcal{S}$, $\mathcal{F}(\mathcal{S}') = \mathcal{S}'$, and $\mathcal{F}(\mathcal{O}') = \mathcal{L}$, where $\mathcal{F}$ denotes Fourier transform.*
4. *Let p be a polynomial. The equation*

$$p(\partial)y = u$$

is uniquely solvable for $y \in \mathcal{S}$ (resp., $\mathcal{O}$, $\mathcal{S}'$, $\mathcal{O}'$) for any $u \in \mathcal{S}$ (resp., $\mathcal{O}$, $\mathcal{S}'$, $\mathcal{O}'$), if and only if p has no imaginary roots.
5. *Any polynomial p is in $\mathcal{L}$, and the function $p(\imath\zeta)$ of ζ has an inverse in $\mathcal{L}$ if and only if p has no imaginary roots.*

Proof. The last claim is not given explicitly in [4]; note that polynomials are contained in $\mathcal{O} \subseteq \mathcal{L}$. That the inverse of $p(\imath\zeta)$ of ζ is in $\mathcal{L}$ if and only if p has no imaginary roots is implicit in the argument of [4, sec. 1.1.5]; we provide an argument based on their explicit results. By claim 4, the polynomial p has no imaginary roots if and only if the convolutional equation $p(\partial)y = u$ is uniquely solvable for $y \in \mathcal{O}'$ for any $u \in \mathcal{O}'$. By [4, Thm. 1.1.5], this holds if and only if there is a fundamental solution $E \in \mathcal{O}'$ with $(p(\partial)\delta) * E = E * (p(\partial)\delta) = \delta$, where δ is the Dirac delta and $*$ denotes convolution. By Fourier transform (use claim 3), this is equivalent to invertibility of $p(\imath\zeta)$ in $\mathcal{L}$. ∎

5.2. Convolution operators on $\mathcal{S}_+$, $\mathcal{S}'_+$. Still following [4], we now introduce the spaces $\mathcal{C}^{(s)}_{(l)+}$ of all functions in $\mathcal{C}^{(s)}_{(l)}$ with support in $\mathbb{R}^n_+$ and define $\mathcal{S}_+$, $\mathcal{S}'_+$, $\mathcal{S}'_\oplus$, etc., as before. $\mathcal{O}_+$ and $\mathcal{O}'_+$ are defined analogously. We have that $\mathcal{S}'_\oplus$ is the dual space of $\mathcal{S}_+$ and $\mathcal{S}'_+$ is dual to $\mathcal{S}_\oplus$.

Now for any Banach space B of functions $(\nu, \sigma) \in \mathbb{R}^{(n-1)+1}$ with norm $\phi \mapsto |\phi|_B$, denote by B^+ the space of functions f of $(\nu, \xi) \in \mathbb{R}^{n-1} \times \mathbb{C}$, $\xi = \sigma + \imath\rho$, with the following properties:

1. For each $\rho \leq 0$, the functions $f_\rho = f(\cdot, \cdot + \imath\rho)$ are in B, and the map $(-\infty, 0] \mapsto B$, $\rho \mapsto f_\rho$ is continuous.
2. For each $\nu \in \mathbb{R}^{n-1}$, the functions $f_\nu = f(\nu, \cdot)$ are functions holomorphic in $\mathbb{C}_-$.
3. The norm $\sup_{\rho<0} |f_\rho|_B$ is finite.

A space B^- may be defined analogously by changing the sign of ρ in conditions 1 and 3 and changing $\mathbb{C}_-$ for $\mathbb{C}_+$ in condition 2. Note that if $f(s) \in B^+$, then $F(-s) \in B^-$ and vice versa, provided that B is preserved by the same operation.

Now we define

$$\mathcal{S}^+ := \bigcap_{s,l} \mathcal{C}^{(s)+}_{(l)}, \tag{26}$$

$$\mathcal{L}^+ := \bigcap_{s} \bigcup_{l} \mathcal{C}^{(s)+}_{(l)}. \tag{27}$$

Spaces $\mathcal{S}^-$ and $\mathcal{L}^-$ may be defined analogously, and again if $f(s) \in \mathcal{L}^+$, then $f(-s) \in \mathcal{L}^-$ and vice versa.

The interest in these spaces comes from the following collection of points, from [4] except where proof is given below.

LEMMA 5.3.

1. $\mathcal{L}^+$ *is closed under multiplication.*
2. $\mathcal{S}^+$ *is closed under multiplication by elements of* $\mathcal{L}^+$ *and this multiplication rule is associative.*
3. $\mathcal{F}(\mathcal{S}_+) = \mathcal{S}^+$ *and* $\mathcal{F}(\mathcal{O}'_+) = \mathcal{L}^+$.
4. *Let* p *be a polynomial. The equation*

$$p(\partial)y = u$$

is uniquely solvable for $y \in \mathcal{S}'_+$ *for any* $u \in \mathcal{S}'_+$ *if and only if* p *has no roots in* $\mathcal{X}^+ = \mathbb{R}^{n-1} \times \overline{\mathbb{C}_+}$.
5. *The equation*

$$p(\partial)y = u$$

is uniquely solvable for $y \in \mathcal{S}'_+$ *for any* $u \in \mathcal{S}'_+$ *if and only if* p *has no roots in* $\mathbb{R}^{n-1} \times \overline{\mathbb{C}_+}$.
6. *Any polynomial* p *is in* $\mathcal{L}^+$ *and the function* $1/p(\imath\zeta)$ *of* ζ *is in* $\mathcal{L}^+$ *if and only if* p *has no roots in* $\mathcal{X}^+$.

Proof.

1. Note that multiplication is defined since the elements of $\mathcal{L}^+$ are functions. It follows from Lemma 5.1 that if $f, g \in \mathcal{C}^{(s)+}_{(l/2)}$ for some $l \in \mathbb{R}$, then $fg \in \mathcal{C}^{(s)+}_{(l/2)}$. Hence if $f, g \in \mathcal{L}^+$, then for any s there exists l with $f, g \in \mathcal{C}^{(s)+}_{(l)}$, and so $fg \in \mathcal{C}^{(s)+}_{(l)}$. Therefore $fg \in \mathcal{L}^+$.
2. The first point is given in [4, sec. 1.2.2], and the second is immediate as $\mathcal{L}^+$, $\mathcal{S}^+ \subseteq \mathcal{C}^\infty$.
3. These identities are (21) and (24) in [4, sec. 1].
4. This is Theorem 2(i) in [4, sec. 1.2.5].
5. This is Theorem 2(ii) in [4, sec. 1.2.5].
6. That polynomials are in $\mathcal{L}^+$ follows from claim 3, as polynomials are Fourier transforms of distributions with support at 0, which are therefore in $\mathcal{O}'_+$. Also from [4, Thm. 2(i), sec. 1.2.5], p has no roots in $\mathcal{X}^+$ if and only if $p(\partial)y = u$ is uniquely solvable over $\mathcal{O}'_+$. By [4, Thm. 1, sec. 1.2.5], this is equivalent to the condition that $p(\imath\zeta)^{-1} \in \mathcal{F}(\mathcal{O}'_+) = \mathcal{L}^+$. □

Clearly all the claims of Lemma 5.3 can be "time-reversed" to give corresponding results for $\mathcal{L}^-, \mathcal{S}_-, \mathcal{X}^-$, etc.

Claims 4 and 5 of Lemma 5.3 are our first input/output stability results. Claim 4, for example, states that if the input is both spatially and temporally rapidly decreasing (the first condition being a reasonable prior assumption on physical signals and the second meaning that it is "stable") and has zero past, then there exists a causal response with the same properties if and only if a certain condition (in fact CV) holds on p. Moreover, by time-autonomy (which can be assumed a priori), there cannot be any different causal system response; i.e., all causal responses are stable. Our main goal in what follows will be to generalize this result to the general system case.

5.3. Ideal convexity. In order to generalize the results in the previous section to systems, we will need certain properties of polynomials with respect to the set $\mathcal{X}^+$ to extend to ideals.

DEFINITION 5.4. *We call a set $S \subseteq \mathbb{C}^n$* codimension $k-$convex, $k = 1, \ldots, n$, *if for any codimension k prime ideal J we have*

$$\mathcal{V}(J) \cap S = \emptyset \Rightarrow \exists\, f \in J : \mathcal{V}(f) \cap S = \emptyset. \tag{28}$$

We say that S is ideal-convex if property (28) *holds for any (not necessarily prime) ideal J.*

The first of these properties was introduced in [32], in which it is shown that the closed unit polydisc is codimension $k-$convex for all k. Codimension 1–convexity is trivial, since any prime ideal of codimension 1 in $\mathbb{C}[s]$ is principal (it must contain an irreducible polynomial and so must be equal to the codimension 1 prime ideal generated by that polynomial). It was also observed in [32] that if $\mathcal{S}$ is codimension $k-$convex for $k = 1, \ldots, n$, then S is ideal convex; this is essentially due to claim 2 of the following simple but important result.

THEOREM 5.5. *Let $S \subseteq \mathbb{C}^n$ be one of the sets: $\imath\mathbb{R}^n, \mathcal{X}^+, \mathcal{X}^-$. We have*

1. *S is codimension $n-$convex.*
2. *If the minimal prime divisors of an ideal satisfy* (28)*, then the ideal itself also does.*
3. *For $n = 2$, S is ideal-convex.*

Proof.

1. Suppose that J is a maximal ideal with variety not intersecting $\mathcal{X}^+$. Both real and complex coefficient polynomial rings are covered by the following argument. Let $(\alpha_1, \ldots, \alpha_{n-1}, \beta)$ be a point in $\mathcal{V}(J)$. Then either $\Re(\alpha_i) \neq 0$ for some $i = 1, \ldots, n-1$, or else $\Re(\beta) < 0$. In the first case, $f := (s_i - \alpha_i)(s_i - \overline{\alpha_i})$ is a suitable polynomial in J, where $\bar{\cdot}$ denotes complex conjugate. In the second case, $f := (s_n - \beta)(s_n - \overline{\beta})$ may be chosen. The argument for $\mathcal{X}^-$ is symmetric and that for $\imath\mathbb{R}^n$ similar but simpler.
2. This property holds for arbitrary S. For suppose that I is a given ideal such that $\mathcal{V}(I) \cap S = \emptyset$, and that $J_1, \ldots, J_l$ are prime divisors of I and satisfy property (28). Then the product f of the corresponding polynomials $f_i \in J_i$ is contained in the intersection of the J_i, which equals the radical of I, and therefore some power f^k of f is in I, and we have $\mathcal{V}(f^k) \cap S = \emptyset$ as required.
3. For $n = 2$, any nonzero prime ideal has codimension either 1 or n; using claim 1 we have that property (28) holds for all prime ideals. By claim 2, S is ideal-convex. ∎

An important open question is whether the sets S in the preceding theorem are actually ideal-convex for all n. We will see in section 7 that this question has major implications for input/output stability. Ideal-convexity can also be expressed (see also Proposition 1 in [28]) as a "(weak) Nullstellensatz"-type result for the localized ring $U^{-1}\mathbb{C}[s]$, where U is the multiplicatively closed set of all polynomials which do not vanish anywhere in the given domain.

In [33], Valcher characterizes stability of autonomous behaviors for "square" behaviors (those admitting a square kernel representation matrix) and finite-dimensional behaviors (those for which the corresponding system module has codimension n). These seem to correspond to principal modules/ideals and maximal ideals, respectively, and we suspect that the generalization of [33] to the nD case hits a problem analogous to proving ideal-convexity here.

6. Structure theory over $\mathcal{S}$, $\mathcal{S}'$, $\mathcal{S}_+$, $\mathcal{S}'_\oplus$. In this section, we will develop some basic but highly nontrivial results concerning the structure of behaviors over $\mathcal{S}$ and $\mathcal{S}'$, and then $\mathcal{S}_+$ and $\mathcal{S}'_\oplus$. The second set of results and proofs follow the structure

of the first. The results for $\mathcal{S}$ and $\mathcal{S}'$ are also included for independent interest. Many of the results in this section and section 7 are restricted to the case $n = 2$; this is due only to the lack of a proof that the sets S in Theorem 5.5 are ideal-convex for all n. The material in this section and section 7 is entirely new, barring Theorem 6.1.

We begin by recalling a result on (Willems) closure from [36], which, however, we state in the most general form as indicated in that paper.

THEOREM 6.1. *Let $\mathcal{W}$ be a module over $\mathbb{C}[s]$ and $\mathcal{N}$ a submodule of $\mathbb{C}[s]^{1\times q}$ for some q. Suppose that $\mathcal{N} = \cap_{i=1}^{t}\mathcal{N}_i$ is an irredundant decomposition of $\mathcal{N}$ in $\mathbb{C}[s]^{1\times q}$, where $\mathcal{N}_i$ is a J_i-primary submodule of $\mathbb{C}[s]^{1\times q}$ for some prime ideals $J_1, \dots, J_t$. Let the components be ordered so that for some $r \in 0, \dots, t$, $J_1, \dots, J_r$ each annihilate some nonzero element of $\mathcal{W}$, but $J_{r+1}, \dots, J_t$ do not. Assume that each of the primes J_i, $i = 1, \dots, r$, is contained in some prime K_i which also annihilates some nonzero element of $\mathcal{W}$, and for which*

$$(0 : K_i^{\infty})_{\mathcal{W}} := \{w \in \mathcal{W} \mid K_i^l w = 0 \text{ for some } l \in \mathcal{N}\}$$

is an injective $\mathbb{C}[s]$-module. Then the closure of $\mathcal{N}$ with respect to $\mathcal{W}$ is given by

$$\mathcal{N}^{\perp\perp} = \bigcap_{i=1}^{r} \mathcal{N}_i. \tag{29}$$

The assumption holds in particular whenever $\mathcal{W}$ is itself injective.

Proof. This result is a small refinement of [36, Thm. 3.9], as indicated in the comments following that result. For completeness, we sketch the necessary modifications to the proof given in [36] to achieve this refinement.

As shown in the original proof, the inclusion "$\supseteq$" in (29) does not require any assumption on the primes $J_1, \dots, J_r$. For the reverse inclusion, set $\mathcal{D} = \mathbb{C}[s]$, $\mathcal{M} = \mathcal{D}^{1\times q}/\mathcal{N}$, and $\mathcal{L} = \cap_{i=1}^{r}\mathcal{M}_i$, $\mathcal{M}_i = \mathcal{N}_i/\mathcal{N}$, as in the original proof. As in the original proof, if $x \in \mathcal{M}\backslash\mathcal{L}$, then there exist $i \in 1, \dots, r$ and $a \in \mathcal{D}$ with $\mathcal{D}(ax + \mathcal{M}_i) \cong \mathcal{D}/J_i$. Now by assumption, J_i is contained in K_i with $(0 : K_i^{\infty})_{\mathcal{W}}$ injective and also nonzero as K_i annihilates some nonzero element of $\mathcal{W}$. As in the argument for P_j in the original proof, K_i must be contained in some prime Q with a copy of the injective hull $E(\mathcal{D}/Q)$ of $\mathcal{D}/Q$ embedded in $(0 : K_i^{\infty})_{\mathcal{W}}$. We therefore have a sequence of maps of the form

$$\mathcal{D}(ax + \mathcal{M}_i) \xrightarrow{\cong} \mathcal{D}/J_i \xrightarrow{\rho_1} \mathcal{D}/K_i \xrightarrow{\rho_2} \mathcal{D}/Q \to E(\mathcal{D}/Q) \to (0 : K_i^{\infty})_{\mathcal{W}},$$

where ρ_1 and ρ_2 are the natural projections. Call the composition map w_1. As in the original proof, if $w_1(ax + \mathcal{M}_i) = 0$, then we must have $(\rho_2 \circ \rho_2)(1 + J_i) = 0$ in $\mathcal{D}/Q$, which entails that $1 \in Q$. This is impossible, so $w_1(ax + \mathcal{M}_i) \neq 0$ and w_1 is not the zero map. As in the original proof, w_1 can then be extended using injectivity of $(0 : K_i^{\infty})_{\mathcal{W}}$ to a nonzero map $w : \mathcal{M} \to \mathcal{W}$, which further is nonzero at x. This is enough to complete the proof, as explained in [36].

The final claim follows since injectivity of $\mathcal{W}$ guarantees injectivity of $(0 : I^{\infty})_{\mathcal{W}}$ for any ideal I [14]. ∎

6.1. Structure theory over $\mathcal{S}$, $\mathcal{S}'$. In particular, we may derive from Theorem 6.1 the characterization by Shankar of closure with respect to $\mathcal{S}'$ [29] (this latter result, however, preceded and motivated the theorem above). $\mathcal{S}'$ is injective, as proved originally by Malgrange [12] and more recently by Shankar [31], and therefore

Theorem 6.1 applies; a prime J kills some nonzero element of $\mathcal{S}'$ if and only if it vanishes at some imaginary point. We repeat Shankar's result from [29] here, together with an alternative form for a special case.

COROLLARY 6.2. *Suppose that $\mathcal{N} = \cap_{i=1}^{t} \mathcal{N}_i$ is an irredundant decomposition of a submodule $\mathcal{N}$ of $\mathbb{C}[s]^{1\times q}$, where $\mathcal{N}_i$ is J_i-primary. Let the components be ordered so that for some $r \in 0, \ldots, t$, $J_1, \ldots, J_r$ each vanish at some point $\imath\mathbb{R}^n$ but $J_{r+1}, \ldots, J_t$ do not. Then the closure of $\mathcal{N}$ with respect to $\mathcal{W}$ is equal to $\cap_{i=1}^{r} \mathcal{N}_i$. Moreover, in the case where the primes are minimal in the set $J_{r+1}, \ldots, J_t$ and each has codimension 1 or n (e.g., when $n \leq 2$), there exists a polynomial f with no imaginary roots such that the closure with respect to $\mathcal{W}$ may be written*

$$\mathcal{N}^{\perp\perp} = \{v \in \mathbb{C}[s]^{1\times q} \mid \exists k \in \mathbb{N},\ f^k v \in \mathcal{N}\}. \tag{30}$$

Proof. The equality

$$\mathcal{N}^{\perp\perp} = \bigcap_{i=1}^{r} \mathcal{N}_i \tag{31}$$

is already given [29]. Now suppose that the primes are minimal among $J_{r+1}, \ldots, J_t$ and each has codimension 1 or n. By Theorem 5.5 there then exists $f \in \cap_{i=r+1}^{t} J_i$ with no imaginary roots. Note that f is contained in $J_{r+1}, \ldots, J_t$ but not in any of $J_1, \ldots, J_r$ (since these do have imaginary roots). Hence by [3, Prop. 3.13], the module

$$H^0_{(f)}(\mathcal{M}) := \{v + \mathcal{N} \in \mathcal{M} \mid f^k v \in \mathcal{N} \text{ for some } k \in \mathcal{N}\}$$

agrees with $\cap_{i=1}^{r} \mathcal{N}_i/\mathcal{N}$, where $\mathcal{M} = \mathbb{C}[s]^{1\times q}/\mathcal{N}$. This together with (31) gives us the required result. ∎

Note from the proof that the assumptions in the second claim of the corollary may be dropped if it can be shown that $\imath\mathbb{R}^n$ is ideal-convex for any n. From the corollary we can give conditions for one behavior over $\mathcal{S}'$ to be contained in another (when the assumptions of Corollary 6.2 hold).

COROLLARY 6.3. *Let $\mathcal{B}_1 = \ker_{\mathcal{S}'} R_1$ and $\mathcal{B}_2 = \ker_{\mathcal{S}'} R_2$ be two behaviors contained in $(\mathcal{S}')^q$ for some q. Let $\mathcal{N}_1$ be the row span of R_1 over $\mathbb{C}[s]$. Suppose that the primes, minimal among those associated primes of the module $\mathbb{C}[s]^{1\times q}/\mathcal{N}_1$ with no imaginary points in their varieties, each have codimension 1 or n; this occurs, in particular, when $n \leq 2$. Then $\mathcal{B}_1 \subseteq \mathcal{B}_2$ if and only if there exist a polynomial f with no imaginary roots and a polynomial matrix L, such that $fR_2 = LR_1$.*

Proof. Suppose that such an f and L exist with $fR_2 = LR_1$. Then for any $w \in \ker_{\mathcal{S}'} R_1$, $f(\partial)(R_2(\partial)w) = 0$. Since f has no imaginary roots, by Corollary 6.2 the closure with respect to $\mathcal{S}'$ of the ideal generated by f is equal to $\mathbb{C}[s]$; i.e., f kills no nonzero element of $\mathcal{S}'$. Hence $R_2(\partial)w = 0$ as required.

Conversely, suppose that $\mathcal{B}_1 \subseteq \mathcal{B}_2$. Writing $\mathcal{N}_1$ and $\mathcal{N}_2$ for the row spans of R_1 and R_2, respectively, we have

$$\mathcal{N}_2 \subseteq \mathcal{N}_2^{\perp\perp} \subseteq \mathcal{N}_1^{\perp\perp},$$

where $\cdot^{\perp\perp}$ denotes closure with respect to $\mathcal{S}'$. By the second claim of Corollary 6.2 (which requires the given assumptions on $\mathbb{C}[s]^{1\times q}/\mathcal{N}_1$ or on n), there is a polynomial f with no imaginary roots such that for each row of R_2, f^k times that row is in $\mathcal{N}_1$ for sufficiently large k. Note that f^k also has no imaginary roots. The converse follows. ∎

Again, the assumptions in Corollary 6.3 are needed only because it has not been proven yet that $\imath\mathbb{R}^n$ is ideal-convex. These assumptions prove an obstacle to the next result, which for this reason only is restricted to the special cases when P is a polynomial or $n \leq 2$.

THEOREM 6.4. *Let*

$$\mathcal{B} := \{(u, y) \in (\mathcal{D}')^{m+p} \mid P(\partial)y = Q(\partial)u\}$$

be a behavior with the indicated input/output structure and corresponding transfer matrix G. Suppose that either P is a single polynomial or $n \leq 2$. Then the following are equivalent:

1. *The variables u are free over $\mathcal{S}$ in $\mathcal{B}$.*
2. $\ker_{\mathcal{S}'} P^* \subseteq \ker_{\mathcal{S}'} Q^*$.
3. *There exist a polynomial r with no imaginary roots and a polynomial matrix L, such that $G = \frac{1}{r}L$.*
4. *$\mathcal{B}$ has no imaginary controllable poles.*
5. $G(\imath\zeta) \in \mathcal{L}^{p\times m}$.

Proof. 5 $\Rightarrow$ 1: Suppose $G^*(\imath\zeta) \in \mathcal{L}^{p\times m}$. Then for any $u \in \mathcal{S}^m$ we have that $v(\zeta) := G^*(\imath\zeta)\hat{u}(\zeta) \in \mathcal{S}^p$, where $\hat{\cdot}$ denotes the Fourier transform, using claim 2 of Lemma 5.2. Now $P(\imath\zeta)v(\zeta) = Q(\imath\zeta)\hat{u}(\zeta)$, so the inverse Fourier transform of v is a solution y to $P(\partial)y = Q(\partial)u$.

1 $\Rightarrow$ 2: If the variables u are free over $\mathcal{S}$, then $\mathrm{im}_{\mathcal{S}}Q \subseteq \mathrm{im}_{\mathcal{S}}P$. Dualizing (in the distributional sense), we have condition 2.

2 $\Rightarrow$ 3: Suppose that condition 2 holds. If P is a single polynomial, then the associated primes of the system module of $P^*(\partial)$ all have codimension 1; thus in either this case or when $n \leq 2$, the conditions of Corollary 6.3 hold. Thus by this corollary, there exist a polynomial r with no imaginary roots and a polynomial matrix L, with $r^*Q^* = L^*P^*$, and hence $P(L/r) = Q$. Since P has full column rank, L/r equals G.

3 $\Rightarrow$ 4: This is immediate from Lemma 2.3.

4 $\Rightarrow$ 5: Suppose $\mathcal{B}$ has no imaginary controllable poles, and let r be the least common denominator of G. So, in particular, $G = (1/r)L$ for some polynomial matrix L, and r has no imaginary zeros. By claim 5 of Lemma 5.2, $(1/r)I_p \in \mathcal{L}^{p\times p}$, and now by claim 1 of Lemma 5.2, $G = (1/r)L \in \mathcal{L}^{p\times m}$. ∎

Similarly, we can also produce the following lemma.

LEMMA 6.5. *With the notation of Theorem 6.4 (but no assumptions on n or P are needed), the following are equivalent:*

1. *The variables u are free over $\mathcal{C}_0^\infty$ in $\mathcal{B}$.*
2. *The variables u are free over $\mathcal{E}'$ in $\mathcal{B}$.*
3. *$\mathcal{B}$ has no controllable poles or, equivalently, the outputs are observable from the inputs in $\mathcal{B}^c$.*
4. *G is a polynomial matrix.*

Proof. For either case $\mathcal{C}_0^\infty$, $\mathcal{E}'$, we use the same proof structure as for Theorem 6.4; dualizing condition 1 or 2 gives a condition for inclusion of behaviors over $\mathcal{D}'$ or over $\mathcal{C}^\infty$, which results in G being a polynomial matrix. Conversely, if G is a polynomial matrix, then with input $u \in (\mathcal{C}_0^\infty)^m$ or $u \in (\mathcal{E}')^m$ we can simply choose $y = G(\partial)u$. ∎

For completeness, we offer a result without prior assumption of an input/output structure.

COROLLARY 6.6. *Let a behavior $\mathcal{B}$ be given by*

$$\mathcal{B} := \{(u, y) \in (\mathcal{D}')^{m+p} \mid P(\partial)y = Q(\partial)u\},$$

where (u, y) is an arbitrary partition of the system variables. Then the following hold:

1. *The variables u are free over either $\mathcal{C}_0^\infty$ or $\mathcal{E}'$ in $\mathcal{B}$ if and only if there exists a polynomial matrix L with $PL = Q$. The maximum number of free variables over $\mathcal{C}_0^\infty$ or $\mathcal{E}'$ of a behavior $\mathcal{B}$ with kernel representation R equals the number of systems variables minus the minimum number of columns of R needed to generate column span over $\mathbb{C}[s]$.*
2. *Suppose $n \leq 2$. The variables u are free over $\mathcal{S}$ if and only if there exist a polynomial r and a polynomial matrix L, such that $PL = Qr$. The maximum number of free variables over $\mathcal{S}$ of a behavior $\mathcal{B}$ with kernel representation R equals the number of system variables minus the number of columns of R needed to generate the column span over the localized ring $U^{-1}\mathbb{C}[s]$, where U is the set of polynomials with no imaginary roots.*

Proof. We prove the second claim only, that u is free over $\mathcal{S}$ if and only if given that L, r exist follows entirely from the arguments $1 \Rightarrow 2 \Rightarrow 3,\ 5 \Rightarrow 1$, which do not require prior existence of an input/output structure. Finally, note that a given set of columns forming a submatrix P of R generates the column span of R over $U^{-1}\mathbb{C}[s]$ if and only if there exists a matrix K over $U^{-1}\mathbb{C}[s]$ with $PK = Q$, where Q consists of the complementary columns of R. This is clearly equivalent to the given condition for freeness over $\mathcal{S}$ of the variables corresponding to the columns of Q; the result follows. ∎

Recall that over $\mathcal{C}^\infty, \mathcal{D}'$, or $\mathcal{S}'$, a set of variables are free if and only if the corresponding elements in the system module are linearly independent over $\mathbb{C}[s]$ [35, Lem. 5.3]; this is due to injectivity of these three spaces. Equivalently, with the notation of Corollary 6.6, there exists a rational function matrix G with $PG = Q$. For $n \leq 2$, the corollary above completes the description of free variables over the classical spaces; the full generalization to $n > 2$ for $\mathcal{S}$ rests on proving ideal-convexity of $\imath\mathbb{R}^n$.

6.2. Structure theory over $\mathcal{S}_+, \mathcal{S}'_\oplus$. Our next aim is to repeat the pattern of Corollaries 6.2 and 6.3 and Theorem 6.4, but for $\mathcal{S}'_\oplus$ and $\mathcal{S}_+$ instead of $\mathcal{S}'$ and $\mathcal{S}$. This will give us results which we can then interpret in terms of stability. Notice that $\mathcal{S}'_\oplus := \mathcal{S}'/\mathcal{S}_-$ may be interpreted as the restriction of tempered distributions to $\mathbb{R}^n_+$.

To begin, we demonstrate that usual results on adjoint operators apply to the space $\mathcal{S}_+$ and dual $\mathcal{S}'_\oplus$.

LEMMA 6.7. *If $R \in \mathbb{C}[s]^{g\times q}$, then for any $f \in (\mathcal{S}'_\oplus)^q, \phi \in (\mathcal{S}_+)^g$ we have*

$$\langle R(\partial)f, \phi\rangle = \langle f, R^*(\delta)\phi\rangle. \tag{32}$$

Hence $f \in \ker'_{\mathcal{S}_\oplus} R$ if and only if f kills $\operatorname{im}_{\mathcal{S}_+} R^*$.

Proof. Let R, f, and ϕ be as given; denote by ι the natural inclusion $\mathcal{S}_+ \to \mathcal{S}$. Let $\overline{f} \in (\mathcal{S}')^q$ be some element such that its equivalence class in $\mathcal{S}'_+$ agrees with f, and let $\overline{h} = R(\partial)\overline{f} \in (\mathcal{S}')^g$; then $\overline{h} + (\mathcal{S}'_+)^g = R(\partial)f$ by definition of differentiation on a factor space such as $\mathcal{S}'_\oplus$. We now have

$$\begin{aligned}
\langle R(\partial)f, \phi\rangle &:= \langle \overline{h}, \iota(\phi)\rangle \\
&= \langle R(\partial)\overline{f}, \iota(\phi)\rangle \\
&= \langle \overline{f}, R^*(\partial)\iota(\phi)\rangle \\
&= \langle \overline{f}, \iota(R^*(\partial)\phi)\rangle \\
&= \langle f, R^*(\partial)\phi\rangle.
\end{aligned}$$

The second claim is immediate. ∎

We now provide some elementary results on the structure of $\mathcal{S}'_\oplus$ as a differential module.

LEMMA 6.8. *Let r be a polynomial. Then the differential operator $r(\partial)$ kills some nonzero element of $\mathcal{S}'_\oplus$ if and only if r has a root in $\mathcal{X}^-$.*

Proof. Suppose that r has no roots in $\mathcal{X}^-$ and let $\overline{w} = w + \mathcal{S}'_- \in \mathcal{S}'_\oplus$ be such that $r(\partial)\overline{w} = 0$. Then $r(\partial)w \in \mathcal{S}'_-$, so $r(\imath\zeta)\hat{w}(\zeta) \in \mathcal{F}(\mathcal{S}'_-)$, where $\mathcal{F}$ denotes Fourier transform, and by claim 3 of Lemma 5.3 this space equals $\mathcal{L}^-$. By claim 6 of Lemma 5.3, $1/r(\imath\zeta) \in \mathcal{L}^-$, and so by claim 1 of Lemma 5.3, $\hat{w}(\zeta) \in \mathcal{L}^- = \mathcal{F}(\mathcal{S}'_-)$ also. Hence $\overline{w} = 0$.

Conversely, suppose that r does not have a root $\zeta \in (\imath\mathbb{R}^{n-1} \times \overline{\mathbb{C}_-})$. Corresponding to this is a nonzero exponential trajectory u of frequency ζ which lies in $\mathcal{C}^\infty$. Multiplying it by a suitable "cut-off" function, we have a trajectory $w \in L^\infty \subseteq \mathcal{S}'$ which agrees with u on $\mathbb{R}^n_+$. As u is killed by $r(\partial)$, the support of $r(\partial)w$ lies in $\mathbb{R}^n_-$, and therefore $r(\partial)(w + \mathcal{S}'_-) = 0$, whereas $w + \mathcal{S}^+_- \neq 0$ because u is nonzero on $\mathbb{R}^n_+$. ∎

Next we generalize this to ideals, which requires ideal-convexity, and so is restricted to special cases, e.g., $n \leq 2$.

LEMMA 6.9. *If an ideal I kills some nonzero element of $\mathcal{S}'_\oplus$, then $\mathcal{V}(I) \cap \mathcal{X}^- \neq \emptyset$. The converse holds when the minimal primes containing I each have codimension 1 or n (e.g., when $n \leq 2$).*

Proof. Suppose that I vanishes at some point $\zeta \in (\imath\mathbb{R}^{n-1} \times \overline{\mathbb{C}_-})$. As in the proof of Lemma 6.8, we can construct a nonzero trajectory $w \in \mathcal{S}'_\oplus$ which is killed by the maximal ideal corresponding to ζ, and therefore killed by I.

Conversely, suppose that $\mathcal{V}(I) \cap \mathcal{X}^- = \emptyset$. By Theorem 5.5 (using the assumption on I), there exists $f \in I$ which has no roots in $\imath\mathbb{R}^{n-1} \times \overline{\mathbb{C}_-}$. Now if $Iw = 0$ for some $w \in \mathcal{S}'_\oplus$, then $f(\partial)w = 0$ in particular, so by Lemma 6.8, $w = 0$. ∎

Lemmas 6.8 and 6.9 are basic results on closure with respect to $\mathcal{S}'_\oplus$. In order to extend them to more general cases, we need some injectivity properties of certain modules associated to $\mathcal{S}'_\oplus$.

LEMMA 6.10. *For any maximal ideal I, $(0 : I^\infty)_{\mathcal{S}'_\oplus}$ is injective.*

Proof. In the case where I annihilates no nonzero element of $\mathcal{S}'_\oplus$, $(0 : I^\infty)_{\mathcal{S}'_\oplus} = 0$ is trivially injective. So suppose that I does annihilate some nonzero element of $\mathcal{S}'_\oplus$, which by Lemma 6.9 means that $\exists\, \zeta \in \mathcal{V}(I) \cap \mathcal{X}^-$; indeed we must have $\mathcal{V}(I) = \{\zeta\}$.

Now let $\Psi : (0 : I^\infty)_{\mathcal{D}'} \to (0 : I^\infty)_{\mathcal{D}'_\oplus}$ be the module homomorphism induced by the natural projection $\mathcal{D}' \to \mathcal{D}'_\oplus$. Since $(0 : I^\infty)_{\mathcal{D}'} = (0 : I^\infty)_{\mathcal{C}^\infty}$ is the set of polynomial exponentials with frequency ζ [18, 37], Ψ is clearly injective. Consider an arbitrary element $w \in (0 : I^\infty)_{\mathcal{D}'}$. We can write $w = pw'$, where p is a polynomial function and w' is an exponential trajectory of frequency ζ. Letting $f \in \mathcal{C}^\infty(\mathbb{R})$ be a "cut-off" function with $f(t) = 1$ for $t \geq 0$ and $f(t) = 0$ for $t \ll 0$, by the location of ζ we have that $f(t)w'(x,t) \in L^\infty \subseteq \mathcal{S}'$. Since by the Fourier transformation $\mathcal{S}'$ is closed under multiplication by polynomials, $fw = pfw' \in \mathcal{S}'$ also. Now fw agrees with w on $\mathbb{R}^n_+$, so $\Phi(w) = \Phi(fw) = fw + \mathcal{D}'_- \in \mathcal{D}'_\oplus$. It follows that

$$\text{im } \Psi \subseteq \frac{\mathcal{S}' + \mathcal{D}'_-}{\mathcal{D}'_-} \cong \frac{\mathcal{S}'}{\mathcal{S}' \cap \mathcal{D}'_-} = \mathcal{S}'_\oplus.$$

Hence we can construct another injective module homomorphism $\Phi : (0 : I^\infty)_{\mathcal{D}'} \to (0 : I^\infty)_{\mathcal{S}'_\oplus}$. The module $(0 : I^\infty)_{\mathcal{D}'}$ is injective because $\mathcal{D}'$ is injective [14], so it remains to show that Φ is surjective, or equivalently that the image of Ψ is $(0 : I^\infty)_{\mathcal{W}}$, where $\mathcal{W} = (\mathcal{S}' + \mathcal{D}'_-)/\mathcal{D}'$. However, if $\overline{w} = w + \mathcal{D}'$ with $w \in \mathcal{S}'$ and $I^k\overline{w} = 0$ for some k, we have that the support of $I^k w$ is in $\mathbb{R}^n_-$; i.e., w is a solution in $\mathbb{R}^n_+$ to the equations

of I^k. However, any distributional solution in $\mathbb{R}^n_+$ to the equations of I^k necessarily agrees in that region with a polynomial exponential of degree k, i.e., an element $\hat{w}$ of $(0 : I^k)_{\mathcal{D}'}$, as can be seen by induction on k (the base case $k = 1$ is clear as it reduces to the ODE case). Hence $\Psi(\hat{w}) = w$ as required. ∎

Lemma 6.10 leaves the important open question as to whether $\mathcal{S}'_\oplus$ itself is an injective module. We now provide a technical result, which sets up the conditions necessary to apply Theorem 6.1 over $\mathcal{S}'_\oplus$.

COROLLARY 6.11. *Any prime ideal which vanishes at some point of $\mathcal{X}^-$ is contained in a prime I which has the same property, and for which $(0 : I^\infty)_{\mathcal{S}'_\oplus}$ is injective.*

Proof. Let J be a prime for which $\mathcal{V}(J) \cap \mathcal{X}^- \neq \emptyset$. Thus there exists a maximal ideal I which vanishes at some point of $\mathcal{X}^-$ and for which $\mathcal{V}(I) \subseteq \mathcal{V}(J)$, so $J \subseteq I$. Lemma 6.10 now completes the proof. ∎

In certain special cases, e.g., $n \leq 2$, we can now characterize closure over $\mathcal{S}'_\oplus$.

COROLLARY 6.12. *Suppose that $\mathcal{N} = \cap_{i=1}^t \mathcal{N}_i$ is an irredundant primary decomposition of a submodule $\mathcal{N}$ of $\mathbb{C}[s]^{1\times q}$, where $\mathcal{N}_i$ is J_i-primary. Let the components be ordered so that for some $r \in 0, \ldots, t$, $J_1, \ldots, J_r$ each vanishes at some point of $\mathcal{X}^-$ but $J_{r+1}, \ldots, J_t$ do not. Suppose further that the primes $J_{r+1}, \ldots, J_t$ each have codimension 1 or n (e.g., $n \leq 2$). Then the closure of $\mathcal{N}$ with respect to $\mathcal{S}'_\oplus$ is equal to $\cap_{i=1}^r \mathcal{N}_i$, and, furthermore, there exists a polynomial f with no roots in $\mathcal{X}^-$ such that the closure with respect to $\mathcal{S}'_\oplus$ may be written*

$$\mathcal{N}^{\perp\perp} = \{v \in \mathbb{C}[s]^{1\times q} \mid \exists\, k \in \mathbb{N},\ f^k v \in \mathcal{N}\}. \tag{33}$$

Proof. Suppose that the primes $J_{r+1}, \ldots, J_t$ each has codimension 1 or n. Then by Lemma 6.9, $J_1, \ldots, J_r$ each annihilate a nonzero element of $\mathcal{S}'_\oplus$, whereas $J_{r+1}, \ldots, J_t$ do not. Using also Corollary 6.11, we have that $\mathcal{N}^{\perp\perp} = \cap_{i=1}^r \mathcal{N}_i$ by Theorem 6.1. Now by Theorem 5.5 there exists $f \in \cap_{i=r+1}^t J_i$ with no roots in $\mathcal{X}^-$. As in the proof of Corollary 6.2, we can now establish (33). ∎

Note that the codimension conditions in Corollary 6.12 in particular hold when $\mathcal{N}$ is equal to $\mathbb{C}[s] \cdot p(s)$ for p a single polynomial.

COROLLARY 6.13. *Let $\mathcal{B}_1 = \ker_{\mathcal{S}'_\oplus} R_1$ and $\mathcal{B}_2 = \ker_{\mathcal{S}'_\oplus} R_2$ be two behaviors contained in $(\mathcal{S}'_\oplus)^q$ for some q. Let $\mathcal{N}_1$ be the row span of R_1 over $\mathbb{C}[s]$. Suppose that those associated primes of $\mathbb{C}[s]^{1\times q}/\mathcal{N}_1$ with varieties not intersecting $\mathcal{X}^-$ each have codimension 1 or n (for example, $\mathbb{C}[s]^{1\times q}/\mathcal{N}_1$ is principal or $n \leq 2$). Then $\mathcal{B}_1 \subseteq \mathcal{B}_2$ if and only if there exist a polynomial f with no roots in $\mathcal{X}^-$ and a polynomial matrix L, such that $fR_2 = LR_1$.*

Proof. The proof uses exactly the same argument as for Corollary 6.3. ∎

It is the failure to prove ideal-convexity of $\mathcal{X}^-$ which prevents generalization of Corollaries 6.12 and 6.13 to the general nD case.

7. Stable input/output structures. Having finally done all the necessary groundwork, we can now consider input/output stability. We consider this only for input/output structures which are a priori causal (with respect to $\mathcal{C}^\infty$ or $\mathcal{D}'$ according to the type of stability required); thus, in particular, we assume that $\mathcal{B}_{0,y}$ is time-autonomous.

DEFINITION 7.1. *Let $\mathcal{B}$ be a behavior with associated input/output structure (u, y), where $\mathcal{B}_{0,y}$ is time-autonomous and the input/output structure is causal. Call this input/output structure* stable (with respect to $\mathcal{S}$ (resp., $\mathcal{S}'$)) *if for any $u \in (\mathcal{S}_+)^m$ (resp., $u \in (\mathcal{S}_+)^m$) and $y \in (\mathcal{C}^\infty_+)^p$ (resp., $y \in (\mathcal{D}'_+)^p$) for which $(u, y) \in \mathcal{B}$, we must have $y \in (\mathcal{S}_+)^p$ (resp., $y \in (\mathcal{S}'_+)^p$).*

Thus, roughly speaking, an input/output structure is stable if any causal output response to a stable input is itself stable. Since it is reasonable to assume a priori that our input/output structure is causal, the existence of a $y \in (\mathcal{C}_+^\infty)^p$ corresponding to a $u \in (\mathcal{S}_+)^m$ is guaranteed. Moreover, if $\mathcal{B}_{0,y}$ is a priori time-autonomous, this y is unique, and so in this case the input/output structure is stable with respect to $\mathcal{S}$ (resp., $\mathcal{S}'$) if and only if the variables u are free over the signal space $\mathcal{S}$ (resp., $\mathcal{S}'$). Fortunately, using methods analogous to those in the proof of Theorem 6.4, we can now characterize freeness of variables over $\mathcal{S}_+$ using the structure theory developed in the last section. The results for $\mathcal{S}_+$ are, however, restricted to the special cases when $n \leq 2$ or P is a single polynomial (i.e., there is a single system equation), due to the difficulty of proving ideal-convexity for $n > 2$, whereas those for $\mathcal{S}'_+$ give sufficient conditions for freeness only.

THEOREM 7.2. *Let*

$$\mathcal{B} := \{(u, y) \in (\mathcal{D}')^{m+p} \mid P(\partial)y = Q(\partial)u\}$$

be a behavior with given input/output structure, and transfer matrix G. Suppose that either P is a single polynomial or $n \leq 2$. Then the following are equivalent.

1. *The variables u are free over $\mathcal{S}_+$ in $\mathcal{B}$.*
2. $\ker_{\mathcal{S}'_\oplus} P^* \subseteq \ker_{\mathcal{S}'_\oplus} Q^*$.
3. *There exist a polynomial r with no roots in $\mathcal{X}^+$ and a polynomial matrix L, such that $G = \frac{1}{r}L$.*
4. *$\mathcal{B}$ has no controllable poles in $\mathcal{X}^+$.*
5. $G(\imath\zeta) \in (\mathcal{L}^+)^{p+m}$.

Proof. The proof has the same structure of that of Theorem 6.4.

5 $\Rightarrow$ 1: Suppose $G(\imath\zeta) \in (\mathcal{L}^+)^{p\times m}$. Then for any $u \in (\mathcal{S}_+)^m$ we have that $v(\zeta) := G(\imath\zeta)\hat{u}(\zeta) \in (\mathcal{S}^+)^p$, where $\hat{\cdot}$ denotes the Fourier transform, using claims 2 and 3 of Lemma 5.3. Now using claims 2 and 6 of Lemma 5.3, we have that $P(\imath\zeta)v(\zeta) = Q(\imath\zeta)\hat{u}(\zeta)$, so the inverse Fourier transform of v is a solution y to $P(\partial)y = Q(\partial)u$. By claim 3 of Lemma 5.3, $y \in (\mathcal{S}_+)^p$.

1 $\Rightarrow$ 2: Suppose the variables u are free over $\mathcal{S}_+$. Then $\mathrm{im}_{\mathcal{S}_+} Q \subseteq \mathrm{im}_{\mathcal{S}_+} P$. Using Lemma 6.7, we find the dual condition 2.

2 $\Rightarrow$ 3: Suppose that condition 2 holds. If P is a single polynomial, then the associated primes of the system module of $P^*(\partial)$ all have codimension 1; thus in either this case or when $n \leq 2$, the conditions of Corollary 6.13 hold. Thus by this corollary there exist a polynomial r^* with no roots in $\mathcal{X}^-$ and a polynomial matrix L^*, with $r^*Q^* = L^*P^*$. Hence $P(L/r) = Q$, where r has no roots in $\mathcal{X}^+$. Since P has full column rank, L/r equals G.

3 $\Rightarrow$ 4: This part of the proof is again immediate from the fact that the controllable pole variety of $\mathcal{B}$ is the variety of the least common denominator of G.

4 $\Rightarrow$ 5: Suppose $\mathcal{B}$ has no controllable poles in $\mathcal{X}^+$, and let r be the least common denominator of G, so in particular $G = (1/r)L$ for some polynomial matrix L, and r has no zeros in $\mathcal{X}^+$. By claim 6 of Lemma 5.3, $(1/r)I_p \in (\mathcal{L})^{p\times p}$, and now by claim 1 of Lemma 5.3, $G = (1/r)L \in (\mathcal{L}^+)^{p\times m}$. ∎

Note. A topic for future research is to seek further characterizations for a principal module, other than computing its associated primes. This could well lead to generalizations of Theorems 6.4 and 7.2 with less restrictive assumptions.

Let us consider the case of single polynomials in the equivalent conditions $2 \equiv 3 \equiv 5$ in Theorem 7.2; take q and $p \neq 0$ rather than their adjoints for ease of notation. We have that $\ker_{\mathcal{S}'_\oplus} p \subseteq \ker_{\mathcal{S}'_\oplus} q$ if and only if $p/\gcd(p, q)$ has no roots in $\mathcal{X}^-$ if and

only if $(q/p)(\imath\zeta) \in \mathcal{L}^-$. From this we may define an action of $\mathcal{L}^-$ on $\mathcal{S}'_\oplus$: given any $u \in \mathcal{S}'_\oplus$ we can choose an arbitrary $v \in \mathcal{S}'_\oplus$ with $u = p(\partial)v$; this is possible as $\mathcal{S}'_\oplus$ is a divisible $\mathbb{C}[s]$-module due to divisibility of $\mathcal{S}'$. Now $y := q(\partial)v \in \mathcal{S}_{\oplus'}$ is uniquely determined by u, due to the condition $\ker_{\mathcal{S}'_\oplus} p \subseteq \ker_{\mathcal{S}'_\oplus} q$, and we have $p(\partial)y = q(\partial)u$. Thus we have an extension of the operator ring on $\mathcal{S}'_\oplus$ from $\mathbb{C}[s]$ to $\mathcal{L}^-$, analogous to the construction of Glüsing-Lüerssen in [5] for delay-differential systems (however, in the current case, the extended ring is a localization, unlike in [5]). It may be profitable, as in the delay-differential case, to consider systems defined as equations over this extended ring. Analogous remarks apply to the signal space $\mathcal{S}'$ and ring $\mathcal{L}$.

We now give a generalization of claim 5 of Lemma 5.3, which gives sufficient conditions for freeness of variables over $\mathcal{S}'_+$. We suspect that these conditions are also necessary.

COROLLARY 7.3. *Let*

$$\mathcal{B} := \{(u, y) \in (\mathcal{D}')^{m+p} \mid P(\partial)y = Q(\partial)u\}$$

be a behavior with given input/output structure and transfer matrix G, and suppose that the denominators of G have no roots in $\mathbb{R}^{n-1} \times \mathbb{C}_+$. Then the variables u are free over $\mathcal{S}'_+$ in $\mathcal{B}$.

Proof. Let $\mathcal{B}, P, Q$, and G be as given. Let d be the least common denominator of G, so that $Gd = N$, a polynomial matrix, and d has no roots in $\mathbb{R}^{n-1} \times \mathbb{C}_+$. Let $u \in (\mathcal{S}'_+)^m$ be arbitrary. Then by claim 5 of Lemma 5.3, there exists $y \in (\mathcal{S}'_+)^p$ satisfying

$$d(\partial)y = N(\partial)u.$$

We now have that $P(\partial)d(\partial)y = P(\partial)N(\partial)u = Q(\partial)d(\partial)u$. Hence $d(\delta)$ kills $P(\partial)y - Q(\partial)u$. However, $d(\delta)$ can kill no elements of $\mathcal{S}'_+$, again by claim 5 of Lemma 5.3, and so $P(\partial)y = Q(\partial)u$. This proves that the variables u are free over $\mathcal{S}'_\oplus$. ∎

Note that the conditions of Corollary 7.3 are particularly met when the equivalent conditions of Theorem 7.2 are satisfied. One consequence of this corollary is that when G is as specified, given any input u which is a Dirac delta in one component and zero in the others (and so in $(\mathcal{S}'_+)^m$), there is a corresponding causal output in $(\mathcal{S}'_+)^p$. If we assume time-autonomy of $\mathcal{B}_{0,y}$, then these causal outputs are unique, and we may collect them into a matrix called the *impulse response matrix* H_{imp}. When G is further stable, the input-to-output map over $\mathcal{S}_+$, which exists due to Theorem 7.2, is then given by applying H_{imp} as a convolution operator. However, as shown in the proof $5 \Rightarrow 1$ of Theorem 7.2, it can also be given by Fourier transformation, multiplication by $G(\imath\zeta)$, and inverse Fourier transformation. Thus H_{imp} is indeed the inverse Fourier transform of the transfer matrix G. Moreover, by Lemma 5.3 we have $\mathcal{L}^+ = \mathcal{F}(\mathcal{O}'_+)$, so we have $H_{imp} \in (\mathcal{O}'_+)^{p \times m}$.

Our next result shows that, as for causality, when stability with respect to $\mathcal{S}$ of an input/output structure is defined, it is characterized purely in terms of the transfer matrix. This result is, however, restricted to the cases $n \leq 2$ or P is a single polynomial. For the case of stability with respect to $\mathcal{S}'$, no such restriction is needed, but only a sufficient condition is obtained.

THEOREM 7.4. *Let $\mathcal{B}$ be a behavior with a given input/output structure (u, y), such that $\mathcal{B}_{0,y}$ is time-autonomous, and associated transfer matrix G. If G is weakly stable, then (u, y) is causal with respect to $\mathcal{D}'$ and stable with respect to $\mathcal{S}'$. Moreover, suppose that either $n \leq 2$ or $\mathcal{B}_{0,y}$ is defined by a single polynomial. Then (u, y) is*

both causal with respect to $\mathcal{C}^\infty$ and stable with respect to $\mathcal{S}$ if and only if G is stable or, equivalently, if and only if $G(\imath\zeta) \in (\mathcal{L}^+)^{p\times m}$.

Proof. Let $\mathcal{B}$ be given with $\mathcal{B}_{0,y}$ time-autonomous. Suppose first that G is weakly stable or stable (the equivalent characterization of stability of G in terms of $\mathcal{L}^+$ is immediate from Theorem 7.2). Let d be the least common denominator of G, which satisfies the WCV condition and therefore obeys the Gårding condition (20). As $\mathcal{B}_{0,y}$ is time-autonomous, $(\mathcal{B}^c)_{0,y}$ is time-autonomous also by Corollary 4.5. Since $\mathcal{V}(d)$ is equal to the characteristic variety of $(\mathcal{B}^c)_{0,y}$ (Lemma 2.3), $(0, 0, \ldots, 1)$ is also a non-characteristic direction for $d(\partial)$; hence, being a single polynomial, d is hyperbolic by Theorem 3.3. In other words, G is causal. Now by Theorem 4.7, the input/output structure on $\mathcal{B}$ is causal with respect to both $\mathcal{C}^\infty$ and $\mathcal{D}'$.

When G is stable, then by Theorem 7.2 the variables u are free over $\mathcal{S}_+$, so each such input there corresponds to some output $y \in (\mathcal{S}_+)^p$, and by time-autonomy there cannot exist a different causal response, so all causal responses to u are in $(\mathcal{S}_+)^p$. In other words, (u, y) is stable with respect to $\mathcal{S}$. The same argument establishes stability with respect to $\mathcal{S}'$ on the condition that G is weakly stable.

Conversely, suppose the input/output structure is both causal with respect to $\mathcal{C}^\infty$ and stable with respect to $\mathcal{S}$. This shows that u is free over $\mathcal{S}_+$, so by Theorem 7.2 (this being the only point where we need assumptions on $\mathcal{B}_{0,y}$ or on n), G is stable, as required. ∎

Note that Theorem 7.4 effectively states that an input/output structure is both causal and stable (with respect to $\mathcal{C}^\infty$ and $\mathcal{S}$, respectively) if and only if the zero-input behavior $(\mathcal{B}^c)_{0,y}$ of the controllable part satisfies the CV condition, i.e., if and only if $\mathcal{B}$ has no controllable unstable poles. It is pleasing that input/output stability is determined by the poles of the system, as in the 1D case, and that the condition for input/output stability is precisely that which has been proposed for stability of the autonomous behavior $(\mathcal{B}^c)_{0,y}$. Also, observe that stability with respect to $\mathcal{S}$ is stronger than stability with respect to $\mathcal{S}'$.

While we have taken time-autonomy as a prior condition for the definition of causal and therefore stable (with respect to $\mathcal{S}$) input/output structures, it is in fact a consequence of these two properties. For if $y \in \mathcal{B}_{0,y}$ has support in $\mathbb{R}^n_+$, then by stability (0 being a stable input!), $y \in (\mathcal{S}_+)^p \subseteq \mathcal{S}^p$. If P is a kernel matrix representation matrix of $\mathcal{B}_{0,y}$, then it has a nonzero highest order minor r, and now we find that $r(\partial)y = 0$, which as $y \in \mathcal{S}$ necessitates $y = 0$ (e.g., by taking Fourier transforms). Thus $\mathcal{B}_{0,y}$ is time-autonomous.

As is the case for causality, Theorem 7.4 in particular implies that stability is determined by the properties of a single polynomial d, the least common denominator of the transfer matrix. To ascertain stability of the input/output structure with respect to $\mathcal{S}$, we need only test whether d obeys the CV condition, i.e., whether the roots of d intersect the set $\mathcal{X}^+$. This test amounts to checking whether a set of real algebraic equations and inequalities has a solution and so may be solved by quantifier elimination theory (e.g., [1]). An important open question is whether a simpler algorithm may be developed, making special use of the structure of $\mathcal{X}^+$.

We conclude with a final result which is a sufficient condition only but drops the restrictions $n \leq 2$ or P a single polynomial.

COROLLARY 7.5. *If $\mathcal{B}$ is a behavior with given input/output structure such that $\mathcal{B}_{0,y}$ is time-autonomous and G is stable, then the input/output structure is causal and stable with respect to $\mathcal{S}$. This occurs in particular when $\mathcal{B}_{0,y}$ is time-autonomous and satisfies the CV condition.*

Proof. The proof in Theorem 7.4 that G is stable implies (u, y) is causal and stable depends only on the proof $5 \Rightarrow 1$ of Theorem 7.2, which in turn does not require the assumptions $n \leq 2$ or P a single polynomial. The final claim is immediate from Lemma 2.4. ∎

8. Conclusions. We have defined causality and stability for input/output structures with the a priori property that $\mathcal{B}_{0,y}$ be time-autonomous. This property means that any output is determined by its own past together with the input. When this property does not hold (e.g., for the heat equation), our definitions do not apply, and we believe that an entirely different approach will be necessary to define and characterize these properties in that case.

We have shown that both causality and stability are characterized by properties of the system transfer matrix or, more precisely, its least common denominator d. Presupposing time-autonomy, the input/output structure is causal (with respect to $\mathcal{C}^\infty$) if and only if d is hyperbolic, and is stable (with respect to $\mathcal{S}$) if and only if d satisfies the CV condition. The CV condition is precisely that which we have proposed for stability of an autonomous behavior; establishing (or disproving) that the CV condition is equivalent to stability (in a suitable sense) for an autonomous behavior is an important open question.

The obstacle to generalizing the stability results for $\mathcal{S}$ to the general $n > 2$ case is in proving that the set $\mathcal{X}^-$ has the property of being ideal-convex. This seems difficult to establish, particularly since this set is noncompact (and so it is not clear that we can approximate holomorphic functions uniformly on $\mathcal{X}^-$ by polynomials). However, we note that $\mathcal{X}^-$ may be bilinearly transformed into the bounded set $(S^1)^{n-1} \times D$, where S^1 is the unit circle and D the unit disc. The stability results for S^1 require no prior assumptions but give sufficient conditions only.

Another open problem is the generalization of these results to the case of equations with real coefficients. We have used only complex coefficients here since we have applied many results from the theory of PDEs which have been developed for complex coefficients; but the real coefficient case will probably require close examination of the PDE literature.

Finally, we have developed some structure theory results for the sets $\mathcal{S}, \mathcal{S}', \mathcal{S}_+$, and $\mathcal{S}'_\oplus$. The use of algebraic "local cohomology" may prove a useful approach to characterizing (Willems) closures in other situations, particularly since when we apply this tool we immediately get, as a corollary, concrete characterizations of the inclusion of one behavior in another, as in Corollaries 6.3 and 6.13. We have also been able to characterize freeness of variables over $\mathcal{S}$ for $n \leq 2$ (and over $\mathcal{C}_0^\infty, \mathcal{E}'$ also); generalization to nD depends on establishing ideal-convexity of $\imath\mathbb{R}^n$. The question as to whether the space $\mathcal{S}'_\oplus$ is injective (and whether the space $\mathcal{S}_+$, to which it is dual, is flat) is also open.

REFERENCES

[1] B. F. Caviness and J. R. Johnson, eds., *Quantifier Elimination and Cylindrical Algebraic Decomposition. Texts and Monographs in Symbolic Computation*, Springer-Verlag, Vienna, 1998.

[2] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts in Applied Mathematics 21, Springer-Verlag, New York, 1995.

[3] D. Eisenbud, *Commutative Algebra with a View Toward Algebraic Geometry*, Graduate Texts in Mathematics 150, Springer-Verlag, New York, 1995.

[4] S. G. Gindikin and L. R. Volevich, *The Cauchy problem*, in Partial Differential Equations III Yu V. Egorov and M. A. Shubin (eds.), Encyclopedia of Mathematical Sciences 32, Chapter I, Springer-Verlag, Berlin, 1991.

[5] H. Glüsing-Lürssen, *Linear Delay-Differential Systems with Commensurate Delays: An Algebraic Approach*, Number 1770 in Lecture Notes in Mathematics, Springer-Verlag, Berlin, 2002.
[6] L. Hörmander, *On the theory of general partial differential operators*, Acta Math., 94 (1955), pp. 161–248.
[7] L. Hörmander, *An Introduction to Complex Analysis in Several Variables*, 2nd ed., North-Holland/American Elsevier, New York, 1973.
[8] L. Hörmander, *The Analysis of Linear Partial Differential Operators* I: *Distribution Theory and Fourier Analysis*, Springer-Verlag, Berlin, 1983.
[9] L. Hörmander, *The Analysis of Linear Partial Differential Operators* II: *Differential Operators With Constant Coefficients*, Springer-Verlag, Berlin, 1983.
[10] F. John, *Partial Differential Equations*, Springer-Verlag, New York, 1971.
[11] T. Y. Lam, *Lectures on Modules and Rings*, Graduate Texts in Mathematics 189, Springer-Verlag, New York, 1999.
[12] B. Malgrange, *Division des distributions*, Seminaire L. Schwartz, pp. Exposes 21–25, 1960.
[13] B. Malgrange, *Systemes differentiels a coefficients constants*, in Seminiare Bourbaki, Soc. Math. France, 246 (1995), pp. 79–89.
[14] E. Matlis, *Divisible modules*, Proc. Amer. Math. Soc., 11 (1960), pp. 385–391.
[15] S. Matsuura, *On general systems of partial differential operators with constant coefficients*, J. Math. Soc. Japan, 13 (1961), pp. 94–103.
[16] M. Nacinovich, *Cauchy problem for overdetermined systems*, Ann. Mat. Pura Appl., 156 (1990), pp. 265–321.
[17] U. Oberst, *Multidimensional constant linear systems*, Acta. Appl. Math., 20 (1990), pp. 1–175.
[18] U. Oberst, *Variations on the fundamental principle for linear systems of partial differential and difference equations with constant coefficients*, Appl. Algebra Engrg., Comm. Comput., 6 (1995), pp. 211–243.
[19] V. P. Palamodov, *Linear Differential Operators with Constant Coefficients*, Springer-Verlag, New York, 1970.
[20] H. Pillai and S. Sankar, *A behavioral approach to control of distributed systems*, SIAM J. Control Optim., 37 (1999), pp. 388–408.
[21] H. Pillai, J. Wood, and E. Rogers, *On homomorphisms of nD behaviors*, IEEE Trans. Circuits Systems I. Fund. Theory Appl. (special issue on multidimensional systems), 49 (2002), pp. 732–742.
[22] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Texts in Applied Mathematics 26, Springer-Verlag, New York, 1998.
[23] M. Renardy and R. C. Rogers, eds., *An Introduction to Partial Differential Equation*, Texts in Applied Mathematics 13, Springer-Verlag, New York, 1993.
[24] A. J. Sasane, *Time-autonomy and time-controllability of $\mathcal{W}_{\mathcal{S}}$-behaviors*, SIAM J. Control Optim., (2002), submitted.
[25] A. J. Sasane, E. G. F. Thomas, and J. C. Willems, *Time-autonomy versus time-controllability*, Systems Control Lett., 45 (2002), pp. 145–153.
[26] A. J. Sasane, *On the Willems closure with respect to $\mathcal{W}_{\mathcal{S}}$*, IMA J. Math. Inform. Control, 20 (2003), pp. 217–232.
[27] A. J. Sasane, *Time-autonomy and time-controllability of 2-D behaviors that are tempered in the spatial direction*, Multidimens. Systems Signal Process., 15 (2004), pp. 97–116.
[28] S. Shankar, *An obstruction to the simultaneous stabilization of two n-D plants*, Acta Appl. Math., 36 (1994), pp. 289–301.
[29] S. Shankar, *The Nullstellensatz for systems of PDE*, Adv. Appl. Math., 23 (1999), pp. 360–374.
[30] S. Shankar. *Can one control the vibrations of a drum?* Multidimens. Systems Signal Process., 11 (2000), pp. 67–81.
[31] S. Shankar, *Geometric completeness of distribution spaces*, Acta Appl. Math., 77 (2003), pp. 163–180.
[32] S. Shankar and V. R. Sule, *Algebraic geometric aspects of feedback stabilization.* SIAM J. Control Optim., 30 (1992), pp. 11–30.
[33] M. E. Valcher, *Characteristic cones and stability of two-dimensional autonomous behaviors*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 47 (2000), pp. 290–294.
[34] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.
[35] J. Wood, *Modules and behaviors in nD systems theory*, Multidimens. Systems Signal Process., 11 (2000), pp. 11–48.

[36] J. Wood, *Key problems in the extension of module-behavior duality*, Linear Algebra Appl., 351/352 (2002), pp. 761–798.

[37] J. Wood, U. Oberst, E. Rogers, and D. H. Owens, *A behavioral approach to the pole structure of one-dimensional and multidimensional linear systems*, SIAM J. Control Optim., 38 (2000), pp. 627–661.

[38] S. Zampieri, *Causal input/output representation of 2D systems in the behavioral approach*, SIAM J. Control Optim., 36 (1998), pp. 1133–1146.

[39] P. Zaris, J. Wood, and E. Rogers, *Controllable and uncontrollable poles and zeros of nD systems*, Math. Control Signals Systems, 14 (2001), pp. 281–298.

[40] E. Zerz, *Extension modules in behavioral linear systems theory*, Multidimens. Systems Signal Process., 12 (2001), pp. 309–327.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 4, pp. 1521–1534

WHAT IS THE BETTER SIGNAL SPACE FOR DISCRETE-TIME SYSTEMS: $\ell_2(\mathbb{Z})$ OR $\ell_2(\mathbb{N}_0)$?*

BIRGIT JACOB†

Abstract. In this paper a system is considered as a (possibly unbounded) linear operator between Hilbert spaces. As signal space we consider the spaces $\ell_2(\mathbb{Z})$ and $\ell_2(\mathbb{N}_0)$. Whereas the case $\ell_2(\mathbb{N}_0)$ has been well studied in the literature, the case $\ell_2(\mathbb{Z})$ has hardly been studied, and one goal of this paper is to study these systems from first principle. Further, we compare these two mathematical formalisms and show that the stabilizable systems and the stabilizing controllers are the same in both mathematical formalisms, provided that a suitable definition of stabilizability is used for systems over the signal space $\ell_2(\mathbb{Z})$.

Key words. linear systems, time-invariant systems, signal spaces, stabilizability, feedback systems

AMS subject classifications. 93D15, 93B28, 93D25

DOI. 10.1137/S0363012902412471

1. Introduction.

In this paper we consider an input-output approach to linear time-invariant discrete-time systems and, moreover, we do this quite generally, that is, without assuming specific system representations such as differential or integral equations. We have only two further basic requirements: First we assume that the input and output space are finite-dimensional, that is, input and output functions take values in $\mathbb{C}^p$, respectively, $\mathbb{C}^m$. Second, we assume that input and output functions are elements of Hilbert spaces. Under these assumptions a system can be seen as a (possibly) unbounded operator on Hilbert spaces. For the choice of the Hilbert spaces there are two natural possibilities: $\ell_2(\mathbb{Z})^n$ and $\ell_2(\mathbb{N}_0)^n$, depending on whether we consider as time axis $\mathbb{N}_0$ or $\mathbb{Z}$.

Systems with signal space $\ell_2(\mathbb{N}_0)$ have been studied from first principle in detail by Georgiou and Smith [5]. They discussed different system descriptions such as operators, transfer functions, and coprime factorizations, proved in which sense these representations are interchangeable, and discussed the question of how physically motivated properties such as causality and stabilizability should be defined in this mathematical formalism. Some of their results are reviewed in section 3. The input-output approach to systems is not new; it has been studied in [23], [21], [18], [2], [4], [22], and other works. However, systems are considered only as mappings, and equivalent conditions for some basic properties such as stabilizability and causality are missing. There are alternative approaches to systems, such as the algebraic/coprime factor theory approach [1], [19]. Using this approach a system is represented by its transfer function, which is in our case an element of the quotient field of H_∞. However, alternative approaches pay no attention to the signal space.

There are few results available concerning systems over the signal space $\ell_2(\mathbb{Z})$. Georgiou and Smith [5], [6] started to study these systems, and they discovered intrinsic difficulties: A causal system could have a noncausal closure, and a well-known stabilizable system seemed not to be stabilizable. The problem was analyzed further

*Received by the editors August 1, 2002; accepted for publication (in revised form) April 16, 2004; published electronically January 27, 2005.
http://www.siam.org/journals/sicon/43-4/41247.html
†Fachbereich Mathematik, University of Dortmund, D-44221 Dortmund, Germany (birgit.jacob@math.uni-dortmund.de).

by Mäkilä, who primarily studied in a series of papers [11], [12], [13] the question of whether the graphs of linear systems are in fact closable. In Jacob and Partington [10] and Jacob [7] these systems are considered in more detail: In these papers transfer functions and symbols are studied, and equivalent conditions for causality are given. One goal of this paper is to solve the problem concerning stabilizability for systems over the signal space $\ell_2(\mathbb{Z})$. Using an alternative definition of stabilizability in the mathematical formalism we are able to guarantee that all—from a physical point of view—stabilizable systems are actually stabilizable in the mathematical formalism. Further, we give equivalent conditions for stabilizability.

Once these two mathematical formalisms are available, the question arises which signal space—$\ell_2(\mathbb{N}_0)$ or $\ell_2(\mathbb{Z})$—should be used to model a system. The study of this question is the second goal of this paper. We show that the stabilizable systems and the stabilizing controllers are the same, and thus it does not really matter whether we work with $\ell_2(\mathbb{N}_0)$ or $\ell_2(\mathbb{Z})$ as signal space. Note that the results of this paper can also be found in the author's habilitation thesis [8], and for the single input-single output case in the conference paper [9].

We proceed as follows. First we review results on systems over the signal space $\ell_2(\mathbb{N}_0)$ (section 3) and $\ell_2(\mathbb{Z})$ (section 4). In section 5 we introduce and study an adapted notion of stabilizability for systems over the signal space $\ell_2(\mathbb{Z})$. Finally, in section 6 we compare these two mathematical formalisms.

2. Preliminaries. We introduce the following notation. We define $\mathbb{T} := \{z \in \mathbb{C} \mid |z| = 1\}$, and $\mathbb{D} := \{z \in \mathbb{C} \mid |z| < 1\}$. $H_\infty(\mathbb{D})$ denotes the Hardy space of bounded holomorphic function f on $\mathbb{D}$, and $L_\infty(\mathbb{T})$ denotes the space of all (equivalence classes of) measurable and essentially bounded functions on $\mathbb{T}$. Fatou's theorem (see, for example, Duren [3]) shows that to every $f \in H_\infty(\mathbb{D})$ there corresponds a function $\tilde{f} \in L_\infty(\mathbb{T})$, defined a.e. by $\tilde{f}(e^{it}) := \lim_{r\to 1} f(re^{it})$, $t \in [0, 2\pi]$. Note that $\tilde{f}(z) = 0$ on a set of positive measure implies that f equals the zero function. By means of the boundary function we can consider $H_\infty(\mathbb{D})$ as a subspace of $L_\infty(\mathbb{T})$. By $L_2(\mathbb{T})$ we denote the space of all (equivalence classes of) measurable and square integrable functions on $\mathbb{T}$, and $H_2(\mathbb{D})$ denotes the Hardy space of holomorphic functions f on $\mathbb{D}$ satisfying $\sup_{r\in(0,1)} \|f(r\cdot)\|_{L_2(\mathbb{T})} < \infty$. A similar argument as above shows that $H_2(\mathbb{D})$ can be considered as a subspace of $L_2(\mathbb{T})$. By $L_\infty(\mathbb{T})^{m\times n}$ we denote the set of all $m\times n$-matrices with elements in $L_\infty(\mathbb{T})$, and similar notations are used for $H_\infty(\mathbb{D})$, $L_2(\mathbb{T})$, and $H_2(\mathbb{D})$. If n equals 1, then we shorten this to $L_\infty(\mathbb{T})^m$, $H_\infty(\mathbb{D})^m$, $L_2(\mathbb{T})^m$, and $H_2(\mathbb{D})^m$.

By $\ell_2(\mathbb{Z})^n$, $n \in \mathbb{N}$, we denote the space of all vector valued square summable sequences on $\mathbb{Z}$. Similarly, we define $\ell_2(\mathbb{N}_0)^n$. We consider $\ell_2(\mathbb{N}_0)^n$ as a subset of $\ell_2(\mathbb{Z})^n$ by extending $x \in \ell_2(\mathbb{N}_0)^n$ to $\ell_2(\mathbb{Z})^n$ by defining the sequence to be zero outside $\mathbb{N}_0$. Moreover, $x \in \ell_2(\mathbb{Z})^n$ is an element of $\ell_2(\mathbb{N}_0)^n$ if $x(j) = 0$ for $j < 0$.

By S we denote the *right shift* on $\ell_2(\mathbb{Z})^n$, which is given by $(Sx)(j) := x(j-1)$, $j \in \mathbb{Z}$, as well as the *right shift* on $\ell_2(\mathbb{N}_0)^n$, which is given by $(Sx)(j) := x(j-1)$, $j \in \mathbb{N}$ and $(Sx)(0) := 0$. Thus the right shift on $\ell_2(\mathbb{Z})^n$ is bijective, whereas the right shift on $\ell_2(\mathbb{N}_0)^n$ is injective, but not surjective. By e_k, $k \in \mathbb{Z}$, we denote the kth unit vector of $\ell_2(\mathbb{Z})$, namely, $e_k(j) := \delta_{k,j}$, as well as the kth unit vector of $\mathbb{C}^n$, $n \geq k$. However, from the context it is always clear which space is meant. Finally, by $\hat{\cdot}$ we denote the *z-transform*, which is given by

$$\hat{u}(z) := \sum_{j\in\mathbb{Z}} u(j) z^j, \qquad u \in \ell_2(\mathbb{Z})^n.$$

The z-transform is an isometric isomorphism from $\ell_2(\mathbb{N}_0)^n$ onto $H_2(\mathbb{D})^n$, and from $\ell_2(\mathbb{Z})^n$ onto $L_2(\mathbb{T})^n$.

3. Review on LTI($\mathbb{N}_0$)-systems. Systems over the signal space $\ell_2(\mathbb{N}_0)$ are well studied in the literature; see [5] and the references therein. Transfer functions of these systems, which are matrices with entries in the quotient field of $H_\infty(\mathbb{D})$, are studied in even more detail. Following [5] we define a linear time-invariant system as follows.

DEFINITION 3.1. *A linear operator $P : D(P) \subset \ell_2(\mathbb{N}_0)^m \to \ell_2(\mathbb{N}_0)^p$ is called an* LTI$(\mathbb{N}_0)^{p\times m}$-system *if P is* shift-invariant, *i.e. $SG(P) \subset G(P)$, and $G(P)$ is a closed subspace of $\ell_2(\mathbb{N}_0)^{m+p}$.*

Here $D(P)$ denotes the *domain* of P, and $G(P) \subset \ell_2(\mathbb{N}_0)^{m+p}$ denotes the *graph* of P. If P is an LTI$(\mathbb{N}_0)^{p\times m}$-system and the dimensions of the input and output space are not important, we just say P is an LTI$(\mathbb{N}_0)$-*system*. An LTI$(\mathbb{N}_0)^{p\times m}$-system is called *stable* if $D(P) = \ell_2(\mathbb{N}_0)^m$. We stated the definition of an LTI$(\mathbb{N}_0)$-system in the time-domain. Using the z-transform and the fact that the z-transform is an isometric isomorphism from $\ell_2(\mathbb{N}_0)$ onto $H_2(\mathbb{D})$, we can interchangeably use the description of P in the frequency domain. By $\hat{P} : D(\hat{P}) \subset H_2(\mathbb{D})^p \to H_2(\mathbb{D})^m$ we denote the z-transform of P, that is, $\hat{P}\hat{u} = \widehat{Pu}$, $\hat{u} \in D(\hat{P}) = \widehat{D(P)}$.

DEFINITION 3.2. *We call an LTI$(\mathbb{N}_0)^{p\times m}$-system P* maximal *if for any LTI$(\mathbb{N}_0)^{p\times m}$-system $\tilde{P}$ with $G(P) \subset G(\tilde{P})$ we actually have $G(P) = G(\tilde{P})$. Further, an LTI$(\mathbb{N}_0)^{p\times m}$-system P is called* causal *if for every $u \in \ell_2(\mathbb{N}_0)^m$ and every $k \in \mathbb{N}_0$ with $S^k u \in D(P)$ there exists $v \in \ell_2(\mathbb{N}_0)^p$ such that $PS^k u = S^k v$.*

Causality guarantees that the previous output does not depend on the present and future input. Note that a stable LTI$(\mathbb{N}_0)$-system is always causal [5]. Next we show that every maximal LTI$(\mathbb{N}_0)^{p\times m}$-system is uniquely described by an element of $R(H_\infty(\mathbb{D}))^{p\times m}$. Here $R(H_\infty(\mathbb{D}))$ denotes the *quotient field of* $H_\infty(\mathbb{D})$, which is the set of equivalent classes of fractions $\frac{n}{m}$, where $n, m \in H_\infty(\mathbb{D})$ and $m \neq 0$.

THEOREM 3.3. *Any maximal LTI$(\mathbb{N}_0)^{p\times m}$-system P uniquely defines a function $\mathcal{P} \in R(H_\infty(\mathbb{D}))^{p\times m}$ which determines the action of P in the frequency domain, that is, $\hat{P}\hat{x} = \mathcal{P}\hat{x}$, $\hat{x} \in D(\hat{P})$. Conversely, any matrix $\mathcal{P} \in R(H_\infty(\mathbb{D}))^{p\times m}$ uniquely specifies a maximal LTI$(\mathbb{N}_0)^{p\times m}$-system P which satisfies $\hat{P}\hat{x} = \mathcal{P}\hat{x}$, $\hat{x} \in D(\hat{P})$.*

For the proof of this theorem we refer the reader to [5]. The function $\mathcal{P} \in R(H_\infty(\mathbb{D}))^{p\times m}$, which is uniquely given by a maximal LTI$(\mathbb{N}_0)^{p\times m}$-system, is called a *transfer function*. Of special interest are left and right coprime factorizations of transfer functions. A function $\theta \in H_\infty(\mathbb{D})^{m\times r}$, $r \leq m$, is called *inner* if $\theta^*(z)\theta(z) = I_r$ for a.e. $z \in \mathbb{T}$ and $\|\theta(z)\| \leq 1$ for every $z \in \mathbb{D}$.

DEFINITION 3.4. *Let $\mathcal{P} \in R(H_\infty(\mathbb{D}))^{p\times m}$. We say $\mathcal{P}$ has a* right coprime factorization (rcf) over $H_\infty(\mathbb{D})$ *if there exist $M \in H_\infty(\mathbb{D})^{m\times m}$ with $\det M \neq 0$, and $N \in H_\infty(\mathbb{D})^{p\times m}$ such that $\mathcal{P} = NM^{-1}$ and $\binom{M}{N}$ is left-invertible over $H_\infty(\mathbb{D})$. If additionally $\binom{M}{N}$ is inner, then we say $\mathcal{P}$ has a* normalized right coprime factorization (normalized rcf) over $H_\infty(\mathbb{D})$. *We say $\mathcal{P}$ has a* left coprime factorization (lcf) over $H_\infty(\mathbb{D})$ *if there exist $\tilde{M} \in H_\infty(\mathbb{D})^{p\times p}$ with $\det \tilde{M} \neq 0$, and $\tilde{N} \in H_\infty(\mathbb{D})^{p\times m}$ such that $\mathcal{P} = \tilde{M}^{-1}\tilde{N}$ and $(\tilde{M}\ \tilde{N})$ is right-invertible over $H_\infty(\mathbb{D})$. If additionally $\tilde{M}\tilde{M}^* + \tilde{N}\tilde{N}^* = I$, then we say $\mathcal{P}$ has a* normalized left coprime factorization (normalized lcf) over $H_\infty(\mathbb{D})$.

DEFINITION 3.5. *Let $\mathcal{P} \in R(H_\infty(\mathbb{D}))^{p\times m}$. We say $\mathcal{P}$ has a* weak right coprime factorization (weak rcf) over $H_\infty(\mathbb{D})$ *if there exist $M \in H_\infty(\mathbb{D})^{m\times m}$ with $\det M \neq 0$, and $N \in H_\infty(\mathbb{D})^{p\times m}$ such that $\mathcal{P} = NM^{-1}$ and $\binom{M}{N}$ is irreducible over $H_\infty(\mathbb{D})$. Here a matrix Q with elements in $H_\infty(\mathbb{D})$ is* irreducible over $H_\infty(\mathbb{D})$ *if one (and hence all) gcd of all highest order minors of M are invertible in $H_\infty(\mathbb{D})$. We say $\mathcal{P}$ has a* weak

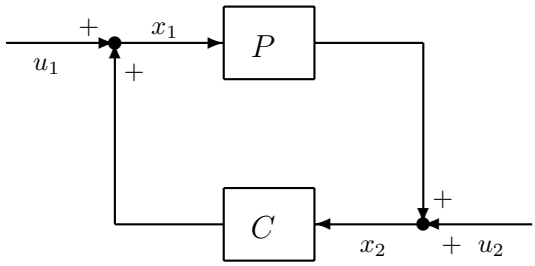


FIG. 3.1. *Standard feedback configuration.*

left coprime factorization (weak lcf) over $H_\infty(\mathbb{D})$, *if there exist* $\tilde{M} \in H_\infty(\mathbb{D})^{p\times p}$ *with* $\det \tilde{M} \neq 0$, *and* $\tilde{N} \in H_\infty(\mathbb{D})^{p\times m}$ *such that* $\mathcal{P} = \tilde{M}^{-1}\tilde{N}$ *and* $(\tilde{M}\ \tilde{N})$ *is irreducible over* $H_\infty(\mathbb{D})$.

It should be noted that not every transfer function of a maximal LTI$(\mathbb{N}_0)$-system possesses an rcf and lcf over $H_\infty(\mathbb{D})$. In general there exist only a weak rcf and weak lcf; see [5]. However, if P is a maximal LTI$(\mathbb{N}_0)^{p\times m}$-system and NM^{-1} is a (weak) rcf of the corresponding transfer function, then the graph of P is given by $G(P) = \binom{M}{N} H_2(\mathbb{D})^m$ (see [5]).We now consider feedback systems as given in Figure 3.1, which is the standard feedback configuration used in system and control theory; see Vidyasagar [19] for more details. We say that such a feedback system is stable if all the paths in the loop in Figure 3.1 are stable, or, more precisely, that Definition 3.6 holds.

DEFINITION 3.6. *Let* P *be an LTI*$(\mathbb{N}_0)^{p\times m}$*-system and let* C *be an LTI*$(\mathbb{N})^{m\times p}$*-system. We say that the feedback system* $[P, C]$, *as given by Figure* 3.1, *is stable if the mapping*

$$F_{[P,C]} := \begin{pmatrix} I & C \\ P & I \end{pmatrix} : D(P) \times D(C) \to \ell_2(\mathbb{N}_0)^{m+p}$$

has a bounded inverse, that is, if the operators $u_i \mapsto x_j$, $i, j = 1, 2$, *are well-defined and bounded. Further,* P *is called* stabilizable *if there is an LTI*$(\mathbb{N}_0)^{m\times p}$*-system* C *such that the feedback system* $[P, C]$ *is stable.*

If $[P, C]$ is stable, then $F^{-1}_{[P,C]}$ is a stable and causal LTI$(\mathbb{N}_0)^{(p+m)\times(p+m)}$-system. In [5], equivalent conditions for stable feedback systems as well as for stabilizable LTI$(\mathbb{N}_0)^{p\times m}$-systems are given. The following result will be of particular interest for us.

THEOREM 3.7. *Let* P *be an LTI*$(\mathbb{N}_0)^{p\times m}$*-system. Then* P *is stabilizable if and only if* P *is maximal and the transfer function* $\mathcal{P}$ *of* P *possesses (normalized) rcf over* $H_\infty(\mathbb{D})$.

The proof can be found in [5]. Further, it is shown in [5] that for every stabilizable LTI$(\mathbb{N}_0)$-system P there exists a causal system C which stabilizes P.

4. LTI$(\mathbb{Z})$-systems. In contrast to systems over the signal space $\ell_2(\mathbb{N}_0)$, systems with signals in $\ell_2(\mathbb{Z})$ are not that well studied. Following [7] we define a linear time-invariant system as follows.

DEFINITION 4.1. *A linear operator* $P : D(P) \subset \ell_2(\mathbb{Z})^m \to \ell_2(\mathbb{Z})^p$ *is called an* LTI$(\mathbb{Z})^{p\times m}$-system *if* P *is* shift-invariant, *i.e.,* $SG(P) = G(P)$, *and* $\overline{D(P)} = \ell_2(\mathbb{Z})^m$.

As above, $D(P)$ denotes the *domain* of P, and $G(P)$ denotes the *graph* of P. If P is an LTI$(\mathbb{Z})^{p\times m}$-system and the dimensions of the input and output space are not

important, we just say P is an *LTI($\mathbb{Z}$)-system*. Note that we do not assume that the graph of P is closed, because a lot of simple systems of relevance do not satisfy this assumption [6]. However, all those systems are densely defined [8] and thus we include this as a requirement in the definition.

DEFINITION 4.2. *We say an LTI($\mathbb{Z}$)$^{p\times m}$-system P is* closed *if the operator P is closed, that is, if $G(P)$ is a closed subspace of $\ell_2(\mathbb{Z})^{m+p}$, and we say an LTI($\mathbb{Z}$)$^{p\times m}$-system P is* closable *if the operator P is closable, i.e., if for every sequence $\{u_n\} \subseteq D(P)$ which tends to 0 and for which Pu_n tends to a function $y \in \ell_2(\mathbb{Z})^p$ we have $y = 0$. Further, P is called* causal *if $u \in \ell_2(\mathbb{N}_0)^m \cap D(P)$ implies $Pu \in \ell_2(\mathbb{N}_0)^p$.*

Closability means that there exists a closed LTI($\mathbb{Z}$)$^{p\times m}$-system $T : D(T) \subset \ell_2(\mathbb{Z})^m \to \ell_2(\mathbb{Z})^p$ such that $D(P) \subset D(T)$ and $Tu = Pu$ for every $u \in D(P)$. If P is closable, then the closure $\overline{P}$ of P is the smallest closed LTI($\mathbb{Z}$)$^{p\times m}$-system, which extends P. We stated the definition of an LTI($\mathbb{Z}$)-system in the time-domain. Using the z-transform and the fact that the z-transform is an isometric isomorphism, we can interchangeably use the description of P in the frequency domain. The z-transform of an LTI($\mathbb{Z}$)-system P is denoted by $\hat{P}$. In general, a causal LTI($\mathbb{Z}$)-system can have a noncausal closure; see [6], [7], [10]. This phenomenon is called the Georgiou–Smith paradox. In [7], closable causal systems with causal closure are studied and described via equivalent conditions. Next we introduce the notion of a transfer function of an LTI($\mathbb{Z}$)-system. The *total ring of fractions of* $L_\infty(\mathbb{T})$, denoted by $R(L_\infty(\mathbb{T}))$, is defined to be the set of equivalent classes of fractions $\frac{n}{m}$, where $n, m \in L_\infty(\mathbb{T})$ and $m(z) \neq 0$ a.e.

DEFINITION 4.3. *Let P be a closed LTI($\mathbb{Z}$)$^{p\times m}$-system. We call a function $\mathcal{P} \in R(L_\infty(\mathbb{T}))^{p\times m}$ the* transfer function *of P if $\hat{P}\hat{u} = \mathcal{P}\hat{u}$ for $\hat{u} \in D(\hat{P})$.*

A transfer function describes the action of the system P in the frequency domain. In [7] it is shown that every closed LTI($\mathbb{Z}$)-system possesses a transfer function, and a transfer function is uniquely determined. Further, a function $\mathcal{P} \in R(L_\infty(\mathbb{T}))^{p\times m}$ completely describes a closed LTI($\mathbb{Z}$)-system [7]. We say a matrix $M \in L_\infty(\mathbb{T})^{m\times m}$ is regular if $\det M(z) \neq 0$ for a.e. $z \in \mathbb{T}$. Of special interest are left and right coprime factorizations of transfer functions.

DEFINITION 4.4. *Let $\mathcal{P} \in R(L_\infty(\mathbb{T}))^{p\times m}$. We say $\mathcal{P}$ has a* right coprime factorization (rcf) over $L_\infty(\mathbb{T})$ *if there exist $M \in L_\infty(\mathbb{T})^{m\times m}$ regular and $N \in L_\infty(\mathbb{T})^{p\times m}$ such that $\mathcal{P} = NM^{-1}$ and $\binom{M}{N}$ is left-invertible over $L_\infty(\mathbb{T})$. If additionally $M^*M + N^*N = I$ holds, then we say that $\mathcal{P}$ has a* normalized rcf over $L_\infty(\mathbb{T})$. *We say $\mathcal{P}$ has a* left coprime factorization (lcf) over $L_\infty(\mathbb{T})$ *if there exist $\tilde{M} \in L_\infty(\mathbb{T})^{p\times p}$ regular and $\tilde{N} \in L_\infty(\mathbb{T})^{p\times m}$ such that $\mathcal{P} = \tilde{M}^{-1}\tilde{N}$ and $(\,\tilde{M}\ \tilde{N}\,)$ is right-invertible over $L_\infty(\mathbb{T})$. If additionally $\tilde{M}\tilde{M}^* + \tilde{N}\tilde{N}^* = I$ holds, then we say $\mathcal{P}$ has a* normalized lcf over $L_\infty(\mathbb{T})$.

In the following proposition it is shown that every $\mathcal{P} \in R(L_\infty(\mathbb{T}))^{p\times m}$ possesses normalized coprime factorizations over $L_\infty(\mathbb{T})$.

PROPOSITION 4.5. *Every $\mathcal{P} \in R(L_\infty(\mathbb{T}))^{p\times m}$ possesses a normalized lcf and a normalized rcf over $L_\infty(\mathbb{T})$. Moreover, an lcf and an rcf are unique up to an invertible element of $L_\infty(\mathbb{T})^{m\times m}$.*

Proof. It is easy to see that there exist matrices $N_1 \in L_\infty(\mathbb{T})^{p\times m}$ and $M_1 \in L_\infty(\mathbb{T})^{m\times m}$, M_1 diagonal and regular, such that $\mathcal{P} = N_1M_1^{-1}$. We define $G_1 \in L_\infty(\mathbb{T})^{(m+p)\times m}$ by $G_1 := \binom{M_1}{N_1}$. Using polar factorization (see, for example, Weidmann [20]), there exist $T \in L_\infty(\mathbb{T})^{m\times m}$ and $G \in L_\infty(\mathbb{T})^{(m+p)\times m}$ such that $G_1 = GT$, and $G^*G = I$. Actually, T is given by the square root of $G_1^*G_1$. Since M_1 is regular, T has to be regular as well. We split G as $G = \binom{M}{N}$ with $M \in L_\infty(\mathbb{T})^{m\times m}$ and

$N \in L_\infty(\mathbb{T})^{p\times m}$. Clearly, M is regular, $G^*G = I$, and $\mathcal{P} = NM^{-1}$. Thus $\mathcal{P}$ possesses a normalized rcf over $L_\infty(\mathbb{T})$.

Next we show the uniqueness of the rcf over $L_\infty(\mathbb{T})$. Let us consider two rcf's over $L_\infty(\mathbb{T})$ of $\mathcal{P}$, that is, $\mathcal{P} = N_1M_1^{-1} = N_2M_2^{-1}$ with $\binom{M_1}{N_1}, \binom{M_2}{N_2} \in L_\infty(\mathbb{T})^{(m+p)\times m}$ left-invertible over $L_\infty(\mathbb{T})$, and M_1, M_2 regular. This implies $\binom{M_1}{N_1} = \binom{M_2}{N_2}M_2^{-1}M_1$ and $\binom{M_2}{N_2} = \binom{M_1}{N_1}M_1^{-1}M_2$. Now Proposition 3.9 of [7] shows $M_2^{-1}M_1, M_1^{-1}M_2 \in L_\infty(\mathbb{T})^{m\times m}$, and since $(M_1^{-1}M_2)^{-1} = M_2^{-1}M_1$, the rcf over $L_\infty(\mathbb{T})$ is unique up to multiplication by an invertible matrix in $L_\infty(\mathbb{T})^{m\times m}$.

In order to show that $\mathcal{P}$ also possesses an lcf over $L_\infty(\mathbb{T})$, we proceed as follows. It is easy to see that we can write $\mathcal{P}$ as $\mathcal{P} = \tilde{M}_1^{-1}\tilde{N}_1$ with $\tilde{M}_1 \in L_\infty(\mathbb{T})^{p\times p}$ diagonal and regular, and $\tilde{N}_1 \in L_\infty(\mathbb{T})^{p\times m}$. We define G_1 by $G_1 := \binom{\tilde{M}_1^T}{\tilde{N}_1^T}$. Applying the same procedure as above we get that there exist matrices $\tilde{M}$ and $\tilde{N}$ of the required form, and that the lcf over $L_\infty(\mathbb{T})$ is unique up to multiplication by an invertible matrix in $L_\infty(\mathbb{T})^{m\times m}$. □

An LTI$(\mathbb{Z})^{p\times m}$-system P is *stable* if P is closed and $D(P) = \ell_2(\mathbb{Z})^m$. Using the closed graph theorem we see that P is stable if and only if P is a linear bounded operator from $\ell_2(\mathbb{Z})^m$ to $\ell_2(\mathbb{Z})^p$. In terms of the transfer function, a closed LTI$(\mathbb{Z})^{p\times m}$-system is stable if and only if the transfer function $\mathcal{P}$ satisfies $\mathcal{P} \in L_\infty(\mathbb{T})^{p\times m}$. We will see later on that a stable LTI$(\mathbb{Z})$-system is not automatically causal, as is the case for stable LTI$(\mathbb{N}_0)$-systems. Here a stable LTI$(\mathbb{Z})^{p\times m}$-system is causal if and only if its transfer function is an element of $H_\infty(\mathbb{D})^{p\times m}$.

We now consider feedback systems as given in Figure 3.1, which is the standard feedback configuration [19]. We say that such a feedback system is stable if all the paths in the loop in Figure 3.1 are stable, or, more precisely, that Definition 4.6 holds.

DEFINITION 4.6. *Let P be an LTI$(\mathbb{Z})^{p\times m}$-system, and let C be an LTI$(\mathbb{Z})^{m\times p}$-system. We say that the feedback system $[P, C]$, as given in Figure* 3.1, *is* stable *if*

$$F_{[P,C]} := \begin{pmatrix} I & C \\ P & I \end{pmatrix} : D(P) \times D(C) \to \ell_2(\mathbb{Z})^{p+m} : \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix} \to \begin{pmatrix} u_1 \\ -u_2 \end{pmatrix}$$

has a bounded inverse, that is, the operators $u_i \to x_j$, $i, j = 1, 2$, are well defined and bounded. If $[P, C]$ is stable, then we denote the inverse of $F_{[P,C]}$ by $H_{[P,C]}$.

In Figure 3.1 we have

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + \begin{pmatrix} 0 & C \\ P & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

If the feedback system is stable, it is easy to see that the operators $I - PC : D(C) \to \ell_2(\mathbb{Z})^p$ and $I - CP : D(P) \to \ell_2(\mathbb{Z})^m$ are boundedly invertible, and that the inverse of $F_{[P,C]}$ is given by

$$\begin{aligned} H_{[P,C]} \begin{pmatrix} u_1 \\ -u_2 \end{pmatrix} &= \begin{pmatrix} (I-CP)^{-1} & -(I-CP)^{-1}C \\ -(I-PC)^{-1}P & (I-PC)^{-1} \end{pmatrix} \begin{pmatrix} u_1 \\ -u_2 \end{pmatrix} \\ &= \begin{pmatrix} (I-CP)^{-1} & -(I-CP)^{-1}C \\ -P(I-CP)^{-1} & I + P(I-CP)^{-1}C \end{pmatrix} \begin{pmatrix} u_1 \\ -u_2 \end{pmatrix}. \end{aligned} \tag{4.1}$$

The stability of $[P, C]$ implies that $H_{[P,C]}$ is a stable LTI$(\mathbb{Z})^{(m+p)\times(m+p)}$-system; that is, the transfer function of $H_{[P,C]}$, denoted by $\mathcal{H}_{[P,C]}$, satisfies $\mathcal{H}_{[P,C]} \in L_\infty(\mathbb{T})^{(m+p)\times(m+p)}$.

Further, we get that $\mathcal{H}_{[P,C]}$ is given by

$$\begin{aligned}\mathcal{H}_{[P,C]} &= \begin{pmatrix} (I-\mathcal{C}\mathcal{P})^{-1} & -(I-\mathcal{C}\mathcal{P})^{-1}\mathcal{C} \\ -(I-\mathcal{P}\mathcal{C})^{-1}\mathcal{P} & (I-\mathcal{P}\mathcal{C})^{-1} \end{pmatrix} \\ &= \begin{pmatrix} (I-\mathcal{C}\mathcal{P})^{-1} & -(I-\mathcal{C}\mathcal{P})^{-1}\mathcal{C} \\ -\mathcal{P}(I-\mathcal{C}\mathcal{P})^{-1} & I+\mathcal{P}(I-\mathcal{C}\mathcal{P})^{-1}\mathcal{C} \end{pmatrix}. \end{aligned} \tag{4.2}$$

The *inverse graph* of an LTI($\mathbb{Z}$)-system P is defined by $G^I(P) := \binom{P}{I} D(P)$. We have the following necessary conditions for stability of feedback systems. The proof follows Georgiou and Smith [5].

PROPOSITION 4.7. *Let P be an LTI$(\mathbb{Z})^{p\times m}$-system and let C be an LTI$(\mathbb{Z})^{m\times p}$-system. If $[P,C]$ is stable, then P and C are closed systems.*

Proof. Let $\{v_n\}_n \subset D(P)$ be a sequence which converges to v in $\ell_2(\mathbb{Z})^m$ and let $\{Pv_n\}_n$ converge to y in $\ell_2(\mathbb{Z})^p$. As n tends to ∞ in

$$\begin{pmatrix} v_n \\ 0 \end{pmatrix} = H_{[P,C]} \begin{pmatrix} v_n \\ Pv_n \end{pmatrix}$$

we get

$$\begin{pmatrix} v \\ 0 \end{pmatrix} = H_{[P,C]} \begin{pmatrix} v \\ y \end{pmatrix},$$

which implies $v \in D(P)$ and $y = Pv$. Thus P is closed. Similarly, it can be proved that C is closed. □

Next we characterize stable feedback systems $[P,C]$ by means of equivalent conditions. The proof of the equivalence of parts 1 and 2 follows Georgiou and Smith [5]. The equivalence of parts 1, 3, 4, 5, and 6 is standard using the coprime factorization approach (see Vidyasagar [19]), and the proof is based on these known results.

THEOREM 4.8. *Let P be a closed LTI$(\mathbb{Z})^{p\times m}$-system and let C be a closed LTI$(\mathbb{Z})^{m\times p}$-system. $\mathcal{P}$ denotes the transfer function of P with rcf $\mathcal{P} = NM^{-1}$ over $L_\infty(\mathbb{T})$ and lcf $\mathcal{P} = \tilde{M}^{-1}\tilde{N}$ over $L_\infty(\mathbb{T})$, and $\mathcal{C}$ denotes the transfer function of C with rcf $\mathcal{C} = TS^{-1}$ over $L_\infty(\mathbb{T})$ and lcf $\mathcal{C} = \tilde{S}^{-1}\tilde{T}$ over $L_\infty(\mathbb{T})$. Then the following statements are equivalent.*

1. *$[P,C]$ is stable.*
2. *$G(P) \cap G^I(C) = \{0\}$, and $G(P) + G^I(C) = \ell_2(\mathbb{Z})^{m+p}$.*
3. *$\left(\begin{smallmatrix} M & T \\ N & S \end{smallmatrix}\right)$ is invertible over $L_\infty(\mathbb{T})$.*
4. *$\left(\begin{smallmatrix} I & \mathcal{C} \\ \mathcal{P} & I \end{smallmatrix}\right)^{-1} \in L_\infty(\mathbb{T})^{(m+p)\times(m+p)}$.*
5. *$\tilde{S}M - \tilde{T}N$ is invertible over $L_\infty(\mathbb{T})$.*
6. *$\tilde{M}S - \tilde{N}T$ is invertible over $L_\infty(\mathbb{T})$.*

It is easy to see that the inverse of $\left(\begin{smallmatrix} I & \mathcal{C} \\ \mathcal{P} & I \end{smallmatrix}\right)$, if it exists, is given by $\mathcal{H}_{[P,C]}$.

Proof.

1⇒2 The stability of $[P,C]$ implies $G(P) \cap G^I(C) = \{0\}$ and $G(P) + G^I(C) = \ell_2(\mathbb{Z})^{p+m}$.

2⇒3 We get

$$L_2(\mathbb{T})^{m+p} = G(\hat{P}) + G^I(\hat{C}) = \begin{pmatrix} M \\ N \end{pmatrix} L_2(\mathbb{T})^m + \begin{pmatrix} T \\ S \end{pmatrix} L_2(\mathbb{T})^p. \tag{4.3}$$

We define the multiplication operator $X : L_2(\mathbb{T})^m \to L_2(\mathbb{T})^m$ by $Xu := \left(\begin{smallmatrix} M & T \\ N & S \end{smallmatrix}\right) u$, $u \in L_2(\mathbb{T})^m$. Equation (4.3) shows that X as well as $\left(\begin{smallmatrix} M & T \\ N & S \end{smallmatrix}\right)(z)$

is surjective for a.e. $z \in \mathbb{T}$, and hence $\left(\begin{smallmatrix} M & T \\ N & S \end{smallmatrix}\right)(z)$ is injective for a.e. $z \in \mathbb{T}$. This shows that X is injective. Finally, the invertibility of X (and hence the invertibility of $\left(\begin{smallmatrix} M & T \\ N & S \end{smallmatrix}\right)$ over $L_\infty(\mathbb{T})$) is implied by the open mapping theorem.

3⇒1 The invertibility of $\left(\begin{smallmatrix} M & T \\ N & S \end{smallmatrix}\right)$ over $L_\infty(\mathbb{T})$ implies that $G(P) \cap G^I(C) = \{0\}$ and $G(P) + G^I(C) = \ell_2(\mathbb{Z})^{m+p}$. This shows that $F_{[P,C]}$ is injective and surjective. Thus the inverse of $F_{[P,C]}$ exists, and it remains only to show that the inverse is bounded. Using the closed graph theorem, this holds if $F_{[P,C]}$ has a closed graph. The graph of $F_{[P,C]}$ is given by

$$\left\{ \begin{pmatrix} x_1 \\ -x_2 \\ x_1 - Cx_2 \\ Px_1 - x_2 \end{pmatrix} \;\middle|\; x_1 \in D(P), x_2 \in D(C) \right\},$$

and it is closed since P and C are closed.

3⇔4 The equation

$$\begin{pmatrix} I & \mathcal{C} \\ \mathcal{P} & I \end{pmatrix}^{-1} = \begin{pmatrix} M & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} M & T \\ N & S \end{pmatrix}^{-1}$$

shows that part 3 implies part 4. If $\left(\begin{smallmatrix} I & \mathcal{C} \\ \mathcal{P} & I \end{smallmatrix}\right)$ is invertible, we have

$$\begin{pmatrix} M & 0 \\ 0 & S \\ M & T \\ N & S \end{pmatrix} \begin{pmatrix} M & T \\ N & S \end{pmatrix}^{-1} = \begin{pmatrix} \begin{pmatrix} I & \mathcal{C} \\ \mathcal{P} & I \end{pmatrix}^{-1} \\ \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \end{pmatrix} \in L_\infty(\mathbb{T})^{(2m+2p)\times(m+p)}.$$

Using the fact that $\binom{M}{N}$ and $\binom{S}{T}$ are left-invertible over $L_\infty(\mathbb{T})$, Proposition 3.9 of [7] shows that

$$\begin{pmatrix} M & 0 \\ 0 & S \\ 0 & T \\ N & 0 \end{pmatrix}$$

is left-invertible over $L_\infty(\mathbb{T})$. Thus the matrix

$$\begin{pmatrix} M & 0 \\ 0 & S \\ M & T \\ N & S \end{pmatrix} = \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ I & 0 & I & 0 \\ 0 & I & 0 & I \end{pmatrix} \begin{pmatrix} M & 0 \\ 0 & S \\ 0 & T \\ N & 0 \end{pmatrix}$$

is left-invertible over $L_\infty(\mathbb{T})$, and hence $\left(\begin{smallmatrix} M & T \\ N & S \end{smallmatrix}\right)$ is invertible over $L_\infty(\mathbb{T})$.

4⇔5 We define $D := \tilde{S}M - \tilde{T}N$. Clearly, $D = \tilde{S}(I - \mathcal{C}\mathcal{P})M$. Thus $D^{-1} = M^{-1}(I - \mathcal{C}\mathcal{P})^{-1}\tilde{S}^{-1}$ or, equivalently, $(I - \mathcal{C}\mathcal{P})^{-1} = MD^{-1}\tilde{S}$. We first assume that part 4 holds. An easy calculation shows that $\left(\begin{smallmatrix} I & \mathcal{C} \\ \mathcal{P} & I \end{smallmatrix}\right)^{-1}$ equals the matrix $\mathcal{H}_{[P,C]}$, given by (4.2). Thus the functions $(I - \mathcal{C}\mathcal{P})^{-1}$, $(I - \mathcal{C}\mathcal{P})^{-1}\mathcal{C}$, $\mathcal{P}(I - \mathcal{C}\mathcal{P})^{-1}$, and $\mathcal{P}(I - \mathcal{C}\mathcal{P})^{-1}\mathcal{C}$ are measurable and essentially bounded on $\mathbb{T}$. Using $(I - \mathcal{C}\mathcal{P})^{-1} = MD^{-1}\tilde{S}$, this shows that

$$MD^{-1}\tilde{S}, \quad MD^{-1}\tilde{T}, \quad ND^{-1}\tilde{S}, \quad \text{and} \quad ND^{-1}\tilde{T}$$

are measurable and essentially bounded on $\mathbb{T}$. By Proposition 3.9 in [7] we get that

$$D^{-1}\tilde{S} \quad \text{and} \quad D^{-1}\tilde{T}$$

are measurable and essentially bounded on $\mathbb{T}$. Using again Proposition 3.9 in [7] we see that $D^{-1} \in L_\infty(\mathbb{T})^{m\times m}$. Next we assume that part 5 holds. Thus D is invertible over $L_\infty(\mathbb{T})$. Using $(I-\mathcal{C}\mathcal{P})^{-1} = MD^{-1}\tilde{S}$, we get that

$$\begin{pmatrix} I & \mathcal{C} \\ \mathcal{P} & I \end{pmatrix}^{-1} \in L_\infty(\mathbb{T})^{(m+p)\times(m+p)}.$$

4⇔6 The proof of this equivalence is similar to the proof of 4⇔5. □

Beside stability another important property of a feedback system is causality, which is defined as follows.

DEFINITION 4.9. *Let P be an LTI$(\mathbb{Z})^{p\times m}$-system and let C be an LTI$(\mathbb{Z})^{m\times p}$-system such that the feedback system $[P,C]$ is stable. We say the feedback system $[P,C]$ is* causal *if $H_{[P,C]}$, as given in* (4.1)*, is causal.*

Equivalently, $[P,C]$ is causal if and only if $\mathcal{H}_{[P,C]} \in H_\infty(\mathbb{D})^{(m+p)\times(m+p)}$. Considering LTI$(\mathbb{N}_0)$-systems, we see that every stable feedback system is automatically causal [5]. Unfortunately, this is not the case for LTI$(\mathbb{Z})$-systems, as the following example shows. Note that in the example both systems P and C are causal, whereas the feedback system is not causal.

Example 4.10. We consider a feedback system $[P,C]$, which is given by

$$\begin{aligned} Pu &:= u, \quad u \in \ell_2(\mathbb{Z}), \\ (Cu)(t) &:= u(t) - u(t-1), \quad u \in \ell_2(\mathbb{Z}),\ t \in \mathbb{Z}. \end{aligned}$$

Clearly, P and C are stable, causal LTI$(\mathbb{Z})^{1\times 1}$-systems, and $G(\hat{P}) = \binom{1}{1}L_2(\mathbb{T})$ and $G(\hat{C}) = \binom{1}{1-z}L_2(\mathbb{T})$. Since $D := 1\cdot 1 - (1-z)\cdot 1 = z$ is invertible over $L_\infty(\mathbb{T})$, Theorem 4.8 shows that the feedback system $[P,C]$ is stable. In order to show that the feedback system is not causal, we choose the inputs $u_1 := (\dots,0,1,0,\dots)$, where the 1 stands at position 0, and $u_2 := 0$. This choice implies $x_1 = (\dots,0,1,0,\dots)$, where the 1 stands at position -1, and thus the feedback system is not causal.

Next we give a necessary condition for a feedback system to be causal.

PROPOSITION 4.11. *Let P be an LTI$(\mathbb{Z})^{p\times m}$-system and let C be an LTI$(\mathbb{Z})^{m\times p}$-system such that the feedback system $[P,C]$ is stable and causal. Then the transfer functions of P and C, denoted by $\mathcal{P}$ and $\mathcal{C}$, respectively, satisfy*

$$\mathcal{P} \in R(H_\infty(\mathbb{D}))^{p\times m} \quad \text{and} \quad \mathcal{C} \in R(H_\infty(\mathbb{D}))^{m\times p},$$

and $\mathcal{P}$ as well as $\mathcal{C}$ possesses normalized rcf's and normalized lcf's over $H_\infty(\mathbb{D})$.

Note that the stability of $[P,C]$ implies that the systems P and C are closed; see Proposition 4.7.

Proof. We define $u_1^{(1)}, u_1^{(2)}, \dots, u_1^{(p)} \in H_\infty(\mathbb{D})^m$ by $u_1^{(1)} = \cdots = u_1^{(p)} = 0$ and $u_2^{(1)}, u_2^{(2)}, \dots, u_2^{(p)} \in H_\infty(\mathbb{D})^p$ by $u_2^{(j)} = (0,\dots,0,1,0,\dots,0)^T$, $j \in \{1,\dots,p\}$, where the 1 stands at position j. Thus, the matrix $(u_2^{(1)},\dots,u_2^{(p)})$ is invertible over $H_\infty(\mathbb{D})$. Since the feedback system $[P,C]$ is stable and causal, there exist unique elements $x_1^{(1)},\dots,x_1^{(p)} \in H_2(\mathbb{D})^m$ and $x_2^{(1)},\dots,x_2^{(p)} \in H_2(\mathbb{D})^p$ such that

$$x_1^{(j)} = \mathcal{C}x_2^{(j)}, \qquad j \in \{1,\dots,p\}, \tag{4.4}$$

$$x_2^{(j)} = u_2^{(j)} + \mathcal{P}x_1^{(j)}, \qquad j \in \{1,\dots,p\}. \tag{4.5}$$

Using $x_1^{(j)} = (I - \mathcal{CP})^{-1}\mathcal{C}u_2^{(j)}$, $x_2^{(j)} = (I - \mathcal{PC})^{-1}u_2^{(j)}$, $(I - \mathcal{CP})^{-1}\mathcal{C} \in H_\infty(\mathbb{D})^{p\times m}$, $(I - \mathcal{PC})^{-1} \in H_\infty(\mathbb{D})^{p\times p}$, and $u_2^{(j)} \in H_\infty(\mathbb{D})^p$, we get $x_1^{(j)} \in H_\infty(\mathbb{D})^m$ and $x_2^{(j)} \in H_\infty(\mathbb{D})^p$, $j \in \{1,\dots,p\}$. Next we show that the determinant of $U := (x_2^{(1)},\dots,x_2^{(p)}) \in H_\infty(\mathbb{D})^{p\times p}$ is not zero. We assume that the determinant is zero. Similar to Proposition 3.12 in [7] it can be shown that there are functions $\alpha_1,\dots,\alpha_p \in H_\infty(\mathbb{D})$ such that $|\alpha_1|^2 + \cdots + |\alpha_p|^2 > 0$ and

$$\alpha_1 x_2^{(1)} + \cdots + \alpha_p x_2^{(p)} = 0.$$

Thus equation (4.4) implies

$$\alpha_1 x_1^{(1)} + \cdots + \alpha_p x_1^{(p)} = \mathcal{C}\left(\alpha_1 x_2^{(1)} + \cdots + \alpha_p x_2^{(p)}\right) = 0.$$

This shows, using equation (4.5),

$$\alpha_1 u_2^{(1)} + \cdots + \alpha_p u_2^{(p)} = \alpha_1 x_2^{(1)} + \cdots + \alpha_p x_2^{(p)} - \mathcal{P}\left(\alpha_1 x_1^{(1)} + \cdots + \alpha_p x_1^{(p)}\right) = 0.$$

However, this is in contradiction to the definition of $u_2^{(j)}$, $j = 1,\dots,p$. Thus $\det U \neq 0$, which implies $U^{-1} \in R(H_\infty(\mathbb{D}))^{p\times p}$.

Further, we define $V := \mathcal{C}U = (x_1^{(1)},\dots,x_1^{(j)}) \in H_\infty(\mathbb{D})^{m\times p}$, and thus $\mathcal{C} = VU^{-1} \in R(H_\infty(\mathbb{D}))^{m\times p}$. In a similar manner it can be proved that $\mathcal{P} \in R(H_\infty(\mathbb{D}))^{p\times m}$.

The stability and causality of the feedback system show that the functions $(I-\mathcal{CP})^{-1}$, $(I - \mathcal{CP})^{-1}\mathcal{C}$, $\mathcal{P}(I - \mathcal{CP})^{-1}$, and $\mathcal{P}(I - \mathcal{CP})^{-1}\mathcal{C}$ are holomorphic and bounded on $\mathbb{D}$. Let $\mathcal{P} = NM^{-1}$ be a weak rcf over $H_\infty(\mathbb{D})$ of P and let $\mathcal{C} = \tilde{S}^{-1}\tilde{T}$ be a weak lcf over $H_\infty(\mathbb{D})$ of C. For the proof that every element of $H_\infty(\mathbb{D})^{m\times p}$ possesses a weak lcf and a weak rcf over $H_\infty(\mathbb{D})$ we refer the reader to [5] and [16]. Then we get that

$$M(\tilde{S}M - \tilde{T}N)^{-1}\tilde{S}, \qquad M(\tilde{S}M - \tilde{T}N)^{-1}\tilde{T}, \qquad N(\tilde{S}M - \tilde{T}N)^{-1}\tilde{S},$$
$$\text{and} \quad N(\tilde{S}M - \tilde{T}N)^{-1}\tilde{T}$$

are holomorphic and bounded on $\mathbb{D}$. By Lemma 4 of Smith [16] we get that

$$(\tilde{S}M - \tilde{T}N)^{-1}\tilde{S} \quad \text{and} \quad (\tilde{S}M - \tilde{T}N)^{-1}\tilde{T}$$

are holomorphic and bounded on $\mathbb{D}$. Using again Lemma 4 of Smith [16] we see that $(\tilde{S}M - \tilde{T}N)^{-1} \in H_\infty(\mathbb{D})^{m\times m}$. This shows that $\mathcal{P} = NM^{-1}$ is an rcf over $H_\infty(\mathbb{D})$ of P and that $\mathcal{C} = \tilde{S}^{-1}\tilde{T}$ is an lcf over $H_\infty(\mathbb{D})$ of C. In a similar manner it can be shown that $\mathcal{C}$ possesses an rcf over $H_\infty(\mathbb{D})$ and that $\mathcal{P}$ possesses an lcf over $H_\infty(\mathbb{D})$. Finally, the existence of normalized rcf's (lcf's) is shown in [5], [16]. □

Next we formulate equivalent conditions for a stable feedback system to be causal. The results are based on standard results using the coprime factorization approach (see Vidyasagar [19]). Similar results for systems over the signal space $\ell_2(\mathbb{N}_0)$ can be found in Georgiou and Smith [5].

THEOREM 4.12. *Let P be an $LTI(\mathbb{Z})^{p\times m}$-system and let C be an $LTI(\mathbb{Z})^{m\times p}$-system such that the feedback system $[P, C]$ is stable. We assume that the transfer functions of P and C, denoted by $\mathcal{P}$ and $\mathcal{C}$, respectively, satisfy $\mathcal{P} \in R(H_\infty(\mathbb{D}))^{p\times m}$ and $\mathcal{C} \in R(H_\infty(\mathbb{D}))^{m\times p}$ and that both transfer functions possess rcf's and lcf's over $H_\infty(\mathbb{D})$, denoted by $\mathcal{P} = NM^{-1} = \tilde{M}^{-1}\tilde{N}$ and $\mathcal{C} = TS^{-1} = \tilde{S}^{-1}\tilde{T}$. Then the following statements are equivalent.*

1. $[P, C]$ *is causal.*
2. $\left(\begin{smallmatrix} M & T \\ N & S \end{smallmatrix}\right)$ *is invertible over* $H_\infty(\mathbb{D})$.
3. $\left(\begin{smallmatrix} I & C \\ P & I \end{smallmatrix}\right)^{-1} \in H_\infty(\mathbb{D})^{(m+p)\times(m+p)}$.
4. $\tilde{S}M - \tilde{T}N$ *is invertible over* $H_\infty(\mathbb{D})$.
5. $\tilde{M}S - \tilde{N}T$ *is invertible over* $H_\infty(\mathbb{D})$.

Proof. The equivalence of parts 1 and 3 is easy to see, using the fact that $\left(\begin{smallmatrix} I & \mathcal{P} \\ C & I \end{smallmatrix}\right)^{-1}$ is the transfer function of the feedback system $H_{[P,C]}$. Moreover, the equivalence of parts 2 and 3, parts 3 and 4, and parts 3 and 5 can be proved similarly to Theorem 4.8, using Lemma 4 of Smith [16] instead of Proposition 3.9 of [7]. □

5. Stabilizable LTI($\mathbb{Z}$)-systems. The system C is called a *controller* of P. Note that a controller always is a closed LTI($\mathbb{Z}$)-system. There are different possibilities to define the notion of stabilizability for an LTI($\mathbb{Z}$)-system P. The simplest one would be to require that there exists an LTI($\mathbb{Z}$)-system C such that the feedback system $[P, C]$ is stable. This is the usual definition used in system and control theory. However, in our situation this definition is not suitable, since it would rule out a huge class of important systems from being stabilizable and this definition would not guarantee that the feedback system is causal. Thus we adapt the definition as follows.

DEFINITION 5.1. *An LTI($\mathbb{Z}$)-system P is called* stabilizable *if P is closable and if there exists an LTI($\mathbb{Z}$)-system C such that the feedback system $[\overline{P}, C]$ is stable and causal.*

We obtain the following equivalent condition for stabilizability.

THEOREM 5.2. *Let P be a closable LTI$(\mathbb{Z})^{p\times m}$-system. Then the following statements are equivalent.*

1. *P is stabilizable.*
2. *P is stabilizable by a causal controller.*
3. *The transfer function of $\overline{P}$, denoted by $\mathcal{P}$, possesses a normalized rcf as well as a normalized lcf over $H_\infty(\mathbb{D})$.*
4. *The transfer function of $\overline{P}$, denoted by $\mathcal{P}$, possesses an rcf over $H_\infty(\mathbb{D})$.*

Proof. Clearly, part 3 implies part 4, part 2 implies part 1, and the implication of part 1 to part 3 has been proved in Proposition 4.11.

Next we show that part 4 implies part 1. Part 4 implies that there exists a symbol $\binom{M}{N} \in H_\infty(\mathbb{D})^{(m+p)\times m}$ of $\overline{P}$ which is left-invertible over $H_\infty(\mathbb{D})$. Using Tolokonnikov's lemma (see, for example, Nikol′skiĭ [14, page 293]), there exist matrices $S \in H_\infty(\mathbb{D})^{p\times p}$ and $T \in H_\infty(\mathbb{D})^{m\times p}$ such that

$$X = \begin{pmatrix} M & T \\ N & S \end{pmatrix}$$

is invertible over $H_\infty(\mathbb{D})$. If $\det S \neq 0$, then we can define C via the graph $G(\hat{C}) = \binom{S}{T} L_2(\mathbb{T})^p$. The stability of $[P, C]$ is then implied by Theorem 4.8, since X is invertible over $H_\infty(\mathbb{D})$, and hence over $L_\infty(\mathbb{T})$, and the causality of $[P, C]$ is implied by Theorem 4.12. However it can happen that $\det S = 0$. Clearly, for every $Q \in H_\infty(\mathbb{D})^{m+p}$, the matrix $X_Q := \left(\begin{smallmatrix} M & T+MQ \\ N & S+NQ \end{smallmatrix}\right) = \left(\begin{smallmatrix} M & T \\ N & S \end{smallmatrix}\right)\left(\begin{smallmatrix} I & Q \\ 0 & I \end{smallmatrix}\right)$ is invertible over $H_\infty(\mathbb{D})$, and so it is enough to show that there is a matrix $Q \in H_\infty(\mathbb{D})^{m\times p}$ such that $\det(S + NQ) \neq 0$. We have that $(N\, S)$ is right-invertible over $H_\infty(\mathbb{D})$. Thus there exists at least one $p \times p$-minor of $(N\, S)$ which is nonzero. Let A be such a $p \times p$-submatrix of $(N\, S)$ with the fewest possible columns from N. We now define $Q \in H_\infty(\mathbb{D})^{m\times p}$ as follows. Suppose that we obtain A by excluding columns $j_1, \ldots, j_l$ of S and including columns $i_1, \ldots, i_l$ of N. Let now the i_kth row of Q be equal to the j_kth row of the identity

I_p on $\mathbb{C}^p$, $k = 1, \dots, l$, and be zero otherwise. It is easy to see that $Q \in H_\infty(\mathbb{D})^{m\times p}$. Moreover, since A possesses the fewest possible columns from N, we get that the determinant of $S + NQ = (S\ N)\binom{I_p}{Q}$ is nonzero. This shows that there is a controller which stabilizes P and the feedback system is causal; see Theorem 4.12.

Finally, it remains to show that part 1 implies part 2. Let $\mathcal{P}$ be the transfer function of $\overline{P}$, let $\mathcal{P} = NM^{-1}$ be an rcf over $H_\infty(\mathbb{D})$ of $\mathcal{P}$, and let C be a stabilizing controller with transfer function TS^{-1}. It remains to show that there exists a matrix $Q \in H_\infty(\mathbb{D})^{m\times p}$ such that $S + NQ$ is invertible over $H_\infty(\mathbb{D})$, because then P is causally stabilized by the controller C' with transfer function $(T+MQ)(S+NQ)^{-1} \in H^\infty(\mathbb{D})^{m\times p}$. By Theorem 4.12 the matrix $X = \left(\begin{smallmatrix} M & T \\ N & S \end{smallmatrix}\right)$ is invertible over $H_\infty(\mathbb{D})$. In particular, $(N\ \ S)$ is right-invertible over $H_\infty(\mathbb{D})$. This implies

$$\inf_{z\in\mathbb{D}} (\|N(z)\| + \|S(z)\|) > 0.$$

Now Quadrat [15] (see also Treil [17]) shows the existence of a matrix $Q \in H_\infty(\mathbb{D})^{m\times p}$ with the required properties, and thus the theorem is proved. □

The algorithm to stabilize an LTI($\mathbb{Z}$)-system can also be used for stable noncausal LTI($\mathbb{Z}$)-systems. In this case we obtain a stable and causal system.

6. Comparison: LTI($\mathbb{Z}$)-systems versus LTI($\mathbb{N}_0$)-systems. Every closable LTI($\mathbb{Z}$)-system can be *restricted* to an LTI($\mathbb{N}_0$)-system in the following way. Let P be a closable LTI$(\mathbb{Z})^{p\times m}$-system. Then we define $P_\mathbb{N} : D(P_\mathbb{N}) \subset \ell_2(\mathbb{N}_0)^m \to \ell_2(\mathbb{N}_0)^p$ as the closure of the operator $T : D(T) \subset \ell_2(\mathbb{N}_0)^m \to \ell_2(\mathbb{N}_0)^m$, which is given by

$$\begin{aligned} Tu &:= Pu, \quad u \in D(T), \\ D(T) &:= \{u \in D(P) \cap \ell_2(\mathbb{N}_0)^m \mid Pu \in \ell_2(\mathbb{N}_0)^p\}. \end{aligned}$$

By this definition $P_\mathbb{N}$ is an LTI$(\mathbb{N}_0)^{p\times m}$-system. Next we study the relation between the transfer function of $\overline{P}$ and $P_\mathbb{N}$. We show that if $P_\mathbb{N}$ is stabilizable, then P is stabilizable and both systems have the same transfer function. Moreover, if P is closed and stabilizable, then $P_\mathbb{N}$ is stabilizable and both systems have the same transfer function.

THEOREM 6.1. *Let P be a closable LTI$(\mathbb{Z})^{p\times m}$-system. If $P_\mathbb{N}$ is stabilizable, then P is stabilizable. Moreover, both systems have the same transfer function and they are stabilized by the same controllers.*

Proof. Let $\mathcal{P}_\mathbb{N}$ and $\mathcal{P}$ denote the transfer function of $P_\mathbb{N}$ and $\overline{P}$, respectively. Since $P_\mathbb{N}$ is stabilizable, $\mathcal{P}_\mathbb{N}$ possesses a normalized rcf $N_\mathbb{N}M_\mathbb{N}^{-1}$ over $H_\infty(\mathbb{D})$. Further, $\mathcal{P}$ possesses a normalized rcf NM^{-1} over $L_\infty(\mathbb{T})$. Since

$$\begin{pmatrix} M_\mathbb{N} \\ N_\mathbb{N} \end{pmatrix} H_2(\mathbb{D})^m = G(P_\mathbb{N}) \subset G(\overline{P}) = \begin{pmatrix} M \\ N \end{pmatrix} L_2(\mathbb{T})^m,$$

there exists a matrix $Q \in L_2(\mathbb{T})^{m\times p}$ such that

$$\begin{pmatrix} M_\mathbb{N} \\ N_\mathbb{N} \end{pmatrix} = \begin{pmatrix} M \\ N \end{pmatrix} Q.$$

$\binom{M}{N}^* \binom{M}{N} = I$ now implies $Q \in L_\infty(\mathbb{T})^{m\times m}$. Since $\binom{M_\mathbb{N}}{N_\mathbb{N}}$ is left-invertible over $H_\infty(\mathbb{D})$, there is a $Q_L \in L_\infty(\mathbb{T})^{m\times m}$ such that $Q_L Q = I$. Thus Q is invertible over $L_\infty(\mathbb{T})$, and $G(\overline{P}) = \binom{M_\mathbb{N}}{N_\mathbb{N}} L_2(\mathbb{T})^m$. This proves that P is stabilizable and $\mathcal{P} = N_\mathbb{N}M_\mathbb{N}^{-1} = NM^{-1}$.

That the sets of controllers coincide follows from Theorem 4.12 and Georgiou and Smith [5, Lemma 1]. □

THEOREM 6.2. *Let P be a closed LTI$(\mathbb{Z})^{p\times m}$-system. Then P is stabilizable if and only if $P_{\mathbb{N}}$ is stabilizable. Moreover, both systems have the same transfer function and they are stabilized by the same controllers.*

Proof. In view of the previous theorem it remains to show that when P is stabilizable it implies that $P_{\mathbb{N}}$ is stabilizable. As above, let $\mathcal{P}_{\mathbb{N}}$ and $\mathcal{P}$ denote the transfer function of $P_{\mathbb{N}}$ and P, respectively. Since P is stabilizable, $\mathcal{P}$ possesses a normalized rcf NM^{-1} over $H_\infty(\mathbb{D})$. Since

$$\begin{pmatrix} M \\ N \end{pmatrix} H_2(\mathbb{D})^m \subset G(P_{\mathbb{N}}) \subset G(P) = \begin{pmatrix} M \\ N \end{pmatrix} L_2(\mathbb{T})^m,$$

we get that NM^{-1} is the transfer function of a stabilizable and maximal LTI$(\mathbb{Z})^{p\times m}$-system T (see [5]) with the property $G(T) \subset G(P_{\mathbb{N}})$. Thus $T = P_{\mathbb{N}}$, and the theorem is proved. □

Conclusions. We studied discrete-time systems with signal space $\ell_2(\mathbb{Z})$ from first principal. For these systems Georgiou and Smith [5], [6] discovered that a well-known stabilizable system is not stabilizable when using the canonical definition of stabilizability. Using an adapted notion of stabilizability in the mathematical formalism we are able to guarantee that all—from a physical point of view—stabilizable systems are actually stabilizable in the mathematical formalism. Further, we give equivalent conditions for stabilizability.

In this paper we work with discrete-time LTI-systems only, and now is a good time to discuss how things change when we move to continuous-time LTI-systems. In fact, every result for discrete-time LTI-systems does have a counterpart for continuous-time LTI-systems, and the proofs are basically the same. Instead of using the z-transform, which is an isometric isomorphism between $\ell_2(\mathbb{Z})$ and $L_2(\mathbb{T})$, we use the Laplace transform, which is an isometric isomorphism between $L_2(\mathbb{R})$ and $L_2(i\mathbb{R})$. The subspace $L_2(0,\infty)$ then corresponds to the Hardy class $H_2(\mathbb{C}_+)$.

REFERENCES

[1] C. A. DESOER, R.-W. LIU, J. MURRAY, AND R. SAEKS, *Feedback system design: The fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, 25 (1980), pp. 399–412.

[2] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[3] P. L. DUREN, *Theory of H^p Spaces*, Pure and Applied Mathematics 38, Academic Press, San Diego, 1970.

[4] A. FEINTUCH AND R. SAEKS, *System Theory: A Hilbert Space Approach*, Academic Press, New York, 1975.

[5] T. T. GEORGIOU AND M. C. SMITH, *Graphs, causality and stabilizability: Linear, shift-invariant systems on $L_2[0,\infty)$*, Math. Control Signals Systems, 6 (1993), pp. 195–223.

[6] T. T. GEORGIOU AND M. C. SMITH, *Intrinsic difficulties in using the double-infinite time axis for input-output control theory*, IEEE Trans. Automat. Control, 40 (1995), pp. 516–518.

[7] B. JACOB, *An operator theoretical approach towards systems over the signal space $\ell_2(\mathbb{Z})$*, Integral Equations Operator Theory, 46 (2003), pp. 189–214.

[8] B. JACOB, *Stabilizability and Causality of Discrete-Time Systems over the Signal Space $\ell_2(\mathbb{Z})$*, Habilitation thesis, University of Dortmund, Dortmund, Germany, 2001.

[9] B. JACOB, *Stabilizability of systems on $\ell_2(\mathbb{Z})$*, in Proceedings of the 15th International Symposium of Mathematical Theory of Networks and Systems (MTNS 2002), Notre Dame, IN, 2002, CD-ROM.

[10] B. Jacob and J. R. Partington, *Graphs, closability, and causality of linear time-invariant discrete-time systems*, Internat. J. Control, 73 (2000), pp. 1051–1060.
[11] P. M. Mäkilä, *Puzzles in systems and control*, in Robustness in Identification and Control, Lecture Notes in Control and Inform. Sci. 245, Springer-Verlag, London, 1999, pp. 242–257.
[12] P. M. Mäkilä, *On three puzzles in robust control*, IEEE Trans. Automat. Control, 45 (2000), pp. 552–556.
[13] P. M. Mäkilä, *When is a linear convolution system stabilizable?*, Systems Control Lett., 46 (2002), pp. 371–378.
[14] N. K. Nikoĺskiĭ, *Treatise on the Shift Operator*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1986.
[15] A. Quadrat, *On a general structure of the stabilizing controllers based on stable range*, SIAM J. Control Optim., 42 (2004), pp. 2264–2285.
[16] M. C. Smith, *On stabilization and the existence of coprime factorizations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.
[17] S. Treil, *The stable rank of the algebra H^∞ equals* 1, J. Funct. Anal., 109 (1992), pp. 130–154.
[18] M. Vidyasagar, *Input-output stability of a broad class of linear time-invariant multivariable feedback systems*, SIAM J. Control, 10 (1972), pp. 203–209.
[19] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
[20] J. Weidmann, *Lineare Operatoren in Hilberträumen*, Teubner, Stuttgart, Germany, 1976.
[21] J. C. Willems, *The Analysis of Feedback Systems*, MIT Press, Cambridge, MA, 1971.
[22] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.
[23] G. Zames, *Realizability conditions for nonlinear feedback systems*, IEEE Trans. Circuit Theory, 11 (1964), pp. 186–194.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1535–1582

GAP METRICS, REPRESENTATIONS, AND NONLINEAR ROBUST STABILITY*

M. R. JAMES†, M. C. SMITH‡, AND G. VINNICOMBE‡

Abstract. Various alternative definitions for the nonlinear H_2-, L_2-, and ν-gap metrics are studied. The concept of β-conjugacy and multiplicative homogeneity are introduced to relate the metrics to each other and to compare the stability margins of nonlinear feedback loops expressed in terms of the norms of complementary parallel projections. Left and right representations for the graph of a nonlinear system are studied. A new definition of "normalized" is introduced for left representations. Formulas for the gap metrics as the norm of the product of left and right representations are derived. The problem of controller synthesis for input-affine nonlinear systems to achieve norm bounds on the parallel projection operators is studied for input-affine nonlinear systems. The duality between the optimization of the two parallel projections is highlighted. State-space realizations for the normalized left and right representations are derived using nonlinear H_∞ synthesis methods.

Key words. robust control, nonlinear systems, gap metric, graph representations, controller synthesis, nonlinear H^∞ control, information states

AMS subject classifications. 93D09, 47J05, 49L20

DOI. 10.1137/S0363012901393067

1. Introduction.

This paper is concerned with the approach to robust stability of nonlinear systems using gap metrics following the work of [1], [11], [30]. The paper considers several related issues surrounding the following basic robustness theorem: feedback stability is preserved if gap perturbations do not exceed the inverse of the norm of a nonlinear parallel projection operator associated with the feedback loop. The paper develops and unifies a number of concepts and results concerned with the definition and computation of gap metrics, graph representations, controller synthesis to achieve norm bounds on the parallel projection operators, and state-space realizations of graph representations.

The gap metric provides a measure of distance between dynamical systems which are not required to be stable in themselves. The original rationale for the (linear) gap metric was to provide a suitable topology in which small errors in the gap in open-loop systems would correspond to small errors in norm in the stable closed loop. In [9] the gap metric was shown to be exactly equal to the solution of a certain H_∞ optimization problem. Building on this work the optimal robustness problem in the gap metric was solved in [10] and connections emerged with H_∞ loop-shaping [20]. The stability margin for uncertainty in the gap metric was shown to be the norm of a certain parallel projection operator in [8]. In [29] the tightest possible metric for this stability margin, called the ν-gap metric, was derived.

An attempt to generalize the gap metric robust stability theory to nonlinear systems began in [5], where two complementary parallel projection operators of a

*Received by the editors July 31, 2001; accepted for publication (in revised form) July 6, 2004; published electronically March 11, 2005. This research was supported in part by ARC (Australia) and EPSRC (UK).

http://www.siam.org/journals/sicon/43-5/39306.html

†Department of Engineering, Faculty of Engineering and Information Technology, Australian National University, Canberra, ACT 0200, Australia (Matthew.James@anu.edu.au).

‡Department of Engineering, University of Cambridge, Trumpington St., Cambridge, CB2 1PZ, United Kingdom (mcs@eng.cam.ac.uk, gv@eng.cam.ac.uk).

nonlinear feedback system were highlighted. In [11] the inverse of the induced norm of one of these projections was shown to define a guaranteed radius of stability for uncertainty in a nonlinear generalization of the gap metric. In [30] it was shown how to get a tighter version of the robustness theorem using ideas drawing on the linear ν-gap metric. These results showed in principle that the linear gap metric theory was capable of a powerful generalization to nonlinear systems. It is the purpose of this paper to refine and develop several aspects of this theory.

In [11], [30] several definitions of the gap and ν-gap were given which generalize definitions of the H_2-, L_2-, and ν-gap from the linear case. These, and some new definitions, will form a family of "δ-type" and "ρ-type" metrics which will be studied in section 4. Various versions of the main robustness theorem will be given involving the two complementary parallel projections of a feedback loop. In particular, the "δ-type" (respectively, "ρ-type") gap metrics are needed for results involving the parallel projection onto the plant (respectively, controller) graph. The δ-type and ρ-type gaps are shown to be equal subject to a certain conjugacy transformation on one of the systems (Lemma 4.1). Under similar conditions, the norms of the two complementary parallel projections are shown to be equal (Theorem 4.7).

The connection between gap metrics and representations of the graph is developed in this paper. Operators whose image (respectively, kernel) generates the graph are termed right (respectively, left) representations of the system. These operators, called graph symbols, generalize the usual notions of right and left coprime factorizations. As usual, we call a right representation normalized if the symbol is inner. We introduce a new definition of "normalized" for left representations, which requires that the "amplification" or "gain" of any L_2-function by the symbol is equal to the minimal distance to the graph (section 3.2). This allows formulas to be derived (Theorem 4.10), involving norms of products of the left and right symbols, for the various ρ-type gap metrics. This also allows versions of the robustness theorem to be obtained involving ∞-norm errors between the graph symbols to account for system uncertainty (Theorems 4.12 and 4.13).

The paper also considers the problem of controller synthesis to achieve norm bounds on each of the two complementary parallel projection operators. This problem is at the heart of H_∞ loop-shaping for linear systems and represents one of the most important "special" problems of H_∞ optimization. This problem is solved for the class of input-affine nonlinear systems in section 5. Applying the approach of [14] to the two corresponding generalized plants exposes notable simplifications in the information state dynamics and control laws compared to the usual case of H_∞-synthesis and highlights the duality between the two problems.

The circle of ideas explored in this paper is completed by considering state-space realizations for the graph representations defined in section 3. In the development of the theory of coprime factors and graph representations, state-space realizations have been important for computation and in studying the connection between existence of representations and stabilizability. Section 6.1 derives state-space realizations for the right graph representations in terms of the solution of a Hamilton–Jacobi–Bellman (HJB) equation and a state feedback. Realizations for the left graph representations are derived in the form of an information state system in section 6.2. A brief review of relevant facts from nonlinear H_∞ control which are relevant for sections 5 and 6 is given in the appendix.

We conclude the introduction by summarizing the main contributions of the paper.

- A unified set of definitions for the nonlinear gap metric is provided which highlights the natural linkage with the two fundamental parallel projection operators. Several versions of the fundamental robustness result are given in terms of these metrics. The notion of β-conjugacy is introduced to relate the norms of the two complementary parallel projection operators.
- A new definition of "normalized" for left graph representations is given, and robustness results are given in terms of the graph representations.
- A solution to the controller synthesis problem for a feedback system which optimizes the two fundamental parallel projection operators is given for input-affine nonlinear systems.
- A new realization theory for graph representations is given.

2. Preliminaries. This section provides the background on signal spaces, systems, and stability, which forms the basis for the rest of the paper. In the definitions to follow, n denotes the generic dimension of the range of the signals. Write

$$L_2 = L_2(-\infty,\infty) = \{w : (-\infty,\infty) \to \mathbf{R}^n \mid \| w \|_2 < \infty\}$$

for the Lebesgue space of signals on the doubly infinite time axis, where

$$\| w \|_2 = \| w \|_{L_2(-\infty,\infty)} \triangleq \left(\int_{-\infty}^{\infty} |w(s)|^2 ds \right)^{1/2}.$$

The signals in the following space are zero before a finite time and square integrable on each finite interval:

$$L_{2,ce} = L_{2,ce}(-\infty,\infty) = \{w : \quad (-\infty,\infty) \to \mathbf{R}^n \mid \mathbf{T}_T w = 0 \text{ for some } T \in (-\infty,\infty) \text{ and } \| \mathbf{T}_T w \|_2 < \infty \ \forall T > -\infty\},$$

where

$$(\mathbf{T}_T w)(s) = \begin{cases} w(s), & s \le T, \\ 0, & s > T. \end{cases}$$

The space of signals defined for positive times and square integrable on each finite interval is

$$L_{2,e} = L_{2,e}[0,\infty) = \{w : [0,\infty) \to \mathbf{R}^n \mid \| \mathbf{T}_T w \|_2 < \infty \ \forall T \ge 0\}.$$

We can regard $L_{2,e}$ as a subset of $L_{2,ce}$ by defining elements of $L_{2,e}$ to be 0 before time 0. For $w \in L_{2,ce}$ we write

$$\| w \|_T = \| \mathbf{T}_T w \|_2 .$$

The signal spaces $\mathcal{U}$, $\mathcal{Y}$, $\mathcal{W}$, etc., will be $L_{2,ce}$ spaces of suitable range dimension. The plant P and controller K will be operators defined on these spaces. We will also consider restrictions of these operators to $L_{2,e}$ and L_2.

We write

$$\begin{aligned} \| P \|_\infty &= \limsup_{T\to\infty} \sup_{u \in L_{2,ce}(-\infty,\infty), \|u\|_T \neq 0} \frac{\| Pu \|_T}{\| u \|_T} \\ &= \limsup_{T\to\infty} \sup_{u \in L_{2,e}(-\infty,\infty), \|u\|_T \neq 0} \frac{\| Pu \|_T}{\| u \|_T} \qquad (1) \\ &= \sup_{u \in L_2[0,\infty), u \neq 0} \frac{\| Pu \|_2}{\| u \|_2} \end{aligned}$$

for the induced norm for a causal, time-invariant operator $P : L_{2,ce} \to L_{2,ce}$; this is often called the H_∞ *norm* of P even if P is nonlinear. An operator P is defined to be *stable* if $\| P \|_\infty$ is finite. We sometimes use finite time restrictions $P|_{[T_0,T]}$ of operators, or operators defined on intervals $[T_0, T]$ (we can take $T_0 = -\infty$), and we define

$$\| P|_{[T_0,T]} u \|_{[T_0,T],\infty} = \sup_{u \in L_2[T_0,T], u \neq 0} \frac{\| P|_{[T_0,T]} u \|_T}{\| u \|_T}. \tag{2}$$

The feedback configuration of Figure 1 is denoted by $[P, K]$ and consists of a *plant* $P : \mathcal{U} \to \mathcal{Y}$ and a *controller* $K : \mathcal{Y} \to \mathcal{U}$, both causal, time-invariant maps defined on $L_{2,ce}$ signal spaces $\mathcal{U}$, $\mathcal{Y}$ and which satisfy $P0 = 0$ and $K0 = 0$.

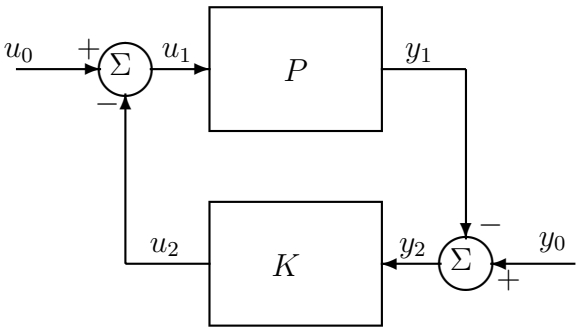


FIG. 1. *Feedback configuration* $[P, K]$.

In Figure 1, $u_i \in \mathcal{U}$ and $y_i \in \mathcal{Y}$ $(i = 0, 1, 2)$, and we write $\mathcal{W} = \mathcal{U} \times \mathcal{Y}$. The feedback system $[P, K]$ is assumed to be well-posed. Namely, for any $w = (u_0, y_0) \in \mathcal{W}$ there exist unique $u_1, u_2 \in \mathcal{U}$, $y_1, y_2 \in \mathcal{Y}$ such that the following closed-loop equations hold:

$$\begin{aligned} u_0 &= u_1 + u_2, \\ y_0 &= y_1 + y_2, \\ y_1 &= P u_1, \\ u_2 &= K y_2, \end{aligned} \tag{3}$$

and moreover, the map

$$\begin{aligned} H_{P,K} &: \mathcal{W} \to \mathcal{W} \times \mathcal{W}, \\ &: \begin{pmatrix} u_0 \\ y_0 \end{pmatrix} \mapsto \left(\begin{pmatrix} u_1 \\ y_1 \end{pmatrix} , \begin{pmatrix} u_2 \\ y_2 \end{pmatrix} \right) \end{aligned} \tag{4}$$

is causal.

Throughout the paper, it will be assumed that both the nominal feedback system as shown in Figure 1 and any perturbed feedback system (with P replaced by a perturbed plant P_1) are well-posed. The feedback system $[P, K]$ is defined to be *stable* if the operator from (u_0, y_0) to (u_1, y_1, u_2, y_2) has finite H^∞ norm, i.e., $\| H_{P,K} \|_\infty < \infty$.

It is convenient to consider *graphs* of operators. The graph of the plant P is

$$\mathcal{G}_P = \left\{ \begin{pmatrix} u \\ Pu \end{pmatrix} \ : \ u \in \mathcal{U}, Pu \in \mathcal{Y} \right\} \subset \mathcal{W},$$

and the graph of the controller K is

$$\mathcal{G}_K = \left\{ \begin{pmatrix} Ky \\ y \end{pmatrix} \ : \ y \in \mathcal{Y}, Ky \in \mathcal{U} \right\} \subset \mathcal{W}.$$

We often write

$$\mathcal{M} = \mathcal{G}_P, \quad \mathcal{N} = \mathcal{G}_K.$$

Of central importance to robustness of the feedback system $[P, K]$ are the *parallel projection* operators (see [5], [11])

$$\Pi_{\mathcal{M}\|\mathcal{N}} = \Pi_1 H_{P,K}, \quad \Pi_{\mathcal{N}\|\mathcal{M}} = \Pi_2 H_{P,K},$$

where $\Pi_i : \mathcal{W} \times \mathcal{W} \to \mathcal{W}$ denote the natural projections ($i = 1, 2$). The operators $\Pi_{\mathcal{M}\|\mathcal{N}}$, $\Pi_{\mathcal{N}\|\mathcal{M}}$ both enjoy the defining *parallel projection* property

$$\Pi(\Pi w_1 + (I - \Pi)w_2) = \Pi w_1 \tag{5}$$

for any w_1, $w_2 \in \mathcal{W}$, where I denotes the identity operator, and the following identities hold:

$$H_{P,K} = (\Pi_{\mathcal{M}\|\mathcal{N}}, \Pi_{\mathcal{N}\|\mathcal{M}}) \quad \text{and} \quad \Pi_{\mathcal{M}\|\mathcal{N}} + \Pi_{\mathcal{N}\|\mathcal{M}} = I. \tag{6}$$

Consequently, stability of the feedback system $[P, K]$, i.e., the finiteness of $\| H_{P,K} \|_\infty$, is equivalent to the stability of either parallel projection [11].

We note that the parallel projection operators can be represented as generalized plant and controller configurations as in Figures 2 and 3, where

$$\mathcal{P}_1 : \begin{pmatrix} u_0 \\ y_0 \\ u_2 \end{pmatrix} \mapsto \begin{pmatrix} u_1 \\ y_1 \\ y_2 \end{pmatrix} \quad \text{and } \mathcal{P}_2 : \begin{pmatrix} u_0 \\ y_0 \\ u_2 \end{pmatrix} \mapsto \begin{pmatrix} u_2 \\ y_2 \\ y_2 \end{pmatrix} \tag{7}$$

so that

$$\| \Pi_{\mathcal{M}\|\mathcal{N}} \|_\infty = \| [\mathcal{P}_1, K] \|_\infty \quad \text{and} \quad \| \Pi_{\mathcal{N}\|\mathcal{M}} \|_\infty = \| [\mathcal{P}_2, K] \|_\infty \ . \tag{8}$$

Here, $[\mathcal{P}_1, K]$ and $[\mathcal{P}_2, K]$ refer to the closed-loop configurations of Figures 2 and 3, respectively. As we shall see in section 4, there are guaranteed bounds for robust stability in terms of the norms of these parallel projection operators. The synthesis problem of finding controllers to minimize either quantity is solved in section 5.

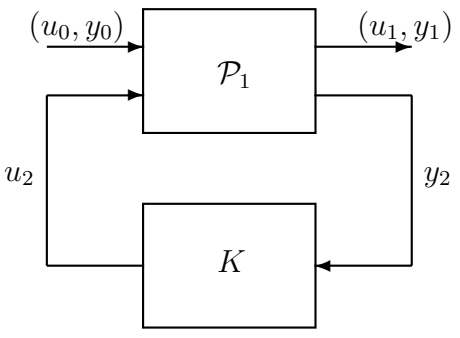


FIG. 2. *The parallel projection* $\Pi_{\mathcal{M}\|\mathcal{N}}$ *and generalized plant* $\mathcal{P}_1$.

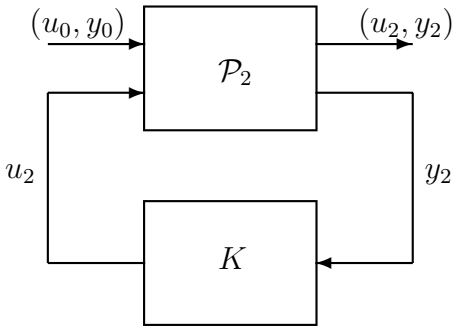


FIG. 3. *The parallel projection* $\Pi_{\mathcal{N}\|\mathcal{M}}$ *and generalized plant* $\mathcal{P}_2$.

3. Graph representations. The subject of right and left coprime factorizations has been fundamental in the development of the (linear) gap metric robustness theory. From an operator viewpoint these factorizations may be viewed as graph "symbols." A number of generalizations for nonlinear operators have been considered in the literature; see [24], [22], [27], [19]. In this section we propose a new approach which focuses on the essential energy balances related to a graph representation. This allows definitions to be made for both finite and infinite time horizons. More importantly, it allows a new definition of "co-innerness" for the left representations which appears to have significant advantages; in particular, it allows the left representations to be used for robustness analysis (section 4.5). The definition requires that the "amplification" of any L_2-function by the symbol be equal to the minimal distance of that function to the graph.

3.1. Right representations. Roughly speaking, we wish to generalize the idea that $P = NM^{-1}$ has a normalized coprime factorization, with

$$R = \begin{bmatrix} M \\ N \end{bmatrix}$$

defining a range representation of the graph of P:

$$\{w \in \mathcal{W} \mid w = R\psi, \text{ some } \psi \in \mathcal{U}\} = \mathcal{G}_P,$$

together with notions of the inner and coprime properties.

1. For any $T \in (-\infty, \infty)$ a causal operator

$$\mathsf{R}_{T,P} \ : \ \mathcal{U} \to \mathcal{W} \cap L_2(-\infty, T]$$

is called a *finite horizon right representation* of P provided it maps onto the graph:

$$\text{range}\, \mathsf{R}_{T,P}|_{[T_0,T]} = \mathcal{G}_P \cap L_2[T_0, T] \quad \forall T_0 \in (-\infty, T]. \tag{9}$$

(a) $\mathsf{R}_{T,P}$ is *contractive* if

$$\| \mathsf{R}_{T,P}\psi \|_T \leq \| \psi \|_T \quad \forall \psi \in \mathcal{U}; \tag{10}$$

(b) $\mathsf{R}_{T,P}$ is *finite time inner* (or *normalized*) if

$$\| \mathsf{R}_{T,P}\psi \|_T = \| \psi \|_T \tag{11}$$

for all $\psi \in \mathcal{U}$.

(c) A causal operator

$$\mathsf{R}_{T,P}^{-L} \ : \ \mathcal{W} \to \mathcal{U} \cap L_2(-\infty, T]$$

is called a *left inverse* of $\mathsf{R}_{T,P}$ if

$$\mathsf{R}_{T,P}^{-L}\mathsf{R}_{T,P} = I. \tag{12}$$

$\mathsf{R}_{T,P}^{-L}$ is norm bounded if there is some $c > 0$ such that

$$\| \mathsf{R}_{T,P}^{-L} w \|_T \leq c \, \| w \|_T \quad \forall w \in \mathcal{W}. \tag{13}$$

$\mathsf{R}_{T,P}$ is *finite time right coprime* if there exists a causal operator $\mathsf{R}_{T,P}^{-L}$ which is a left inverse of $\mathsf{R}_{T,P}$ and norm bounded.

2. A causal operator

$$\mathsf{R}_{eP} \ : \ \mathcal{U} \to \mathcal{W}$$

is called a *right representation* of P provided it maps onto the graph

$$\begin{aligned} \operatorname{range} \mathsf{R}_{eP} &= \mathcal{G}_P, \\ \operatorname{range} \mathsf{R}_{eP}|_{L_2} &= \mathcal{G}_P \cap L_2, \\ \operatorname{range} \mathsf{R}_{eP}|_{L_2[0,\infty)} &= \mathcal{G}_P \cap L_2[0,\infty). \end{aligned} \tag{14}$$

(a) R_{eP} is *contractive* if

$$\| \mathsf{R}_{eP}\psi \|_T \leq \| \psi \|_T \quad \forall \ T \in (-\infty, \infty), \psi \in \mathcal{U}; \tag{15}$$

(b) R_{eP} is *inner* (or *normalized*) if

$$\| \mathsf{R}_{eP}\psi \|_2 = \| \psi \|_2 \quad \forall \psi \in \mathcal{U} \cap L_2. \tag{16}$$

(c) A causal operator

$$\mathsf{R}_{eP}^{-L} \ : \ \mathcal{W} \to \mathcal{U}$$

is called a *left inverse* of R_{eP} if

$$\mathsf{R}_{eP}^{-L}\mathsf{R}_{eP} = I. \tag{17}$$

R_{eP}^{-L} is norm bounded if there is some $c > 0$ such that

$$\| \mathsf{R}_e{}^{-L} w \|_T \leq c \, \| w \|_T \quad \forall T \in (-\infty, \infty), \ w \in \mathcal{W}. \tag{18}$$

R_{eP} is *right coprime* if there exists a causal operator R_{eP}^{-L} which is a left inverse of R_{eP} and norm bounded.

3.2. Left representations. We next generalize the idea that $P = \tilde{M}^{-1}\tilde{N}$ has a normalized left coprime factorization, with

$$L = [-\tilde{N}, \ \tilde{M}]$$

defining a kernel representation of P:

$$\{w \in \mathcal{W} \mid Lw = 0\} = \mathcal{G}_P,$$

satisfying co-inner and coprime properties.

1. A causal operator

$$\mathsf{L}_{tvP} \ : \ \mathcal{W} \cap L_{2,e} \to \mathcal{Y} \cap L_{2,e}$$

is called a *finite horizon left representation* of P provided its kernel is the graph

$$\ker \mathsf{L}_{tvP} = \mathcal{G}_P \cap L_{2,e}. \tag{19}$$

(a) L_{tvP} is *contractive* if

$$\| \mathsf{L}_{tvP} w \|_T \leq \| w \|_T \quad \forall \ T \geq 0, w \in \mathcal{W} \cap L_{2,e}. \tag{20}$$

(b) L_{tvP} is *positive time co-inner* (or *normalized*) if

$$\| \mathsf{L}_{tvP} w \|_T = \inf \{ \| w - \tilde{w} \|_T \ | \ \tilde{w} \in \mathcal{G}_P \cap L_2[0,T] \} \tag{21}$$

for all $w \in \mathcal{W} \cap L_2[0,T]$, $T \geq 0$, and

$$\| \mathsf{L}_{tvP} w \|_2 = \inf \{ \| w - \tilde{w} \|_2 \ | \ \tilde{w} \in \mathcal{G}_P \cap L_2[0,\infty) \} \tag{22}$$

for all $w \in \mathcal{W} \cap L_2[0,\infty)$.

(c) A causal operator

$$\mathsf{L}_{tvP}^{-R} \ : \ \mathcal{Y} \cap L_{2,e} \to \mathcal{W} \cap L_{2,e}$$

is called a *right inverse* of L_{tvP} if

$$\mathsf{L}_{tvP} \mathsf{L}_{tvP}^{-R} = I. \tag{23}$$

L_{tvP}^{-R} is norm bounded if there is some $c > 0$ such that

$$\| \mathsf{L}_{tvP}^{-R} \phi \|_T \leq c \, \| \phi \|_T \quad \forall \phi \in \mathcal{Y} \cap L_{2,e}. \tag{24}$$

L_{tvP} is *finite time left coprime* if there exists a causal operator L_{tvP}^{-R} which is a right inverse of R_{tvP} and norm bounded.

2. A causal operator

$$\mathsf{L}_{eP} \ : \ \mathcal{W} \to \mathcal{Y}$$

is called a *left representation* of P provided its kernel is the graph

$$\ker \mathsf{L}_{eP} = \mathcal{G}_P. \tag{25}$$

(a) L_{eP} is *contractive* if

$$\| \mathsf{L}_{eP} w \|_T \leq \| w \|_T \quad \forall \ T \in (-\infty, \infty), w \in \mathcal{W}. \tag{26}$$

(b) L_{eP} is *co-inner* if

$$\| \mathsf{L}_{eP} w \|_T = \inf \{ \| w - \tilde{w} \|_T \ | \ \tilde{w} \in \mathcal{G}_P \} \ \forall T \in (-\infty, \infty), \tag{27}$$

for all $w \in \mathcal{W}$, $T \in (-\infty, \infty)$, and

$$\| \mathsf{L}_{eP} w \|_2 = \inf \{ \| w - \tilde{w} \|_2 \ | \ \tilde{w} \in \mathcal{G}_P \cap L_2 \} \tag{28}$$

for all $w \in L_2$.

(c) A causal operator

$$\mathsf{L}_{eP}^{-R}\colon \mathcal{Y} \to \mathcal{W}$$

is called a *right inverse* of L_{eP} if

$$\mathsf{L}_{eP}\mathsf{L}_{eP}^{-R} = I. \tag{29}$$

L_{eP}^{-R} is norm bounded if there is some $c > 0$ such that

$$\| \mathsf{L}_{eP}^{-R}\phi \|_T \leq c \| \phi \|_T \quad \forall\, T \in (-\infty, \infty), \phi \in \mathcal{Y}. \tag{30}$$

L_{eP} is *left coprime* if there exists a causal operator L_{eP}^{-R} which is a right inverse of L_{eP} and norm bounded.

Remark 3.1.

1. The definition proposed for the coinner property in 2b above agrees with the usual definition for linear, time-invariant (LTI) systems. In this case, L_{eP} can be interpreted as a normalized Kalman filter for P, with $\| \mathsf{L}_{eP} w \|_T$ the residual.
2. It is interesting to note that range and kernel representations do not behave in a completely analogous way with regard to the concept of coprimeness. If P is a finite-dimensional LTI system and G is *any* LTI system whose range is the graph, then G will be left invertible over H_∞, which is equivalent to coprimeness. The same is not true for kernel representations $\tilde{G}$, which may be freely multiplied by noninvertible factors on the left without changing the kernel. This situation does not change when the properties of inner/co-inner are added.
3. Representations satisfying the above properties can be constructed for input-affine systems subject to appropriate regularity assumptions. This is done in section 6.

4. Gap metrics and robust stability. The gap metric was introduced into the control literature by Zames and El-Sakkary [32], where it was defined as the norm of the difference of two orthogonal projection operators in Hilbert space. The result of Georgiou [9] that the gap metric was exactly equal to the solution of a certain H_∞ optimization problem opened up direct connections between uncertainty in gap metric and coprime fraction uncertainty and indeed proved the following formula for the (linear, directed) gap metric:

$$\vec{\rho}_{H_2}(P, P_1) = \inf_{\Delta_N, \Delta_M \in H_\infty} \left\{ \begin{bmatrix} \Delta_M \\ \Delta_N \end{bmatrix} \mid P_1 = (N + \Delta_N)(M + \Delta_M)^{-1} \right\},$$

where $P = NM^{-1}$ is a normalized right coprime factorization of the plant transfer-function. The basic gap metric robustness theorem can be expressed as follows [10]: if K stabilizes the nominal plant P, then it stabilizes all P_1 with $\vec{\rho}_{H_2}(P, P_1) < \gamma$ if and only if $\gamma \leq \| \Pi_{\mathcal{M}\|\mathcal{N}} \|_\infty^{-1}$.

A smaller distance function can be defined by

$$\vec{\rho}_{L_2}(P, P_1) = \inf_{\Delta_N, \Delta_M \in L_\infty} \left\{ \begin{bmatrix} \Delta_M \\ \Delta_N \end{bmatrix} \mid P_1 = (N + \Delta_N)(M + \Delta_M)^{-1} \right\}$$

but proximity here does not guarantee robust stability by itself. However, if P and P_1 can both be approximated arbitrarily closely in the gap metric by finite-dimensional systems and $\vec{\rho}_{L_2}(P, P_1) < \| \Pi_{\mathcal{M}\|\mathcal{N}} \|_\infty^{-1}$, then K stabilizes P_1 if and only

if $\operatorname{wno}\det(N_1^*N + M_1^*M) = 0$, where $P = NM^{-1}$ and $P_1 = N_1M_1^{-1}$ are any continuous right coprime factorizations (here, "wno" refers to the winding number). The ν-gap metric is obtained by setting it equal to $\vec{\rho}_{L_2}$ when this winding number test is satisfied and 1 otherwise, and is consequently less conservative than the gap metric [29].

For nonlinear systems $\vec{\rho}_{H_2}$ can be perfectly well defined (see below) but does not appear to give any robust stability guarantees. We shall consider two alternatives which do. The winding number test in the ν-gap definition can be viewed as implicitly checking for the existence of a homotopy linking P and P_1. In the nonlinear case this check appears in this form in Theorem 4.6.

4.1. Definitions and basic properties. There are a number of definitions for gap metrics in the literature. We now give the definitions of some of these together with some new ones.

We write $\mathcal{G}_0 = \mathcal{G}_{P_0}$, $\mathcal{G}_1 = \mathcal{G}_{P_1}$, for two plants P_0 and P_1. Define the δ-type "gap metrics"as follows:

(31)

$$\vec{\delta}_0(P_0,P_1) = \limsup_{T\to\infty} \sup_{x_1\in\mathcal{G}_1\cap L_{2,e}[0,\infty),x_1\neq 0} \inf_{x_0\in\mathcal{G}_0\cap L_{2,e}[0,\infty),x_0\neq 0} \frac{\| x_1 - x_0 \|_T}{\| x_0 \|_T},$$

$$\vec{\delta}_g(P_0,P_1) = \limsup_{T\to\infty} \sup_{x_1\in\mathcal{G}_1\cap L_{2,ce}(-\infty,\infty),x_1\neq 0} \inf_{x_0\in\mathcal{G}_0\cap L_{2,ce}(-\infty,\infty),x_0\neq 0} \frac{\| x_1 - x_0 \|_T}{\| x_0 \|_T},$$

$$\vec{\delta}_{H_2}(P_0,P_1) = \sup_{x_1\in\mathcal{G}_1\cap L_2[0,\infty),x_1\neq 0} \inf_{x_0\in\mathcal{G}_0\cap L_2[0,\infty),x_0\neq 0} \frac{\| x_1 - x_0 \|_2}{\| x_0 \|_2},$$

$$\vec{\delta}_{L_2}(P_0,P_1) = \sup_{x_1\in\mathcal{G}_1\cap L_2(-\infty,\infty),x_1\neq 0} \inf_{x_0\in\mathcal{G}_0\cap L_2(-\infty,\infty),x_0\neq 0} \frac{\| x_1 - x_0 \|_2}{\| x_0 \|_2},$$

and also define the ρ-gaps

(32)

$$\vec{\rho}_0(P_0,P_1) = \limsup_{T\to\infty} \sup_{x_1\in\mathcal{G}_1\cap L_{2,e}[0,\infty),x_1\neq 0} \inf_{x_0\in\mathcal{G}_0\cap L_{2,e}[0,\infty)} \frac{\| x_1 - x_0 \|_T}{\| x_1 \|_T},$$

$$\vec{\rho}_g(P_0,P_1) = \limsup_{T\to\infty} \sup_{x_1\in\mathcal{G}_1\cap L_{2,ce}(-\infty,\infty),x_1\neq 0} \inf_{x_0\in\mathcal{G}_0\cap L_{2,ce}(-\infty,\infty)} \frac{\| x_1 - x_0 \|_T}{\| x_1 \|_T},$$

$$\vec{\rho}_{H_2}(P_0,P_1) = \sup_{x_1\in\mathcal{G}_1\cap L_2[0,\infty),x_1\neq 0} \inf_{x_0\in\mathcal{G}_0\cap L_2[0,\infty)} \frac{\| x_1 - x_0 \|_2}{\| x_1 \|_2},$$

$$\vec{\rho}_{L_2}(P_0,P_1) = \sup_{x_1\in\mathcal{G}_1\cap L_2(-\infty,\infty),x_1\neq 0} \inf_{x_0\in\mathcal{G}_0\cap L_2(-\infty,\infty)} \frac{\| x_1 - x_0 \|_2}{\| x_1 \|_2}.$$

The definition of $\vec{\delta}_0$ is from [11] and those of $\vec{\delta}_g$ and $\vec{\delta}_{L_2}$ are taken from [30]. As pointed out in [30], for time-invariant systems, $\vec{\delta}_g(P_0,P_1) \le \vec{\delta}_0(P_0,P_1)$, since the extra freedom x_1 has in being shifted to the left is more than offset by a similar freedom in x_0. For the same reason $\vec{\delta}_{L_2}(P_0,P_1) \le \vec{\delta}_{H_2}(P_0,P_1)$, and the corresponding inequalities for the ρ-gaps, hold for time-invariant systems. For linear systems $\vec{\rho}_{H_2}(P_0,P_1)$ is the

usual directed gap over Hilbert spaces and is also equal to $\vec{\delta}_{H_2}(P_0, P_1)$ [11, proof of Proposition 5]. The same argument shows that each respective pair of δ and ρ gap metrics coincides. However, for nonlinear systems they need not be the same, though they are related as follows.

LEMMA 4.1. *Let $\gamma > 1$, and set $\beta = \sqrt{1-\gamma^{-2}}$. For each of the four cases $\vec{\delta} \in \{\vec{\delta}_0, \vec{\delta}_g, \vec{\delta}_{L_2}, \vec{\delta}_{H_2}\}$, with respective $\vec{\rho} \in \{\vec{\rho}_0, \vec{\rho}_g, \vec{\rho}_{L_2}, \vec{\rho}_{H_2}\}$, we have*

$$\vec{\delta}(P_0, P_1) < \gamma^{-1} \quad \textit{iff} \quad \vec{\rho}(P_0, \beta^{-2} P_1 \beta^2) < \gamma^{-1} \quad \textit{iff} \quad \vec{\rho}(\beta^2 P_0 \beta^{-2}, P_1) < \gamma^{-1}. \tag{33}$$

Similarly,

$$\vec{\rho}(P_0, P_1) < \gamma^{-1} \quad \textit{iff} \quad \vec{\delta}(P_0, \beta^2 P_1 \beta^{-2}) < \gamma^{-1} \quad \textit{iff} \quad \vec{\delta}(\beta^{-2} P_0 \beta^2, P_1) < \gamma^{-1}. \tag{34}$$

Proof. We give the proof for the case $\vec{\delta} = \vec{\delta}_{L_2}$ and $\vec{\rho} = \vec{\rho}_{L_2}$. Suppose

$$\vec{\delta}(P_0, P_1) < \gamma^{-1}.$$

Then for any $0 \neq x_1 \in \mathcal{G}_1 \cap L_2(-\infty, \infty)$ there exists $0 \neq x_0 \in \mathcal{G}_0 \cap L_2(-\infty, \infty)$ such that

$$\| x_1 - x_0 \|_2^2 < \gamma^{-2} \| x_0 \|_2^2 . \tag{35}$$

However, by expanding and completion of squares, this holds if and only if for any $0 \neq x_1 \in \mathcal{G}_1 \cap L_2(-\infty, \infty)$ there exists $0 \neq x_0 \in \mathcal{G}_0 \cap L_2(-\infty, \infty)$ such that

$$\| \beta^{-2} x_1 - x_0 \|_2^2 < \gamma^{-2} \| \beta^{-2} x_1 \|_2^2,$$

where $\beta = \sqrt{1-\gamma^{-2}}$. Since $\gamma > 1$, the condition that $x_0 \neq 0$ is redundant. Now $x_1 \in \mathcal{G}_1 \cap L_2(-\infty, \infty)$ iff $\beta^{-2} x_1 \in \tilde{\mathcal{G}}_1 \cap L_2(-\infty, \infty)$, where $\tilde{\mathcal{G}}_1 = \mathcal{G}_{\tilde{P}_1}$ and $\tilde{P}_1 = \beta^{-2} P_1 \beta^2$. Therefore, the last displayed equation holds if and only if for any $0 \neq \tilde{x}_1 \in \tilde{\mathcal{G}}_1 \cap L_2(-\infty, \infty)$, there exists $x_0 \in \mathcal{G}_0 \cap L_2(-\infty, \infty)$ such that

$$\| \tilde{x}_1 - x_0 \|_2^2 < \gamma^{-2} \| \tilde{x}_1 \|_2^2,$$

that is, if and only if

$$\vec{\rho}(P_0, \beta^{-2} P_1 \beta^2) < \gamma^{-1}.$$

Also, (35) holds if and only if for any $0 \neq x_1 \in \mathcal{G}_1 \cap L_2(-\infty, \infty)$, there exists $x_0 \in \mathcal{G}_0 \cap L_2(-\infty, \infty)$ such that

$$\| x_1 - \beta^2 x_0 \|_2^2 < \gamma^{-2} \| x_1 \|_2^2 .$$

Since $\gamma > 1$, then $x_0 \neq 0$. Thus the above equation holds if and only if for any $x_1 \in \mathcal{G}_1$ there exists $\tilde{x}_0 \in \tilde{\mathcal{G}}_0$ such that

$$\| x_1 - \tilde{x}_0 \|_2^2 < \gamma^{-2} \| x_1 \|_2^2,$$

where $\tilde{P}_0 = \beta^2 P_0 \beta^{-2}$. That is, (35) holds if and only if

$$\vec{\rho}(\beta^2 P_0 \beta^{-2}, P_1) < \gamma^{-1}. \qquad \square$$

We have the following result giving a class of systems for which the δ and ρ gaps are the same.

COROLLARY 4.2. *If P_0 and/or P_1 is homogeneous, then for each of the four cases $\vec{\delta} \in \{\vec{\delta}_0, \vec{\delta}_g, \vec{\delta}_{L_2}, \vec{\delta}_{H_2}\}$, with respective $\vec{\rho} \in \{\vec{\rho}_0, \vec{\rho}_g, \vec{\rho}_{L_2}, \vec{\rho}_{H_2}\}$,*

$$\min\left\{\vec{\delta}(P_0, P_1), \vec{\rho}(P_0, P_1)\right\} < 1 \tag{36}$$

implies

$$\vec{\delta}(P_0, P_1) = \vec{\rho}(P_0, P_1). \tag{37}$$

For any $\vec{\delta} \in \{\vec{\delta}_0, \vec{\delta}_g, \vec{\delta}_{L_2}, \vec{\delta}_{H_2}\}$, we define the corresponding symmetric quantities via

$$\delta(P_0, P_1) = \max\left\{\vec{\delta}(P_0, P_1), \vec{\delta}(P_1, P_0)\right\},$$

and similarly for any $\vec{\rho} \in \{\vec{\rho}_0, \vec{\rho}_g, \vec{\rho}_{L_2}, \vec{\rho}_{H_2}\}$, we define

$$\rho(P_0, P_1) = \max\left\{\vec{\rho}(P_0, P_1), \vec{\rho}(P_1, P_0)\right\}.$$

4.2. Robust stability. The $\vec{\delta}$, $\vec{\delta}_0$, $\vec{\delta}_g$, $\vec{\delta}_{L_2}$, and ν-gap metrics introduced in [11], [30], and [29] provide a maximum stability margin expressed as the inverse of the induced norm of the parallel projection operator $\Pi_{\mathcal{M}_0\|\mathcal{N}}$; a result of this type is quoted as Theorem 4.3.

THEOREM 4.3 ([30, Proposition 2.2], [11, Theorem 3]). *Assume $H_{P_0,K}$ is stable. If*

$$\vec{\delta}_g(P_0, P_1) < \| \Pi_{\mathcal{M}_0\|\mathcal{N}} \|_\infty^{-1}, \tag{38}$$

where $\mathcal{M}_0 = \mathcal{G}_{P_0}$ and $\mathcal{N} = \mathcal{G}_K$, then $H_{P_1,K}$ is stable and

$$\| \Pi_{\mathcal{M}_1\|\mathcal{N}} \|_\infty \leq \| \Pi_{\mathcal{M}_0\|\mathcal{N}} \|_\infty \frac{1 + \vec{\delta}_g(P_0, P_1)}{1 - \| \Pi_{\mathcal{M}_0\|\mathcal{N}} \|_\infty \vec{\delta}_g(P_0, P_1)}, \tag{39}$$

where $\mathcal{M}_1 = \mathcal{G}_{P_1}$.

The ρ-type gap metrics introduced above also lead to a natural stability margin, but this time it is expressed in terms of the inverse of the induced norm of the parallel projection operator $\Pi_{\mathcal{N}\|\mathcal{M}_0}$, revealing an interesting duality. This will be further explored below in section 4.5 in connection with uncertainty in the graph representations.

THEOREM 4.4. *Assume $H_{P_0,K}$ is stable. If*

$$\vec{\rho}_g(P_0, P_1) < \| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty^{-1}, \tag{40}$$

where $\mathcal{M}_0 = \mathcal{G}_{P_0}$ and $\mathcal{N} = \mathcal{G}_K$, then $H_{P_1,K}$ is stable and

$$\| \Pi_{\mathcal{N}\|\mathcal{M}_1} \|_\infty \leq \| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty \frac{1 + \vec{\rho}_g(P_0, P_1)}{1 - \| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty \vec{\rho}_g(P_0, P_1)}, \tag{41}$$

where $\mathcal{M}_1 = \mathcal{G}_{P_1}$.

Proof. The proof is a modification of the proof of [11, Theorem 3].

Suppose $\vec{\rho}_g(P_0, P_1) < \alpha$, $\| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty < \gamma$. Let $w \in \mathcal{W}$. Since $[P_1, K]$ is well-posed by assumption, there exists $m_1 \in \mathcal{M}_1$, $n \in \mathcal{N}$ such that $w = m_1 + n$, so $\Pi_{\mathcal{N}\|\mathcal{M}_1}(w) = n$. Our goal is to bound the norm of n in terms of the norm of w.

There exists $T_0 > 0$ such that for all $T \geq T_0$

$$\inf_{x_0 \in \mathcal{M}_0} \frac{\| m_1 - x_0 \|_T}{\| m_1 \|_T} \leq \sup_{x_1 \in \mathcal{M}_1} \inf_{x_0 \in \mathcal{M}_0} \frac{\| x_1 - x_0 \|_T}{\| x_1 \|_T} < \alpha.$$

Hence there exists $m_0 \in \mathcal{M}_0$ (depending on T) such that

$$\| m_1 - m_0 \|_T \leq \alpha \| m_1 \|_T .$$

Now $\Pi_{\mathcal{N}\|\mathcal{M}_0}(m_0 + n) = n$, and

$$\| n \|_T \leq \gamma \| m_0 + n \|_T .$$

Then

$$\begin{aligned} \| m_1 + n \|_T &\geq \| m_0 + n \|_T - \| m_1 - m_0 \|_T \\ &\geq \gamma^{-1} \| n \|_T - \alpha \| m_1 \|_T, \end{aligned}$$

and so

$$\begin{aligned} \gamma^{-1} \| n \|_T &\leq \| m_1 + n \|_T + \alpha \| m_1 \|_T \\ &\leq (1 + \alpha) \| m_1 + n \|_T + \alpha \| n \|_T . \end{aligned}$$

Therefore,

$$(\gamma^{-1} - \alpha) \| n \|_T \leq (1 + \alpha) \| m_1 + n \|_T$$

for all $T \geq T_0$. This is enough to prove the theorem. □

Note that both Theorems 4.3 and 4.4 also hold for $\vec{\delta}_0$ and $\vec{\rho}_0$, respectively, though the above are the stronger forms since $\vec{\delta}_g(P_0, P_1) \leq \vec{\delta}_0(P_0, P_1)$ and $\vec{\rho}_g(P_0, P_1) \leq \vec{\rho}_0(P_0, P_1)$. Interestingly, we are not able to give a robust stability theorem for $\vec{\delta}_{H_2}$ or $\vec{\rho}_{H_2}$ in spite of the fact that each is equal to the original definition of the gap metric for linear systems.

The L_2 gaps are not by themselves sufficient to prove stability. However, if stability is known for a plant P_1 near a nominal P_0, then the norms of the parallel projections can be estimated. This was done in [30, Theorem 2.1] for the $\vec{\delta}_{L_2}$ gap and the $\Pi_{\mathcal{M}\|\mathcal{N}}$ projection. The corresponding result for the $\vec{\rho}_{L_2}$ gap and the $\Pi_{\mathcal{N}\|\mathcal{M}}$ projection is as follows.

THEOREM 4.5. *Assume $H_{P_0,K}$ is stable. If $H_{P_1,K}$ is stable and*

$$\vec{\rho}_{L_2}(P_0, P_1) < \| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty^{-1}, \tag{42}$$

where $\mathcal{M}_0 = \mathcal{G}_{P_0}$ and $\mathcal{N} = \mathcal{G}_K$, then

$$\| \Pi_{\mathcal{N}\|\mathcal{M}_1} \|_\infty \leq \| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty \frac{1 + \vec{\rho}_{L_2}(P_0, P_1)}{1 - \| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty \vec{\rho}_{L_2}(P_0, P_1)}, \tag{43}$$

where $\mathcal{M}_1 = \mathcal{G}_{P_1}$.

Proof. Suppose $\vec{\rho}_{L_2}(P_0, P_1) < \alpha$, $\| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty < \gamma$. Let $w \in \mathcal{W} \cap L_2(-\infty, \infty)$. Since $[P_1, K]$ is stable by assumption, there exists $m_1 \in \mathcal{M}_1 \cap L_2(-\infty, \infty)$, $n \in \mathcal{N} \cap L_2(-\infty, \infty)$ such that $w = m_1 + n$, so $\Pi_{\mathcal{N}\|\mathcal{M}_1}(w) = n$.

Note that there exists $m_0 \in \mathcal{M}_0 \cap L_2(-\infty, \infty)$ such that

$$\| m_1 - m_0 \|_2 \leq \alpha \| m_1 \|_2 .$$

Now $\Pi_{\mathcal{N}\|\mathcal{M}_0}(m_0+n)=n$, and

$$\| n \|_2 \leq \gamma \| m_0+n \|_2 .$$

Then, in a similar way to the proof of Theorem 4.4, we obtain

$$(\gamma^{-1}-\alpha)\| n \|_2 \leq (1+\alpha)\| m_1+n \|_2$$

as required. □

THEOREM 4.6. *Assume $H_{P_0,K}$ is stable. If*

$$\vec{\rho}_{L_2}(P_0,P_1) < \| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty^{-1}, \tag{44}$$

where $\mathcal{M}_0=\mathcal{G}_{P_0}$ and $\mathcal{N}=\mathcal{G}_K$, then $H_{P_1,K}$ is stable if there exists a homotopy of plants $\{P_\lambda \mid 0<\lambda<1\}$ such that

1. *the mapping $\lambda \mapsto P_\lambda$ is ρ_g-continuous for $0\leq\lambda\leq 1$, and*
2. $$\vec{\rho}_{L_2}(P_0,P_\lambda) < \| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty^{-1} - \varepsilon \tag{45}$$

   for some $\varepsilon>0$ and all $0\leq\lambda\leq 1$.

Proof. Assume $[P_{\lambda_0},K]$ is stable for some $\lambda_0\in[0,1]$. Then it must also be the case that $\| \Pi_{\mathcal{N}\|\mathcal{M}_{\lambda_0}} \|_\infty^{-1} > \epsilon' = \epsilon/(1+ \| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty^{-1} -\epsilon)$. By continuity there will exist a neighborhood $\mathcal{S}$ of λ_0, such that $\vec{\delta}_g(P_{\lambda_0},P_\lambda)<\epsilon'$ for all $\lambda\in\mathcal{S}$. It follows that $[P_\lambda,K]$ is stable for all $\lambda\in\mathcal{S}$ and hence that the set of $\lambda\in[0,1]$ for which $[P_\lambda,K]$ is stable is open. Conversely, let $\{\lambda_i\}\to\lambda$ and assume that $[P_{\lambda_i},K]$ is stable for all i. Then there must exist an i such that $\vec{\delta}_g(P_{\lambda_i},P_\lambda)<\epsilon'$, from which it follows that $[P_\lambda,K]$ is stable and hence that the set of $\lambda\in[0,1]$ for which $[P_\lambda,K]$ is stable is both open and closed, and since it includes the point $\lambda=0$ it must also include the whole interval $[0,1]$. □

4.3. Conjugate norm equivalence. For linear systems it is known that $\| \Pi_{\mathcal{M}\|\mathcal{N}} \|_\infty = \| \Pi_{\mathcal{N}\|\mathcal{M}} \|_\infty$, and the inverse of this quantity is the stability margin for uncertainty in the gap metric [10]. This equality has several implications and interpretations. It means that the radius of gap metric uncertainty tolerated by the feedback system is the same for the plant and the controller. It also means that the radius of coprime factor balls of uncertainty tolerated by the feedback system is the same for both left and right coprime factors.

In the general nonlinear case it is known that $\| \Pi_{\mathcal{M}\|\mathcal{N}} \|_\infty \neq \| \Pi_{\mathcal{N}\|\mathcal{M}} \|_\infty$ (see [7] for an example where both $\mathcal{M}$ and $\mathcal{N}$ are piecewise linear static functions). In this section we demonstrate a connection between these quantities under a certain scaling. This type of conjugacy transformation has also appeared in Lemma 4.1 in relating two different types of gap metrics. In this way, a new type of duality between the two complementary parallel projections is uncovered which is perhaps a more fundamental property.

For $\gamma>1$ let

$$\beta=\sqrt{1-\gamma^{-2}}<1. \tag{46}$$

We say that controllers K_1, K_2 are *β-conjugate* if

$$K_1=\beta^2 K_2 \beta^{-2}. \tag{47}$$

THEOREM 4.7. *Any plant P enjoys* conjugate norm equivalence,

$$\| \Pi_{\mathcal{M}\|\mathcal{N}_1} \|_\infty < \gamma \quad \textit{iff} \quad \| \Pi_{\mathcal{N}_2\|\mathcal{M}} \|_\infty < \gamma, \tag{48}$$

whenever K_1, K_2 are β-conjugate, where $\mathcal{M} = \mathcal{G}_P$, $\mathcal{N}_i = \mathcal{G}_{K_i}$ $(i = 1, 2)$.

Proof. We shall first show that, if K_2 satisfies $\| \Pi_{\mathcal{N}_2\|\mathcal{M}} \|_\infty < \gamma$ and K_1 is defined by (47), then $\| \Pi_{\mathcal{M}\|\mathcal{N}_1} \|_\infty < \gamma$. Then we shall show the converse.

Let $w \in \mathcal{W}$. Since $[P, K_1]$ is well-posed, there exists $m \in \mathcal{M}$, $n_1 \in \mathcal{N}_1$ such that $w = m + n_1$, so $\Pi_{\mathcal{M}\|\mathcal{N}_1}(w) = m$. Let $n_2 = \beta^{-2} n_1$ and note that $n_2 \in \mathcal{N}_2$. Now, $\| \Pi_{\mathcal{N}_2\|\mathcal{M}} \|_\infty < \gamma$ implies that $\|\beta^{-2} n_1 + m\|_T \geq \gamma^{-1} \|\beta^{-2} n_1\|_T$ for all T. However, squaring both sides, expanding, and completing the squares shows that this holds if and only if $\|n_1 + m\|_T \geq \gamma^{-1} \|m\|_T$ for all T as required.

Conversely, write $w = m + n_2$, so $\Pi_{\mathcal{N}_2\|\mathcal{M}}(w) = n_2$, and put $n_1 = \beta^2 n_2 \in \mathcal{N}_1$. $\| \Pi_{\mathcal{M}\|\mathcal{N}_1} \|_\infty < \gamma$ implies that $\|n_1 + m\|_T \geq \gamma^{-1} \|m\|_T$ for all T which is equivalent to $\|n_2 + m\|_T \geq \gamma^{-1} \|n_2\|_T$ as above. ∎

COROLLARY 4.8. *We have*

$$\inf_{K_1} \| \Pi_{\mathcal{M}\|\mathcal{N}_1} \|_\infty = \inf_{K_2} \| \Pi_{\mathcal{N}_2\|\mathcal{M}} \|_\infty . \tag{49}$$

Proof. Given any K_2 we can define K_1 beta-conjugate to K_2 as in (47) to achieve the same norm for the complementary parallel projection, and similarly with the roles of K_1 and K_2 interchanged. Hence the infimums must be equal. ∎

An operator P is *(positively, multiplicatively) homogeneous* if

$$\alpha P = P \alpha \tag{50}$$

for all real $\alpha > 0$. Linear systems, of course, enjoy this property.

COROLLARY 4.9. *If P and/or K is homogeneous, then*

$$\| \Pi_{\mathcal{M}\|\mathcal{N}} \|_\infty = \| \Pi_{\mathcal{N}\|\mathcal{M}} \|_\infty . \tag{51}$$

Proof. Suppose $[P, K]$ is stable, $\| \Pi_{\mathcal{N}\|\mathcal{M}} \|_\infty < \gamma$, and K is homogeneous. Write $K_2 = K$, so $\mathcal{N}_2 = \mathcal{N}$, and put $K_1 = \beta^2 K_2 \beta^{-2}$, which gives $K_1 = K$ and $\mathcal{N}_1 = \mathcal{N}$, since K is homogeneous. It follows, from Theorem 4.7, that $\| \Pi_{\mathcal{M}\|\mathcal{N}} \|_\infty < \gamma$. The reverse may be established similarly, writing $K_1 = K$ and letting $K_2 = \beta^{-2} K_1 \beta^2 = K$. This establishes (51).

The same result will hold if P is homogeneous, since (51) is symmetric in $\mathcal{M}$ and $\mathcal{N}$. ∎

4.4. Evaluation. The following theorem shows how the ρ-gaps can be evaluated in terms of the right and left symbols from the plant representations.

THEOREM 4.10.

1. *Assume L_{tv0} is a positive time co-inner left representation for P_0 and $\mathsf{R}_{T,1}$ is a finite time inner right representation for P_1; then*

$$\vec{\rho}_0(P_0, P_1) = \limsup_{T \to \infty} \| \mathsf{L}_{tv0} \mathsf{R}_{T,1} \|_{[0,T],\infty} . \tag{52}$$

2. *Assume L_{e0} is a co-inner left representation for P_0 and $\mathsf{R}_{T,1}$ is a finite time inner right representation for P_1; then*

$$\vec{\rho}_g(P_0, P_1) = \limsup_{T \to \infty} \| \mathsf{L}_{e0} \mathsf{R}_{T,1} \|_{(-\infty,T],\infty} . \tag{53}$$

3. *Assume* L_{tv0} *is a positive time co-inner left representation for* P_0 *and* R_{e1} *is an inner right representation for* P_1*; then*

$$\vec{\rho}_{H_2}(P_0, P_1) = \| \mathsf{L}_{tv0}\mathsf{R}_{e1} \|_\infty . \tag{54}$$

4. *Assume* L_{e0} *is a co-inner left representation for* P_0 *and* R_{e1} *is an inner right representation for* P_1*; then*

$$\vec{\rho}_{L_2}(P_0, P_1) = \| \mathsf{L}_{e0}\mathsf{R}_{e1} \|_\infty . \tag{55}$$

Proof. 1. $\vec{\rho}_0(P_0, P_1)$. Let $T \geq 0$ and let $0 \neq x_1 \in \mathcal{G}_1 \cap L_2[0, T]$. Then by the positive time co-inner property (21)

$$\inf_{x_0 \in \mathcal{G}_0 \cap L_{2,e}} \| x_1 - x_0 \|_T = \| \mathsf{L}_{tv0} x_1 \|_T .$$

Now $x_1|_{[0,T]} = \mathsf{R}_{T,1}\psi$ for some $\psi \in L_2[0, T]$ by the onto property (9), and so the finite time inner property (11) implies

$$\| x_1 \|_T = \| \mathsf{R}_{T,1}\psi \|_T = \| \psi \|_T .$$

Then

$$\inf_{x_0 \in \mathcal{G}_0 \cap L_{2,e}} \frac{\| x_1 - x_0 \|_T}{\| x_1 \|_T} = \frac{\| \mathsf{L}_{tv0}\mathsf{R}_{T,1}\psi \|_T}{\| \psi \|_T},$$

and so

$$\sup_{\substack{x_1 \in \mathcal{G}_1 \cap L_{2,e} \\ \|x_1\|_T \neq 0}} \inf_{x_0 \in \mathcal{G}_0 \cap L_{2,e}} \frac{\| x_1 - x_0 \|_T}{\| x_1 \|_T} = \sup_{\substack{\psi \in L_{2,e} \\ \|\psi\|_T \neq 0}} \frac{\| \mathsf{L}_{tv0}\mathsf{R}_{T,1}\psi \|_T}{\| \psi \|_T} = \| \mathsf{L}_{tv0}\mathsf{R}_{T,1} \|_{[0,T],\infty} .$$

Taking the lim sup as $T \to \infty$ we obtain the formula (52) for $\vec{\rho}_0(P_0, P_1)$.

2. $\vec{\rho}_g(P_0, P_1)$. Let $T \in (-\infty, \infty)$ and let $x_1 \in \mathcal{G}_1$, $\|x_1\|_T \neq 0$. Then by the co-inner property (27),

$$\| \mathsf{L}_{eP} x_1 \|_T = \inf_{x_0 \in \mathcal{G}_0} \| x_1 - x_0 \|_T .$$

By the finite time inner property (11),

$$\| x_1 \|_T = \| \mathsf{R}_{T,1}\psi \|_T = \| \psi \|_T$$

for some $\psi \in \mathcal{U} \cap L_2(-\infty, T]$ (by (9), $T_0 = -\infty$). Then

$$\sup_{\substack{x_1 \in \mathcal{G}_1 \\ \|x_1\|_T \neq 0}} \inf_{x_0 \in \mathcal{G}_0} \frac{\| x_1 - x_0 \|_T}{\| x_1 \|_T} = \sup_{\substack{\psi \in \mathcal{U} \\ \|\psi\|_T \neq 0}} \frac{\| \mathsf{L}_{e0}\mathsf{R}_{T,1}\psi \|_T}{\| \psi \|_T} = \| \mathsf{L}_{e0}\mathsf{R}_{T,1} \|_{(-\infty,T],\infty} .$$

Taking the lim sup as $T \to \infty$ we obtain the formula (53) for $\vec{\rho}_g(P_0, P_1)$.

3. $\vec{\rho}_{H_2}(P_0, P_1)$. Let $0 \neq x_1 \in \mathcal{G}_1 \cap L_2[0, \infty)$. Then by the positive time co-inner property (22),

$$\inf_{x_0 \in \mathcal{G}_0 \cap L_2[0,\infty)} \| x_1 - x_0 \|_2 = \| \mathsf{L}_{tv0} x_1 \|_2 .$$

Now by the inner property (16)

$$\| x_1 \|_2 = \| \mathsf{R}_{e1}\psi \|_2 = \| \psi \|_2$$

for some $\psi \in \mathcal{U} \cap L_2[0,\infty)$, by (14). Then

$$\sup_{\substack{x_1 \in \mathcal{G}_1 \cap L_2[0,\infty) \\ x_1 \neq 0}} \inf_{x_0 \in \mathcal{G}_0 \cap L_2[0,\infty)} \frac{\| x_1 - x_0 \|_2}{\| x_1 \|_2} = \sup_{\substack{\psi \in L_2[0,\infty) \\ \psi \neq 0}} \frac{\| \mathsf{L}_{tv0}\mathsf{R}_{e1}\psi \|_2}{\| \psi \|_2} = \| \mathsf{L}_{tv0}\mathsf{R}_{e1} \|_\infty .$$

This gives the formula (54) for $\vec{\rho}_{H_2}(P_0, P_1)$.

4. $\vec{\rho}_{L_2}(P_0, P_1)$. Let $0 \neq x_1 \in \mathcal{G}_1 \cap L_2$. Then by the co-inner property (28)

$$\inf_{x_0 \in \mathcal{G}_0 \cap L_2} \| x_1 - x_0 \|_2 = \| \mathsf{L}_{e0}x_1 \|_2 .$$

Now by the inner property (16)

$$\| x_1 \|_2 = \| \mathsf{R}_{e1}\psi \|_2 = \| \psi \|_2$$

for some $\psi \in \mathcal{U} \cap L_2$, by (14). Then

$$\sup_{\substack{x_1 \in \mathcal{G}_1 \cap L_2 \\ x_1 \neq 0}} \inf_{x_0 \in \mathcal{G}_0 \cap L_2} \frac{\| x_1 - x_0 \|_2}{\| x_1 \|_2} = \sup_{\substack{\psi \in L_2 \\ \psi \neq 0}} \frac{\| \mathsf{L}_{e0}\mathsf{R}_{e1}\psi \|_2}{\| \psi \|_2} = \| \mathsf{L}_{e0}\mathsf{R}_{e1} \|_\infty .$$

This gives the formula (55) for $\vec{\rho}_{L_2}(P_0, P_1)$. □

4.5. Robust stability and representation uncertainty. The subject of coprime fraction uncertainty, and robustness for nonlinear systems has previously been considered in [1] and [25] and some partial results obtained. In this section we give two dual results relating uncertainty in the graph representations to gap balls of uncertainty for the new definitions given in this paper.

LEMMA 4.11.

1. *Let L_{e0} be a co-inner left representation of P_0, and let L_{e1} be a left representation of P_1 (not necessarily coinner); then*

$$\vec{\rho}_{L_2}(P_0, P_1) \leq \| \mathsf{L}_{e0} - \mathsf{L}_{e1} \|_\infty . \tag{56}$$

2. *Let R_{e0} be an inner right representation of P_0, and let R_{e1} be a right representation of P_1 (not necessarily inner) satisfying $\mathsf{R}_{e0}0 = 0$; then*

$$\vec{\delta}_{L_2}(P_0, P_1) \leq \| \mathsf{R}_{e0} - \mathsf{R}_{e1} \|_\infty . \tag{57}$$

Proof. Let $x_1 \in \mathcal{G}_1 \cap L_2(-\infty,\infty), x_1 \neq 0$. Then $\mathsf{L}_{e1}x_1 = 0$, and so

$$\mathsf{L}_{e0}x_1 + (\mathsf{L}_{e1} - \mathsf{L}_{e0})x_1 = 0.$$

Since L_{e0} is coinner, by (28) we have

$$\begin{aligned} \inf_{x_0 \in \mathcal{G}_0 \cap L_2(-\infty,\infty)} \| x_1 - x_0 \|_2 &= \| \mathsf{L}_{e0}x_1 \|_2 \\ &= \| (\mathsf{L}_{e1} - \mathsf{L}_{e0})x_1 \|_2 . \end{aligned}$$

Therefore,

$$\begin{aligned}\vec{\rho}_{L_2}(P_0,P_1) &= \sup_{x_1\in\mathcal{G}_1\cap L_2(-\infty,\infty),x_1\neq 0}\frac{\|\ (\mathsf{L}_{e1}-\mathsf{L}_{e0})x_1\ \|_2}{\|\ x_1\ \|_2}\\ &\leq \sup_{x_1\in L_2(-\infty,\infty),x_1\neq 0}\frac{\|\ (\mathsf{L}_{e1}-\mathsf{L}_{e0})x_1\ \|_2}{\|\ x_1\ \|_2}\\ &=\|\ \mathsf{L}_{e0}-\mathsf{L}_{e1}\ \|_\infty,\end{aligned}$$

establishing (56).

Again let $x_1 \in \mathcal{G}_1 \cap L_2(-\infty,\infty), x_1 \neq 0$. Then by (14) there exists $\psi_1 \in \mathcal{U}\cap L_2(-\infty,\infty)$, $\psi_1 \neq 0$, such that $x_1 = \mathsf{R}_{e1}\psi_1$ and $\psi_1 \neq 0$ since $\mathsf{L}_{e0}0 = 0$. Since R_{e0} is inner, by (16) we have

$$\|\ x_0\ \|_2=\|\ \psi_1\ \|_2,$$

where $x_0 = \mathsf{R}_{e0}\psi_1 \neq 0$, and also we have

$$\|\ x_1 - x_0\ \|_2=\|\ (\mathsf{R}_{e1}-\mathsf{R}_{e0})\psi_1\ \|_2\ .$$

Hence

$$\begin{aligned}\frac{\|\ x_1-x_0\ \|_2}{\|\ x_0\ \|_2} &= \frac{\|\ (\mathsf{R}_{e1}-\mathsf{R}_{e0})\psi_1\ \|_2}{\|\ \psi_1\ \|_2}\\ &\leq\|\ \mathsf{R}_{e0}-\mathsf{R}_{e1}\ \|_\infty,\end{aligned}$$

from which (57) follows. □

THEOREM 4.12. *Assume $H_{P_0,K}$ is stable. Let L_{e0} be a co-inner left representation of P_0, and let L_{e1} be a left representation of P_1 (not necessarily co-inner). If*

$$\|\ \mathsf{L}_{e0}-\mathsf{L}_{e1}\ \|_\infty<\|\ \Pi_{\mathcal{N}\|\mathcal{M}_0}\ \|_\infty^{-1}, \tag{58}$$

where $\mathcal{M}_0 = \mathcal{G}_{P_0}$ and $\mathcal{N} = \mathcal{G}_K$, then $H_{P_1,K}$ is stable,

$$\vec{\rho}_{L_2}(P_0,P_1) <\|\ \Pi_{\mathcal{N}\|\mathcal{M}_0}\ \|_\infty^{-1}, \tag{59}$$

and the norm bound (43) *of Theorem* 4.5 *applies.*

Proof. Assume $\|\ \mathsf{L}_{e0}-\mathsf{L}_{e1}\ \|_\infty< \alpha$, $\|\ \Pi_{\mathcal{N}\|\mathcal{M}_0}\ \|_\infty< \gamma$, and $\alpha\gamma < 1$. As in the proof of Theorem 4.4, let $w \in \mathcal{W}$. Since $[P_1, K]$ is well-posed by assumption, there exists $m_1 \in \mathcal{M}_1$, $n \in \mathcal{N}$ such that $w = m_1 + n$, so $\Pi_{\mathcal{N}\|\mathcal{M}_1}(w) = n$.

There exists $T_0 > 0$ such that for all $T \geq T_0$

$$\sup_{x\in L_{2,ce}(-\infty,\infty),x\neq 0}\frac{\|\ (\mathsf{L}_{e0}-\mathsf{L}_{e1})x\ \|_T}{\|\ x\ \|_T} < \alpha.$$

Then

$$\frac{\|\ (\mathsf{L}_{e0}-\mathsf{L}_{e1})m_1\ \|_T}{\|\ m_1\ \|_T} < \alpha,$$

and hence

$$\frac{\|\ \mathsf{L}_{e0}m_1\ \|_T}{\|\ m_1\ \|_T} < \alpha,$$

because $\mathsf{L}_{e1}m_1 = 0$ implies $\mathsf{L}_{e0}m_1 = (\mathsf{L}_{e0} - \mathsf{L}_{e1})m_1$. Since L_{e0} is co-inner, by (27) there exists $m_0 \in \mathcal{G}_0$ (depending on T) such that

$$\frac{\| m_1 - m_0 \|_T}{\| m_1 \|_T} < \alpha.$$

Also, we have $\Pi_{\mathcal{N}\|\mathcal{M}_0}(m_0 + n) = n$, and following the proof of Theorem 4.4, we find that

$$(\gamma^{-1} - \alpha) \| n \|_T \leq (1 + \alpha) \| m_1 + n \|_T$$

for all $T \geq T_0$. This is enough to prove $\| \Pi_{\mathcal{N}\|\mathcal{M}_0} \|_\infty < \infty$. Note that once stability of $H_{P_1,K}$ has been established, the bound (43) follows from Theorem 4.5 and Lemma 4.11. □

The corresponding result for right representations is not perfectly dual to Theorem 4.12 in that a left inverse for the representation appears to be required. This is perhaps not completely unexpected in the light of Remark 3.1(2).

THEOREM 4.13. *Assume $H_{P_0,K}$ is stable. Let R_{e0} be an inner, contractive right representation of P_0, with causal left inverse R_{e0}^{-L}, and let R_{e1} be a right representation of P_1 (not necessarily inner). If*

$$\| \mathsf{R}_{e0} - \mathsf{R}_{e1} \|_\infty < \| \Pi_{\mathcal{M}_0\|\mathcal{N}} \|_\infty^{-1}, \tag{60}$$

where $\mathcal{M}_0 = \mathcal{G}_{P_0}$ and $\mathcal{N} = \mathcal{G}_K$, then $H_{P_1,K}$ is stable,

$$\vec{\delta}_{L_2}(P_0, P_1) < \| \Pi_{\mathcal{M}_0\|\mathcal{N}} \|_\infty^{-1}, \tag{61}$$

and by [30, Theorem 2.1] *the norm bound*

$$\| \Pi_{\mathcal{M}_1\|\mathcal{N}} \|_\infty \leq \| \Pi_{\mathcal{M}_0\|\mathcal{N}} \|_\infty \frac{1 + \vec{\delta}_{L_2}(P_0, P_1)}{1 - \| \Pi_{\mathcal{M}_0\|\mathcal{N}} \|_\infty \vec{\delta}_{L_2}(P_0, P_1)} \tag{62}$$

holds, where $\mathcal{M}_1 = \mathcal{G}_{P_1}$.

Proof. Assume $\| \mathsf{R}_{e0} - \mathsf{R}_{e1} \|_\infty < \alpha$, $\| \Pi_{\mathcal{M}_0\|\mathcal{N}} \|_\infty < \gamma$, and $\alpha\gamma < 1$. As in the proof of [11, Theorem 3], let $w \in \mathcal{W}$. Since $[P_1, K]$ is well-posed by assumption, there exists $m_1 \in \mathcal{M}_1$, $n \in \mathcal{N}$ such that $w = m_1 + n$, so $\Pi_{\mathcal{M}_1\|\mathcal{N}}(w) = m_1$.

Define an operator $\tilde{\Pi}_{\mathcal{M}_0\|\mathcal{N}} : \mathcal{W} \to \mathcal{U}$ by

$$\tilde{\Pi}_{\mathcal{M}_0\|\mathcal{N}} = \mathsf{R}_{e0}^{-L} \Pi_{\mathcal{M}_0\|\mathcal{N}},$$

so that $\mathsf{R}_{e0}^{-L}\mathsf{R}_{e0}\tilde{\Pi}_{\mathcal{M}_0\|\mathcal{N}} = \mathsf{R}_{e0}^{-L}\Pi_{\mathcal{M}_0\|\mathcal{N}}$. We can check that $\mathsf{R}_{e0}^{-L}w_1 = \mathsf{R}_{e0}^{-L}w_2$, for $w_1, w_2 \in \mathcal{M}_0$ implies $w_1 = w_2$, which gives $\Pi_{\mathcal{M}_0\|\mathcal{N}} = \mathsf{R}_{e0}\tilde{\Pi}_{\mathcal{M}_0\|\mathcal{N}}$. Then, since R_{e0} is inner, (16) implies

$$\| \tilde{\Pi}_{\mathcal{M}_0\|\mathcal{N}} \|_\infty = \| \Pi_{\mathcal{M}_0\|\mathcal{N}} \|_\infty < \gamma.$$

Since R_{e1} is a right representation of P_1, by (14) there exists $\psi_1 \in \mathcal{U}$ such that $m_1 = \mathsf{R}_{e1}\psi_1$. Set $m_0 = \mathsf{R}_{e0}\psi_1$, and note that $m_0 = \Pi_{\mathcal{M}_0\|\mathcal{N}}(m_0 + n)$. Then $\psi_1 = \tilde{\Pi}_{\mathcal{M}_0\|\mathcal{N}}(m_0 + n)$, and for sufficiently large T

$$\begin{aligned} \| \psi_1 \|_T &\leq \gamma \| m_0 + n \|_T \\ &\leq \gamma \| m_1 + n \|_T + \gamma \| m_0 - m_1 \|_T \\ &= \gamma \| m_1 + n \|_T + \gamma \| (\mathsf{R}_{e0} - \mathsf{R}_{e1})\psi_1 \|_T \\ &\leq \gamma \| m_1 + n \|_T + \gamma\alpha \| \psi_1 \|_T \end{aligned}$$

so that

$$\| \psi_1 \|_T \leq \frac{\gamma}{1-\alpha\gamma} \| w \|_T .$$

Next, since R_{e0} is contractive (see (15)) and from above we have

$$\begin{aligned} \| m_1 \|_T &\leq \| m_0 \|_T + \| m_1 - m_0 \|_T \\ &\leq \| \psi_1 \|_T + \alpha \| \psi_1 \|_T \\ &\leq \tfrac{\gamma(1+\alpha)}{1-\alpha\gamma} \| w \|_T, \end{aligned}$$

and the theorem follows. □

Other relationships among the ρ-gaps and representation uncertainties are as follows (L_{e0} and L_{tv} are co-inner):

$$\begin{aligned} \vec{\rho}_0(P_0, P_1) &\leq \| \mathsf{L}_{tv0} - \mathsf{L}_{tv1} \|_\infty, \\ \vec{\rho}_g(P_0, P_1) &\leq \| \mathsf{L}_{e0} - \mathsf{L}_{e1} \|_\infty, \\ \vec{\rho}_{H_2}(P_0, P_1) &\leq \| \mathsf{L}_{tv0} - \mathsf{L}_{tv1} \|_\infty . \end{aligned} \tag{63}$$

Similarly, for the δ-gaps we have (R_{e0} and $\mathsf{R}_{T,0}$ are inner)

$$\begin{aligned} \vec{\delta}_0(P_0, P_1) &\leq \| \mathsf{R}_{T,0} - \mathsf{R}_{T,1} \|_{[0,T],\infty}, \\ \vec{\delta}_g(P_0, P_1) &\leq \| \mathsf{R}_{T,0} - \mathsf{R}_{T,1} \|_{(-\infty,T],\infty}, \\ \vec{\delta}_{H_2}(P_0, P_1) &\leq \| \mathsf{R}_{e0} - \mathsf{R}_{e1} \|_\infty . \end{aligned} \tag{64}$$

It is possible to establish representation uncertainty stability results using these relationships, though we do not do so explicitly here.

5. Controller synthesis. We turn our attention to controller synthesis for the two H^∞ control problems which minimize the norms of the two parallel projections in (8). These two problems are identical for linear systems, and their special structure has made them amenable to computation and particularly useful in controller design. For this reason it is of interest to try to generalize them to nonlinear systems. This will be done here for input-affine systems.

We consider a plant $P : \mathcal{U} \to \mathcal{Y}$ with the following state space model:

$$P \ : \ \begin{cases} \dot{x} = A(x) + B(x)u \ \text{ on } (-\infty, \infty), \ x(-\infty) = 0, \\ y = C(x), \end{cases} \tag{65}$$

where $u \in \mathbf{R}^m$, $y \in \mathbf{R}^p$. We use the notation

$$P = \begin{bmatrix} A & B \\ C & 0 \end{bmatrix}$$

to mean P has a state space realization with matrix-valued functions A, B, and C.

Assumption 5.1. We assume A, B, C are of class C^1 with globally bounded derivatives, B is globally bounded, and 0 is an equilibrium: $A(0) = 0$, $C(0) = 0$.

The meaning of (65) is as follows. If $u \in \mathcal{U}$, then $u(t) = 0$ for all $t \leq T_0$, for some finite T_0, and since 0 is an equilibrium we must have $x(t) = 0$ for all $t \leq T_0$. For $t \geq T_0$, the differential equation is integrated as usual.

Our approach is to apply the information state framework of [17], [14] to this problem (see Appendix A for background information). We shall see that a single H_2 equation underlies the controllers:

$$\nabla V A - \tfrac{1}{2} \nabla V B B' \nabla V' + \tfrac{1}{2} |C|^2 = 0. \tag{66}$$

We denote by V_+ the unique smooth solution of (66) satisfying $V_+ \geq 0$, $V_+(0) = 0$, and

$$(A - BB'\nabla V_+') \quad \text{asymptotically stable,} \tag{67}$$

and by V_- the unique smooth solution of (66) satisfying $V_- \leq 0$, $V_-(0) = 0$, and

$$-(A - BB'\nabla V_-') \quad \text{asymptotically stable.} \tag{68}$$

Before applying this theory it will be useful to carry out some preliminary transformations on the generalized plants corresponding to the two parallel projection operators.

5.1. Generalized plant transformations. In this section we transform the generalized plants $\mathcal{P}_1$ and $\mathcal{P}_2$ in equation (7) into modified forms $\tilde{\mathcal{P}}_1$ and $\tilde{\mathcal{P}}_2$ which have "zero D_4 terms" in the standard terminology; see, e.g., [13], [1], [31] for background on the use of such transformations.

In terms of the state-space model (65) for the plant P, the generalized plants $\mathcal{P}_1$, $\mathcal{P}_2$ in (7) are given by

$$\mathcal{P}_1 = \left[\begin{array}{c|cc|c} A & B & 0 & -B \\ \hline 0 & I & 0 & -I \\ C & 0 & 0 & 0 \\ \hline -C & 0 & I & 0 \end{array}\right] \quad \text{and} \quad \mathcal{P}_2 = \left[\begin{array}{c|ccc} A & B & 0 & -B \\ \hline 0 & 0 & 0 & I \\ -C & 0 & I & 0 \\ \hline -C & 0 & I & 0 \end{array}\right]. \tag{69}$$

Now consider the following transformed plants:

$$\tilde{\mathcal{P}}_1 : \begin{pmatrix} \tilde{u}_0 \\ \tilde{y}_0 \\ \tilde{u}_2 \end{pmatrix} \mapsto \begin{pmatrix} \tilde{u}_1 \\ \tilde{y}_1 \\ y_2 \end{pmatrix} \quad \text{and} \quad \tilde{\mathcal{P}}_2 : \begin{pmatrix} \hat{u}_0 \\ \hat{y}_0 \\ u_2 \end{pmatrix} \mapsto \begin{pmatrix} \hat{u}_2 \\ \hat{y}_2 \\ \tilde{y}_2 \end{pmatrix},$$

defined by the block diagrams in Figures 4 and 5, where

$$\Theta_1 = \gamma^{-1} \left(\begin{array}{cc|cc} \gamma^{-1}I & 0 & \beta I & 0 \\ 0 & 0 & 0 & I \\ \hline -\beta I & 0 & \gamma^{-1}I & 0 \\ 0 & -I & 0 & 0 \end{array}\right) \quad \text{and} \quad \Theta_2 = \gamma^{-1} \left(\begin{array}{cc|cc} 0 & 0 & I & 0 \\ 0 & \gamma^{-1}I & 0 & \beta I \\ \hline -I & 0 & 0 & 0 \\ 0 & -\beta I & 0 & \gamma^{-1}I \end{array}\right), \tag{70}$$

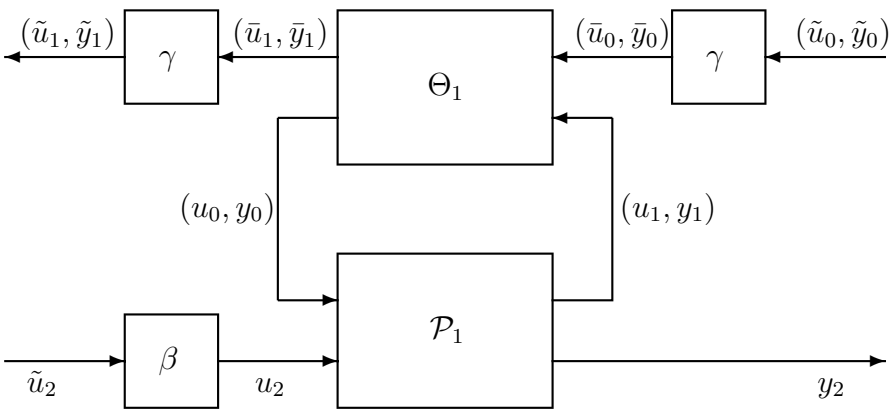


FIG. 4. *Construction of* $\tilde{\mathcal{P}}_1$.

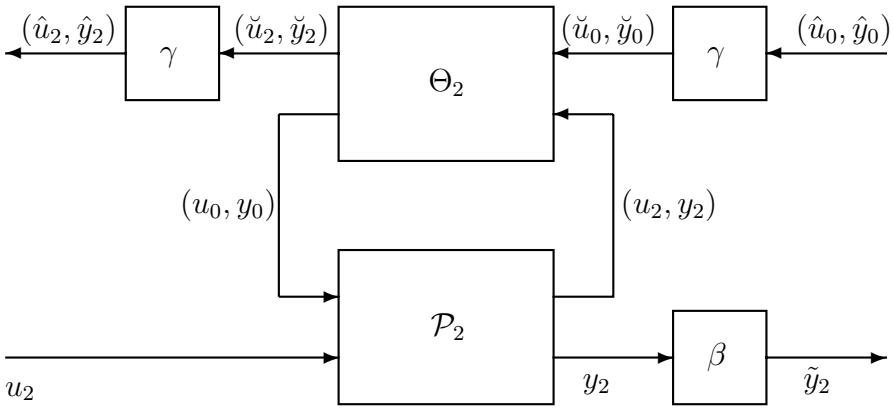


FIG. 5. *Construction of* $\tilde{\mathcal{P}}_2$.

and $\beta = \sqrt{1-\gamma^{-2}}$. It can be verified (see Lemma 5.2 below) that the plants have the following state-space realizations:

(71)
$$\tilde{\mathcal{P}}_1 = \left[\begin{array}{c|cc|c} A & -\beta^{-1}B & 0 & -\beta^{-1}B \\ \hline 0 & 0 & 0 & -I \\ C & 0 & 0 & 0 \\ \hline -C & 0 & -I & 0 \end{array}\right] \quad \text{and} \quad \tilde{\mathcal{P}}_2 = \left[\begin{array}{c|cc|c} A & -B & 0 & -B \\ \hline 0 & 0 & 0 & I \\ -\beta^{-1}C & 0 & 0 & 0 \\ \hline -\beta^{-1}C & 0 & -I & 0 \end{array}\right].$$

We also define the controllers K_1 and K_2:

$$\tilde{K}_1 = \beta^{-1}K_1, \quad \tilde{K}_2 = K_2\beta^{-1}. \tag{72}$$

The constructions of the transformed plants $\tilde{\mathcal{P}}_1$, $\tilde{\mathcal{P}}_2$ take the same form as in [13, p. 162] and yield the following result.

LEMMA 5.2. *The plants* $\tilde{\mathcal{P}}_1$ *and* $\tilde{\mathcal{P}}_1$ *defined in Figures* 4 *and* 5 *have the state-space realizations given by* (71) *and satisfy*

$$\begin{array}{l} \| (\mathcal{P}_1, K_1) \|_\infty < \gamma \quad \textit{iff} \quad \| (\tilde{\mathcal{P}}_1, \tilde{K}_1) \|_\infty < \gamma, \quad \textit{and} \\ \| (\mathcal{P}_2, K_2) \|_\infty < \gamma \quad \textit{iff} \quad \| (\tilde{\mathcal{P}}_2, \tilde{K}_2) \|_\infty < \gamma. \end{array} \tag{73}$$

Proof. The proof is substantially as in [13, p. 162], and we give some details here for completeness.

The plant $\mathcal{P}_1$ in state-space form reads (recall (69)):

$$\mathcal{P}_1 \ : \ \left\{ \begin{array}{l} \dot{x} = A(x) + B(x)(u_0 - u_2), \\ u_1 = u_0 - u_2, \\ y_1 = C(x), \\ y_2 = -C(x) + y_0. \end{array} \right. \tag{74}$$

We also have the relations

$$\begin{array}{r} u_0 = -\frac{\gamma^{-1}}{\beta}\bar{u}_0 - \frac{\gamma^{-2}}{\beta^2}u_2, \\ \bar{u}_1 = -\frac{\gamma^{-1}}{\beta}u_2, \\ u_0 - u_2 = -\frac{\gamma^{-1}}{\beta}\bar{u}_0 - \frac{1}{\beta^2}u_2. \end{array} \tag{75}$$

Then (74) and (75) imply

$$\begin{aligned} \dot{x} &= A(x) + B(x)(-\tfrac{\gamma^{-1}}{\beta}\bar{u}_0 - \tfrac{1}{\beta^2}u_2), \\ \bar{u}_1 &= -\tfrac{\gamma^{-1}}{\beta}u_2, \\ \bar{y}_1 &= \gamma^{-1}C(x), \\ y_2 &= -C(x) + y_0. \end{aligned} \tag{76}$$

Then by scaling (refer to Figure 4) we get

$$\tilde{\mathcal{P}}_1 \ : \ \begin{cases} \dot{x} = A(x) + B(x)(-\frac{1}{\beta}\tilde{u}_0 - \frac{1}{\beta}\tilde{u}_2), \\ \tilde{u}_1 = -\tilde{u}_2, \\ \tilde{y}_1 = C(x), \\ y_2 = -C(x) - \tilde{y}_0. \end{cases} \tag{77}$$

The construction of $\mathcal{P}_2$ is similar.

We claim that K_1 achieves

$$\| u_1 \|_{2,T}^2 + \| y_1 \|_{2,T}^2 \leq \gamma^2(\| u_0 \|_{2,T}^2 + \| y_0 \|_{2,T}^2) \tag{78}$$

for all $T \geq 0$ if and only if $\tilde{K}_1$ achieves

$$\| \tilde{u}_1 \|_{2,T}^2 + \| \tilde{y}_1 \|_{2,T}^2 \leq \gamma^2(\| \tilde{u}_0 \|_{2,T}^2 + \| \tilde{y}_0 \|_{2,T}^2) \tag{79}$$

for all $T \geq 0$.

Since $\Theta_1^*\Theta_1 = \gamma^{-2}I$ we have

$$\begin{aligned} &\| \bar{u}_1 \|_{2,T}^2 + \| \bar{y}_1 \|_{2,T}^2 + \| u_0 \|_{2,T}^2 + \| y_0 \|_{2,T}^2 \\ &\quad = \gamma^{-2}(\| \bar{u}_0 \|_{2,T}^2 + \| \bar{y}_0 \|_{2,T}^2 + \| u_1 \|_{2,T}^2 + \| y_1 \|_{2,T}^2). \end{aligned} \tag{80}$$

Then (78) holds if and only if

$$\| \bar{u}_1 \|_{2,T}^2 + \| \bar{y}_1 \|_{2,T}^2 \leq \gamma^{-2}(\| \bar{u}_0 \|_{2,T}^2 + \| \bar{y}_0 \|_{2,T}^2), \tag{81}$$

which is equivalent to (79)

Similarly, it can be checked that K_2 achieves

$$\| u_2 \|_{2,T}^2 + \| y_2 \|_{2,T}^2 \leq \gamma^2(\| u_0 \|_{2,T}^2 + \| y_0 \|_{2,T}^2) \tag{82}$$

for all $T \geq 0$ if and only if $\tilde{K}_2$ achieves

$$\| \hat{u}_2 \|_{2,T}^2 + \| \hat{y}_2 \|_{2,T}^2 \leq \gamma^2(\| \hat{u}_0 \|_{2,T}^2 + \| \hat{y}_0 \|_{2,T}^2) \tag{83}$$

for all $T \geq 0$. We omit the remaining details. □

5.2. Synthesis for $\Pi_{\mathcal{N}\|\mathcal{M}}$. We are now ready to apply the information state framework of [17] and [14] to the controller synthesis problem for $\Pi_{\mathcal{N}\|\mathcal{M}}$. Given $\gamma > 0$, we wish to find, if possible, a controller K_2 such that

$$\| \Pi_{\mathcal{N}_2\|\mathcal{M}} \|_\infty = \| (\mathcal{P}_2, K_2) \|_\infty < \gamma, \tag{84}$$

where $\mathcal{P}_2$ is given by (69). By Lemma 5.2, it is enough to find a controller $\tilde{K}_2$ such that

$$\| (\tilde{\mathcal{P}}_2, \tilde{K}_2) \|_\infty < \gamma, \tag{85}$$

and this is technically simpler since the D_{11} term is zero in $\tilde{\mathcal{P}}_2$.

The information state for $\tilde{\mathcal{P}}_2$ is denoted by p_{2t} and satisfies (from [14, eq. (3.9)] and (195) below)

$$\begin{aligned}\tfrac{\partial}{\partial t}p_2 \quad &= -\nabla p_2(A - Bu_2) + \gamma^{-2}\tfrac{1}{2}\nabla p_2 BB'\nabla p_2' - \gamma^2\tfrac{1}{2}|C|^2 \\ &\quad +\tfrac{1}{2}|u_2|^2 - \gamma^2\tfrac{1}{2}|\tilde{y}_2|^2 - \gamma^2\beta^{-1}C'\tilde{y}_2,\end{aligned} \tag{86}$$

or in shorthand,

$$\dot{p}_2 = \tilde{F}_2(p_2, u_2, \tilde{y}_2), \tag{87}$$

where $\tilde{F}_2(p_2, u_2, \tilde{y}_2)$ is the differential operator defined by the right-hand side of (86). The Hamilton–Jacoby–Bellman–Isaacs (HJBI) equation for the value function $W_2(p_2)$ is (from [14, eq. (4.16)] and (199) below)

$$\inf_{u_2\in\mathbf{R}^m}\sup_{\tilde{y}_2\in\mathbf{R}^p}\left\{\nabla W_2(p_2)[\tilde{F}_2(p_2, u_2, \tilde{y}_2)]\right\} = 0, \tag{88}$$

from which the optimal information state feedback function is (by [14, eq. (4.23)] and (200) below)

$$u_2^*(p_2) = -\nabla W_2(p_2)[B'\nabla p_2']. \tag{89}$$

The stationary information state for the control attractor p_{2e} is (from [14, eq. (3.28)] and (197) below)

$$0 = -\nabla p_{2e}A + \gamma^{-2}\tfrac{1}{2}\nabla p_{2e}BB'\nabla p_{2e}' - \gamma^2\tfrac{1}{2}|C|^2. \tag{90}$$

Now this must satisfy $p_{2e} \le 0$, $p_{2e}(0) = 0$, and

$$-(A - \gamma^{-2}BB'\nabla p_{2e}') \quad \text{asymptotically stable}, \tag{91}$$

and so we have, by uniqueness,

$$p_{2e} = \gamma^2 V_-. \tag{92}$$

Significantly, this says that the control attractor for the present H_∞ problem can be obtained from the antistable solution to an H_2 equation (66), which is straightforward to find since (66) is sign-definite and independent of γ (cf. [12], [1]).

The optimal information state controller $\tilde{K}_2^* : \tilde{y}_2 \mapsto u_2$ is given by (from [14, sec. 4.1] and (201) below)

$$\tilde{K}_2^* \;:\; \begin{cases} \dot{p}_2 = \tilde{F}_2(p_2, u_2, \tilde{y}_2), \quad p_{20} = \gamma^2 V_-, \\ u_2 = -\nabla W_2(p_2)[B'\nabla p_2']. \end{cases} \tag{93}$$

By a simple scaling, we obtain the desired controller $K_2^* : y_2 \mapsto u_2$ for $\mathcal{P}_2$ by

$$K_2^* = \tilde{K}_2^*\beta,$$

given by

$$K_2^* \;:\; \begin{cases} \dot{p}_2 = F_2(p_2, u_2, y_2), \quad p_{20} = \gamma^2 V_-, \\ u_2 = -\nabla W_2(p_2)[B'\nabla p_2'], \end{cases} \tag{94}$$

where

$$
\begin{aligned}
F_2(p_2,u_2,y_2) &= \tilde{F}_2(p_2,u_2,\beta y_2)\\
&= -\nabla p_2(A-Bu_2)+\gamma^{-2}\tfrac{1}{2}\nabla p_2 BB'\nabla p_2' \\
&\quad +\tfrac{1}{2}|u_2|^2+\tfrac{1}{2}|y_2|^2-\gamma^2\tfrac{1}{2}|y_2+C|^2.
\end{aligned}
\tag{95}
$$

In this form, it can be seen that $p_{2t}(x)$ has the representation

$$
\begin{aligned}
p_{2t}(x) = \sup_{v(\cdot)}\Big\{ & p_{20}(\xi(0))+\tfrac{1}{2}\int_0^t [|u_2(s)|^2+|y_2(s)|^2]ds \\
& -\gamma^2\tfrac{1}{2}\textstyle\int_0^t[|v(s)|^2+|y_2(s)+C(\xi(s))|^2]ds, \\
& \dot{\xi}=A(\xi)+B(\xi)(u_2+v),\ 0\le s\le t, \\
& \xi(t)=x\Big\}
\end{aligned}
\tag{96}
$$

and as a consequence is finite for all $t\ge 0$ and all $x\in\mathbf{R}^n$ for any finite $p_{20}\le 0$ and any pair $(u_2,y_2)\in\mathcal{W}$; this is again a consequence of the special structure of the present H_∞ problem.

The HJBI equation (88) can be written as

$$
\inf_{u_2\in\mathbf{R}^m}\sup_{y_2\in\mathbf{R}^p}\{\nabla W_2(p_2)[F_2(p_2,u_2,y_2)]\}=\inf_{u_2\in\mathbf{R}^m}\sup_{\tilde{y}_2\in\mathbf{R}^p}\left\{\nabla W_2(p_2)[\tilde{F}_2(p_2,u_2,\tilde{y}_2)]\right\}=0,
$$

or more explicitly, as

$$
\begin{aligned}
&\nabla W_2(p_2)[-\nabla p_2 A+\gamma^{-2}\tfrac{1}{2}\nabla p_2 BB'\nabla p_2'-\gamma^2\tfrac{1}{2}|C|^2] \\
&\quad -\tfrac{1}{2}|\nabla W_2(p_2)[\nabla p_2 B]|^2+\gamma^2\beta^{-2}|\nabla W_2(p_2)[C]|^2=0,
\end{aligned}
\tag{97}
$$

which is an infinite-dimensional nonlinear Riccati equation.

Of interest is the smallest value of $\gamma>1$ for which there exists a controller K_2 with the objective (84) satisfied. Let

$$
\gamma_2^*=\inf\left\{\gamma>1\ :\ p_{2e}\in\{p\in\mathcal{X}\ :\ \exists\varepsilon>0 \text{ with } p+\varepsilon|\cdot|^2\in \mathrm{dom}W_2\}\right\}. \tag{98}
$$

Then, for $\gamma>\gamma_2^*$, under the hypotheses of the sufficiency theorems of [14, Chapter 4], the controller K_2^* achieves the H_∞ norm objective (84). One of the key hypotheses concerning the so-called "quadratic upper limiting" property of the control attractor p_{2e} follows from a strengthened form of the antistability condition for V_- (68) (e.g., "incremental L_2 exponential antistability") and negative definiteness: $V_-(x)\le -c_-|x|^2$ ($c_->0$). Indeed, under these circumstances it will always be the case that

$$
p_{2t}\in\{p\in\mathcal{X}\ :\ \exists\varepsilon>0 \text{ with } p+\varepsilon|\cdot|^2\in\mathrm{dom}W_2\} \tag{99}
$$

whenever $(u_2,y_2)\in\mathcal{W}\cap L_2[0,\infty)$ (this is a coupling condition).

The state feedback HJBI equation for the state feedback H_∞ problem specified by $\mathcal{P}_2$ is (by [14, eq. (5.45)] and (193) below)

$$
\nabla V_2 A-\tfrac{1}{2}\beta^2\nabla V_2 BB'\nabla V_2'+\tfrac{1}{2}\beta^{-2}|C|^2=0 \tag{100}
$$

with $V_2\ge 0$, $V_2(0)=0$, and

$$
(A-\beta^2 BB'\nabla V_2') \quad \text{asymptotically stable.} \tag{101}
$$

Then by uniqueness,

$$V_2 = \beta^{-2} V_+. \tag{102}$$

Again, the present special structure leads to a considerable simplification since the state feedback H_∞ value function V_2 can be solved in terms of the stabilizing solution V_+ of the H_2 equation (66).

The certainty equivalence controller K_2^{ce} (under the standard certainty equivalence assumptions [4]) is given by

$$K_2^{ce}\colon \begin{cases} \dot{p}_2 = F_2(p_2, u_2, y_2), \quad p_{20} = \gamma^2 V_-, \\ u_2 = \beta^{-2} B(\bar{x})' \nabla V_+(\bar{x})', \end{cases} \tag{103}$$

where

$$\bar{x}(t) = \operatorname*{argmax}_{x \in \mathbf{R}^n} \{p_{2t}(x) + V_2(x)\}. \tag{104}$$

Remark 5.3. The results of this section are closely related to [1] and [25], which treated coprime factor uncertainty.

5.3. Synthesis for $\boldsymbol{\Pi_{\mathcal{M}\|\mathcal{N}}}$. For this second projection, given $\gamma > 0$, a controller K_1 can be found such that

$$\| \Pi_{\mathcal{M}\|\mathcal{N}_1} \|_\infty = \| (\mathcal{P}_1, K_1) \|_\infty < \gamma \tag{105}$$

using the controller K_2^* obtained above via conjugacy (see (47))

$$K_1^* = \beta^2 K_2^* \beta^{-2}. \tag{106}$$

However, it is instructive to do this directly, as we now show.

By Lemma 5.2, as above, it is enough to find a controller $\tilde{K}_1$ such that

$$\| (\tilde{\mathcal{P}}_1, \tilde{K}_1) \|_\infty < \gamma. \tag{107}$$

The information state equation for $p_{1t}(x) = p_1(x,t)$ is (from [14, eq. (3.9)] and (195) below)

$$\begin{aligned} \tfrac{\partial}{\partial t} p_1 &= -\nabla p_1 (A - \beta^{-1} B \tilde{u}_2) + \gamma^{-2} \beta^{-2} \tfrac{1}{2} \nabla p_2 B B' \nabla p_2' - \gamma^2 \beta^2 \tfrac{1}{2} |C|^2 \\ &\quad + \tfrac{1}{2} |\tilde{u}_2|^2 - \gamma^2 \tfrac{1}{2} |y_2|^2 - \gamma^2 C' y_2, \\ \dot{p}_1 &= \tilde{F}_1(p_1, \tilde{u}_2, y_2). \end{aligned} \tag{108}$$

The HJBI equation for the value function $W_1(p_1)$ is (from[14, eq. (4.16)] and (199) below)

$$\inf_{\tilde{u}_2 \in \mathbf{R}^m} \sup_{y_2 \in \mathbf{R}^p} \left\{ \nabla W_1(p_1) [\tilde{F}_1(p_1, \tilde{u}_2, y_2)] \right\} = 0, \tag{109}$$

from which the optimal information state feedback function is (by [14, eq. (4.23)] and (200) below)

$$\tilde{u}_2^*(p_1) = -\beta^{-1} \nabla W_1(p_1) [B' \nabla p_1']. \tag{110}$$

The stationary information state for the control attractor p_{1e} is (from [14, eq. (3.28)] and (197) below)

$$0 = -\nabla p_{1e} A + \gamma^{-2} \beta^{-2} \tfrac{1}{2} \nabla p_{1e} B B' \nabla p_{1e}' - \gamma^2 \beta^2 \tfrac{1}{2} |C|^2. \tag{111}$$

Now $p_{1e} \leq 0$, $p_{1e}(0) = 0$, and

$$-(A - \gamma^{-2}\beta^{-2}BB'\nabla p_{1e}') \quad \text{asymptotically stable,} \tag{112}$$

and so we have, by uniqueness,

$$p_{1e} = \gamma^2\beta^2 V_-. \tag{113}$$

The optimal information state controller $\tilde{K}_1^* : y_2 \mapsto \tilde{u}_2$ is given by (from [14, sec. 4.1] and (201) below)

$$\tilde{K}_1^* \ : \ \begin{cases} \dot{p}_1 = \tilde{F}_1(p_1, \tilde{u}_2, y_2), \quad p_{10} = \gamma^2\beta^2 V_-, \\ \tilde{u}_2 = -\beta^{-1}\nabla W_1(p_1)[B'\nabla p_1']. \end{cases} \tag{114}$$

The state feedback HJBI equation is (by [14, eq. (5.45)] and (193) below)

$$\nabla V_1 A - \tfrac{1}{2}\nabla V_1 BB'\nabla V_1' + \tfrac{1}{2}|C|^2 = 0 \tag{115}$$

with $V_1 \geq 0$, $V_1(0) = 0$, and

$$(A - BB'\nabla V_1') \quad \text{asymptotically stable.} \tag{116}$$

Then by uniqueness,

$$V_1 = V_+. \tag{117}$$

The controller $K_1^* : y_2 \mapsto u_2$ for $\mathcal{P}_1$ is given by

$$K_1^* = \beta\tilde{K}_1^*$$

so that

$$K_1^*: \ \begin{cases} \dot{p}_1 = F_1(p_1, u_2, y_2), \quad p_{10} = \gamma^2\beta^2 V_-, \\ u_2 = -\nabla W_1(p_1)[B'\nabla p_1'], \end{cases} \tag{118}$$

where

$$F_1(p_1, u_2, y_2) = \tilde{F}_1(p_1, \beta^{-1}u_2, y_2). \tag{119}$$

Remark 5.4. It can be checked that (106) (β-conjugacy) holds for the controllers K_1^* and K_2^* defined by (118) and (94), respectively.

Let

$$\gamma_1^* = \inf\left\{\gamma > 1 \ : \ p_{1e} \in \{p \in \mathcal{X} \ : \ \exists \varepsilon > 0 \text{ with } p + \varepsilon|\cdot|^2 \in \mathrm{dom} W_1\}\right\}. \tag{120}$$

Then as above we have

$$\gamma_1^* \geq (1 - \bar{\beta}^2)^{-1/2} \tag{121}$$

as in [1, Theorem 4.3].

5.4. A local controller. In this subsection we construct a local solution to the $\Pi_{\mathcal{N}\|\mathcal{M}}$ H_∞ problem using the results of [15] (see also [16], [2], [26], etc.). As in section 5.2, we consider the problem of finding $\tilde{K}_2$ such that (85) holds, in a local sense.

The controller can be obtained from two solutions to the H_2 PDE (66) or slightly more generally, from partial differential inequalities (PDIs) related to (66). Assume as follows.

1. There exists a C^3 positive definite function V_+ defined in a neighborhood of 0 and vanishing at 0 which satisfies
$$\nabla V_+ A - \tfrac{1}{2}\nabla V_+ BB'\nabla V_+' + \tfrac{1}{2}|C|^2 \le 0 \tag{122}$$
in the neighborhood.
2. There exists a C^3 negative definite function V_- defined in a neighborhood of 0 and vanishing at 0 which satisfies
$$\nabla V_- A - \tfrac{1}{2}\nabla V_- BB'\nabla V_-' + \tfrac{1}{2}|C|^2 \le 0 \tag{123}$$
in the neighborhood, and in addition
$$\begin{aligned}\nabla^2\{&-\nabla r_{2e}(A+\gamma^{-2}\beta^{-2}BB'\nabla V_+) + \gamma^{-2}\tfrac{1}{2}\nabla r_e BB'\nabla r_e' \\ &+\tfrac{1}{2}\beta^{-4}\nabla V_+ BB'\nabla V_+' - \gamma^2\tfrac{1}{2}|C|^2\} < 0\end{aligned} \tag{124}$$
in the neighborhood, where
$$r_{2e} = \gamma^2 V_- + \beta^{-2}V_+.$$
3. There exists a C^2 function G_2 such that
$$\nabla r_{2e}(x)G_2(x) = \gamma^2\beta^{-1}C(x)' \tag{125}$$
in a neighborhood of 0.

Consider the controller
$$\tilde{K}_2^\dagger\colon \left\{\begin{array}{l}\dot{\xi} = A(\xi) - B(\xi)B(\xi)'\nabla V_+(\xi)' + G_2(\xi)(\tilde{y} + \beta^{-1}C(\xi)),\\ u_2 = \beta^{-2}B(\xi)'\nabla V_+(\xi)'\end{array}\right. \tag{126}$$
defined locally in a neighborhood of 0 (the intersection defined by the assumptions).

Then by [15, Theorem 3.1], the controller $\tilde{K}_2^\dagger$ achieves local disturbance attenuation (dissipation) with stability. This means that the function
$$U(x,\xi) = \beta^{-2}V_+(x) - \gamma^2 V_-(x-\xi) - \beta^{-2}V_+(x-\xi)$$
is a storage function for the closed-loop system in a neighborhood of 0 and is a local Lyapunov function; in particular, for trajectories $(x(\cdot),\xi(\cdot))$ of $(\tilde{\mathcal{P}}_2, \tilde{K}_2^\dagger)$ staying near 0 one has
$$\begin{aligned}U(x(t),\xi(t)) + \tfrac{1}{2}\textstyle\int_0^t[|\hat{u}_1(s)|^2 + |\hat{y}_1(s)|^2]ds \\ \le U(x_0,\xi_0) + \tfrac{1}{2}\textstyle\int_0^t[|\hat{u}_0(s)|^2 + |\hat{y}_0(s)|^2]ds.\end{aligned} \tag{127}$$

Rescaling, one obtains a local controller $K_2^\dagger$ for $\mathcal{P}_2$, and via conjugation, one for $\mathcal{P}_1$.

6. State-space realizations of graph representations. The generalization of the well-known state-space formulae for normalized coprime factorizations of linear systems (see, e.g., [31] and the references therein) to nonlinear systems has been considered by a number of authors. In [22] and [27] the connection between existence of right representations and the existence of stabilizing feedbacks is explored. In [1], [24], [25], and [19], state-space realizations for image and kernel representations are considered, and definitions of normalized representations are proposed. In this section we rework the theory for the right representations according to our definitions and derive new realizations for normalized left representations.

Consider the state-space model (65) for the plant $P : \mathcal{U} \to \mathcal{Y}$. Using this model, we derive realizations of the right and left representations defined in section 3.

6.1. Right representations.

6.1.1. Preliminaries. Consider the finite horizon HJB:

(128)
$$\begin{aligned}\tfrac{\partial}{\partial t}V_T(t,x) + \nabla V_T(t,x)A(x)\\ -\tfrac{1}{2}\nabla V_T(t,x)B(x)B(x)'\nabla V_T(t,x)' + \tfrac{1}{2}|C(x)|^2 = 0 \ \text{ in } (-\infty,T)\times\mathbf{R}^n,\\ V_T(T,x) = 0 \ \text{ for } x\in\mathbf{R}^n.\end{aligned}$$

In general, this equation will have a unique viscosity solution, not necessarily smooth. However, for our purposes we make the following assumption.

Assumption 6.1. *There exists a unique smooth solution $V_T(t,x)$ to the HJB equation* (128), *with $\nabla V_T(t,x)$ globally Lipschitz in x.*

Define R_{tT}: $\mathcal{U}\cap L_2[t,T] \to \mathcal{W}\cap L_2[t,T]$ by

(129)
$$\mathsf{R}_{tT}\colon \begin{cases} \dot\xi(s) = A(\xi(s)) + B(\xi(s))(v_T^*(s)+\psi(s)), \quad t\le s\le T,\ \xi(t)=0,\\ \begin{bmatrix} u_T(s)\\ y_T(s)\end{bmatrix} = \begin{bmatrix} v_T^*(s)+\psi(s)\\ C(\xi(s))\end{bmatrix},\end{cases}$$

where

(130)
$$v_T^*(s) = -B(\xi(s))'\nabla V_T(s,\xi(s))'.$$

Lemma 6.2. *We have the following.*

1. *The solution of the HJB equation* (128) *is given by*

(131)
$$\begin{aligned}V_T(t,x) = \inf_{v\in\mathcal{U}}\{\tfrac{1}{2}\textstyle\int_t^T [|v(s)|^2 + |C(\xi(s))|^2]\,ds\\ \mid \dot\xi(s) = A(\xi(s)) + B(\xi(s))v(s), \quad t\le s\le T,\\ \xi(t) = x\}.\end{aligned}$$

2. *$V_T(t,x)\ge 0$ for all x and $V_T(t,0) = 0$ for all $t\ge 0$.*
3. *For $\psi\in L_2[t,T]$, the system R_{tT} defined by (129) satisfies the inner property*

(132)
$$\tfrac{1}{2}\int_t^T [|u_T(s)|^2 + |y_T(s)|^2]\,ds = V_T(t,x) + \tfrac{1}{2}\int_t^T |\psi(s)|^2\,ds.$$

Proof. The proof is based on a standard verification argument from optimal control (completion of squares). For any $v \in \mathcal{U}$, let $\xi(\cdot)$ denote the solution of

$$\dot{\xi}(s) = A(\xi(s)) + B(\xi(s))v(s),\ t \le s \le T,\ \xi(t) = x.$$

Then

$$\begin{aligned}
\tfrac{d}{ds}V_T(s,\xi(s)) &= \tfrac{\partial}{\partial s}V_T(s,\xi(s)) + \nabla V_T(s,\xi(s))[A(\xi(s)) + B(\xi(s))v(s)] \\
&= \tfrac{\partial}{\partial s}V_T(s,\xi(s)) + \nabla V_T(s,\xi(s))[A(\xi(s)) + B(\xi(s))v_T^*(s) \\
&\qquad + B(\xi(s))(v(s) - v_T^*(s))] \\
&= \tfrac{\partial}{\partial s}V_T(s,\xi(s)) + \nabla V_T(s,\xi(s))A(\xi(s)) \\
&\qquad - \tfrac{1}{2}\nabla V_T(s,\xi(s))B(\xi(s))B(\xi(s))'\nabla V_T(s,\xi(s))' \\
&\qquad - \tfrac{1}{2}|v_T^*(s)|^2 + \nabla V_T(s,\xi(s))B(\xi(s))(v(s) - v_T^*(s)) \\
&= \tfrac{1}{2}|v(s) - v_T^*(s)|^2 - \tfrac{1}{2}|v(s)|^2 - \tfrac{1}{2}|C(\xi(s))|^2,
\end{aligned}$$

and on integrating from t to T we get

$$\begin{aligned}
V_T(t,x) &= \tfrac{1}{2}\textstyle\int_t^T [|v(s)|^2 + |C(\xi(s))|^2]ds - \tfrac{1}{2}\int_t^T |v(s) - v_T^*(s)|^2 ds. \\
&\le \tfrac{1}{2}\textstyle\int_t^T [|v(s)|^2 + |C(\xi(s))|^2]ds.
\end{aligned} \tag{133}$$

For $v(s) = v_T^*(s)$ this gives the optimal value

$$V_T(t,x) = \tfrac{1}{2}\int_t^T [|v_T^*(s)|^2 + |C(\xi(s))|^2]ds.$$

This proves 1.

From the representation formula (131) it is clear that $V_T(t,x) \ge 0$. If $x = 0$ and $v(s) = 0$ on $[t,T]$, then the corresponding trajectory $\xi(s) = 0$ and $C(\xi(s)) = 0$ on $[t,T]$, and so $V_T(t,0) = 0$. This proves 2.

Now let $\psi \in \mathcal{U} \cap L_2[t,T]$, and define v by

$$v(s) = v_T^*(s) + \psi(s) \ \text{ on } [t,T].$$

Then $v \in \mathcal{U} \cap L_2[t,T]$, and so item 3 follows from the first line of (133). ∎

Consider next the infinite horizon HJB equation

$$\nabla V(x)A(x) - \tfrac{1}{2}\nabla V(x)B(x)B(x)'\nabla V(x)' + \tfrac{1}{2}|C(x)|^2 = 0 \ \text{ in } \mathbf{R}^n. \tag{134}$$

Assumption 6.3. *There exists a unique smooth solution to the HJB equation* (134), *with* $\nabla V(x)$ *globally Lipschitz, such that* $V \ge 0$, $V(0) = 0$, *and* V *stabilizing:*

$$(A - BB'\nabla V') \ \textit{ asymptotically stable.} \tag{135}$$

That is, for any $v \in \mathcal{U} \cap L_2[0,\infty)$, *if* ξ *is the solution on* $[0,\infty)$ *of*

$$\dot{\xi} = A(\xi) - B(\xi)B(\xi)\nabla V(\xi)' + B(\xi)v, \quad \xi(0) = x, \tag{136}$$

then $\xi \in L_2[0,\infty)$ *and* $\xi(t) \to 0$ *as* $t \to \infty$.

Define $\mathsf{R}_{[0,\infty)} : \mathcal{U} \cap L_{2,e} \to \mathcal{W} \cap L_{2,e}$ by

(137)
$$\mathsf{R}_{[0,\infty)}\colon \left\{ \begin{array}{c} \dot{\xi}(s) = A(\xi(s)) + B(\xi(s))(v^*(s) + \psi(s)), \quad 0 \le s < \infty,\ \xi(0) = 0, \\ \left[\begin{array}{c} u(s) \\ y(s) \end{array} \right] = \left[\begin{array}{c} v^*(s) + \psi(s) \\ C(\xi(s)) \end{array} \right], \end{array} \right.$$

where

$$v^*(s) = -B(\xi(s))' \nabla V(\xi(s))'. \tag{138}$$

LEMMA 6.4. *We have the following.*

1. *The solution of the HJB equation* (134) *is given by*

$$\begin{array}{l} V(x) = \inf_{v \in \mathcal{U} \cap L_2} \{ \frac{1}{2} \int_0^\infty [|v(s)|^2 + |C(\xi(s))|^2]\, ds \\ \qquad\qquad \mid \dot{\xi}(s) = A(\xi(s)) + B(\xi(s))v(s), \quad 0 \le s < \infty, \\ \qquad\qquad\quad \xi(0) = x, \ \textit{and}\ \xi \in L_2 \}. \end{array} \tag{139}$$

2. *For $\psi \in L_{2,e}$ and all $t \ge 0$ we have*

$$V(\xi(t)) + \tfrac{1}{2} \int_0^t [|u(s)|^2 + |y(s)|^2]\, ds = V(x) + \tfrac{1}{2} \int_0^t |\psi(s)|^2\, ds \tag{140}$$

for the system $\mathsf{R}_{[0,\infty)}$.

3. *For $\psi \in L_2$ we have*

$$\tfrac{1}{2} \int_0^\infty [|u(s)|^2 + |y(s)|^2]\, ds = V(x) + \tfrac{1}{2} \int_0^\infty |\psi(s)|^2\, ds. \tag{141}$$

Proof. The proof is similar to that of Lemma 6.2 and we indicate only the main changes. For any $v \in \mathcal{U} \cap L_{2,e}$, let $\xi(\cdot)$ denote the solution of

$$\dot{\xi}(s) = A(\xi(s)) + B(\xi(s))v(s), \ s \ge 0, \ \xi(0) = x.$$

Then

$$\frac{d}{ds} V(\xi(s)) = \tfrac{1}{2}|v(s) - v^*(s)|^2 - \tfrac{1}{2}|v(s)|^2 - \tfrac{1}{2}|C(\xi(s))|^2,$$

and on integrating from 0 to t we have

$$V(x) = V(\xi(t)) + \tfrac{1}{2} \int_0^t [|v(s)|^2 + |C(\xi(s))|^2] ds - \tfrac{1}{2} \int_0^t |v(s) - v_T^*(s)|^2 ds. \tag{142}$$

Now if $v \in \mathcal{U} \cap L_2$ is such that the trajectory $\xi(\cdot) \in L_2$ and $\xi(t) \to 0$ as $t \to \infty$, then $V(\xi(t)) \to 0$ as $t \to \infty$ and the integrands in (142) are integrable on $[0, \infty)$; hence

$$\begin{array}{l} V(x) = \frac{1}{2} \int_0^\infty [|v(s)|^2 + |C(\xi(s))|^2] ds - \frac{1}{2} \int_0^\infty |v(s) - v_T^*(s)|^2 ds \\ \qquad \le \frac{1}{2} \int_0^\infty [|v(s)|^2 + |C(\xi(s))|^2] ds. \end{array} \tag{143}$$

The stability property in Assumption 6.3 implies that $v = v^*$ is a valid (admissible) control in the representation (139), and (143) implies

$$V(x) = \tfrac{1}{2} \int_0^\infty [|v^*(s)|^2 + |C(\xi(s))|^2] ds,$$

and so v^* is optimal. This proves item 1.

Let $\psi \in \mathcal{U} \cap L_{2,e}$ and define

$$v(s) = v^*(s) + \psi(s) \quad \text{on } [0, \infty).$$

Then $v \in \mathcal{U} \cap L_{2,e}$, and item 2 follows from (142).

If $\psi \in \mathcal{U} \cap L_2$, Assumption 6.3 and (143) imply item 3. ∎

6.1.2. Finite horizon right representation: $\mathsf{R}_{T,P}$. Define $\mathsf{R}_{T,P} : \mathcal{U} \to \mathcal{W} \cap L_2(-\infty, T]$ by

(144)

$$\mathsf{R}_{T,P}\colon \left\{ \begin{array}{c} \dot{\xi}(s) = A(\xi(s)) + B(\xi(s))(v_T^*(s) + \psi(s)), \quad s \le T,\ \xi(-\infty) = 0, \\ \left[\begin{array}{c} u_T(s) \\ y_T(s) \end{array} \right] = \left[\begin{array}{c} v_T^*(s) + \psi(s) \\ C(\xi(s)) \end{array} \right], \end{array} \right.$$

where $v_T^*(s)$ is given by (130).

THEOREM 6.5. *The operator $\mathsf{R}_{T,P}$ defined by* (144) *is a finite horizon right representation of the plant P given by* (65) *and is contractive and finite time inner. A left inverse of $\mathsf{R}_{T,P}$ is given by*

$$\mathsf{R}_{T,P}^{-L}\colon \left\{ \begin{array}{l} \dot{\xi}(s) = A(\xi(s)) + \left[\begin{array}{cc} B(\xi(s)) & 0 \end{array} \right] \left[\begin{array}{c} u \\ y \end{array} \right], \quad s \le T,\ \xi(-\infty) = 0, \\ \psi(s) = u(s) - v_T^*(s), \end{array} \right. \tag{145}$$

where $v_T^*(s)$ is given by (130).

Proof. We must check that the operator $\mathsf{R}_{T,P}$ defined by (144) satisfies the definitions given in section 3.1, part 1.

Let $(u, y) \in \operatorname{range} \mathsf{R}_{T,P}|_{[T_0,T]}$. Then

$$u(s) = v_T^*(s) + \psi(s) \text{ and } y(s) = C(\xi(s))$$

for some $\psi \in \mathcal{U}$, where

$$\dot{\xi} = A(\xi) + B(\xi)u \text{ on } [T_0, T],\ \xi(T_0) = 0.$$

Then clearly $(u, y) \in \mathcal{G}_P \cap L_2[T_0, T]$, and so $\operatorname{range} \mathsf{R}_{T,P}|_{[T_0,T]} \subset \mathcal{G}_P \cap L_2[T_0, T]$. It can similarly be checked that $\mathcal{G}_P \cap L_2[T_0, T] \subset \operatorname{range} \mathsf{R}_{T,P}|_{[T_0,T]}$, proving that $\mathsf{R}_{T,P}$ is a finite horizon right representation.

Let $\psi \in \mathcal{U}$. Then $\psi(t) = 0$ for all $t \le T_0$, some $-\infty < T_0$. We suppose $T_0 < T$. Now item 3 of Lemma 6.2, i.e., (132), implies

$$\int_{T_0}^{T} |\mathsf{R}_{T,P}(\psi)(s)|^2 ds = \int_{T_0}^{T} |\psi(s)|^2 ds.$$

From this we see that $\mathsf{R}_{T,P}$ is finite time inner and contractive.

Next we check that $\mathsf{R}_{T,P}^{-L} : \mathcal{W} \to \mathcal{U} \cap L_2(-\infty, T]$ defined by (145) is a left inverse of $\mathsf{R}_{T,P}$. Let $\psi \in \mathcal{U} \cap L_2(-\infty, T]$, $\psi(t) = 0$ for all $t \le T_0$, some $-\infty < T_0 < T$, and $(u, y) = \mathsf{R}_{T,P}(\psi)$. Then

$$\dot{\xi} = A(\xi) + B(\xi)(v_T^* + \psi) \text{ on } [T_0, T],\ \xi(T_0) = 0.$$

Now $u = v_T^* + \psi$, $y = C(\xi)$ on $[T_0, T]$, and so this same trajectory satisfies

$$\dot{\xi} = A(\xi) + B(\xi)u \text{ on } [T_0, T],\ \xi(T_0) = 0.$$

Hence we must have

$$\psi = \mathsf{R}_{T,P}^{-L}(u, y),$$

showing that $\mathsf{R}_{T,P}^{-L}$ is a left inverse of $\mathsf{R}_{T,P}$. ∎

6.1.3. Right representation: R_{eP}. Define $\mathsf{R}_{eP} : \mathcal{U} \to \mathcal{W}$ by

$$\mathsf{R}_{eP}\colon \left\{ \begin{array}{rl} \dot{\xi}(s) & = A(\xi(s)) + B(\xi(s))(v^*(s) + \psi(s)),\ \xi(-\infty) = 0, \\ \left[\begin{array}{c} u(s) \\ y(s) \end{array} \right] & = \left[\begin{array}{c} v^*(s) + \psi(s) \\ C(\xi(s)) \end{array} \right], \end{array} \right. \tag{146}$$

where $v^*(s)$ is given by (138).

A state space realization (A, B, C) of a plant P is said to be *L_2-detectable* if $(u, y) \in \mathcal{G}_P \cap L_2$ implies that the corresponding state $\xi \in L_2$.

THEOREM 6.6. *The operator R_{eP} defined by* (146) *is a right representation of the plant P given by* (65) *(assumed L_2-detectable) and is contractive and inner. A left inverse of R_{eP} is given by*

(147)

$$\mathsf{R}_{eP}^{-L}\colon \left\{ \begin{array}{l} \dot{\xi}(s) = A(\xi(s)) + \left[\begin{array}{cc} B(\xi(s)) & 0 \end{array} \right] \left[\begin{array}{c} u \\ y \end{array} \right], \quad s \geq -\infty,\ \xi(-\infty) = 0, \\ \psi(s) = u(s) - v^*(s), \end{array} \right.$$

where $v^*(s)$ is given by (138). *The system R_{eP}^{-L} has asymptotically stable zero dynamics.*

Proof. We must check that the operator R_{eP} defined by (144) satisfies the definitions given in section 3.1, part 2.

Let $(u, y) \in \operatorname{range} \mathsf{R}_{eP}$. Then

$$u(s) = v^*(s) + \psi(s) \text{ and } y(s) = C(\xi(s))$$

for some $\psi \in \mathcal{U}$, where

$$\dot{\xi} = A(\xi) + B(\xi)u \text{ on } (-\infty, \infty),\ \xi(-\infty) = 0.$$

Then clearly $(u, y) \in \mathcal{G}_P$, and so $\operatorname{range} \mathsf{R}_{eP} \subset \mathcal{G}_P$. If $(u, y) \in \operatorname{range} \mathsf{R}_{eP}|_{L_2}$, then $\psi \in \mathcal{U} \cap L_2$. The proof is similar for the $L_2[0, \infty)$ case.

Now suppose $(u, y) \in \mathcal{G}_P$. Reversing the above argument, with $\psi = u - v^* \in \mathcal{U}$, we see that $(u, y) \in \operatorname{range} \mathsf{R}_{eP}$. If $(u, y) \in \mathcal{G}_P \cap L_2$, L_2-detectability ensures $\psi = u - v^* \in \mathcal{U} \cap L_2$. Similarly for the $L_2[0, \infty)$ case. This shows that R_{eP} is a right representation of P.

The contractive and inner properties of R_{eP} follow from Lemma 6.4, items 2 and 3, respectively.

Next we check that $\mathsf{R}_{eP}^{-L} : \mathcal{W} \to \mathcal{U}$ defined by (147) is a left inverse of R_{eP}. Let $\psi \in \mathcal{U}$ and $(u, y) = \mathsf{R}_{eP}(\psi)$. Then

$$\dot{\xi} = A(\xi) + B(\xi)(v^* + \psi) \text{ on } (-\infty, \infty),\ \xi(-\infty) = 0.$$

Now $u = v^* + \psi$, $y = C(\xi)$, and so this same trajectory satisfies

$$\dot{\xi} = A(\xi) + B(\xi)u \text{ on } (-\infty, \infty),\ \xi(-\infty) = 0.$$

Hence we must have

$$\psi = \mathsf{R}_{eP}^{-L}(u, y),$$

showing that R_{eP}^{-L} is a left inverse of R_{eP} on $\mathcal{U}$.

If $\psi \in \mathcal{U} \cap L_2$, Assumption 6.3 ensures $(u, y) = \mathsf{R}_{eP}(\psi) \in \mathcal{W} \cap L_2$. The argument just done for the case $\psi \in \mathcal{U}$ implies here that $\psi = \mathsf{R}_{eP}^{-L}(u, y)$. (We do not use or assert norm boundedness of R_{eP}^{-L} on $\mathcal{W}$.) So R_{eP}^{-L} is a left inverse of R_{eP} on $\mathcal{U} \cap L_2$. The case $\psi \in \mathcal{U} \cap L_2[0, \infty)$ is similar.

Finally, let $(u, y) \in \ker \mathsf{R}_{eP}^{-L}$ and consider the associated dynamics of (147). Then $\psi \equiv 0$, so $u(s) = v^*(s)$ for all $s > 0$. By Assumption 6.3, we see that the trajectory $\xi(\cdot)$ of (147) tends to zero asymptotically for any initial condition. This proves the stability of the zero dynamics. □

6.2. Left representations.

6.2.1. Preliminaries. Let $\mathcal{X}$ be the Banach space of real-valued continuous functions on $\mathbf{R}^n$ with at most quadratic growth with norm

$$\| p \|_{\mathcal{X}} = \sup_{x \neq 0} \frac{|p(x)|}{1 + |x|^2}.$$

Write

$$\langle p \rangle = \sup_{x \in \mathbf{R}^n} \{p(x)\}.$$

Let $p_0 \in \mathcal{X}$ be strictly negative definite:

$$p_0(x) < -c|x|^2 \ \forall x \in \mathbf{R}^n, \tag{148}$$

with $c = c[p_0] > 0$ and $p_0(0) = 0$.

Define a system evolving in $\mathcal{X}$ for $t \geq 0$ with state $p_t(x) = p(t, x)$ by

$$\mathsf{L}(0, p_0)\colon \begin{cases} \dot{p} = F(p, u, y), \quad p|_{t=0} = p_0, \\ \phi = C(\hat{x}) + \begin{bmatrix} 0 & -I \end{bmatrix} \begin{bmatrix} u \\ y \end{bmatrix}, \end{cases} \tag{149}$$

where $\dot{p} = \frac{\partial}{\partial t} p$, and $F(p, u, y)$ is the differential operator

$$F(p, u, y) = -\nabla p(A + Bu) + \tfrac{1}{2} \nabla p B B' \nabla p' - \tfrac{1}{2} |y - C|^2, \tag{150}$$

and

$$\hat{x} \in [[p]] \stackrel{\triangle}{=} \operatorname*{argmax}_{x \in \mathbf{R}^n} \{p(x)\} \tag{151}$$

is specified by

$$\tfrac{1}{2} |y - C(\hat{x})|^2 = \min_{x \in [[p]]} \{\tfrac{1}{2} |y - C(x)|^2\}. \tag{152}$$

We will also have occasion to use functions belonging to an extension $\mathcal{X}_e$ of the space $\mathcal{X}$. For instance,

$$\mathbf{\Delta}_{x_0}(x) = \begin{cases} 0 & \text{if } x = x_0, \\ -\infty & \text{if } x \neq x_0. \end{cases} \tag{153}$$

This function belongs to $\mathcal{X}_e$ and can be viewed as a limit $\mathbf{\Delta}_{x_0}(x) = \lim_{N \to \infty} -N|x - x_0|^2$. More generally, elements of $\mathcal{X}_e$ can be viewed as max-plus measures [14].

We make the following representation and regularity assumption concerning the PDE in the system (149).

Assumption 6.7. *Let the initial condition p_0 belong to $\mathcal{X}$ and satisfy* (148), $p_0(0) = 0$, *or let $p_0 = \mathbf{\Delta}_0$. For $(u, y) \in \mathcal{W}$ we assume as follows.*

1. *The HJB equation in* (149) *has a unique solution given by*

$$\begin{aligned} p_t(x) = \sup_{v \in \mathcal{U} \cap L_{2,e}} \{ & p_0(\xi(0)) - \tfrac{1}{2} \textstyle\int_0^t [|v(s)|^2 + |y(s) - C(\xi(s))|^2]\, ds \\ & \mid \dot{\xi}(s) = A(\xi(s)) + B(\xi(s))(u(s) + v(s)), \quad 0 \le s \le t, \\ & \xi(t) = x \}. \end{aligned} \tag{154}$$

2. *The map $t \mapsto \langle p_t \rangle$ is absolutely continuous, and*

$$\frac{d}{dt}\langle p_t \rangle = - \inf_{\hat{x} \in [[p_t]]} \{ \tfrac{1}{2} |y(t) - C(\hat{x})|^2 \} \quad a.e. \tag{155}$$

Remark 6.8. If the information state p_t were smooth, the representation would follow from classical optimal control, and the derivative (155) would follow from Danskin's Theorem [4, Theorem 10.1] (see also [14, Theorem 10.3.4]). To gain some heuristic understanding of (155), suppose the maximum $\hat{x}(t)$ in (151) is unique for each t, and then

$$\begin{aligned} \tfrac{d}{dt}\langle p_t \rangle &= \tfrac{\partial}{\partial t} p_t(\hat{x}(t)) + \nabla p_t(\hat{x}(t)) \dot{\hat{x}}(t) \\ &= \tfrac{\partial}{\partial t} p_t(\hat{x}(t)) \\ &= F(p_t, u(t), y(t))(\hat{x}(t)) \\ &= -\tfrac{1}{2} |y(t) - C(\hat{x}(t))|^2 \end{aligned}$$

using $\nabla p_t(\hat{x}(t)) = 0$ (because by definition $\hat{x}(t)$ maximizes $p_t(\cdot)$).

We also need an assumption concerning the existence of an equilibrium for the PDE in (149), and stability of this system:

Assumption 6.9. *There exists a unique smooth solution $p_e \in \mathcal{X}$ of the stationary HJB equation*

$$0 = F(p_e, 0, 0) \tag{156}$$

such that p_e is strictly negative definite ($p_e(x) \le -c_e |x|^2$, $c_e > 0$), $p_e(0) = 0$, and antistable (essentially $-(A - BB'\nabla p_e')$ asymptotically stable) in the following sense:

1. *For any $v \in L_2[0, \infty)$, if $\xi_t(\cdot)$ is the solution on $[0, t]$ of*

$$\dot{\xi}_t(s) = A(\xi_t(s)) - B(\xi_t(s))B(\xi_t(s))\nabla p_e(\xi_t(s))' + B(\xi_t(s))v(s), \quad 0 \le s \le t, \ \xi(t) = x, \tag{157}$$

then

$$\int_0^t |\xi_t(s)|^2 ds \le c_0 |x|^2 + c_1 \int_0^t |v(s)|^2 \, ds \quad \forall t \ge 0. \tag{158}$$

2. *For any $(u, y) \in \mathcal{W}$, the information state p_t given by* (149) *with $p_0 = p_e$ or $p_0 = \mathbf{\Delta}_0$ satisfies*

$$p_t \to p_e + c(u, y) \tag{159}$$

as $t \to \infty$, where $c(u, y) \in \mathbf{R}$ is a constant depending on (u, y).

3. *In addition, p_t is uniformly bounded above as follows (tight):*

$$p_t(x) < -c_2|x|^2 + c_3 \ \forall x \in \mathbf{R}^n, \tag{160}$$

with $c_2 = c_2(u,y) > 0$, $c_3 = c_3(u,y) \geq 0$.

4. *There exists $T(x) \geq 0$ and $k(x) \geq 0$ such that for any $w = (u,y) \in \mathcal{W} \cap L_2[0,\infty)$, the information state p_t given by (149) with $p_0 = p_e$ or $p_0 = \mathbf{\Delta}_0$ satisfies*

$$-k(x)(1+ \parallel w \parallel^2_{L_2[0,\infty)}) \leq p_t(x) \quad \forall t \geq T(x). \tag{161}$$

The function p_e is given by

$$\begin{aligned} p_e(x) &= \sup_{v\in\mathcal{U}}\{-\lim_{T\to\infty} \tfrac{1}{2}\int_{-T}^0 [|v(s)|^2 + |C(\xi(s))|^2]\,ds \\ &\qquad | \ \dot{\xi}(s) = A(\xi(s)) + B(\xi(s))v(s), \quad -T \leq s \leq 0, \\ &\qquad \lim_{T\to\infty} \xi(-T) = 0, \quad \xi(0) = x\} \\ &= \lim_{T\to\infty} \sup_{v\in\mathcal{U}\cap L_2[-T,0]}\{-\tfrac{1}{2}\int_{-T}^0 [|v(s)|^2 + |C(\xi(s))|^2]\,ds \\ &\qquad | \ \dot{\xi}(s) = A(\xi(s)) + B(\xi(s))v(s), \quad -T \leq s \leq 0, \\ &\qquad \xi(-T) = 0, \quad \xi(0) = x\}. \end{aligned} \tag{162}$$

Remark 6.10. Assumption 6.9 is essentially an H_2 filter assumption together with stability consequences [14, section 11.4], [24, section 4.2]. Item 4 also involves controllability.

LEMMA 6.11. *Given $(u,y) \in \mathcal{W} \cap L_{2,e}$,*

1. *the set $[[p_t]]$ is compact for each $t \geq 0$;*
2. *for each $t \geq 0$ we have*

$$\langle p_t \rangle + \tfrac{1}{2}\int_0^t |\phi(s)|^2\,ds = \langle p_0 \rangle = 0; \tag{163}$$

3. *for some constant $c(u,y)$,*

$$\lim_{t\to\infty} \langle p_t \rangle = c(u,y) = -\tfrac{1}{2}\int_0^\infty |\phi(s)|^2\,ds, \tag{164}$$

whenever $(u,y) \in \mathcal{W} \cap L_2[0,\infty)$.

Proof. Item 1 follows from the upper bound (160) by Assumption 6.9. The equality (163) of item 2 follows by integrating the derivative (155) by Assumption 6.7, and the definition of ϕ, (149).

We turn now to item 3; this will follow from Assumption 6.9. Let $t > 0$ and combine the bound (161) with item 2, (163), to give

$$-k(x)(1+ \parallel w \parallel^2_{L_2[0,\infty)}) + \tfrac{1}{2}\int_0^t |\phi(s)|^2 ds \leq 0.$$

This shows that the monotone nondecreasing map $t \mapsto \frac{1}{2}\int_0^t |\phi(s)|^2 ds$ is bounded, and hence the following limits exist and are equal:

$$-\lim_{t\to\infty} \langle p_t \rangle = \tfrac{1}{2}\int_0^\infty |\phi(s)|^2 ds. \tag{165}$$

This proves item 3, (164). In fact, it follows from [14, Appendix C] that $\lim_{t\to\infty} \langle p_t \rangle = \langle p_e + c(u,y) \rangle = c(u,y)$. ∎

6.2.2. Finite horizon left representation: L_{tvP}. We set $p_0 = \mathbf{\Delta}_0$ to define $\mathsf{L}_{tvP} : \mathcal{W} \cap L_{2,e} \to \mathcal{Y} \cap L_{2,e}$ by

$$\mathsf{L}_{tvP}\colon \begin{cases} \dot{p} = F(p,u,y), \quad p|_{t=0} = \mathbf{\Delta}_0, \\ \phi = C(\hat{x}) + \begin{bmatrix} 0 & -I \end{bmatrix} \begin{bmatrix} u \\ y \end{bmatrix}. \end{cases} \tag{166}$$

THEOREM 6.12. *The system L_{tvP} defined by* (166) *is a finite time left representation for the plant P given by* (65) *and is contractive and co-inner.*

Proof. We must check that the operator L_{tvP} defined by (166) satisfies the definitions given in section 3.2, part 1.

Let $(u,y) \in \mathcal{G}_P \cap L_{2,e}$, so that $y = Pu$. Let ξ be the solution of

$$\dot{\xi} = A(\xi) + B(\xi)u, \quad \xi(0) = 0,$$

on $[0,\infty)$. Then $y = C(\xi)$ and by the representation (154), we have

$$p_t(\xi(t)) = 0 \quad \text{for all} \quad t \geq 0$$

using $v(\cdot) = 0$. This implies $\hat{x}(t) = \xi(t)$, $t \geq 0$, and $\phi(\cdot) = 0$. Hence $(u,y) \in \ker \mathsf{L}_{tvP}$.

Let $(u,y) \in \ker \mathsf{L}_{tvP}$. Then

$$\mathsf{L}_{tvP}(u,y) = \phi = 0,$$

and by (163) we have

$$p_t(\hat{x}(t)) = \langle p_t \rangle = 0 \tag{167}$$

for all $t \geq 0$. In the representation (154), let $v_t^*(\cdot)$ and $\xi_t^*(\cdot)$ be optimal on $[0,t]$ with end condition $x = \hat{x}(t)$, so that

$$\dot{\xi}_t^*(s) = A(\xi_t^*(s)) + B(\xi_t^*(s))(u(s) + v_t^*(s)), \quad 0 \leq s \leq t, \quad \xi_t^*(t) = \hat{x}(t).$$

Then (167) and (154) imply

$$0 = p_t(\hat{x}(t)) = p_0(\xi_t^*(0)) - \tfrac{1}{2}\int_0^t [|v_t^*(s)|^2 + |y(s) - C(\xi_t^*(s))|^2] ds,$$

which implies

$$\begin{aligned} p_0(\xi_t^*(0)) = 0, &\quad \text{which implies} \quad \xi_t^*(0) = 0; \\ \textstyle\int_0^t |v_t^*(s)|^2 ds = 0, &\quad \text{which implies} \quad v_t^*(\cdot) = 0; \\ \textstyle\int_0^t |y(s) - C(\xi_t^*(s))|^2 ds = 0, &\quad \text{which implies} \quad y(\cdot) = C(\xi_t^*(\cdot)). \end{aligned}$$

This implies $y = Pu$ on $[0,\infty)$ and so $(u,y) \in \mathcal{G}_P \cap L_{2,e}$. Consequently, L_{tvP} is a finite horizon left representation of P.

Next, let $w = (u,y) \in \mathcal{W} \cap L_{2,e}$. By (163), Lemma 6.11, we have

$$\tfrac{1}{2}\int_0^T |\phi(s)|^2 \, ds = -\langle p_T \rangle,$$

which implies, by (154) and the definition (153) of $p_o = \mathbf{\Delta}_0$,

$$\begin{aligned} \tfrac{1}{2}\textstyle\int_0^T |\mathsf{L}_{tvP}(w)(s)|^2 ds = \inf_{v \in \mathcal{U} \cap L_{2,e}} \{ & \tfrac{1}{2}\textstyle\int_0^T [|v(s)|^2 + |y(s) - C(\xi(s))|^2] \, ds \\ & \mid \dot{\xi}(s) = A(\xi(s)) + B(\xi(s))(u(s) + v(s)), \quad 0 \leq s \leq T, \\ & \xi(0) = 0\}. \end{aligned} \tag{168}$$

Let $\tilde{w} = (\tilde{u}, \tilde{y}) \in \mathcal{G}_P \cap L_{2,e}$, so that $\tilde{y} = P\tilde{u}$, with state $\tilde{\xi}$, and

$$\dot{\tilde{\xi}} = A(\tilde{\xi}) + B(\tilde{\xi})\tilde{u}, \quad \tilde{\xi}(0) = 0, \quad \tilde{y} = C(\tilde{\xi})$$

on $[0, \infty)$. Set $v = \tilde{u} - u$ in (168) to obtain

$$\tfrac{1}{2}\int_0^T |\mathsf{L}_{tvP}(w)(s)|^2 ds \le \tfrac{1}{2}\int_0^T \left[|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2\right] ds.$$

This holds for all $(\tilde{u}, \tilde{y}) \in \mathcal{G}_P \cap L_{2,e}$, and so

(169)

$$\tfrac{1}{2}\int_0^T |\mathsf{L}_{tvP}(w)(s)|^2 ds \le \inf_{(\tilde{u},\tilde{y})\in\mathcal{G}_P\cap L_{2,e}} \left\{ \tfrac{1}{2}\int_0^T \left[|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2\right] ds \right\}.$$

Setting $(\tilde{u}, \tilde{y}) = (0, 0)$ in (169) gives the contractive property (20).

Given any $v \in \mathcal{U} \cap L_{2,e}$, set $\tilde{u} = u + v$, $\tilde{y} = P\tilde{u}$, and we see from (168) that equality holds in (169):

(170)

$$\tfrac{1}{2}\int_0^T |\mathsf{L}_{tvP}(w)(s)|^2 ds = \inf_{(\tilde{u},\tilde{y})\in\mathcal{G}_P\cap L_{2,e}} \left\{ \tfrac{1}{2}\int_0^T \left[|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2\right] ds \right\};$$

this establishes (21) of the positive time co-inner property.

Now suppose $w = (u, y) \in \mathcal{G}_P \cap L_2[0, \infty)$. By Lemma 6.11, item 3, (164), we know that the limits as $T \to \infty$ of both sides of (170) exist and are equal. We need to show that this value is given by the right-hand side of (22), the second part of the positive time co-inner property.

Let $\tilde{w} = (\tilde{u}, \tilde{y}) \in \mathcal{W} \cap L_2[0, \infty)$. Then by (170) we have

$$\tfrac{1}{2}\int_0^T |\mathsf{L}_{tvP}(w)(s)|^2 ds \le \tfrac{1}{2}\int_0^\infty \left[|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2\right] ds$$

and hence

$$\tfrac{1}{2}\int_0^\infty |\mathsf{L}_{tvP}(w)(s)|^2 ds \le \tfrac{1}{2}\int_0^\infty \left[|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2\right] ds.$$

This implies

(171)

$$\tfrac{1}{2}\int_0^\infty |\mathsf{L}_{tvP}(w)(s)|^2 ds \le \inf_{(\tilde{u},\tilde{y})\in\mathcal{G}_P\cap L_2[0,\infty)} \left\{ \tfrac{1}{2}\int_0^\infty \left[|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2\right] ds \right\}.$$

Also, by (170) we have

$$\tfrac{1}{2}\int_0^\infty |\mathsf{L}_{tvP}(w)(s)|^2 ds \ge \inf_{(\tilde{u},\tilde{y})\in\mathcal{G}_P\cap L_{2,e}} \left\{ \tfrac{1}{2}\int_0^T \left[|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2\right] ds \right\}$$

for all $T \ge 0$. Let $\varepsilon > 0$. Then for each $T \ge 0$, select $\tilde{w}^T = (\tilde{u}^T, \tilde{y}^T) \in \mathcal{G}_P \cap L_2[0, T]$ such that

$$\tfrac{1}{2}\int_0^\infty |\mathsf{L}_{tvP}(w)(s)|^2 ds \ge \tfrac{1}{2}\int_0^T |w(s) - \tilde{w}^T(s)|^2 \, ds - \varepsilon.$$

Set $v^T = \tilde{u}^T - u$ on $[0,T]$, and $v^T = 0$ on (T,∞). Then we have

$$\tfrac{1}{2}\int_0^\infty |\mathsf{L}_{tvP}(w)(s)|^2 ds \geq \tfrac{1}{2}\int_0^T [|v^T(s)|^2 + |y(s) - \tilde{y}^T(s)|^2]\, ds - \varepsilon \tag{172}$$

and

$$\begin{aligned} \dot{\tilde{\xi}}^T &= A(\tilde{\xi}^T) + B(\tilde{\xi}^T)(u + v^T), \quad \tilde{\xi}^T(0) = 0, \\ \tilde{y}^T &= C(\xi^T) \end{aligned} \tag{173}$$

on $[0,T]$. Now inequality (172) implies that $\{v^T\}_{T>0}$ is bounded in $L_2[0,\infty)$. By weak compactness, let $\{v^{T_i}\}$ denote a weakly convergent subsequence, with weak limit $v^* \in L_2[0,\infty)$. Similarly, $\{\tilde{y}^T\}_{T>0}$ is bounded in $L_2[0,\infty)$, with weak limit $\tilde{y}^* \in L_2[0,\infty)$ (use the same index T_i for the convergent subsequence, without loss of generality).

Let $S > 0$ be arbitrary. Then for any $T \geq S$ inequality (172) implies

$$\tfrac{1}{2}\int_0^\infty |\mathsf{L}_{tvP}(w)(s)|^2 ds \geq \tfrac{1}{2}\int_0^S [|v^T(s)|^2 + |y(s) - \tilde{y}^T(s)|^2]\, ds - \varepsilon. \tag{174}$$

Let $T = T_i$ in (174) and send $i \to \infty$, to obtain

$$\tfrac{1}{2}\int_0^\infty |\mathsf{L}_{tvP}(w)(s)|^2 ds \geq \tfrac{1}{2}\int_0^S [|v^*(s)|^2 + |y(s) - \tilde{y}^*(s)|^2]\, ds - \varepsilon \tag{175}$$

by weak lower semicontinuity. Let $S \to \infty$ in (175) to obtain

$$\tfrac{1}{2}\int_0^\infty |\mathsf{L}_{tvP}(w)(s)|^2 ds \geq \tfrac{1}{2}\int_0^\infty [|v^*(s)|^2 + |y(s) - \tilde{y}^*(s)|^2]\, ds - \varepsilon. \tag{176}$$

Next set $\tilde{w}^* = (\tilde{u}^*, \tilde{y}^*) = (u + v^*, \tilde{y}^*) \in L_2[0,\infty)$, so that

$$\tfrac{1}{2}\int_0^\infty |\mathsf{L}_{tvP}(w)(s)|^2 ds \geq \tfrac{1}{2}\int_0^\infty |w(s) - \tilde{w}^*(s)|^2\, ds - \varepsilon. \tag{177}$$

It remains to check that $\tilde{w}^* \in \mathcal{G}_P$. Again let $S > 0$ be arbitrary and consider the system (173) on $[0,S]$ for $T = T_i > S$. Let $\tilde{\xi}^*$ be the solution of

$$\dot{\tilde{\xi}}^* = A(\tilde{\xi}^*) + B(\tilde{\xi}^*)(u + v^*) \tag{178}$$

on $[0,S]$. Then by standard ODE estimates (using the Holder and Jensen inequalities), for $t \in [0,S]$,

$$|\tilde{\xi}^{T_i}(t) - \tilde{\xi}^*(t)|^2 \leq K\int_0^t |\tilde{\xi}^{T_i}(s) - \tilde{\xi}^*(s)|^2 ds + K\left|\int_0^t B(\tilde{\xi}^*(s))(v^{T_i}(s) - v^*(s))ds\right|^2 \tag{179}$$

for a suitable constant $K > 0$. Note that by weak convergence, the second term in the right-hand side of (179) tends to zero as $i \to \infty$ (since $s \mapsto B(\tilde{\xi}^*(s)) \in L_2[0,\infty)$ by our hypotheses on B). Then an application of Gronwall's inequality allows us to conclude that $\tilde{\xi}^{T_i}(t) \to \tilde{\xi}^*(t)$ as $i \to \infty$ uniformly in $t \in [0,S]$. Moreover, we have $\tilde{y}^*(t) = C(\tilde{\xi}^*(t))$ for $t \in [0,S]$, and hence $\tilde{w}^* \in \mathcal{G}_P \cap L_2[0,S]$. But $S > 0$ was arbitrary, and we saw earlier that $\tilde{w}^* \in L_2[0,\infty)$, so we have shown that $\tilde{w}^* \in \mathcal{G}_P \cap L_2[0,\infty)$.

Therefore

$$\tfrac{1}{2}\int_0^\infty |\mathsf{L}_{tvP}(w)(s)|^2 ds \geq \inf_{(\tilde{u},\tilde{y})\in\mathcal{G}_P\cap L_2[0,\infty)} \left\{\tfrac{1}{2}\int_0^\infty |w(s) - \tilde{w}^*(s)|^2\, ds\right\} - \varepsilon. \tag{180}$$

Inequalities (171) and (180) now imply (22). ∎

6.2.3. Left representation: L_{eP}. We set $p_0 = p_e$ to define $\mathsf{L}_{eP} : \mathcal{W} \to \mathcal{Y}$ by

$$\mathsf{L}_{eP}\colon \begin{cases} \dot{p} = F(p,u,y), \quad p|_{t=-\infty} = p_e, \\ \phi = C(\hat{x}) + \begin{bmatrix} 0 & -I \end{bmatrix} \begin{bmatrix} u \\ y \end{bmatrix}. \end{cases} \tag{181}$$

THEOREM 6.13. *The system L_{eP} defined by* (181) *is a left representation of the plant P given by* (65) *and is contractive and co-inner.*

Proof. We must check that the operator L_{eP} defined by (166) satisfies the definitions given in section 3.2, part 2.

Let $(u,y) \in \mathcal{G}_P$, so that $y = Pu$. Since $u \in L_{2,ce}$, there exists T_0 finite such that $u(t) = 0$, $y(t) = 0$, and $p_t = p_e$ for all $t \le T_0$. Let ξ be the solution of

$$\dot{\xi} = A(\xi) + B(\xi)u, \quad \xi(T_0) = 0,$$

on $[T_0, \infty)$. Then $y = C(\xi)$ and by the representation (154), which here takes the form

$$\begin{aligned} p_t(x) = \sup_{v \in \mathcal{U} \cap L_2[T_0,t]} \{ & p_e(\xi(T_0)) - \tfrac{1}{2}\textstyle\int_{T_0}^{t} [|v(s)|^2 + |y(s) - C(\xi(s))|^2]\, ds \\ & \mid \dot{\xi}(s) = A(\xi(s)) + B(\xi(s))(u(s) + v(s)), \quad T_0 \le s \le t, \\ & \xi(t) = x \}, \end{aligned} \tag{182}$$

we have

$$p_t(\xi(t)) = 0 \quad \text{for all} \quad t \ge T_0$$

using $v(\cdot) = 0$. This implies $\hat{x}(t) = \xi(t)$, $t \ge T_0$, and $\phi(\cdot) = 0$. Hence $(u,y) \in \ker \mathsf{L}_{eP}$.

Let $(u,y) \in \ker \mathsf{L}_{eP}$. Since $u \in L_{2,ce}$, there exists T_0 finite such that $u(t) = 0$, $y(t) = 0$, and $p_t = p_e$ for all $t \le T_0$. Then

$$\mathsf{L}_{eP}(u,y) = \phi = 0,$$

and by (163) (with 0 replaced by T_0) we have

$$p_t(\hat{x}(t)) = \langle p_t \rangle = 0 \tag{183}$$

for all $t \ge T_0$. In the representation (182), let $v_t^*(\cdot)$ and $\xi_t^*(\cdot)$ be optimal on $[T_0, t]$ with end condition $x = \hat{x}(t)$, so that

$$\dot{\xi}_t^*(s) = A(\xi_t^*(s)) + B(\xi_t^*(s))(u(s) + v_t^*(s)), \quad T_0 \le s \le t, \quad \xi_t^*(t) = \hat{x}(t).$$

Then (183) and (182) imply

$$0 = p_t(\hat{x}(t)) = p_e(\xi_t^*(T_0)) - \tfrac{1}{2}\int_{T_0}^{t} [|v_t^*(s)|^2 + |y(s) - C(\xi_t^*(s))|^2] ds,$$

which implies

$$\begin{aligned} p_e(\xi_t^*(T_0)) = 0, \quad & \text{which implies} \quad \xi_t^*(T_0) = 0; \\ \textstyle\int_{T_0}^{t} |v_t^*(s)|^2 ds = 0, \quad & \text{which implies} \quad v_t^*(\cdot) = 0; \\ \textstyle\int_{T_0}^{t} |y(s) - C(\xi_t^*(s))|^2 ds = 0, \quad & \text{which implies} \quad y(\cdot) = C(\xi_t^*(\cdot)). \end{aligned}$$

This implies $y = Pu$ on $[T_0, \infty)$ and so $(u,y) \in \mathcal{G}_P$. Consequently, L_{eP} is a left representation of P.

Next, let $w = (u, y) \in \mathcal{W}$. Since $w \in L_{2,ce}$, there exists T_0 finite such that $u(t) = 0$, $y(t) = 0$ and $p_t = p_e$, $\phi(t) = 0$ for all $t \le T_0$. By (163), Lemma 6.11 (with 0 replaced by T_0, and $p_{T_0} = p_e$), we have

$$\tfrac{1}{2}\int_{T_0}^{T} |\phi(s)|^2\, ds = -\langle p_T \rangle,$$

which implies, by (182),

(184)

$$\tfrac{1}{2}\textstyle\int_{T_0}^{T} |\mathsf{L}_{eP}(w)(s)|^2 ds = \inf_{v\in\mathcal{U}\cap L_2[T_0,T]}\{-p_e(\xi(T_0)) + \tfrac{1}{2}\int_{T_0}^{T}[|v(s)|^2 + |y(s) - C(\xi(s))|^2]\, ds \\ |\ \dot{\xi}(s) = A(\xi(s)) + B(\xi(s))(u(s) + v(s)),\quad T_0 \le s \le T\}$$

(the infimum is over all possible trajectories ξ and controls v, with no end point constraint). Let $\tilde{w} = (\tilde{u}, \tilde{y}) \in \mathcal{G}_P$, so that $\tilde{y} = P\tilde{u}$. Since $\tilde{w} \in L_{2,ce}$, there exists T_1 finite such that $\tilde{u}(t) = 0$, $\tilde{y}(t) = 0$ for all $t \le T_1$. Let $T_2 = \min(T_1, T_0)$. Now (184) holds with T_0 replaced by T_2:

(185)

$$\tfrac{1}{2}\textstyle\int_{T_2}^{T} |\mathsf{L}_{eP}(w)(s)|^2 ds = \inf_{v\in\mathcal{U}\cap L_2[T_0,T]}\{-p_e(\xi(T_2)) + \tfrac{1}{2}\int_{T_2}^{T}[|v(s)|^2 + |y(s) - C(\xi(s))|^2]\, ds \\ |\ \dot{\xi}(s) = A(\xi(s)) + B(\xi(s))(u(s) + v(s)),\quad T_2 \le s \le T\}.$$

The state $\tilde{\xi}$ corresponding to $\tilde{w}$ satisfies

$$\dot{\tilde{\xi}} = A(\tilde{\xi}) + B(\tilde{\xi})\tilde{u},\quad \tilde{\xi}(T_2) = 0,\quad \tilde{y} = C(\tilde{\xi})$$

on $[T_2, \infty)$. Set $v = \tilde{u} - u$ in (185) to obtain

$$\tfrac{1}{2}\int_{T_2}^{T} |\mathsf{L}_{eP}(w)(s)|^2 ds \le \tfrac{1}{2}\int_{T_2}^{T} [|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2]\, ds$$

(since $p_e(\xi(T_2)) = 0$). This holds for all $(\tilde{u}, \tilde{y}) \in \mathcal{G}_P$, and so

(186)

$$\tfrac{1}{2}\int_{-\infty}^{T} |\mathsf{L}_{eP}(w)(s)|^2 ds \le \inf_{(\tilde{u},\tilde{y})\in\mathcal{G}_P}\left\{\tfrac{1}{2}\int_{-\infty}^{T}[|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2]\, ds\right\}.$$

Setting $(\tilde{u}, \tilde{y}) = (0, 0)$ in (186) gives the contractive property (26).

In order to prove the co-inner property (27), we must show that (186) holds with equality. To this end, let $\varepsilon > 0$. Let v_1 and ξ_1 be ε-optimal in (184), so that

$$\tfrac{1}{2}\int_{T_0}^{T} |\mathsf{L}_{eP}(w)(s)|^2 ds \ge -p_e(\xi_1(T_0)) + \tfrac{1}{2}\int_{T_0}^{T} [|v_1(s)|^2 + |y(s) - C(\xi_1(s))|^2]\, ds - \varepsilon,$$

where

$$\dot{\xi}_1(s) = A(\xi_1(s)) + B(\xi_1(s))(u(s) + v_1(s)),\quad T_0 \le s \le T.$$

Next, using the representation (162) for $p_e(x)$ with $x = \xi_1(T_0)$, select $S \ge 0$ and v_2 so that

$$-p_e(\xi_1(T_0)) \ge \tfrac{1}{2}\int_{-S}^{T_0} [|v_2(s)|^2 + |C(\xi_2(s))|^2]\, ds - \varepsilon,$$

where

$$\dot{\xi}_2(s) = A(\xi_2(s)) + B(\xi_2(s))v_2(s), \quad -S \le s \le T_0, \quad \xi_2(T_0) = \xi_1(T_0).$$

Set

$$v(s) = \begin{cases} v_1(s) & \text{on } [T_0, T], \\ v_2(s) & \text{on } [-S, T_0), \\ 0 & \text{on } (-\infty, -S), \end{cases}$$

with associated state trajectory ξ. Then combining the above, we see that

$$\begin{aligned} \textstyle\int_{T_0}^{T} |\mathsf{L}_{eP}(w)(s)|^2 ds &\ge \tfrac{1}{2} \textstyle\int_{-S}^{T} [|v(s)|^2 + |y(s) - C(\xi(s))|^2]\, ds - 2\varepsilon \\ &\ge \inf_{(\tilde{u},\tilde{y}) \in \mathcal{G}_P} \left\{ \tfrac{1}{2} \textstyle\int_{-\infty}^{T} [|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2]\, ds \right\} - 2\varepsilon \end{aligned}$$

with $\tilde{u} = u + v$, etc. This proves equality in (186),

(187)
$$\int_{-\infty}^{T} |\mathsf{L}_{eP}(w)(s)|^2 ds = \inf_{(\tilde{u},\tilde{y}) \in \mathcal{G}_P} \left\{ \tfrac{1}{2} \int_{-\infty}^{T} [|u(s) - \tilde{u}(s)|^2 + |y(s) - \tilde{y}(s)|^2]\, ds \right\},$$

and hence (27).

The second part (28) of the co-inner property can be proved using (184), (187), and the techniques used in the proof of Theorem 6.12. □

6.2.4. Special case. To relate our state-space realizations (166), (181) for the left representations with other results in the nonlinear literature [24, Theorem 13 and equation (40)], [25, Theorem 2.1 and equation (9)], we present the following discussion.

Consider the information state p_t, the state in the realizations (166), (181) with suitable initialization. Let us suppose that $\hat{x}(t)$, the maximizer of $p_t(\cdot)$, is unique. Then given appropriate differentiability, we can perform the following calculations. Since $\hat{x}(t)$ maximizes $p_t(\cdot)$, we have (as mentioned earlier)

$$\nabla p_t(\hat{x}(t)) = 0 \quad \text{for all} \quad t > 0.$$

Differentiate this with respect to t to obtain

$$\frac{d}{dt} p_t(\hat{x}(t)) = \frac{\partial}{\partial t} \nabla p_t(\hat{x}(t)) + \nabla^2 p_t(\hat{x}(t)) \dot{\hat{x}}(t) = 0.$$

Now we suppose that the Hessian is positive definite:

$$-\nabla^2 p_t(\hat{x}(t)) > 0$$

for all $t > 0$. Then

$$\dot{\hat{x}}(t) = [-\nabla^2 p_t(\hat{x}(t))]^{-1} \frac{\partial}{\partial t} \nabla p_t(\hat{x}(t)),$$

and taking the gradient of $\dot{p} = F(p, u, y)$ with respect to x and using the fact that $\nabla p_t(\hat{x}(t)) = 0$ we obtain

(188)
$$\begin{aligned} \dot{\hat{x}}(t) \;&= \big(A(\hat{x}(t)) - [-\nabla^2 p_t(\hat{x}(t))]^{-1} \nabla C(\hat{x}(t))' C(\hat{x}(t))\big) \\ &\quad + B(\hat{x}(t))u(t) + [-\nabla^2 p_t(\hat{x}(t))]^{-1} \nabla C(\hat{x}(t))' y(t). \end{aligned}$$

Associated with this equation is the error quantity

$$\phi(t) = C(\hat{x}(t)) - y(t) \tag{189}$$

from (166), (181). The pair of equations (188), (189), is similar to the state space realizations [24, equation (40)], [25, equation (9)] if $[-\nabla^2 p_t(\hat{x}(t))]^{-1}\nabla C(\hat{x}(t))'$ takes the place of $k(x)$ in [24, equation (40)] or if $[-\nabla^2 p_t(\hat{x}(t))]$ replaces $M(x)$ in [25, equation (9)]. We remark that in the linear case $p_t(x)$ is a quadratic form with quadratic term $-\frac{1}{2}x'Y_e^{-1}x$ and so $[-\nabla^2 p_t(\hat{x}(t))] = Y_e^{-1}$ and the state-space realizations (188) and (189) reduce to the well-known linear equations.

We remark also that the co-inner definition introduced in this paper (section 3.2) differs from the one used in [25]; there the co-inner property is defined in terms of a Hamiltonian system of equations [24, Definition 7 and equation (34)], and it appears that these equations are related to the Pontryagin equations associated with the optimal control representation of $p_t(x)$ (recall (154)).

7. Concluding remarks. This paper has been concerned with robustness of nonlinear systems using the closely related gap metric and coprime factor approaches to uncertainty. There have been two main streams of thought running through the paper:

(i) signal-based definitions of the gap metric, robustness results in terms of the parallel projection operators, and synthesis of controllers based on minimizing the induced norms of the parallel projections (sections 4.1, 4.2, 4.3, 4.4, 5);

(ii) graph representations, robustness of feedback systems in terms of perturbations of the graph "symbols," and state-space realizations of graph symbols (sections 3, 4.4, 4.5, 6).

For linear systems, (ii) was the earliest to develop fully, and technical results there provided an important resource for the development of a more operator-based approach along the lines of (i). For nonlinear systems there have been several approaches to (ii) which potentially pave the way for a full robustness theory. Our contention here is that the complete working out of (i) in this paper has shed new light on how to develop (ii) in a way that closely matches the needs of the robustness theory.

The two streams of thought have by no means been independent in this paper—they touch one another most closely in section 4.4 and in the commonality of the nonlinear H_∞ control techniques used in sections 5 and 6. It is interesting to note that the realization theory for graph representations has been presented as the final topic of the paper (section 6), with its importance being a proof of existence of the graph representations defined here. In contrast, this topic was an essential early building block in the development of the linear coprime factorization theory.

Appendix A. Nonlinear H_∞ control. We will now review the nonlinear H_∞ controller synthesis theory developed in [17], [14] for the class of nonlinear systems or *generalized plants*:

$$\begin{aligned} \dot{x} &= A(x) + B_1(x)w + B_2(x)u, & x(0) &= x_0, \\ z &= C_1(x) + D_{12}u, & D_{12}'D_{12} &= I_m, \\ y &= C_2(x) + D_{21}w, & D_{21}D_{21}' &= I_q. \end{aligned} \tag{190}$$

In these equations $x(t) \in \mathbf{R}^n$ denotes the state of the system and $y(t) \in \mathbf{R}^q$ is the measurement signal. The output to be regulated is $z(t) \in \mathbf{R}^p$. The control input is $u(t) \in \mathbf{R}^m$, while $w(t) \in \mathbf{R}^l$ is an exogenous disturbance input. We assume that all the problem data are smooth functions of x with bounded first derivatives, that B_1

and B_2 are bounded, and that the origin is an equilibrium state: $A(0) = 0, C_1(0) = 0$, and $C_2(0) = 0$. In brief, the generalized plant is

$$\left[\begin{array}{c|c|c} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ \hline C_2 & D_{21} & 0 \end{array}\right].$$

The (output feedback) controller K is assumed to be a causal mapping $y \in \mathcal{Y} \mapsto u \in \mathcal{U}$, where $\mathcal{U}$ and $\mathcal{Y}$ are signal spaces as described in section 2. Such a controller is called admissible if the closed-loop equations associated with K and (190) are well-defined in the sense that they have unique solutions in $L_{2,e}$.

A controller K is said to solve the H_∞ control problem provided the closed-loop system is γ-dissipative and internally stable. The closed-loop system is γ-*dissipative* if

$$\tfrac{1}{2}\int_0^T |z(s)|^2 ds \le \tfrac{1}{2}\gamma^2 \int_0^T |w(s)|^2 ds + \beta(x(0)) \tag{191}$$

for some nonnegative function β with $\beta(0) = 0$ and every $w \in L_2[0,T]$ for all $T \ge 0$. *Internal stability* means that if $w \in L_2$, then $u(\cdot), y(\cdot), z(\cdot), x(\cdot) \in L_2$, and consequently $x(t) \to 0$ as $t \to \infty$.

A.1. The state feedback problem. Let us consider static state feedback controllers of the form

$$u = u(x),$$

where $u : \mathbf{R}^n \to \mathbf{R}^m$. Combining this with (190) gives the following closed-loop system:

$$\begin{aligned} \dot{x} &= A(x) + B_2(x)u(x) + B_1(x)w, \\ z &= C_1(x) + D_{12}(x)u(x). \end{aligned} \tag{192}$$

If there exists a static state feedback controller $u(x)$ such that this closed-loop system is γ-dissipative, we know that there exists a storage function $V(x) \ge 0$ which satisfies $V(0) = 0$ and the HJBI equation

$$\nabla V(A - B_2E_1^{-1}D_{12}'C_1) - \tfrac{1}{2}\nabla V(B_2E_2^{-1}B_2' - \gamma^{-2}B_1B_1')\nabla V' + \tfrac{1}{2}C_1'(I - D_{12}E_1^{-1}D_{12}')C_1 = 0 \tag{193}$$

(here $E_1 = D_{12}'D_{12} > 0$); the function $V(x)$ need not necessarily be smooth, in which case the PDE (193) can be interpreted in the viscosity sense [23, Theorem 3.4]; see also [14, section 5.4].

Conversely, if (193) has a smooth solution $V(x) > 0$ if $x \ne 0$, $V(0) = 0$, then the state feedback controller

$$u^*_{state}(x) = -(D_{12}'C_1(x) + B_2(x)'\nabla V(x)') \tag{194}$$

renders the closed loop γ-dissipative. Since the control law depends on ∇V, we assume that $V(\cdot)$ is differentiable in some sense (C^1 with globally Lipschitz derivative is enough). The stability of the closed loop follows from strict positive definiteness of V and the zero state detectability of $((A + B_2u^*_{state})(x), (C_1 + D_{12}u^*_{state})(x))$; see, e.g., [26, section 6.1].

A.2. The information state. We will now consider the case of output feedback controllers of the form $K : y \mapsto u$. We transform the output feedback problem to a new state feedback problem using the information state $p_t(x) = p(x,t)$ defined by the equation

$$\begin{aligned}\frac{\partial}{\partial t}p = &-\nabla p(A + B_1 D_{21}'(y - C_2) + B_2 u) + \gamma^{-2}\tfrac{1}{2}\nabla p B_1(I - D_{21}' D_{21})B_1'\nabla p' \\ &-\gamma^2\tfrac{1}{2}|y - C_2|^2 + \tfrac{1}{2}|C_1 + D_{12}u|^2.\end{aligned} \tag{195}$$

In shorthand notation, we write $p_t(\cdot) = p(\cdot, t)$ and regard p_t as the state of a new dynamical system with state equations

$$\dot{p} = F(p, u, y), \tag{196}$$

where $F(p,u,y)$ is the nonlinear differential operator defined in (195). The state space is an appropriate function space $\mathcal{X}$ (e.g., the Banach space of continuous functions with at most quadratic growth with the norm $\| p \|_{\mathcal{X}} = \sup_{x\in\mathbf{R}^n} \frac{|p(x)|}{1+|x|^2}$). It is known (see [14, Section 3.1]) that if a controller $K : y \mapsto u$ exists such that the closed loop is γ-dissipative, then (195) has a solution. If the solution is not smooth, it can be interpreted in the viscosity sense [3], [6]. However, we will assume it to be smooth.

In the equilibrium (or steady-state) case ($u \equiv 0$ and $y \equiv 0$), this equation reduces to

$$\nabla p_\infty(A - B_1 D_{21}' C_2) - \gamma^{-2}\tfrac{1}{2}\nabla p_\infty B_1(I - D_{21}' D_{21})B_1'\nabla p_\infty' + \gamma^2\tfrac{1}{2}|C_2|^2 - \tfrac{1}{2}|C_1|^2 = 0, \tag{197}$$

or in shorthand, $F(p_\infty, 0, 0) = 0$. A particular equilibrium solution p_e of (197), called a *control attractor*, is important. The present context is nonsingular in the sense of [14], meaning p_e is everywhere finite. The control attractor satisfies $p_e \le 0$, $p_e(0) = 0$, and

$$-(A - B_1 D_{21}' C_2 - \gamma^{-2} B_1[I - D_{21}' D_{21}]B_1'\nabla p_e') \quad \text{asymptotically stable} \tag{198}$$

(it is the unique such function). Convergence of p_t (when driven by $L_2[0,\infty)$ signals u, y) to a solution $p_\infty = p_e + c$ of (197) as $t \to \infty$ is discussed in [14] ($c = c(u,y) \in \mathbf{R}$ is a constant independent of x but depending on the signals u, y). To be precise, the particular type of stability (198) required is referred to in [14] as *incremental L_2 exponential antistability*.

A.3. Information state controller. In the output feedback case, the transformation afforded by the information state gives rise to a nonlinear PDE on an infinite-dimensional (Banach) space $\mathcal{X}$ (the analog of the PDE (193)), viz.

$$\inf_{u\in\mathbf{R}^m} \sup_{y\in\mathbf{R}^p} \{\nabla W(p)[F(p,u,y)]\} = 0, \tag{199}$$

where $\nabla W(p)$ is a linear operator (Frechet derivative). It is known that there exists a value function $W(p)$ solving this equation in a suitable sense [17], [18], [14]. In general it cannot be expected that W is smooth, and in fact at present there is no adequate theory for PDEs of the type (199); however, it can be interpreted in an integrated form, and various notions of smoothness of solution have been considered [18], [14]. We assume that smoothness facilitates system-theoretic interpretation and remark that smoothness issues do not arise in discrete time [17]. The value function

W should satisfy a number of properties in addition to (199): (i) $W(p) \geq \sup_x\{p(x)\}$, (ii) $p_1 \geq p_2$ implies $W(p_1) \geq W(p_2)$, and (iii) $W(p+c) = W(p)+c$ for all $c \in \mathbf{R}$. Also, in general, W has a nontrivial domain of finiteness dom$W$. For full details, see [18], [14]. Here, it is sufficient to note that a smooth solution of (199) defines an output feedback controller K^* via the optimal feedback function obtained by evaluating the infimum in (199) (cf. (194)):

$$u^*(p) = \nabla W(p)[-D_{12}'C_1 + B_2'\nabla p']. \tag{200}$$

The optimal information state controller is given by

$$K^*: \begin{cases} \dot{p} = F(p,u,y), & p_0 = p_e, \\ u = u^*(p). \end{cases} \tag{201}$$

This controller feeds back the information state, initialized at the control attractor p_e, and produces a γ-dissipative closed loop. For precise statements and stability results, see [14]. It turns out that $p_e \in \text{dom}W$ (in fact, $W(p_e) = 0$), and the antistabilizing property of p_e ensures that $p_t \in \text{dom}W$ for all $t \geq 0$ when driven by $L_2[0,\infty)$ signals u, y, and this will be the case provided p_t remains in a region relating to the smoothness of W [14, Chapter 4].

Further, the function $W(p)$ is related to the state feedback function $V(x)$ by

$$\max_x\{p(x) + V(x)\} \leq W(p), \tag{202}$$

and in particular

$$\max_x\{p_e(x) + V(x)\} \leq W(p_e) = 0 < +\infty. \tag{203}$$

Equation (203) is a necessary condition which holds irrespective of the smoothness of $W(p)$, and in fact does not depend on the existence of the optimal feedback $u^*(p)$ [14].

A.4. Certainty equivalence. Under certain conditions, the optimal information state controller simplifies to a controller which is equivalent to the certainty equivalence controller of [4]. Let $u^*_{state}(x)$ denote the optimal state feedback control, given by (194). Assume that the minimum stress state estimate

$$\bar{x}(t) = \arg\max_x\{p(x,t) + V(x)\} \tag{204}$$

is unique. Then the certainty equivalence controller exists and is given by

$$u^{ce}(t) = u^*_{state}(\bar{x}(t)), \tag{205}$$

giving a γ-dissipative closed loop [4], [18], [14]. If a certainty equivalence controller does not exist, it is necessary to solve (199) for the optimal information state controller (200). When the certainty equivalence assumptions are valid, essentially one can take

$$\hat{W}(p) = \max_x\{p(x) + V(x)\} \tag{206}$$

as a solution of (199) [18], [14].

REFERENCES

[1] B. D. O. Anderson, M. R. James, and D. J. N. Limebeer, *Robust stabilization of nonlinear systems via normalized coprime factor representations*, Automatica J. IFAC, 34 (1998), pp. 1593–1599.

[2] J. A. Ball, J. W. Helton, and M. L. Walker, *H^∞ control for nonlinear systems with output feedback*, IEEE Trans. Automat. Control, 38 (1993), pp. 546–559.

[3] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhauser, Boston, 1997.

[4] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, 2nd ed., Birkhauser, Boston, 1995.

[5] J. C. Doyle, T. T. Georgiou, and M. C. Smith, *The parallel projection operators of a nonlinear feedback system*, Systems Control Lett., 20 (1993), pp. 79–85.

[6] W. H. Fleming and H. M. Soner, *Controlled Markov Process and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[7] T. Fliegner, *Some Remarks on Parallel Projection Operator Norms in L_p-Spaces*, Technical Report, CUED/F-INFENG/TR 316, Cambridge University Engineering Department, Cambridge, UK, 1998.

[8] C. Foias, T. T. Georgiou, and M. C. Smith, *Robust stability of feedback systems: A geometric approach using the gap metric*, SIAM J. Control Optim., 31 (1993), pp. 1518–1537.

[9] T. T. Georgiou, *On the computation of the gap metric*, Systems Control Lett., 11 (1988), pp. 253–257.

[10] T. T. Georgiou and M. C. Smith, *Optimal robustness in the gap metric*, IEEE Trans. Automat. Control, 35 (1990), pp. 673–686.

[11] T. T. Georgiou and M. C. Smith, *Robustness analysis of nonlinear feedback systems: An input-output approach*, IEEE Trans. Automat. Control, 42 (1997), pp. 1200–1220.

[12] K. Glover and D. MacFarlane, *Robust stabilisation of normalised coprime factor plant descriptions with $\mathcal{H}_\infty$ bounded uncertainty*, IEEE Trans. Automat. Control, 34 (1989), pp. 821–830.

[13] M. Green and D. J. N. Limebeer, *Linear Robust Control*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[14] J. W. Helton and M. R. James, *Extending H^∞ Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives*, Advances in Design and Control, SIAM, Philadelphia, 1999.

[15] A. Isidori, *H_∞ control via measurement feedback for affine nonlinear systems*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 553–574.

[16] A. Isidori and A. Astolfi, *Disturbance attenuation and H_∞-control via measurement feedback in nonlinear systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1283–1293.

[17] M. R. James and J. S. Baras, *Robust H_∞ output feedback control for nonlinear systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1007–1017.

[18] M. R. James and J. S. Baras, *Partially observed differential games, infinite-dimensional Hamilton–Jacoby–Isaacs equations, and nonlinear H_∞ control*, SIAM J. Control Optim., 34 (1996), pp. 1342–1364.

[19] A. D. B. Paice and A. J. van der Schaft, *The class of stabilizing nonlinear plant controller pairs*, IEEE Trans. Automat. Control, 41 (1996), pp. 634–645.

[20] D. C. McFarlane and K. Glover, *Robust Controller Design Using Normalized Coprime Factor Plant Descriptions*, Lecture Notes in Control and Inform. Sci., Springer-Verlag, New York, 1990.

[21] J. A. Sefton and R. J. Ober, *On the gap metric and coprime factor perturbations*, Automatica J. IFAC, 29 (1993), pp. 723–734.

[22] E. D. Sontag, *Smooth stabilization implies coprime factorization* IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.

[23] P. Soravia, *$\mathcal{H}_\infty$ control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.

[24] J. M. A. Scherpen and A. J. van der Schaft, *Normalized coprime factorizations and balancing for unstable nonlinear systems*, Internat. J. Control, 60 (1994), pp. 1193–1222.

[25] A. J. van der Schaft, *Robust stabilization of nonlinear systems via stable kernel representations with L_2-gain bounded disturbances*, Systems Control Lett., 24 (1995), pp. 75–81.

[26] A. J. van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*, Springer-Verlag, New York, 1996.

[27] M. S. Verma and L. R. Hunt, *Right coprime factorizations and stabilization for nonlinear-systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 222–231.

[28] M. Vidyasagar and H. Kimura, *Robust controllers for uncertain linear multivariable systems*, Automatica J. IFAC, 22 (1986), pp. 85–94.

[29] G. Vinnicombe, *Frequency domain uncertainty and the graph topology*, IEEE Trans. Automat. Control, 38 (1993), pp. 1371–1383.

[30] G. Vinnicombe, *A ν-gap distance for uncertain and nonlinear systems*, in Proceedings of the 38th IEEE CDC, Phoenix, AZ, 1999.

[31] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1996.

[32] G. Zames and A. K. El-Sakkary, *Unstable systems and feedback: The gap metric*, in Proceedings of the Allerton Conference, 1980, pp. 380–385.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1583–1615

AN INVERSE PROBLEM FOR A PARABOLIC VARIATIONAL INEQUALITY ARISING IN VOLATILITY CALIBRATION WITH AMERICAN OPTIONS*

YVES ACHDOU†

Abstract. In finance, the price of an American option is obtained from the price of the underlying asset by solving a parabolic variational inequality. The free boundary associated with this variational inequality can be interpreted as the price for which the option should be exercised. The calibration of volatility from the observations of the prices of an American option yields an inverse problem for the previously mentioned parabolic variational inequality. After studying the variational inequality and the exercise price, we give results concerning the sensitivity of the option price and of the exercise price with respect to the variations of the volatility. The inverse problem is addressed by a least square method, with suitable regularization terms. We give necessary optimality conditions involving an adjoint state for a simplified inverse problem and we study the differentiability of the cost function. Optimality conditions are also given for the genuine calibration problem.

Key words. variational inequalities, inverse problems, American options, calibration in finance

AMS subject classifications. 35K85, 35R35, 49J40, 49K20, 49K40, 49N45

DOI. 10.1137/S0363012903424423

1. Introduction.

A *European vanilla call (resp., put) option is a contract giving its owner the right to buy (resp., sell) a share of a specific common stock at a fixed price K at a certain date T*. The specific stock is called *the underlying asset*. The fixed price K is termed *the strike*, and T is termed *the maturity*. The term *vanilla* is used to notify that this kind of option is the simplest one: indeed, there may be more complicated contracts. The price of the underlying asset at time t will be referred to as the *spot price* and will be noted x_t. Assuming that the market rules out arbitrage (the possibility to make an instantaneous risk-free benefit), it is very easy to see that the price of a call (resp., put) option at maturity is $u_\circ(x_T) = (x_T - K)_+$ (resp., $u_\circ(x_T) = (K - x_T)_+$). The function $u_\circ$ is called the payoff function.

In order to price the option before maturity, some assumptions have to be made on the spot price x_t: the Black–Scholes model involves the above mentioned underlying asset and a risk-free asset whose price at time t is $S_t^0 = e^{rt}$, where r is the interest rate; it assumes that the price of the risky asset is a solution to the following stochastic differential equation:

$$dx_t = x_t(\mu dt + \sqrt{\eta_t} dB_t), \tag{1.1}$$

where B_t is a standard Brownian motion on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Here η_t is a positive number, and $\sqrt{\eta_t}$ is called the *volatility*. With the Black–Scholes assumptions, it is possible to prove that the option's price at time t is given by

$$P_t = u_e(t, x_t) \equiv \mathbb{E}^*(e^{-r(T-t)} u_\circ(x_T) | F_t), \tag{1.2}$$

*Received by the editors March 13, 2003; accepted for publication (in revised form) June 7, 2004; published electronically March 11, 2005.
http://www.siam.org/journals/sicon/43-5/42442.html
†UFR Mathématiques, Université Paris 7, Case 7012, 75251 Paris Cedex 05, France, and Laboratoire Jacques-Louis Lions, Université Paris 6, 75252 Paris Cedex 05, France (achdou@math.jussieu.fr).

where the expectation $\mathbb{E}^*$ is taken with respect to the so-called risk-neutral probability $\mathbb{P}^*$ (equivalent to $\mathbb{P}$ and under which $dx_t = x_t(rdt + \sqrt{\eta_t}dW_t)$, W_t being a standard Brownian motion under $\mathbb{P}^*$ and F_t being the natural filtration of W_t). Assuming that $\eta_t = \eta(t, x_t)$, where η is a smooth enough function, it can be seen that the pricing function u_e solves the parabolic PDE

$$\frac{\partial u_e}{\partial t} + \frac{\eta(t,x)x^2}{2}\frac{\partial^2 u_e}{\partial x^2} + rx\frac{\partial u_e}{\partial x} - ru_e = 0. \tag{1.3}$$

In contrast with European options, American options can be exercised any time before maturity: *An American vanilla call (resp., put) option is a contract giving its owner the right to buy (resp., sell) a share of a specific common stock at a fixed price K before a certain date T.* More generally, for a payoff function $u_\circ$, the American option with payoff $u_\circ$ and maturity T can be exercised at any $t < T$, yielding the payoff $u_\circ(x_t)$. Since the American option gives its owner more rights than the corresponding European option, its price should be higher.

Consider an American vanilla put. If its price P_t were less than $K - x_t$, then one could buy a put and a share of the underlying asset and exercise immediately the option, making a risk-free immediate benefit of $K - x_t - P_t > 0$: this is ruled out by the no arbitrage assumption, so we see that $P_t \geq (K - x_t)_+$. For a general payoff $u_\circ$, we have $P_t > u_\circ(x_t)$. Using the notion of strategy with consumption, the Black–Scholes model leads to the following formula for pricing an American option with payoff $u_\circ$: under the risk-neutral probability,

$$P_t = u(t, x_t) \equiv \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}^* \left(e^{-r(\tau - t)} u_\circ(x_\tau) \middle| F_t \right), \tag{1.4}$$

where $\mathcal{T}_{t,T}$ denotes the set of stopping times in $[t, T]$ (see [23] for the proof of this formula). It can be seen that for an American vanilla call, formula (1.4) coincides with (1.2), so American and European vanilla calls have the same price. This means that an American vanilla call should not be exercised before maturity.

It can be shown that the price u of the American put of strike K and maturity T is given as a solution to

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\eta x^2}{2}\frac{\partial^2 u}{\partial x^2} + rx\frac{\partial u}{\partial x} - ru \leq 0, \quad u(t,x) \geq (K-x)_+, \quad t \in [0,T),\ x > 0, \\ &\left(\frac{\partial u}{\partial t} + \frac{\eta x^2}{2}\frac{\partial^2 u}{\partial x^2} + rx\frac{\partial u}{\partial x} - ru \right)(u - (K-x)_+) = 0, \quad t \in [0,T),\ x > 0, \\ &u(t = T, x) = (K - x)_+; \end{aligned} \tag{1.5}$$

see [31] for a formal derivation of (1.5).

By using suitable Sobolev spaces, it is possible to use now well-known variational procedures for (1.5) (see Kinderlehrer and Stampacchia [19]) and prove the existence and uniqueness of u, under mild assumptions on η.

The volatility is the difficult parameter of the Black–Scholes model. It is convenient to take it to be constant, but then the computed options' prices do not match the prices given by the market. There are essentially three ways to improve on the Black-Scholes model with a constant volatility:

- Use a local volatility; i.e., assume that the volatility is a function of time and of the stock price. Then one has *to calibrate the volatility* from the market data, i.e., to find a volatility function which permits one to recover the prices of the options available on the market.

- Assume that the volatility is itself a stochastic process; see, for example, [14, 11].
- Generalize the Black–Scholes model by assuming that the spot price is, for example, a Lévy process; see [9] and references therein.

There is much discussion among specialists in finance on comparing the merits of the three kinds of models above. Here, we will focus on the first one and we will deal with the calibration of the local volatility.

In mathematical finance there have been a number of valuable studies on calibration with European options for which it is difficult to make a complete account: Avellaneda et al. [3] proposed a procedure based on the maximization of an entropy function via dynamic programming. Lagnado and Osher [21, 22], Jackson, Süli, and Howison [17], Coleman, Li, and Verma [8], and Achdou and Pironneau [2] used, as we do here, a least square fit to the financial data. Calibration with European options is made easier thanks to the linear character of the equations; see [10].

Concerning calibration with American options, we are not aware of any published articles; in this paper, we focus on American vanilla put options and study theoretically the least square optimization problem arising when one tries to calibrate the volatility from the observed prices of the option. Mathematically, it is an inverse problem in order to find the coefficient of a parabolic variational inequality.

The aim is to find a function $\eta(t,x)$ such that the price of the option computed by (1.5) fits some observations of the price of the American put for different maturities T and stroke K (the number of observed prices is finite). In this paper, we will first replace this inverse problem by a simpler one, i.e., with a single maturity T, and assume that the observation is made for all times t in $[0,T]$ and every price x (complete observation). From the theoretical point of view, this inverse problem retains the essential difficulties of the original one. We are going to use a least square approach, and the problem will somehow resemble optimal control problems studied in [28, 16, 5]. Let us mention especially the work of Hintermüller [15] on an inverse problem for an elliptic variational inequality, which has inspired the present work. Finally, we will investigate the genuine inverse problem for the calibration of volatility with a discrete set of observations.

The main results of the paper are:

1. Theorem 6.4 in section 6.2, where first order necessary optimality conditions are given for the least square inverse problem essentially under the assumption that no observed prices lie on the free boundary, i.e., the curve limiting the zone where the option should be exercised.
2. Theorem 5.4 in section 5.3, which states that the least square functional is differentiable with respect to η, under more restrictive assumptions, and which gives the derivative. This result is stated for the simpler least square problem mentioned above, but it holds for the genuine one with obvious modifications.
3. A careful study of the free boundary; see section 3.

The numerical counterpart of the present work has been carried out successfully (see [1]), and software has been written for the calibration of American options: it uses a discrete analogue of Theorem 5.4 for computing descent directions. Alternatively, an optimization method based on a least square version of the optimality conditions given by Theorem 6.4 is possible; see [15].

The paper is organized as follows. In section 2, we study carefully the variational inequality satisfied by the price of the American option. Section 3 is devoted to

the free boundary: under some assumptions on the volatility, we prove that the free boundary is the graph of a continuous function of time. In section 4, we investigate the sensitivity of u and the free boundary with respect to variations of the volatility. An inverse problem (simpler than the original one) is introduced in section 5, and we give some necessary optimality conditions. We also address the differentiability issue: is the cost function differentiable with respect to the volatility, and can we compute a gradient? We give a partial positive answer for volatilities such that the corresponding state function and free boundary are regular enough. Section 6 is devoted to the genuine calibration problem: necessary optimality conditions are given.

2. The variational inequality. Changing t into the time to maturity $T-t$, (1.5) becomes

(2.1)
$$\begin{aligned} \frac{\partial u}{\partial t} - \frac{\eta x^2}{2}\frac{\partial^2 u}{\partial x^2} - rx\frac{\partial u}{\partial x} + ru \geq 0, \qquad u(t,x) &\geq u_\circ(x), \qquad t\in(0,T],\ x>0,\\ \left(\frac{\partial u}{\partial t} - \frac{\eta x^2}{2}\frac{\partial^2 u}{\partial x^2} - rx\frac{\partial u}{\partial x} + ru\right)(u-u_\circ) &= 0, \qquad t\in(0,T],\ x>0,\\ u(t=0,x) &= u_\circ(x), \qquad x>0, \end{aligned}$$

where $\eta : [0,T]\times\mathbb{R}_+ \to \mathbb{R}_+$ is a positive function and

$$u_\circ(x) = (K-x)_+. \tag{2.2}$$

The pricing function u_e for the corresponding European put solves $u_e|_{t=0} = u_\circ$ and for $0 < t \leq T$,

$$\frac{\partial u_e}{\partial t} - \frac{\eta x^2}{2}\frac{\partial^2 u_e}{\partial x^2} - rx\frac{\partial u_e}{\partial x} + ru_e = 0. \tag{2.3}$$

Although (2.1) may be tackled with the results of [12], we prefer to study it completely in order to make the presentation self-contained. For what follows, we introduce the weighted Sobolev space

$$V = \left\{ v\in L^2(\mathbb{R}_+), x\frac{\partial v}{\partial x}\in L^2(\mathbb{R}_+)\right\}, \tag{2.4}$$

which is a Hilbert space for the norm $\|v\|_V = \left(\|v\|^2_{L^2(\mathbb{R}_+)} + \|x\frac{\partial v}{\partial x}\|^2_{L^2(\mathbb{R}_+)}\right)^{\frac{1}{2}}$. The space V is clearly separable, and $\mathcal{D}(\mathbb{R}_+)$ is a dense subspace of V.

Assumption 1. We assume that for two fixed positive constants $\underline{\eta}$ and $\overline{\eta}$, $0 < \underline{\eta} \leq \eta \leq \overline{\eta}$ a.e. in $(0,T)\times\mathbb{R}_+$, and for a positive constant M, $|x\frac{\partial \eta}{\partial x}| \leq M$ a.e in $(0,T)\times\mathbb{R}_+$.

Remark 1. The assumption $|x\frac{\partial\eta}{\partial x}| \leq M$ is necessary for the variational formulation of (2.1): using viscosity solutions techniques (see the article by Friedman and Shen [13] on a closely related variational inequality and the references therein), one can do without it. It is also possible to obtain many results on (2.1) from the probability theory; see [18]. With Assumption 1, we can consider the family of linear operators $A(t) : V \to V'$ defined, for $v, w \in V$, by

$$\langle A(t)v, w\rangle = \int_{\mathbb{R}_+}\left(\frac{\eta(t)}{2}x^2\frac{\partial v}{\partial x}\frac{\partial w}{\partial x} + \left(\eta(t) + \frac{x}{2}\frac{\partial\eta}{\partial x} - r\right)x\frac{\partial v}{\partial x}w + rvw\right)dx. \tag{2.5}$$

From the assumptions on η, there exists a nonnegative constant α such that, for a.e. $t \in (0,T)$ and for any $v \in V$,

$$\langle A(t)v, v\rangle \geq \frac{1}{4}\underline{\eta}\|v\|_V^2 - \alpha\|v\|^2_{L^2(\mathbb{R}_+)}. \tag{2.6}$$

We introduce the closed subset of $L^2(0,T;V)$:

$$\mathcal{K} = \{v \in L^2(0,T;V), v \geq u_\circ \text{ a.e. in } (0,T)\times\mathbb{R}_+\}. \tag{2.7}$$

The variational formulation of (2.1) is as follows: find $u \in \mathcal{K} \cap C^0([0,T];L^2(\mathbb{R}_+))$, with $\frac{\partial u}{\partial t} \in L^2(0,T;V')$ such that for all $v \in \mathcal{K}$,

$$\left\langle \frac{\partial u}{\partial t} + A(t)u, v-u \right\rangle \geq 0 \quad \text{for } t>0 \quad \text{and} \quad u(t=0) = u_\circ. \tag{2.8}$$

In order to prove the existence for u, we make two observations:

1. The price of the American put is larger than that of the European put given by (2.3), which is itself positive for any $t \in (0,T]$ (we shall justify this by using the maximum principle; see below).
2. The function $u_\circ$ satisfies, for $t \in [0,T]$,

$$\left(\frac{\partial u_\circ}{\partial t} + A(t)u_\circ\right)\Big|_{x<K} = rK, \qquad \left(\frac{\partial u_\circ}{\partial t} + A(t)u_\circ\right)\Big|_{x>K} = 0. \tag{2.9}$$

Therefore, following [19], we introduce the penalized problem: find u_ϵ such that

$$\begin{aligned} \frac{\partial u_\epsilon}{\partial t} - \frac{\eta x^2}{2}\frac{\partial^2 u_\epsilon}{\partial x^2} - rx\frac{\partial u_\epsilon}{\partial x} + ru_\epsilon - rK1_{\{x<K\}}\mathcal{V}_\epsilon(u_\epsilon - u_\circ) &= 0 \quad \begin{cases} t \in (0,T], \\ x>0, \end{cases} \\ u_\epsilon(t=0,x) &= u_\circ(x), \end{aligned} \tag{2.10}$$

where $\mathcal{V}_\epsilon(u) = \mathcal{V}(\frac{u}{\epsilon})$ and $\mathcal{V}$ is a smooth nonincreasing convex function such that

$$\mathcal{V}(0) = 1, \qquad \mathcal{V}(u) = 0 \quad \text{for } u \geq 1, \qquad 0 \geq \mathcal{V}'(u) \geq -2 \quad \text{for } 0 \leq u \leq 1. \tag{2.11}$$

THEOREM 2.1. *Under Assumption* 1, (2.10) *has a unique weak solution* $u_\epsilon \in L^2(0,T;V) \cap C^0([0,T];L^2(\mathbb{R}_+))$ *which belongs to* $\mathcal{K}$. *Also,* $x\frac{\partial u_\epsilon}{\partial x}$ *and* $\frac{\partial u_\epsilon}{\partial x}$ *belong to* $L^2(0,T;V) \cap C^0([0,T];L^2(\mathbb{R}_+))$, *and* u_ϵ *is continuous in* $[0,T]\times[0,+\infty]$. *Moreover,* $\frac{\partial u_\epsilon}{\partial t} \in L^2((0,T)\times\mathbb{R}_+)$. *The quantities* $\|u_\epsilon\|_{L^2(0,T;V)}$, $\|u_\epsilon\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$, $\|x\frac{\partial u_\epsilon}{\partial x}\|_{L^2(0,T;V)}$, $\|\frac{\partial u_\epsilon}{\partial x}\|_{L^2(0,T;V)}$, $\|x\frac{\partial u_\epsilon}{\partial x}\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$, $\|\frac{\partial u_\epsilon}{\partial x}\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$, *and* $\|\frac{\partial u_\epsilon}{\partial t}\|_{L^2((0,T)\times\mathbb{R}_+)}$ *are bounded independently of* ϵ.

Furthermore, there exists a function $\hat{z}_\epsilon \in L^2(0,T;V) \cap C^0([0,T];L^2(\mathbb{R}_+))$ *such that*

$$\|\hat{z}_\epsilon\|_{L^2(0,T;V)} + \|\hat{z}_\epsilon\|_{L^\infty(0,T;L^2(\mathbb{R}_+))} \to 0$$

and

$$-1 - \hat{z}_\epsilon \leq \frac{\partial u_\epsilon}{\partial x} \leq 0 \quad \forall t \in (0,T], \text{ a.e. } x>0. \tag{2.12}$$

Finally, u_ϵ *is greater than* u_e, *the price of the European put (given by* (2.3)*).*

Proof. By using results on parabolic equations with monotone operators [26, page 156], it is possible to prove that (2.10) has a unique weak solution in $L^2(0,T;V)\cap C^0([0,T];L^2(\mathbb{R}_+))$, with $\frac{\partial u_\epsilon}{\partial t}\in L^2(0,T;V')$, and that $\|u_\epsilon\|_{L^2(0,T;V)}$ and $\|u_\epsilon\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$ are bounded independently of ϵ. Furthermore, the weak maximum principle for parabolic equations leads to

$$0\le u_\epsilon(t,x)\le K\quad \forall t\in(0,T],\ \text{a.e. } x>0. \tag{2.13}$$

With u_e given by (2.3), we have, again by the maximum principle,

$$0<u_e(t,x)\le u_\epsilon(t,x)\quad \forall t\in(0,T],\ \text{a.e. } x>0, \tag{2.14}$$

and we can also compare the function u_ϵ with the function $K-x$:

$$K-x\le u_\epsilon(t,x)\quad \forall t\in(0,T],\ \text{a.e. } x>0. \tag{2.15}$$

From estimates (2.14) and (2.15), $u_\epsilon\in\mathcal{K}$.

By taking the derivative of (2.10) with respect to x, we obtain that $z_\epsilon=\frac{\partial u_\epsilon}{\partial x}$ satisfies

$$\begin{aligned}&\frac{\partial z_\epsilon}{\partial t}-\frac{\partial}{\partial x}\left(\frac{\eta x^2}{2}\frac{\partial z_\epsilon}{\partial x}\right)-rx\frac{\partial z_\epsilon}{\partial x}-rK1_{\{x<K\}}\mathcal{V}'_\epsilon(u_\epsilon-u_\circ)(z_\epsilon+1)\\&=-rK\mathcal{V}_\epsilon(u_\epsilon(K))\delta_{x=K},\\&z_\epsilon(t=0,x)=-1_{\{x<K\}},\end{aligned} \tag{2.16}$$

which also has a variational formulation in $L^2(0,T;V)$ because

$$-\frac{2}{\epsilon}\le\mathcal{V}'_\epsilon(u_\epsilon-u_\circ)\le 0$$

and $-rK\mathcal{V}_\epsilon(u_\epsilon(K))\delta_{x=K}\in L^2((0,T),V')$; we can easily obtain that

$$\frac{\partial u_\epsilon}{\partial x}\in L^2(0,T,V)\cap C^0([0,T];L^2(\mathbb{R}_+)),$$

with no additional assumption on η. This proves that u_ϵ is actually continuous and that (2.13), (2.14), (2.15) hold pointwise. From (2.16), it is possible to prove that $\|\frac{\partial u_\epsilon}{\partial x}\|_{L^2(0,T;V)}$ and $\|\frac{\partial u_\epsilon}{\partial x}\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$ are bounded independently of ϵ, and that $\frac{\partial u_\epsilon}{\partial x}\le 0$ for all t and a.e. in x. From (2.16), we deduce that $z_\epsilon(t=0,x)+1=1_{\{x\ge K\}}$ and for $t\in(0,T],\ x>0$,

$$\begin{aligned}&\frac{\partial(z_\epsilon+1)}{\partial t}-\frac{\partial}{\partial x}\left(\frac{\eta x^2}{2}\frac{\partial(z_\epsilon+1)}{\partial x}\right)-rx\frac{\partial(z_\epsilon+1)}{\partial x}-rK1_{\{x<K\}}\mathcal{V}'_\epsilon(u_\epsilon-u_\circ)(z_\epsilon+1)\\&\qquad=-rK\mathcal{V}_\epsilon(u_\epsilon(K))\delta_{x=K}.\end{aligned} \tag{2.17}$$

The function $z_\epsilon+1$ is the sum of two functions $\tilde z_\epsilon+1$ and $\hat z_\epsilon$ which satisfy, respectively,

$$\begin{aligned}&\frac{\partial(\tilde z_\epsilon+1)}{\partial t}-\frac{\partial}{\partial x}\left(\frac{\eta x^2}{2}\frac{\partial(\tilde z_\epsilon+1)}{\partial x}\right)-rx\frac{\partial(\tilde z_\epsilon+1)}{\partial x}-rK1_{\{x<K\}}\mathcal{V}'_\epsilon(u_\epsilon-u_\circ)(\tilde z_\epsilon+1)\\&\qquad=0,\quad t>0,\\&\tilde z_\epsilon(t=0,x)+1=1_{\{x\ge K\}}\end{aligned} \tag{2.18}$$

and

$$\begin{aligned}\frac{\partial \hat{z}_\epsilon}{\partial t} - \frac{\partial}{\partial x}\left(\frac{\eta x^2}{2}\frac{\partial \hat{z}_\epsilon}{\partial x}\right) - rx\frac{\partial \hat{z}_\epsilon}{\partial x} - rK1_{\{x<K\}}\mathcal{V}'_\epsilon(u_\epsilon - u_\circ)\hat{z}_\epsilon \\ = -rK\mathcal{V}_\epsilon(u_\epsilon(K))\delta_{x=K}, \quad t>0, \\ \hat{z}_\epsilon(t=0,x) = 0.\end{aligned} \tag{2.19}$$

Since $u_\epsilon \geq u_e$, we know that $\lim_{\epsilon\to 0}\|\mathcal{V}_\epsilon(u_\epsilon(K))\delta_{x=K}\|_{L^2(0,T;V')} = 0$.

Thus, $\lim_{\epsilon\to 0}\|\hat{z}_\epsilon\|_{L^2(0,T;V)} = 0$. On the other hand, $\tilde{z}_\epsilon \geq -1$. Thus, $\frac{\partial u_\epsilon}{\partial x} \geq -1 - \hat{z}_\epsilon$ for all t and for a.e. x.

Finally, we can check that $w_\epsilon = x\frac{\partial u_\epsilon}{\partial x}$ satisfies $w_\epsilon(t=0,x) = -x1_{\{x<K\}}$ and for $t\in(0,T]$, $x>0$,

$$\begin{aligned}\frac{\partial w_\epsilon}{\partial t} - \frac{\partial}{\partial x}\left(\frac{\eta x^2}{2}\frac{\partial w_\epsilon}{\partial x}\right) + (\eta - r)x\frac{\partial w_\epsilon}{\partial x} + \left(r + \frac{x}{2}\frac{\partial \eta}{\partial x}\right)w_\epsilon \\ -rK1_{\{x<K\}}\mathcal{V}'_\epsilon(u_\epsilon - u_\circ)(w_\epsilon + x) = -rK^2\mathcal{V}_\epsilon(u_\epsilon(K))\delta_{x=K},\end{aligned} \tag{2.20}$$

so $x\frac{\partial u_\epsilon}{\partial x} \in L^2(0,T;V)\cap C^0([0,T],L^2(\mathbb{R}_+))$, and it is possible to prove that $\|x\frac{\partial u_\epsilon}{\partial x}\|_{L^2(0,T;V)}$, $\|x\frac{\partial u_\epsilon}{\partial x}\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$, and thus $\|x^2\frac{\partial^2 u_\epsilon}{\partial x^2}\|_{L^2((0,T)\times\mathbb{R}_+)}$ are bounded independently of ϵ. Therefore, so is $\|\frac{\partial u_\epsilon}{\partial t}\|_{L^2((0,T)\times\mathbb{R}_+)}$, from (2.10). ∎

THEOREM 2.2. *With η satisfying Assumption* 1*, problem* (2.8) *has a unique solution u which belongs to $C^0([0,T]\times[0,+\infty))$ with $u(t,0) = K$, for all $t\in[0,T]$, and is such that $x\frac{\partial u}{\partial x}, \frac{\partial u}{\partial x} \in L^2(0,T;V)$, $\frac{\partial u}{\partial x}$ and $x\frac{\partial u}{\partial x}$ belong to $C^0([0,T];L^2(\mathbb{R}_+))$, and $\frac{\partial u}{\partial t}\in L^2((0,T)\times\mathbb{R}_+)$. The function u is also greater than u_e, the price of the European put, solution to* (2.3).

The quantities $\|u\|_{L^2(0,T;V)}$, $\|u\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$, $\|x\frac{\partial u}{\partial x}\|_{L^2(0,T;V)}$, $\|\frac{\partial u}{\partial x}\|_{L^2(0,T;V)}$, $\|x\frac{\partial u}{\partial x}\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$, $\|\frac{\partial u}{\partial x}\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$, and $\|\frac{\partial u}{\partial t}\|_{L^2((0,T)\times\mathbb{R}_+)}$ are bounded independently of η in the class defined in Assumption 1.

We have that

$$-1 \leq \frac{\partial u}{\partial x} \leq 0 \quad \forall t\in(0,T],\ \text{a.e. } x>0. \tag{2.21}$$

Proof. The proof consists of passing to the limit in (2.10) as $\epsilon\to 0$ and proving that the solution to (2.8) is unique. We skip it since it is rather classical: it can be found in [19]. The entire sequence u_ϵ converges weakly to u in $L^2(0,T,V)$. All the bounds on u_ϵ hold for u and they are independent of η in the class defined by Assumption 1. The bounds (2.21) are obtained by passing to the limit in (2.12). From the fact that $\frac{\partial u}{\partial t} + A(t)u \in L^2((0,T)\times\mathbb{R}_+)$, we deduce that, for $w = x\frac{\partial u}{\partial x}$,

$$\frac{\partial w}{\partial t} - \frac{\partial}{\partial x}\left(\frac{\eta x^2}{2}\frac{\partial w}{\partial x}\right) + (\eta - r)x\frac{\partial w}{\partial x} + \left(r + \frac{x}{2}\frac{\partial \eta}{\partial x}\right)w \in L^2(0,T;V'),$$

and $w(t=0)\in L^2(\mathbb{R}_+)$. So from classical results on linear parabolic equations [27], we know that $w = x\frac{\partial u}{\partial x} \in C^0([0,T];L^2(\mathbb{R}_+))$. This implies that the function xu is continuous in $[0,T]\times[0,+\infty)$ and the bounds $K - x \leq u(t,x) \leq K$ hold pointwise for $x>0$, so we have $u(t,x=0) = K$ and u is continuous in $[0,T]\times[0,+\infty)$. Therefore the region where $u = u_\circ$ is a closed subset of $[0,T]\times[0,+\infty)$. ∎

PROPOSITION 2.3. *For any $t \geq 0$ and any $x \in [0, +\infty)$,*

$$u_\epsilon(t,x) - \epsilon \leq u(t,x) \leq u_\epsilon(t,x), \tag{2.22}$$

and u_ϵ converges to u in $C^0([0,T] \times [0,+\infty))$.

Proof. Let $\epsilon' < \epsilon$ be two positive numbers, and consider u_ϵ and $u_{\epsilon'}$ the two solutions to (2.10) corresponding to ϵ and ϵ'. Calling $e_{\epsilon,\epsilon'} = u_{\epsilon'} - u_\epsilon$,

$$\begin{aligned} &\frac{\partial e_{\epsilon,\epsilon'}}{\partial t} + A(t) e_{\epsilon,\epsilon'} - rK1_{x<K}(\mathcal{V}_{\epsilon'}(u_{\epsilon'} - u_\circ) - \mathcal{V}_{\epsilon'}(u_\epsilon - u_\circ)) \\ &= - rK1_{x<K}(\mathcal{V}_\epsilon(u_\epsilon - u_\circ) - \mathcal{V}_{\epsilon'}(u_\epsilon - u_\circ)) \leq 0. \end{aligned}$$

Thanks to the nonincreasing character of $\mathcal{V}_{\epsilon'}$, we can use the weak maximum principle and prove that $e_{\epsilon,\epsilon'} \leq 0$, a.e. Then passing to the limit as $\epsilon' \to 0$ (for a converging subsequence), we deduce that $u \leq u_\epsilon$.

Calling v_ϵ the function $u + \epsilon$, we have that $v_\epsilon(t=0) = u_\circ + \epsilon \geq u_\epsilon$ and, with $\mu = \frac{\partial u}{\partial t} + A(t)u$, that

$$\frac{\partial v_\epsilon}{\partial t} + A(t) v_\epsilon - rK1_{x<K}\mathcal{V}_\epsilon(v_\epsilon - u_\circ) = r\epsilon + \mu \geq 0,$$

because $\mathcal{V}_\epsilon(v_\epsilon - u_\circ) = 0$. The maximum principle shows that $v_\epsilon \geq u_\epsilon$. We have proved (2.22) and therefore the convergence of u_ϵ to u in $C^0([0,T] \times [0,+\infty))$. ∎

LEMMA 2.4. *Let u be the solution to (2.8). There exists a function $\gamma : (0,T] \to [0,K)$, such that for all $t \in (0,T)$, $\{x \text{ s.t. } u(t,x) = u_\circ(x)\} = [0, \gamma(t)]$.*

Calling

$$\mu = \frac{\partial u}{\partial t} + A(t)u, \tag{2.23}$$

we have a.e.

$$0 \leq \mu \leq rK1_{\{u=u_\circ\}} = rK1_{\{x \leq \gamma(t)\}}. \tag{2.24}$$

Proof. The function $x\frac{\partial u}{\partial x}$ belongs to $C^0([0,T]; L^2(\mathbb{R}_+))$, and, for any t, we have that $x\frac{\partial u}{\partial x}(t,x) \geq -x$ a.e. in x. Let us prove Lemma 2.4 by contradiction: if the set $\{x; u(t,x) = u_\circ(x)\}$ were not connected, then there would exist an interval where $\frac{\partial u}{\partial x}(t,x) < -1$ a.e., but this contradicts the bound on $x\frac{\partial u}{\partial x}(t,\cdot)$.

On the other hand, we know that $u(t,0) = K = u_\circ(0)$.

This proves that at each time t, the set where $u(t,x)$ coincides with $u_\circ$ is an interval $[0, \gamma(t)]$. Note that $\gamma(t) \leq K$ for $t > 0$ because $u(t,K) \geq u_e(t,K) > 0 = u_\circ(K)$.

With $\mu \in L^2((0,T) \times \mathbb{R}_+)$ given by (2.23), we have $\mu = 0$ a.e. in the open region where $u > u_\circ$ and $\mu = rK$ in the interior of the region where $u = u_\circ$. Now, μ is the weak limit of $rK1_{x<K}\mathcal{V}_\epsilon(u_\epsilon - u_\circ)$ in $L^2((0,T) \times \mathbb{R}_+)$. From (2.22), we deduce that $rK1_{x<K}\mathcal{V}_\epsilon(u_\epsilon - u_\circ) \leq rK1_{x<K}\mathcal{V}_\epsilon(u - u_\circ)$, and $1_{x<K}\mathcal{V}_\epsilon(u - u_\circ)$ converges pointwise to $1_{\{u=u_\circ\}}$. Therefore, $\mu \leq rK1_{\{u=u_\circ\}}$. ∎

The free boundary $\{x = \gamma(t),\ t \in (0,T)\}$ is called the exercise boundary.

Remark 2. Consider a sequence of penalty parameters $(\epsilon_n)_{n\in\mathbb{N}}$ with $\epsilon_n \to 0$ as $n \to +\infty$, a function η, and a sequence of functions $(\eta_n)_{n\in\mathbb{N}}$ in the class defined by Assumption 1 such that $\|\eta_n - \eta\|_{L^\infty((0,T)\times\mathbb{R}_+)} \to 0$ as $n \to +\infty$. Then it can be

proved (the proof is exactly the same as that of Theorem 2.2) that the solution u_n of the problem

$$\frac{\partial u_n}{\partial t} - \frac{\eta_n x^2}{2}\frac{\partial^2 u_n}{\partial x^2} - rx\frac{\partial u_n}{\partial x} + ru_n - rK1_{\{x<K\}}\mathcal{V}_{\epsilon_n}(u_n - u_\circ) = 0, \quad \begin{cases} t \in (0,T], \\ x > 0, \end{cases}$$

$$u_n(t=0,x) = u_\circ(x)$$

converges weakly in $L^2(0,T;V)$ to the solution u of (2.1), and that $\frac{\partial u_n}{\partial x}$ (resp., $x\frac{\partial u_n}{\partial x}$) converges weakly to $\frac{\partial u}{\partial x}$ (resp., $x\frac{\partial u}{\partial x}$) in $L^2(0,T;V)$. We also have that $\frac{\partial u_n}{\partial t}$ converges weakly to $\frac{\partial u}{\partial t}$ in $L^2((0,T)\times\mathbb{R}_+)$.

LEMMA 2.5. *There exists a real number $q_{\max} > 2$ such that, for any real q, $2 \le q < q_{\max}$, and for any interval (x_1,x_2), $0 < x_1 < x_2 < +\infty$, there exists a constant C independent of η in the class defined by Assumption* 1 *such that*

$$\left\|\frac{\partial u}{\partial t}\right\|_{L^q((0,T)\times(x_1,x_2))} + \left\|\frac{\partial^2 u}{\partial x^2}\right\|_{L^q((0,T)\times(x_1,x_2))} \le C. \tag{2.25}$$

Proof (sketch of the proof). There exists $q_{\max} > 2$ such that, for any real q, $2 \le q < q_{\max}$, $u(t=0) \in W^{2-\frac{2}{q},q}(\mathbb{R}_+)$. Moreover, from (2.24), the left-hand side of (2.23) belongs to $L^\infty((0,T)\times\mathbb{R}_+)$. Then the result follows by applying locally [20, Theorem 9.1, page 340]. ∎

3. The free boundary. With stronger assumptions on η, we can prove some additional regularity, thanks to the convexity of the penalty function $\mathcal{V}_\epsilon$.

Assumption 2. We assume for the constant M introduced in Assumption 1 that we also have

$$\left|x^2\frac{\partial^2\eta}{\partial x^2}\right| + \left|\frac{\partial\eta}{\partial t}\right| + \left|x\frac{\partial^2\eta}{\partial x\partial t}\right| \le M \quad \textit{a.e. in } (0,T)\times\mathbb{R}_+.$$

LEMMA 3.1. *Under Assumptions* 1 *and* 2 *(in fact it is enough to assume that $|x^2\frac{\partial^2\eta}{\partial x^2}| \le M$ for a given constant M), the solution of* (2.8) *satisfies $\frac{\partial^2 u}{\partial x^2} \ge 0$ a.e.*

Proof. Calling u_ϵ the solution to (2.10) and $z_\epsilon = \frac{\partial u_\epsilon}{\partial x}$, we recall that $z_\epsilon = \tilde{z}_\epsilon + \hat{z}_\epsilon$, where $\lim_{\epsilon\to 0}\|\hat{z}_\epsilon\|_{L^2(0,T;V)} = 0$ and where $\tilde{z}_\epsilon$ satisfies (2.18). We derive (2.18) with respect to x. Setting $y_\epsilon = \frac{\partial\tilde{z}_\epsilon}{\partial x}$, we have $y_\epsilon(t=0,x) = \delta_{x=K}$, and for $t \in (0,T]$ and $x > 0$,

$$\begin{aligned} &\frac{\partial y_\epsilon}{\partial t} - \frac{\partial^2}{\partial x^2}\left(\frac{\eta x^2}{2}y_\epsilon\right) - \frac{\partial}{\partial x}(rxy_\epsilon) - rK1_{\{x<K\}}\mathcal{V}'_\epsilon(u_\epsilon - u_\circ)y_\epsilon \\ &= rK1_{\{x<K\}}\mathcal{V}''_\epsilon(u_\epsilon - u_\circ)(\tilde{z}_\epsilon + 1)^2 - rK\mathcal{V}'_\epsilon(u_\epsilon - u_\circ)(\tilde{z}_\epsilon + 1)\delta_{x=K}. \end{aligned} \tag{3.1}$$

From Assumptions 1 and 2, the monotonicity and the convexity of $\mathcal{V}_\epsilon$, and the fact that $\tilde{z}_\epsilon + 1 \ge 0$, we obtain that $y_\epsilon \ge 0$. On the other hand, we know that $x^2 y_\epsilon$ converges weakly to $x^2\frac{\partial^2 u}{\partial x^2}$ in $L^2((0,T)\times\mathbb{R}_+)$. By passing to the limit, we obtain that $\frac{\partial^2 u}{\partial x^2} \ge 0$ in the sense of distributions and a.e. since $x^2\frac{\partial^2 u}{\partial x^2} \in L^2((0,T)\times\mathbb{R}_+)$. ∎

LEMMA 3.2. *Under Assumptions* 1 *and* 2, *there exists a constant C_1 such that the solution u to* (2.8) *satisfies, for any τ, $0 < \tau \le T$, $\|\frac{\partial u}{\partial t}\|_{L^\infty([\tau,T];L^2(\mathbb{R}_+))} \le \frac{C_1}{\sqrt{\tau}}$ and $\|\frac{\partial u}{\partial t}\|_{L^2([\tau,T];V)} \le \frac{C_1}{\sqrt{\tau}}$.*

Proof. By calling y_ϵ the time derivative of u_ϵ and deriving (2.10) with respect to t, we obtain that

$$\frac{\partial y_\epsilon}{\partial t} - \frac{\eta x^2}{2}\frac{\partial^2 y_\epsilon}{\partial x^2} - rx\frac{\partial y_\epsilon}{\partial x} + ry_\epsilon - rK1_{\{x<K\}}\mathcal{V}'_\epsilon(u_\epsilon - u_\circ)y_\epsilon = \frac{x^2}{2}\frac{\partial \eta}{\partial t}\frac{\partial^2 u_\epsilon}{\partial x^2}. \tag{3.2}$$

We can check that there exists a constant C independent of ϵ such that, for all $\tau > s > 0$,

$$\int_{\mathbb{R}_+} y_\epsilon(\tau,x)^2 dx + \int_s^\tau \|y_\epsilon(t)\|_V^2 dt \le C\left(\int_{\mathbb{R}_+} y_\epsilon(s,x)^2 dx + 1\right) \tag{3.3}$$

because $\|\frac{\partial \eta}{\partial t}\|_{L^\infty} \le M$, $x^2\frac{\partial^2 u_\epsilon}{\partial x^2}$ is bounded independently of ϵ, and $\mathcal{V}'_\epsilon \le 0$. Integrating with respect to s between 0 and τ, and using the fact that $\|y_\epsilon\|_{L^2((0,T)\times\mathbb{R}_+)}$ is bounded independently of ϵ leads to $\int_{\mathbb{R}_+} y_\epsilon(\tau,x)^2 dx \le \frac{C}{\tau}$, and using again (3.3), $\int_s^T \|y_\epsilon(t)\|_V^2 dt \le \frac{C}{s}$. The same estimates hold for $\frac{\partial u}{\partial t}$, by passing to the limit. ∎

We can also prove that γ is bounded from below by a positive constant depending only on $\bar\eta$. For that, we use a preliminary lemma.

LEMMA 3.3. *For two positive constants $\tilde\eta$ and K, call $\tilde A_t$ the operator $\tilde A_t v = -\frac{\tilde\eta x^2}{2}\frac{\partial^2 v}{\partial x^2} - rx\frac{\partial v}{\partial x} + rv$, and let $u^{\tilde\eta,K}$ be the solution of*

$$\begin{aligned}
&\frac{\partial \tilde u^{\tilde\eta,K}}{\partial t} + \tilde A_t u^{\tilde\eta,K} \ge 0, \qquad \tilde u^{\tilde\eta,K}(t,x) \ge (K-x)_+, \quad t\in(0,T],\ x>0,\\
&(\tfrac{\partial \tilde u^{\tilde\eta,K}}{\partial t} + \tilde A_t u^{\tilde\eta,K})(\tilde u^{\tilde\eta,K} - (K-x)_+) = 0, \quad t\in(0,T],\ x>0,\\
&\tilde u^{\tilde\eta,K}(t=0,x) = (K-x)_+.
\end{aligned} \tag{3.4}$$

The function $\tilde u^{\tilde\eta,K}$ is a nondecreasing function of time and satisfies $\frac{\partial^2 \tilde u^{\tilde\eta,K}}{\partial x^2} \ge 0$. The function $\tilde\gamma : [0,T] \mapsto [0,K]$ such that $\{x>0 : \tilde u^{\tilde\eta,K}(t,x) = (K-x)_+\} = [0,\tilde\gamma]$ is nonincreasing and continuous and

$$\frac{\partial \tilde u^{\tilde\eta,K}}{\partial t} + \tilde A_t u^{\tilde\eta,K} = rK1_{\{\tilde u^{\tilde\eta,K} = (K-x)_+\}}. \tag{3.5}$$

Proof. Since the coefficient $\tilde\eta$ is constant, we know that $\frac{\partial^2 \tilde u^{\tilde\eta,K}}{\partial x^2} \ge 0$ (from Lemma 3.1). We can also prove that $\tilde u^{\tilde\eta,K}$ is a nondecreasing function of time (one checks first the property on the corresponding penalized problem (analogue to (2.10)) and then passes to the limit). Then, obviously, $\tilde\gamma$ is a nonincreasing function of time. Also, as proved in [19, page 288], $\tilde\gamma$ is continuous. This implies that the boundary of the set $\{t>0, 0\le x\le\gamma(t)\}$ is the graph of $\tilde\gamma$. In other words, the set $\partial\{\tilde u^{\tilde\eta,K} = (K-x)_+\} \cap \{t>0\}$ is the graph of $\tilde\gamma$ so it is measurable and has zero measure. Therefore, $\frac{\partial \tilde u^{\tilde\eta,K}}{\partial t} + \tilde A_t \tilde u^{\tilde\eta,K} = rK1_{x\le\tilde\gamma(t)}$ in $L^2((0,T)\times\mathbb{R}_+)$. ∎

THEOREM 3.4. *Under Assumption 1, there exists $\gamma_0 > 0$ depending only on $\bar\eta$ such that*

$$\gamma(t) \ge \gamma_0 \quad \forall t\in[0,T]. \tag{3.6}$$

Proof. Let $\tilde u = \tilde u^{\bar\eta,K}$ as defined in Lemma 3.3. We have $u \le \tilde u$. Indeed, calling $e = \tilde u - u$,

$$\begin{aligned}
&\frac{\partial e}{\partial t} - \frac{\eta(t,x)x^2}{2}\frac{\partial^2 e}{\partial x^2} - rx\frac{\partial e}{\partial x} + re - rK\left(1_{\{\tilde u\le u_\circ\}} - 1_{\{u\le u_\circ\}}\right)\\
&= \left(\frac{x^2}{2}(\bar\eta-\eta)\frac{\partial^2 \tilde u}{\partial x^2}\right) + \left(rK(1_{u\le u_\circ}) - \mu\right).
\end{aligned}$$

The two terms in the right-hand side are nonnegative, thanks to Lemma 3.1 and to (2.24). Therefore, a weak maximum principle can be applied, thanks to the monotonicity of $z \mapsto 1_{\{z \le u_\circ(x)\}}$, and we see that $e_- = 0$ (everywhere since u and $\tilde{u}$ are continuous).

Since $u \le \tilde{u}$, we know that $\tilde{\gamma} \le \gamma$. Therefore, if there exists $t_0 < T$ such that $\gamma(t_0) = 0$, then $\tilde{\gamma}(t_0) = 0$ and $\tilde{\gamma}(t) = 0$, for $t \ge t_0$, because $\tilde{\gamma}$ is nonincreasing from Lemma 3.3. It follows that $\tilde{u}$ solves the Black–Scholes equation for $t > t_0$, i.e., $\frac{\partial \tilde{u}}{\partial t} - \frac{\bar{\eta} x^2}{2} \frac{\partial^2 \tilde{u}}{\partial x^2} - rx \frac{\partial \tilde{u}}{\partial x} + r\tilde{u} = 0$ and $\tilde{u}(t_0) \le K$: the maximum principle indicates that $\tilde{u}(t) \le K e^{-r(t-t_0)}$ for $t > t_0$. This is in contradiction to the fact that $\tilde{u} \ge u_\circ$. The assertion $\gamma(T) = 0$ is also impossible, because we can always look for $\tilde{u}$ in a larger time interval. Since $\tilde{\gamma}$ is continuous on $[0, T]$, there exists $\gamma_0 > 0$ such that (3.6) is satisfied. ∎

The next results deal with the regularity of the function γ.

LEMMA 3.5. *Under Assumption* 1*, the function* γ *defined in Lemma* 2.4 *is such that for all* $t > 0$*,* $\limsup_{\tau \to t} \gamma(\tau) \le \gamma(t)$*; i.e.,* γ *is upper semicontinuous.*

Proof. The intersection of the epigraph of $t \mapsto -\gamma(t)$ with $[0, T] \times \mathbb{R}_-$ is the region $\{(t, x) \in [0, T] \times \mathbb{R}_-;\, u(t, -x) = u_\circ(-x)\}$. From the continuity of u, this region is closed. This and Theorem 3.4 imply that γ is upper semicontinuous. ∎

LEMMA 3.6. *Under Assumption* 1*, the function* γ *defined in Lemma* 2.4 *is such that for all* $t > 0$*,*

$$\liminf_{\tau \to t^+} \gamma(\tau) = \gamma(t), \quad \textit{and} \quad \liminf_{\tau \to t^-} \gamma(\tau) = \limsup_{\tau \to t^-} \gamma(\tau) \le \gamma(t).$$

Proof. The idea is to compare u for $\tau > t$ with $K - \gamma(t) + \tilde{u}$, where $\tilde{u}(\tau, x) = \tilde{u}^{\bar{\eta}, \gamma(t)}(\tau - t, x)$ and $\tilde{u}^{\bar{\eta}, \gamma(t)}$ is defined in Lemma 3.3. We have from Lemma 3.3 that $\frac{\partial \tilde{u}}{\partial \tau} - \bar{\eta} \frac{x^2}{2} \frac{\partial^2 \tilde{u}}{\partial x^2} - rx \frac{\partial \tilde{u}}{\partial x} + r\tilde{u} = r\gamma(t) 1_{\{\tilde{u}(\tau, x) = (\gamma(t) - x)_+\}}$. Therefore,

$$\begin{aligned} & \frac{\partial \tilde{u}}{\partial \tau} - \frac{\eta x^2}{2} \frac{\partial^2 \tilde{u}}{\partial x^2} - rx \frac{\partial \tilde{u}}{\partial x} + r\tilde{u} \\ = \;& (\bar{\eta} - \eta) \frac{x^2}{2} \frac{\partial^2 \tilde{u}}{\partial x^2} + r\gamma(t) 1_{\{K - \gamma(t) + \tilde{u} = K - \gamma(t) + (\gamma(t) - x)_+\}}, \end{aligned} \tag{3.7}$$

which implies that the error $e = K - \gamma(t) + \tilde{u} - u$ satisfies

$$\begin{aligned} & \frac{\partial e}{\partial \tau} - \eta \frac{x^2}{2} \frac{\partial^2 e}{\partial x^2} - rx \frac{\partial e}{\partial x} + re - rK(1_{\{u+e \le u_\circ\}} - 1_{\{u \le u_\circ\}}) \\ & = r(K - \gamma(t)) 1_{\{u+e > u_\circ\}} + (\bar{\eta} - \eta) \frac{x^2}{2} \frac{\partial^2 \tilde{u}}{\partial x^2} + (rK 1_{\{u \le u_\circ\}} - \mu) \\ & + r\gamma(t) \left(1_{\{u+e \le K - \gamma(t) + (\gamma(t) - x)_+\}} - 1_{\{u+e \le u_\circ\}} \right). \end{aligned} \tag{3.8}$$

From Lemma 3.3 and $\bar{\eta} \ge \eta$, all the terms in the right-hand side of (3.8) are nonnegative, so the maximum principle indicates that $0 \le e$ for $\tau \ge t$. This implies that $\tilde{\gamma}(\tau) \le \gamma(\tau)$ for $\tau > t$.

From $\tilde{\gamma}(\tau) \le \gamma(\tau)$, we deduce that $\lim_{\tau \to t^+} \tilde{\gamma}(\tau) \le \liminf_{\tau \to t^+} \gamma(\tau)$.

But $\lim_{\tau \to t^+} \tilde{\gamma}(\tau) = \gamma(t)$. Therefore $\liminf_{\tau \to t^+} \gamma(\tau) \ge \gamma(t)$, and using Lemma 3.5, we obtain that $\liminf_{\tau \to t^+} \gamma(\tau) = \gamma(t)$. The idea for proving that $\liminf_{\tau \to t^-} \gamma(\tau) = \limsup_{\tau \to t^-} \gamma(\tau)$ is somewhat similar: assume that $s < \limsup_{\tau \to t^-} \gamma(\tau)$, so there exists a sequence of times t_k smaller than t, converging to t and such that $\gamma(t_k) > s$. We construct $\tilde{u}_k(\tau, x) = K - s + \tilde{u}^{\bar{\eta}, s}(\tau - t_k, x)$, where $\tilde{u}^{\bar{\eta}, s}$ is defined in Lemma 3.3, and we define $\tilde{\gamma}_k$ by $\{x : \tilde{u}_k(\tau, x) = K - s + (s - x)_+\} = [0, \tilde{\gamma}_k(\tau)]$. As above,

$\tilde{\gamma}_k$ is a nonincreasing and continuous function of time. From the continuity of $\tilde{\gamma}_k$, we deduce that

$$\lim_{k\to\infty} \inf_{t_k\le\tau\le t} \tilde{\gamma}_k(\tau) = s.$$

The same arguments as above show that for $t_k \le \tau \le t$, $\tilde{u}_k(\tau, x) \ge u(\tau)$ and $\gamma(\tau) \ge \tilde{\gamma}_k(\tau)$. Therefore $\liminf_{\tau\to t_-} \gamma(\tau) \ge \lim_{k\to\infty} \inf_{t_k\le\tau\le t} \tilde{\gamma}_k(\tau) = s$, which shows that $\liminf_{\tau\to t^-} \gamma(\tau) = \limsup_{\tau\to t^-} \gamma(\tau)$. ∎

Remark 3. Lemmas 3.5 and 3.6 indicate that under Assumption 1, the function γ is right continuous in $[0, T)$ and, for each $t \in (0, T]$, γ has a left-limit at t. As a consequence, we have the following theorem.

THEOREM 3.7. *For η satisfying Assumption* 1*, the function $\mu = \frac{\partial u}{\partial t} + A_t u$ is*

$$\mu = rK1_{\{u=u_\circ\}} = rK1_{\{x\le\gamma(t)\}}. \tag{3.9}$$

In other words, a.e. one of the two conditions $u = u_\circ$ and $\mu = 0$ is not satisfied: we see that there is strict complementarity in (2.1).

Proof. For any time t, both $\lim_{\tau<t} \gamma(\tau)$ and $\lim_{\tau>t} \gamma(\tau)$ exist. Therefore, the function γ is the uniform limit of a sequence of piecewise constant functions γ_k (i.e., γ_k is constant on a finite number of intervals). Thus, calling $\mathcal{J}$ (resp., $\mathcal{J}_k$) the set of points where γ (resp., γ_k) jumps, we have $\mathcal{J} \subset \cup_{k\in\mathbb{N}}\mathcal{J}_k$ because of the uniform convergence and $\mathcal{J}_k$ is finite. Thus the set $\mathcal{J}$ is countable.

Consider now the boundary Γ of the set $\{u = u_\circ\} = \{x \le \gamma(t), t \in [0, T]\}$: we have

$$\Gamma = (\Gamma \cap \{(x, t), t \in [0, T]\backslash\mathcal{J}\}) \cup (\Gamma \cap \{(x, t), t \in \mathcal{J}\}).$$

The second set is negligible, since $\mathcal{J}$ is countable. For the first set, we have

$$\Gamma \cap \{(x, t), t \in [0, T]\backslash\mathcal{J}\} = \{(\gamma(t), t), t \in [0, T]\backslash\mathcal{J}\},$$

so it is also negligible.

Therefore, Γ is negligible, and the set $\{u = u_\circ\}$ has the same measure as its interior, on which $\mu = rK$. This proves (3.9). ∎

PROPOSITION 3.8. *Under Assumption* 1*, and with μ defined in* (2.23)*, the sequence*

$$\mu_\epsilon = rK1_{x<K}\mathcal{V}_\epsilon(u_\epsilon - u_\circ), \quad \text{with } u_\epsilon \text{ solution to } (2.10),$$

converges to μ in $L^p((0, T) \times \mathbb{R}_+)$ as $\epsilon \to 0$, for any real number $p \ge 1$.

Proof. We know that the functions μ_ϵ satisfy $0 \le \mu_\epsilon \le rK$ and are supported in $[0, T] \times [0, K]$. We also know that the sequence μ_ϵ converges weakly to μ in L^2. But μ takes only the two values 0 and rK. This implies easily that $\mu_\epsilon \to \mu$ in L^1, and in every $L^p((0, T) \times \mathbb{R}_+)$, $1 \le p < +\infty$, because $\|\mu_\epsilon\|_{L^\infty} \le rK$ and $\|\mu\|_{L^\infty} \le rK$ and the functions are supported in $[0, T] \times [0, K]$. ∎

PROPOSITION 3.9. *Under Assumption* 1*, the sequence $x\frac{\partial u_\epsilon}{\partial x}$ (u_ϵ solution to* (2.10)*) converges to $x\frac{\partial u}{\partial x}$ in $C^0([0, T], L^2(\mathbb{R}_+))$.*

Proof. Calling $w_\epsilon = x\frac{\partial u_\epsilon}{\partial x}$, we have

$$\frac{\partial w_\epsilon}{\partial t} - \frac{\partial}{\partial x}\left(\frac{\eta x^2}{2}\frac{\partial w_\epsilon}{\partial x}\right) + (\eta - r)x\frac{\partial w_\epsilon}{\partial x} + \left(r + \frac{x}{2}\frac{\partial \eta}{\partial x}\right) w_\epsilon = x\frac{\partial \mu_\epsilon}{\partial x},$$

and $w_\epsilon(t = 0) = x\frac{\partial u_\circ}{\partial x} \in L^2(\mathbb{R}_+)$. From Proposition 3.8, $x\frac{\partial \mu_\epsilon}{\partial x}$ converges to $x\frac{\partial \mu}{\partial x}$ strongly in $L^2(0, T; V')$. This implies that $w_\epsilon \to w$ in $C^0([0, T], L^2(\mathbb{R}_+))$, where $w = x\frac{\partial u}{\partial x}$. ∎

It is possible to prove that the function γ is continuous, under stronger assumptions on η as follows.

LEMMA 3.10. *Under Assumptions* 1 *and* 2, *the function* γ *defined in Lemma* 2.4 *is such that* $\lim_{\tau\to t^-}\gamma(\tau)=\gamma(t)$.

Proof. Assume that $s=\lim_{\tau\to t_-}\gamma(\tau)<\gamma(t)$. Then there exists a positive number ϵ such that $u(\tau,x)>u_\circ(x)$ in the region $Q=[t-\epsilon,t)\times(\tilde{s},\gamma(t))$, where $\tilde{s}=\max_{t-\epsilon\le\tau<t}\gamma(\tau)=\gamma(\tilde{\tau})$ (this maximum is attained at $t-\epsilon\le\tilde{\tau}<t$ because γ is upper semicontinuous). In this region, the function satisfies the parabolic PDE

$$\frac{\partial u}{\partial t}+A(t)u=0, \tag{3.10}$$

and u is a regular function. Therefore, we can derive the equation twice with respect to x and we obtain that $y=\frac{\partial^2 u}{\partial x^2}$ satisfies

$$\frac{\partial y}{\partial t}-\frac{\eta x^2}{2}\frac{\partial^2 y}{\partial x^2}+\left(2x\eta+x^2\frac{\partial\eta}{\partial x}-rx\right)\frac{\partial y}{\partial x}-\left(r+\frac{x^2}{2}\frac{\partial^2\eta}{\partial x^2}+2x\frac{\partial\eta}{\partial x}+\eta\right)y=0. \tag{3.11}$$

We have that $y\ge 0$ on the parabolic boundary of Q, whereas $y=0$ on $\tau=t$. The strong maximum principle [30] implies that $y=0$ in Q, so $u(\tau,\cdot)$ is an affine function in Q with respect to the variable x. We can look for $u(\tau,x)$ as $u(\tau,x)=a(\tau)x+\beta(\tau)$. Plugging this into (3.10), we obtain easily that $a(\tau)=-1$ and $b(\tau)=Ke^{r(t-\tau)}$. This means that $u(\tilde{\tau},\tilde{s})=Ke^{r(t-\tilde{\tau})}-\tilde{s}$, so u jumps across the graph of γ, which is in contradiction with the continuity of u. □

From Lemmas 3.5, 3.6, and 3.10, we have proved the following theorem.

THEOREM 3.11. *Under Assumptions* 1 *and* 2, *the function* γ *defined in Lemma* 2.4 *is continuous. The function* μ *defined by* (2.23) *belongs to* $C^0([0,T];L^p((0,T)\times\mathbb{R}_+))$, $1\le p<+\infty$, *and* $\frac{\partial u}{\partial t}\in C^0([0,T];V')$. We define by Γ the free boundary

$$\Gamma=\{(t,\gamma(t)),0<t\le T\} \tag{3.12}$$

and by Ω^+ the set

$$\Omega^+=\{(t,x),x>\gamma(t),0<t\le T\}=\{(t,x)\in(0,T]\times\mathbb{R}_+,u(t,x)>u_\circ(x)\}. \tag{3.13}$$

Remark 4. It is possible, by extending the results of Lamberton [24], to prove that the function $\gamma-K$ is more singular than $\sqrt{t}$ as t tends to 0. Therefore, the function γ is not Lipschitz continuous in $[0,T]$. However, depending on η, γ may be Lipschitz continuous in $[t,T]$, for all $t>0$. In particular, if η is a nondecreasing function of time, it is possible to prove such a regularity by adapting the proofs in [19] for the Stefan problem. Under some additional regularity assumptions on Γ and u, it is possible to prove that $\frac{\partial^2 u}{\partial x^2}$ stays bounded away from zero in $\Omega_t^+\cap\{x<K\}$, for $t>0$.

PROPOSITION 3.12. *Assume that the solution* u *of* (2.1) *is such that for all times* t, $0<t\le T$,

- $\gamma|_{[t,T]}$ *is a Lipschitz function;*
- $\frac{\partial u}{\partial t}\in H^1(\Omega_t^+)$, *where* Ω_t^+ *is the region* $\Omega_t^+=\{u(\tau,x)>u_0,\tau\ge t\}$.

In this case, for t, $0<t<T$, *there exists a positive constant* $m_t>0$ *such that* $\frac{\partial^2 u}{\partial x^2}\ge m_t$ *in the region* $\Omega_t^+\cap\{x<K\}$.

Proof. We take $t>0$ and we call $\Gamma_t=\Gamma\cap\{\tau>t\}$. We know from Theorem 2.2 and Lemma 3.2 that $x\frac{\partial u}{\partial x}\in H^1((t,T)\times\mathbb{R}_+)$, and from Theorem 3.4 that $\frac{\partial u}{\partial x}$ does not jump across Γ_t, so $\frac{\partial u}{\partial x}|_{\Gamma_t}=-1$.

The assumptions also ensure that $\frac{\partial u}{\partial t}|_{\Omega_t^+}$ has a trace on Γ_t. From the fact that $u(\tau,\gamma(\tau)) = K - \gamma(\tau)$ and the Lipschitz regularity of γ, we obtain that

$$\frac{\partial u}{\partial t}(\tau,\gamma(\tau)) + \frac{\partial u}{\partial x}(\tau,\gamma(\tau))\frac{d\gamma}{dt}(\tau) = -\frac{d\gamma}{dt}(\tau) \quad \text{a.e. } \tau > t,$$

which yields $\frac{\partial u}{\partial t}|_{\Gamma_t} = 0$. From these results and the PDE satisfied by u in the region Ω_t^+, we obtain that the trace of $\eta x^2 \frac{\partial^2 u}{\partial x^2})|_{\Omega_t^+}$ on Γ_t is rK.

The function $y = \frac{\partial^2 u}{\partial x^2}$ satisfies the parabolic PDE (3.11) in Ω_+. The coefficients of the PDE are bounded thanks to Assumptions 1 and 2, and the operator $-\frac{\eta x^2}{2}\frac{\partial^2 y}{\partial x^2}$ is uniformly elliptic in Ω_t^+ thanks to Assumption 1 and Theorem 3.4. We also know from Lemma 3.1 that $y \geq 0$. So it is possible to use the strong maximum principle for y in Ω_t^+ (see [30, page 168]): if there exists a point $(t_1,\xi) \in \Omega^+$ such that $y = 0$, then for all $t, 0 < t \leq t_1$, $y = 0$ in $\Omega_t^+ \cap \{\tau \leq t_1\}$. But this is in contradiction with the fact that $y|_{\Gamma_t} = rK$. Therefore, for $t > 0$, the infimum of y in $\Omega_t^+ \cap \{x \leq K\}$ is positive. ∎

4. Sensitivity analysis. Here, we aim at understanding the sensitivity of u with respect to variations of η. For η satisfying Assumption 1, we call $u(\eta)$ the solution to problem (2.8), with $A = A_\eta$ given by (2.5).

PROPOSITION 4.1. *There exists a positive constant C such that, for η and η' satisfying Assumption* 1*,*

(4.1)
$$\|u(\eta) - u(\eta')\|_{L^2(0,T;V)} + \|u(\eta) - u(\eta')\|_{L^\infty(0,T;L^2(\mathbb{R}_+))} \leq C\|\eta - \eta'\|_{L^\infty((0,T)\times\mathbb{R}_+)}$$

and

(4.2)
$$\int_0^T \int_0^K 1_{\{u(\eta)=u_\circ\}}(u(\eta') - u_\circ) + 1_{\{u(\eta')=u_\circ\}}(u(\eta) - u_\circ)dtdx \leq C\|\eta - \eta'\|^2_{L^\infty((0,T)\times\mathbb{R}_+)}.$$

Proof. For all $v \in L^2(0,T,V)$,

$$\begin{aligned}
\left\langle \frac{\partial u(\eta)}{\partial t} + A_\eta(t)u(\eta), v \right\rangle &= rK\langle 1_{\{u(\eta)=u_\circ\}}, v\rangle, \\
\left\langle \frac{\partial u(\eta')}{\partial t} + A_{\eta'}(t)u(\eta'), v \right\rangle &= rK\langle 1_{\{u(\eta')=u_\circ\}}, v\rangle.
\end{aligned}$$

Calling $\delta u = u(\eta) - u(\eta')$, we subtract the two equations above and take $\delta u\, e^{-2\alpha t}$ as a test function, where α is the constant appearing in (2.6):

$$\begin{aligned}
&\left\langle \frac{\partial \delta u}{\partial t} + A_\eta(t)\delta u, \delta u e^{-2\alpha t} \right\rangle \\
&+ rK\int_0^T \int_0^K \left(1_{\{u(\eta)=u_\circ\}}(u(\eta') - u_\circ) + 1_{\{u(\eta')=u_\circ\}}(u(\eta) - u_\circ)\right) e^{-2\alpha t}dtdx \\
&= \int_0^T \langle (A_{\eta'}(t) - A_\eta(t))u(\eta'), \delta u e^{-2\alpha t}\rangle dt.
\end{aligned} \tag{4.3}$$

This implies that, for all $t \in [0,T]$,

$$\begin{aligned}(4.4)\quad &\frac{1}{2}\|\delta u(t)e^{-\alpha t}\|^2_{L^2(\mathbb{R}_+)} + \int_0^T \langle A_\eta(s)\delta u, \delta u\rangle e^{-2\alpha s}ds + \alpha\|\delta u e^{-\alpha t}\|^2_{L^2((0,T)\times\mathbb{R}_+)}\\ &\quad + rK\int_0^T\int_0^K \left(1_{\{u(\eta)=u_\circ\}}(u(\eta')-u_\circ) + 1_{\{u(\eta')=u_\circ\}}(u(\eta)-u_\circ)\right)e^{-2\alpha s}dsdx\\ &\le \int_0^T \langle (A_{\eta'}(s)-A_\eta(s))u(\eta'), \delta u e^{-2\alpha s}\rangle ds\end{aligned}$$

because $\delta u(t=0)=0$. From the Gårding inequality (2.6), we deduce from (4.4) that for all $t \in (0,T]$,

$$\begin{aligned}(4.5)\quad &\frac{1}{2}\|\delta u(t)\|^2_{L^2(\mathbb{R}_+)} + \frac{1}{4}\underline{\eta}\|\delta u\|^2_{L^2(0,T;V)}\\ &\quad + rK\int_0^T\int_0^K \left(1_{\{u(\eta)=u_\circ\}}(u(\eta')-g) + 1_{\{u(\eta')=u_\circ\}}(u(\eta)-u_\circ)\right)dsdx\\ &\le e^{2\alpha T}\int_0^T \langle (A_{\eta'}(s)-A_\eta(s))u(\eta'), \delta u e^{-2\alpha s}\rangle ds.\end{aligned}$$

But

$$\begin{aligned}(4.6)\quad &\int_0^T \left|\langle (A_{\eta'}(s)-A_\eta(s))u(\eta'), \delta u e^{-2\alpha s}\rangle\right| ds\\ &\le \|\eta-\eta'\|_{L^\infty((0,T)\times\mathbb{R}_+)}\left\|x^2\frac{\partial^2 u(\eta')}{\partial x^2}\right\|_{L^2((0,T)\times\mathbb{R}_+)}\|\delta u\|_{L^2((0,T)\times\mathbb{R}_+)}.\end{aligned}$$

Note also that

$$(4.7)\qquad \int_0^T\int_0^K \left(1_{\{u(\eta)=u_\circ\}}(u(\eta')-u_\circ) + 1_{\{u(\eta')=u_\circ\}}(u(\eta)-u_\circ)\right)dsdx \ge 0.$$

The bound (4.1) follows from (4.5), (4.7), and (4.6). The bound (4.2) follows from (4.5), (4.1), and (4.6). □

PROPOSITION 4.2. *Let η and $(\eta_k)_{k\in\mathbb{N}}$ be, respectively, a function and a sequence of functions satisfying Assumption* 1*, such that*

$$(4.8)\qquad \lim_{k\to+\infty}\|\eta_k-\eta\|_{L^\infty((0,T)\times\mathbb{R}_+)} = 0.$$

Then $1_{\{u(\eta_k)=u_\circ\}}$ tends to $1_{\{u(\eta)=u_\circ\}}$ strongly in $L^p((0,T)\times\mathbb{R}_+)$, $1\le p<+\infty$.

Proof. From Proposition 4.1, we know that the sequence $u(\eta_k)$ converges to $u(\eta)$ in $L^2(0,T;V)$ and in $L^\infty(0,T;L^2(\mathbb{R}_+))$. Moreover, we know from Theorem 2.2 that the quantities $\|(x+1)\frac{\partial u(\eta_k)}{\partial x}\|_{L^2(0,T;V)}$, $\|x\frac{\partial u(\eta_k)}{\partial x}\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$, $\|\frac{\partial u(\eta_k)}{\partial t}\|_{L^2((0,T)\times\mathbb{R}_+)}$ are bounded independently of η. Therefore, $u(\eta_k)$ tends to $u(\eta)$ in $L^2(0,T;V)$ and in $L^\infty(0,T;L^2(\mathbb{R}_+))$ strongly, $\frac{\partial u(\eta_k)}{\partial t}$ tends weakly to $\frac{\partial u(\eta)}{\partial t}$ in $L^2((0,T)\times\mathbb{R}_+)$, and $x^2\frac{\partial^2 u(\eta_k)}{\partial x^2}$ tends weakly to $x^2\frac{\partial^2 u(\eta)}{\partial x^2}$ in $L^2((0,T)\times\mathbb{R}_+)$. This implies that $\mu(\eta_k)$ tends weakly to $\mu(\eta)$ in $L^2((0,T)\times\mathbb{R}_+)$, and in every $L^p((0,T)\times\mathbb{R}_+)$, $1\le p<\infty$, because $\mu(\eta_k) = rK1_{\{u(\eta_k)=u_\circ\}}$ is bounded and has a limited support. Thus the sequence $1_{\{u(\eta_k)=u_\circ\}}$ converges weakly in $L^p((0,T)\times\mathbb{R}_+)$, $1\le p<+\infty$, to a function which can take only two values, 0 and 1, and which has a bounded support. Hence, the convergence must be strong. □

PROPOSITION 4.3. *Let η and $(\eta_k)_{k\in\mathbb{N}}$ be, respectively, a function and a sequence of functions satisfying Assumption 1, such that $\lim_{k\to+\infty}\|\eta_k-\eta\|_{L^\infty((0,T)\times\mathbb{R}_+)}=0$. Then for any interval $[x_1,x_2]$, $0<x_1<x_2<+\infty$,*

$$\lim_{k\to+\infty}\|u(\eta_k)-u(\eta)\|_{L^\infty((0,T)\times(x_1,x_2))}=0. \tag{4.9}$$

Proof. Noting $\delta u=u(\eta)-u(\eta_k)$, we have that

$$\frac{\partial\delta u}{\partial t}+A_\eta(t)\delta u=-(\eta-\eta_k)\frac{x^2}{2}\frac{\partial^2u(\eta_k)}{\partial x^2}+rK(1_{\{u(\eta)=u_\circ\}}-1_{\{u(\eta_k)=u_\circ\}}).$$

But from Lemma 2.5, for any compact ω strictly contained in $\mathbb{R}_+$, there exist a real number $q>2$ and a constant C independent of η_k such that

$$\left\|\frac{\partial^2u(\eta_k)}{\partial x^2}\right\|_{L^q((0,T)\times\omega)}\le C.$$

Therefore,

$$\left\|(\eta-\eta_k)\frac{x^2}{2}\frac{\partial^2u(\eta_k)}{\partial x^2}\right\|_{L^q((0,T)\times\omega)}\to 0.$$

From this and (4.8), and by applying locally [20, Theorem 9.1, page 340], we obtain that

$$\lim_{k\to\infty}\left\|\frac{\partial\delta u}{\partial t}\right\|_{L^q((0,T))\times(x_1,x_2)}+\left\|\frac{\partial\delta u}{\partial x}\right\|_{L^q((0,T))\times(x_1,x_2)}+\left\|\frac{\partial^2\delta u}{\partial x^2}\right\|_{L^q((0,T))\times(x_1,x_2)}=0,$$

which implies (4.9) by Sobolev injections. □

Under additional regularity assumptions on Γ and u, it is possible to find estimates on the variation of Γ as follows.

PROPOSITION 4.4. *Let η and η' satisfy Assumptions 1 and 2.*

Assume that the solution $u(\eta)$ of (2.1) is such that, for all times t, $0<t\le T$,

- *$\gamma(\eta)|_{[t,T]}$ is a Lipschitz function;*
- *$\frac{\partial u(\eta)}{\partial t}\in H^1(\Omega_t^+)$, where Ω_t^+ is the region $\Omega_t^+=\{u(\tau,x)>u_0,\tau\ge t\}$.*

In this case, for t, $0<t\le T$, there exists a positive constant $c_t>0$ such that

$$\|(\gamma(\eta')-\gamma(\eta))^+\|^3_{L^3(t,T)}\le c_t\|\eta-\eta'\|^2_{L^\infty((0,T)\times\mathbb{R}_+)}. \tag{4.10}$$

Proof. We know from (4.2) that $\int_0^T\int_0^K 1_{\{u(\eta')=u_\circ\}}(u(\eta)-u_\circ)dtdx\le C\|\eta-\eta'\|^2_{L^\infty((0,T)\times\mathbb{R}_+)}$. But

$$\begin{aligned}&\int_t^T\int_0^K 1_{\{u(\eta')=u_\circ\}}(u(\eta)-u_\circ)d\tau dx\\ =\;&\int_t^T 1_{\{\gamma(\eta')>\gamma(\eta)\}}d\tau\int_{\gamma(\eta)(\tau)}^{\gamma(\eta')(\tau)}dx\int_{\gamma(\eta)(\tau)}^{x}dy\int_{\gamma(\eta)(\tau)}^{y}\frac{\partial^2u}{\partial x^2}(\tau,s)ds.\end{aligned}$$

From Proposition 3.12, there exists a positive constant $m_t>0$ such that $\frac{\partial^2u}{\partial x^2}\ge m_t$ in the region $(t,T]\times[0,K]\cap\{(\tau,x):u(\tau,x)>u_0\}$. Therefore,

$$\int_t^T\int_0^K 1_{\{u(\eta')=u_\circ\}}(u(\eta)-u_\circ)d\tau dx\;\ge\;\frac{m_t}{6}\int_t^T\left((\gamma(\eta')(\tau)-\gamma(\eta)(\tau))^+\right)^3d\tau,$$

which proves the desired result. □

Remark 5. We believe that the estimate

$$\|(\gamma(\eta') - \gamma(\eta))^+\|^3_{L^3(0,T)} \leq c\|\eta - \eta'\|^2_{L^\infty((0,T)\times\mathbb{R}_+)}$$

is true.

5. A first least square problem.

5.1. Description of the problem. Let X be the space

$$X = \left\{\eta \in L^2(\mathbb{R}_+); (x+1)\frac{\partial \eta}{\partial x} \in L^2(\mathbb{R}_+)\right\}. \tag{5.1}$$

We can easily check the Sobolev type embedding $X \subset L^\infty(\mathbb{R}_+) \cap C^0(\mathbb{R}_+)$ with continuous injection. This comes from the facts that $H^1(I) \subset C^0(I)$ for any compact interval and that $x\eta(x) = \int_0^x s\frac{\partial \eta}{\partial x}(s)ds - \int_0^x \eta(s)ds$, and the Cauchy–Schwarz inequality yields that for all $x > 0$, $|\eta(x)|^2 \leq \frac{2}{x}\int_0^x \left(|s\frac{\partial \eta}{\partial x}(s)|^2 + \eta^2(s)\right) ds$. In fact, from this inequality, we deduce that the embedding $X \subset L^\infty(\mathbb{R}_+)$ is also compact, because it is possible to approximate a function $\eta \in X$ by means of a piecewise constant function η_N, with

- $\eta_N(x) = 0$ for $x \geq N$,
- $\eta_N(x) = \frac{1}{N}\int_{i+\frac{j}{N}}^{i+\frac{j+1}{N}} \eta(s)ds$ for $i + \frac{j}{N} \leq x < i + \frac{j+1}{N}$, $0 \leq i,\, j \leq N-1$,

and prove that there exists a positive constant C such that

$$\|\eta - \eta_N\|_{L^\infty(\mathbb{R}_+)} \leq \frac{C}{\sqrt{N}}\|\eta\|_X.$$

We denote by Y the space

$$Y = \left\{\eta : \eta, x\frac{\partial \eta}{\partial x} \in H^1(0,T;X)\right\}. \tag{5.2}$$

We have that $Y \subset \{\eta : \eta, x\frac{\partial \eta}{\partial x} \in L^\infty((0,T)\times\mathbb{R}_+)\}$. We call $\mathcal{H}$ the set of the functions $\eta \in Y$ satisfying Assumption 1. The set $\mathcal{H}$ is a closed and convex subset of Y, compactly embedded in $L^\infty((0,T); W^{1,\infty}(\mathbb{R}_+))$.

Let J_R be a Fréchet differentiable, coercive and convex functional on Y, and let J be a Fréchet differentiable functional defined on $L^2(0,T;V)$.

Remark 6. The discussion below may be generalized to functionals J defined on $C^0((0,T)\times\mathbb{R}_+)$. We are interested in the following minimizing problem:

$$\min_{\substack{\eta\in\mathcal{H}\\ u \text{ satisfies (2.1)}}} \left(J(u) + J_R(\eta)\right). \tag{5.3}$$

For simplifying the discussion below, we make some assumptions on J, but generalizations are possible.

Assumption 3. We assume that $J(u)$ has the form

$$J(u) = \int_{t_1}^{t_2}\int_{x_1}^{x_2} (u(t,x) - u_g(t,x))^2 dtdx, \tag{5.4}$$

where $0 \leq t_1 < t_2 \leq T$ and $0 \leq x_1 < x_2 < \infty$ and $u_g \in L^2((0,T)\times\mathbb{R}_+) \cap L^\infty((0,T)\times\mathbb{R}_+)$. With this special choice,

$$DJ(u) = 2(u - u_g)1_{(t_1,t_2)\times(x_1,x_2)}. \tag{5.5}$$

By using the results of the previous sections, it is fairly easy to prove that (5.3) has solutions.

It is also possible to consider the optimization problem for the penalized problem

$$\min_{\substack{\eta\in\mathcal{H}\\ u \text{ satisfies (2.10)}}} \left(J(u_\epsilon)+J_R(\eta)\right). \tag{5.6}$$

Again, one may prove that this problem has solutions and the following result.

LEMMA 5.1. *Let $(\epsilon_n)_n$ be a sequence of penalty parameters such that $\epsilon_n \to 0$ as $n\to\infty$, and let $\eta^*_{\epsilon_n}, u^*_{\epsilon_n}$ be a solution of the problem* (5.6). *Consider a subsequence such that $\eta^*_{\epsilon_n}\to\eta^*$ weakly in Y and strongly in $L^\infty(0,T,W^{1,\infty}(\mathbb{R}_+))$, and $u^*_{\epsilon_n}\to u^*$ weakly in $L^2(0,T;V)$. Then η^*, u^* is a solution to* (5.3) *and we have for any x_1, x_2, $0<x_1<x_2<+\infty$,*

$$\lim_{n\to\infty} \|u^*_{\epsilon_n}-u^*\|_{L^\infty((0,T)\times(x_1,x_2))}=0. \tag{5.7}$$

Calling $\mu^* = rK1_{\{u^*=u_\circ\}}$,

$$\lim_{n\to\infty} \|rK1_{\{x<K\}}\mathcal{V}_{\epsilon^*_n}(u^*_{\epsilon_n}-u_\circ)-\mu^*\|_{L^p((0,T)\times\mathbb{R}_+)} = 0 \quad \forall p,\ 1\le p<+\infty. \tag{5.8}$$

Finally,

$$\lim_{n\to\infty} \left\| x\frac{\partial u^*_{\epsilon_n}}{\partial x} - x\frac{\partial u^*}{\partial x}\right\|_{L^\infty(0,T;L^2(\mathbb{R}_+))} = 0. \tag{5.9}$$

Proof. The proof that u^* satisfies (2.1) with $\eta=\eta^*$ follows the same lines as that of Theorem 2.2, because $\eta^*_{\epsilon_n}\to\eta^*$ strongly in L^∞; see Remark 2. The convergence of $\eta^*_{\epsilon_n}$ and of $u^*_{\epsilon_n}$ imply that $J(u^*_{\epsilon_n})\to J(u^*)$ and that $J_R(\eta^*)\le \liminf_{n\to\infty} J_R(\eta^*_{\epsilon_n})$.

We have that

$$J(u^*_{\epsilon_n})+J_R(\eta^*_{\epsilon_n}) \le J(u_{\epsilon_n}(\eta))+J_R(\eta) \quad \forall \eta\in\mathcal{H},$$

which yields by passing to the limit

$$J(u^*)+J_R(\eta^*)\le J(u(\eta))+J_R(\eta) \quad \forall\eta\in\mathcal{H}.$$

In order to prove (5.7), we notice that $u(\eta^*_{\epsilon_n})\le u^*_{\epsilon_n}\le u(\eta^*_{\epsilon_n})+\epsilon_n$ from Proposition 2.3 and that $\lim_{n\to\infty}\|u(\eta^*_{\epsilon_n})-u^*\|_{L^\infty((0,T)\times(x_1,x_2))}=0$ from Proposition 4.3. Combining these two results yields (5.7).

The proof of (5.8) is exactly the same as for Proposition 3.8. The proof of (5.9) is exactly the same as for Proposition 3.9. ∎

5.2. First order optimality conditions. It is possible to formulate the variational inequality (2.1) as the identity $\frac{\partial u}{\partial t}+A(t)u=\mu$, with additional constraints on u and μ: $u\ge u_\circ$, $\mu\ge 0$ and $\langle \mu, u-u_\circ\rangle=0$. It is then tempting to use the Lagrange machinery for the least square problem. However, as observed by Bergounioux and

Kunisch [5] for optimal control of obstacle problems, it is generally not possible to find a necessary optimality condition with as many Lagrange multipliers as there are constraints, because the Lagrange system that one would obtain has no solutions. So it is not easy to derive suitable optimality conditions from the variational inequality itself. Instead, we are going to work on the penalized problem (2.10) first, as in [16] and [15].

It is classical to find the first order conditions for a solution $(\eta_\epsilon^*, u_\epsilon^*)$ of (5.6):

- u_ϵ^* satisfies (2.10) for $\eta = \eta_\epsilon^*$;
- there exists an adjoint state $p_\epsilon^* \in L^2(0,T;V) \cap C^0((0,T];L^2(\mathbb{R}_+))$ solution of the problem

(5.10)
$$\left.\begin{array}{l} \dfrac{\partial p_\epsilon^*}{\partial t} + \dfrac{\partial^2}{\partial x^2}\left(\dfrac{\eta_\epsilon^* x^2}{2} p_\epsilon^*\right) - \dfrac{\partial}{\partial x}(rx p_\epsilon^*) - r p_\epsilon^* \\ +rK1_{\{x<K\}} \mathcal{V}_\epsilon'(u_\epsilon^* - u_\circ) p_\epsilon^* \end{array}\right\} = -DJ(u_\epsilon^*), \quad \left\{\begin{array}{l} t \in [0,T), \\ x > 0, \end{array}\right.$$
$$p_\epsilon^*(T) = 0;$$

- $\forall \eta \in \mathcal{H}$,

$$\langle DJ_R(\eta_\epsilon^*), \eta - \eta_\epsilon^* \rangle + \int_0^T \int_{\mathbb{R}_+} \frac{x^2}{2}(\eta - \eta_\epsilon^*) p_\epsilon^* \frac{\partial^2 u_\epsilon^*}{\partial x^2}\, dt\, dx \geq 0. \tag{5.11}$$

We will use the above first order conditions and let ϵ tend to zero, in order to obtain first order conditions for problem (5.3). Such a program has already been applied by Hintermüller [15] and Ito and Künisch [16] for elliptic variational inequalities. At this point, we should also mention Mignot and Puel [28], who applied an elegant method for finding optimality conditions for a special control problem for a parabolic variational inequality.

Before stating the result, we introduce, for a given η satisfying Assumption 1, the Hilbert spaces

$$\begin{aligned} \tilde{Z}_\eta &= \left\{ v \in L^2(0,T;V); \frac{\partial v}{\partial t} - \frac{\eta x^2}{2}\frac{\partial^2 v}{\partial x^2} \in L^2((0,T)\times\mathbb{R}_+) \right\}, \\ Z_\eta &= \left\{ v \in \tilde{Z}_\eta; v(t=0) = 0 \right\}. \end{aligned} \tag{5.12}$$

It is easy to prove that a function $v \in \tilde{Z}_\eta$ belongs to $H^1(0,T;V')$, so the condition $v(t=0)=0$ has a meaning and defines a closed subspace of $\tilde{Z}_\eta$.

It can be shown by studying the properties of $x\frac{\partial v}{\partial x}$ for $v \in Z_\eta$ that

$$x\frac{\partial v}{\partial x} \in L^2(0,T;V) \cap \mathcal{C}^0(0,T;L^2(\mathbb{R}_+)),$$

which implies that both $\frac{\partial v}{\partial t}$ and $\frac{\eta x^2}{2}\frac{\partial^2 v}{\partial x^2}$ belong to $L^2((0,T)\times\mathbb{R}_+)$, and that v is continuous in $[0,T]\times\mathbb{R}_+$. We have that Z_η does not depend on η in the class defined by Assumption 1, so we call this space Z.

THEOREM 5.2. *Let ϵ_n be a sequence of penalty parameters going to zero, and let $(\eta_{\epsilon_n}^*, u_{\epsilon_n}^*)$ be a sequence of solutions to* (5.6) *converging to (η^*, u^*) as in Lemma* 5.1. *There exists a subsequence denoted n_k such that $p_{\epsilon_{n_k}}^*$ converges weakly to p^* in*

$L^2(0,T,V)$. Moreover, there exists a Radon measure α^* such that, in the sense of distributions,

$$\frac{\partial p^*}{\partial t} + \frac{\partial^2}{\partial x^2}\left(\frac{\eta^* x^2}{2} p^*\right) - \frac{\partial}{\partial x}(rxp^*) - rp^* - \alpha^* = -DJ(u^*) \quad \begin{cases} t \in (0,T), \\ x > 0, \end{cases} \tag{5.13}$$

and for all functions $v \in Z$,

$$\int_0^T \int_{\mathbb{R}_+} \left(\frac{\partial v}{\partial t} - \frac{\eta^* x^2}{2}\frac{\partial^2 v}{\partial x^2} - rx\frac{\partial v}{\partial x} + rv\right) p^* + \langle \alpha^*, v\rangle = \langle DJ(u^*), v\rangle. \tag{5.14}$$

Furthermore we have

$$\mu^* |p^*| = 0, \tag{5.15}$$

and, for any $x_1 > 0$ and any function $\phi \in C^0([0,T]\times\mathbb{R}_+)$ such that $\phi(t,x) = 0$ if $x \le x_1$,

$$\langle \alpha^*, |u^* - u_\circ|\phi\rangle = 0. \tag{5.16}$$

Finally, for any $\eta \in \mathcal{H}$,

$$\langle DJ_R(\eta^*), \eta - \eta^*\rangle + \int_0^T \int_{\mathbb{R}_+} \frac{x^2}{2}(\eta - \eta^*) p^* \frac{\partial^2 u^*}{\partial x^2}\, dt\, dx \ge 0. \tag{5.17}$$

Proof. For simplicity, we drop the index n in ϵ_n.

The first thing to do is to obtain estimates on p_ϵ^*: multiplying (5.11) by $p_\epsilon^* e^{\zeta t}$ and integrating leads to

$$\begin{aligned} &\int_{\mathbb{R}_+} \frac{1}{2}(p_\epsilon^*(t,x)e^{\frac{\zeta}{2}t})^2 dx + \int_t^T \int_{\mathbb{R}_+} \frac{\eta_\epsilon^* x^2}{2}\left(\frac{\partial p_\epsilon^*}{\partial x}\right)^2 e^{\zeta s} ds dx \\ &\quad + \int_t^T \int_{\mathbb{R}_+} \left(\frac{\partial}{\partial x}\left(\frac{\eta_\epsilon^* x^2}{2}\right) + rx\right) p_\epsilon^* \frac{\partial p_\epsilon^*}{\partial x} e^{\zeta s} ds dx \\ &\quad + \left(\frac{\zeta}{2} + 2r\right)\int_t^T \int_{\mathbb{R}_+} (p_\epsilon^* e^{\frac{\zeta}{2}s})^2 ds dx - rK \int_t^T \int_0^K \mathcal{V}_\epsilon'(u_\epsilon^* - u_\circ)(p_\epsilon^*)^2 e^{\zeta s} ds dx \\ &= \int_t^T \int_{\mathbb{R}_+} DJ(u_\epsilon^*) p_\epsilon^* e^{\zeta s} ds dx. \end{aligned} \tag{5.18}$$

Taking ζ large enough, one obtains, thanks to the decreasing character of $\mathcal{V}_\epsilon$, that there exists a positive constant C independent of ϵ such that $\|p_\epsilon^*\|_{L^\infty(0,T;L^2(\mathbb{R}_+))} + \|p_\epsilon^*\|_{L^2(0,T;V)} \le C$. Then, for a positive parameter δ, we introduce the nondecreasing function $\rho_\delta : \mathbb{R} \to \mathbb{R}$:

$$\rho_\delta(p) = \left| \begin{array}{cc} -1 & \text{for } p \le -\delta, \\ \frac{p}{\delta} & \text{for } -\delta \le p \le \delta, \\ 1 & \text{for } p \ge \delta. \end{array} \right. \tag{5.19}$$

We also have the nonnegative function $G_\delta(p) = \int_0^p \rho_\delta(q) dq$. For $\bar{K} > K$, we also introduce a smooth cut-off function $\Phi : \mathbb{R}_+ \to [0,1]$, taking the value 1 for $x \le \bar{K}$ and

0 for $x \geq 2\bar{K}$. We multiply (5.18) by $\rho_\delta(p^*_\epsilon)\Phi(x)$ and integrate. We obtain that there exists a constant $C_{\bar{K}}$ independent of δ and ϵ such that

$$\int_{\mathbb{R}_+}\int_t^T\int_{\mathbb{R}_+}\left(\frac{\eta^*_\epsilon x^2}{2}\rho'_\delta(p^*_\epsilon)\left(\frac{\partial p^*_\epsilon}{\partial x}\right)^2\Phi(x) - rK\mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)p^*_\epsilon\rho_\delta(p^*_\epsilon)\right)dsdx \\ + G_\delta(p^*_\epsilon(t,x))\Phi(x)dx \quad \leq C_{\bar{K}}.$$

One can easily check that all the terms in the left-hand side are nonnegative and that a.e. $p^*_\epsilon\rho_\delta(p^*_\epsilon)$ is an increasing sequence with respect to δ, which converges to $|p^*_\epsilon|$. So the Beppo Levi theorem tells us that

$$-rK\int_t^T\int_0^K \mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)p^*_\epsilon\rho_\delta(p^*_\epsilon) \to -rK\int_t^T\int_0^K \mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)|p^*_\epsilon| \quad \text{as} \quad \delta \to 0.$$

Therefore, there exists a positive constant C such that

$$rK\int_0^T\int_0^K |\mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)p^*_\epsilon|dtdx \leq C, \tag{5.20}$$

and it is possible to extract a subsequence ϵ_{n_k} such that $p^*_{\epsilon_{n_k}} \to p^*$ weakly in $L^2(0,T;V)$,

$$-rK1_{\{x\leq K\}}\mathcal{V}'_{\epsilon_{n_k}}(u^*_{\epsilon_{n_k}} - u_\circ)p^*_{\epsilon_{n_k}} \to \alpha^* \text{ weakly}^* \text{ in } (L^\infty((0,T)\times\mathbb{R}_+))^*.$$

Equation (5.13) is satisfied in the sense of distributions, and (5.14) is obtained by passing to the limit.

For proving (5.15), we use the convexity of $\mathcal{V}_\epsilon$ (still dropping the index n_k in ϵ_{n_k}): since $\mathcal{V}_\epsilon(\epsilon) = 0$, we have that for all $u \in [0,\epsilon]$,

$$\mathcal{V}_\epsilon(u) \leq -\mathcal{V}'_\epsilon(u)(\epsilon - u) \leq -\epsilon\mathcal{V}'_\epsilon(u).$$

This implies that $\mathcal{V}_\epsilon(u^*_\epsilon - u_\circ) \leq -\epsilon\mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)$ because we also know that $\mathcal{V}_\epsilon(u^*_\epsilon - u_\circ) = 0$ if $u^*_\epsilon - u_\circ \geq \epsilon$. Thus, calling $\mu^*_\epsilon = rK1_{\{x<K\}}\mathcal{V}_\epsilon(u^*_\epsilon - u_\circ)$, we have that

$$\int_0^T\int_{\mathbb{R}_+}\mu^*_\epsilon|p^*_\epsilon| \leq -\epsilon rK\int_0^T\int_{\mathbb{R}_+}\mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)|p^*_\epsilon| \to 0 \tag{5.21}$$

from (5.20). But we also know that $p^*_\epsilon \to p^*$ weakly in $L^2(0,T;V)$ and that $\mu^*_\epsilon \to \mu^*$ strongly in $L^2((0,T)\times\mathbb{R}_+)$ from Lemma 5.1. Therefore

$$\int_0^T\int_{\mathbb{R}_+}\mu^*_\epsilon|p^*_\epsilon| \to \int_0^T\int_{\mathbb{R}_+}\mu^*|p^*|,$$

and (5.15) is proved.

Let us call $\alpha^*_\epsilon = -rK1_{\{x<K\}}\,\mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)p^*_\epsilon$, let $x_1 < x_2$ be two positive numbers, and let ϕ be a continuous function supported in $[0,T]\times[x_1,+\infty)$,

$$\int_0^T\int_{\mathbb{R}_+}|\alpha^*_\epsilon||(u^*_\epsilon - u_\circ)\phi| \\ \leq rK\left(\int_0^T\int_{\mathbb{R}_+}|\mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)||p^*_\epsilon|^2\right)^{\frac{1}{2}}\left(\int_0^T\int_0^K|\mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)||(u^*_\epsilon - u_\circ)\phi|^2\right)^{\frac{1}{2}}$$

from the Cauchy–Schwarz inequality. But it can be checked that

$$|\mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)||(u^*_\epsilon - u_\circ)\phi|^2 \leq C\epsilon,$$

so $\int_0^T \int_0^K |\mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)||(u^*_\epsilon - u_\circ)\phi|^2 \leq C\epsilon$, and from (5.18),

$$\int_0^T \int_0^K |\mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ)||p^*_\epsilon|^2 \leq C.$$

Therefore, $\int_0^T \int_{\mathbb{R}_+} \alpha^*_\epsilon |u^*_\epsilon - u_\circ|\phi \to 0$ as $\epsilon \to 0$. We know that $\alpha^*_\epsilon \to \alpha^*$ weakly* in $(L^\infty)^*$ and that $|u^*_\epsilon - u_\circ|\phi \to |u^* - u_\circ|\phi$ in $C^0([0,T]\times\mathbb{R}_+)$ from (5.7). We can pass to the limit as $\epsilon \to 0$ and (5.16) is proved.

There remains to prove (5.17). For that, we remark that (5.11) is equivalent to, for all $\eta \in \mathcal{H}$,

$$\begin{aligned} &J_R(\eta^*_\epsilon) - J_R(\eta) + \int_0^T \int_{\mathbb{R}_+} \frac{x^2}{2}(\eta - \eta^*_\epsilon)\frac{\partial p^*_\epsilon}{\partial x}\frac{\partial u^*_\epsilon}{\partial x}\, dt\, dx \\ &+ \int_0^T \int_{\mathbb{R}_+} \frac{x^2}{2}\frac{\partial}{\partial x}(\eta - \eta^*_\epsilon)\frac{\partial u^*_\epsilon}{\partial x}p^*_\epsilon - \int_0^T \int_{\mathbb{R}_+} u^*_\epsilon \frac{\partial}{\partial x}(x(\eta - \eta^*_\epsilon)p^*_\epsilon) \leq 0, \end{aligned}$$

thanks to the convexity of J_R and of $\mathcal{H}$. We can pass to the limit in the first term thanks to the weak convergence of η^*_ϵ to η^* in Y and the fact that J_R is convex and continuous, therefore weakly lower semicontinuous in Y. We can pass to the limit in the other terms from the facts that

- $x\frac{\partial u^*_\epsilon}{\partial x} \to x\frac{\partial u^*}{\partial x}$ in $L^\infty(0,T;L^2(\mathbb{R}_+))$,
- $\eta^*_\epsilon \to \eta^*$ in $L^\infty([0,T];W^{1,\infty}(\mathbb{R}_+))$,
- $p^*_\epsilon \to p^*$ in $L^2(0,T;V)$ weakly,

and we obtain

$$\begin{aligned} &J_R(\eta^*) - J_R(\eta) + \int_0^T \int_{\mathbb{R}_+} \frac{x^2}{2}(\eta - \eta^*)\frac{\partial p^*}{\partial x}\frac{\partial u^*}{\partial x}\, dt\, dx \\ &+ \int_0^T \int_{\mathbb{R}_+} \frac{x^2}{2}\frac{\partial}{\partial x}(\eta - \eta^*)\frac{\partial u^*}{\partial x}p^* - \int_0^T \int_{\mathbb{R}_+} u^* \frac{\partial}{\partial x}(x(\eta - \eta^*)p^*) \leq 0, \end{aligned}$$

which yields (5.17), since J_R is differentiable. □

Remark 7. For a given Radon measure α^*, the problem (5.14) has a unique solution (see the proof of Lemma 6.2 below and also the article by Bergounioux and Kunisch [4]). Under some assumptions, it is possible to give an interpretation of the adjoint problem (5.14), (5.15), (5.16): if η^* satisfies Assumptions 1 and 2, there exists a free boundary Γ^* which can be written $\{x = \gamma^*(t), 0 < t < T\}$, where γ^* is a continuous function from $[0,T]$ with values in $(0,K]$. We call Ω^+ the domain $\Omega^+ = \{(t,x),\ 0 \leq t \leq T,\ x > \gamma^*(t)\}$.

Equation (5.15) tells us that $p^* = 0$ a.e. in the region $\{x < \gamma^*(t)\}$. Equation (5.16) tells us that α^* is supported in the region $\{x \leq \gamma^*(t)\}$. From this, (5.13) tells us that α^* can be written

$$\alpha^* = \alpha^*_s + DJ(u^*)1_{\{x<\gamma^*(t)\}},$$

where the measure α^*_s is a Radon measure supported by Γ (singular with respect to the Lebesgue measure).

5.3. Partial results on differentiability. The aim of this section is to prove that the functional $\eta \mapsto J(u(\eta))$ is Fréchet-differentiable at the points η such that $u(\eta)$ and $\gamma(\eta)$ are smooth enough, and to compute its differential. For that, we will define first an adjoint state, which will be given as a solution of a problem close to (but stronger than) (5.14), (5.15), (5.16). This adjoint state will permit us to build an ansatz for the differential. To prove that this ansatz is actually the Fréchet differential, we will make use of the sensitivity results proved in section 4.

The solution of problem (2.1) is singular at $t = 0$. This causes technical difficulties when trying to see if J is differentiable. To avoid these difficulties, we modify the definition of $\mathcal{H}$: we choose a positive time t_0, $0 < t_0 < T$ (which can be taken as small as desired), and η_0, a smooth bounded function satisfying assumptions 1 and 2. We restrict Y to be

$$Y = \left\{ \eta : \eta, x\frac{\partial \eta}{\partial x}, x^2\frac{\partial^2 \eta}{\partial x^2}, \frac{\partial \eta}{\partial t}, x\frac{\partial^2 \eta}{\partial t \partial x} \in H^1(0,T;X), \qquad \eta|_{t<t_0} = \eta_0 \right\}. \tag{5.22}$$

We call $\mathcal{H}$ the set of the functions $\eta \in Y$ satisfying Assumptions 1 and 2.

Remark 8. If the inequality stated in Remark 5 were proven, the restriction $\eta|_{t<t_0} = \eta_0$ would become unnecessary. We believe that this is indeed unnecessary. For η in $\mathcal{H}$, we call $u(\eta)$ the solution of (2.1), or simply u when no ambiguity is possible. The free boundary Γ is described by the equation $x = \gamma(\tau)$, $0 \le \tau \le T$, with γ a continuous function bounded away from 0. We call Ω^+ the domain $\Omega^+ = \{(\tau, x), 0 \le \tau \le T, x > \gamma(\tau)\}$ and $\Omega_t^+ = \Omega^+ \cap \{\tau \ge t\}$. We call $\Gamma_t = \Gamma \cap \{\tau \ge t\}$.

Consider the backward parabolic problem in the noncylindrical domain Ω^+: find $\tilde{p}$ such that

$$\frac{\partial \tilde{p}}{\partial t} + \frac{\partial^2}{\partial x^2}\left(\frac{\eta x^2}{2}\tilde{p}\right) - \frac{\partial}{\partial x}(rx\tilde{p}) - r\tilde{p} = -DJ(u), \quad (t,x) \in \Omega^+, \tag{5.23}$$

$$\tilde{p} = 0, \quad t = T,\ \gamma(T) < x, \tag{5.24}$$

$$\tilde{p} = 0, \quad x = \gamma(t),\ 0 < t < T. \tag{5.25}$$

Assuming that the function γ is Hölder continuous with power $\frac{1}{2}$, and since η satisfies Assumptions 1 and 2, it is possible to find a weak solution of (5.23)–(5.25); see the work of Brown, Hu, and Lieberman [7]. Following these authors, we call $C_S^1(\Omega^+)$ the space of the functions v whose support is bounded in the variable x and such that v, $\frac{\partial v}{\partial x}$, and $\frac{\partial v}{\partial t}$ are uniformly continuous and v satisfies (5.24), (5.25). We consider the closure $Z(\Omega^+)$ of $C_S^1(\Omega^+)$ with respect to the norm

$$|||v||| = \sup_{t\in(0,T)} \left(\int_{x>\gamma(t)} v(t,x)^2 dx \right)^{\frac{1}{2}} + \left(\int_{\Omega^+} \left| x\frac{\partial v}{\partial x} \right|^2 \right)^{\frac{1}{2}}. \tag{5.26}$$

Following [7], there exists a unique $\tilde{p} \in Z(\Omega_+)$ such that for all $\phi \in C_S^1(\Omega^+)$,

$$\begin{aligned} &- \int_{\Omega_t^+} \tilde{p}\frac{\partial \phi}{\partial t} - \int_{\Omega_t^+} \frac{\eta x^2}{2}\frac{\partial \tilde{p}}{\partial x}\frac{\partial \phi}{\partial x} - \int_{\Omega_t^+} \left(\frac{\partial}{\partial x}\left(\frac{\eta x^2}{2}\right) - rx \right) \tilde{p}\frac{\partial \phi}{\partial x} - r\int_{\Omega_t^+} \tilde{p}\phi \\ &= - \int_{\Omega_t^+} DJ(u)\phi. \end{aligned} \tag{5.27}$$

Remark 9. Among the works on parabolic equations in noncylindrical domains (in any space dimension), let us mention articles by Lions [25] and by Oleĭnik [29] for classes of boundaries smaller than the class considered in [7]. To obtain further results, we need a further assumption on the regularity of the free boundary and u, as in Propositions 3.12 and 4.4.

Assumption 4 (on $\gamma(\eta)$ and $u(\eta)$). For all $t>0$, $\gamma|_{(t,T]}$ is a Lipschitz function, so $\tilde p$ has a trace on Γ_t (because $\frac{\partial \tilde p}{\partial x}\in L^2(\Omega_t^+)$), and we have $\tilde p|_{\Gamma_t}=0$. We also assume that $\frac{\partial u(\eta)}{\partial \tau}\in H^1(\Omega_t^+)$, so the result of Proposition 4.4 holds.

LEMMA 5.3. *For $\eta\in\mathcal{H}$, and under the previous assumption, the solution $\tilde p$ of (5.27) satisfies $x^2\frac{\partial^2\tilde p}{\partial x^2}\in L^2(\Omega_t^+)$ and that $\frac{\partial\tilde p}{\partial\tau}\in L^2(\Omega_t^+)$ for all $t>0$.*

Proof. Since γ is a Lipschitz function, it is possible by a simple change of variables to transform (5.23), (5.24), (5.25) in a problem posed in a cylindrical domain, and then to apply standard regularity results. A similar program has already been carried out in, e.g., [6].

We know that there exists a positive constant γ_0 such that $\gamma_0<\gamma(\tau)\le K$ for all times τ, $0\le\tau\le T$. We make the change of variables $(\tau,x)\mapsto(\tau,\xi)$, with

$$\xi = 2K+\frac{2K-\gamma_0}{2K-\gamma(\tau)}(x-2K) \quad \text{if } \gamma(\tau)\le x\le 2K, \qquad \xi = x \quad \text{if } 2K\le x. \tag{5.28}$$

This a Lipschitz mapping from Ω_t^+ onto $(t,T)\times(\gamma_0,+\infty)$, with a Lipschitz inverse. Its Jacobian matrix is

$$\begin{pmatrix} 1 & \dfrac{(2K-\gamma_0)\gamma'(t)}{(2K-\gamma(\tau))^2}(x-2K) \\ 0 & \dfrac{2K-\gamma_0}{2K-\gamma(\tau)} \end{pmatrix} \quad \text{if } x<2K.$$

As for the inverse mapping, we have

$$x(\tau,\xi)=2K+\frac{2K-\gamma(\tau)}{2K-\gamma_0}(\xi-2K) \quad \text{for} \quad \gamma_0<\xi<2K.$$

With this change of variable, the parabolic equation (5.23) is transformed into a parabolic equation in the variables (τ,ξ):

(5.29)

$$\frac{\partial\hat p}{\partial\tau}-\frac{2K-\xi}{2K-\gamma(\tau)}\gamma'(\tau)\frac{\partial\hat p}{\partial\xi}+\left(\frac{2K-\gamma_0}{2K-\gamma(\tau)}\right)^2\frac{\partial^2}{\partial\xi^2}\left(\frac{\hat\eta x^2}{2}\hat p\right)-\frac{2K-\gamma_0}{2K-\gamma(\tau)}\frac{\partial}{\partial\xi}(rx\hat p)-r\hat p$$
$$=-\widehat{DJ(u)}, \qquad t<\tau<T,\ \gamma_0<\xi<2K,$$

$$\frac{\partial\hat p}{\partial\tau}+\frac{\partial^2}{\partial\xi^2}\left(\frac{\eta\xi^2}{2}\hat p\right)-\frac{\partial}{\partial\xi}(r\xi\hat p)-r\hat p=-DJ(u), \qquad t<\tau<T,\ 2K<\xi,$$

$$\hat p=0, \quad \tau=T,\ \xi>\gamma_0,$$

$$\hat p=0, \quad \xi=\gamma_0,\ t<\tau<T,$$

where for a function u defined in Ω_t^+, the function $\hat u$ is defined in $(t,T)\times(\gamma_0,+\infty)$ by $\hat u(\tau,\xi)=u(t,x)$.

It is possible to use standard results on parabolic equations and prove that problem (5.29) has a unique solution in $L^2((t,T);\hat{V}) \cap C^0([t,T];L^2(\gamma_0,+\infty))$, with $\frac{\partial \hat{p}}{\partial \tau} \in L^2((t,T);\hat{V}')$, where

$$\hat{V} = \left\{ v \in L^2(\gamma_0,+\infty),\ \xi \frac{\partial v}{\partial \xi} \in L^2(\gamma_0,+\infty) \right\}.$$

Furthermore, thanks to Assumptions 1 and 2 and the Lipschitz regularity of γ, we have in fact that $\frac{\partial \hat{p}}{\partial \tau} \in L^2((t,T)\times(\gamma_0,+\infty))$ and $\xi^2 \frac{\partial^2 \hat{p}}{\partial \xi^2} \in L^2((t,T)\times(\gamma_0,+\infty))$. The desired result on $\tilde{p}$ follows easily. □

For any $\tau > 0$, it is possible to extend $\tilde{p}$ by 0 in $(\tau,T)\times\mathbb{R}_+$. Calling p this extension, $\frac{\partial p}{\partial t} \in L^2((\tau,T)\times\mathbb{R}_+)$ (resp., $\frac{\partial p}{\partial x} \in L^2((\tau,T)\times\mathbb{R}_+)$) is the extension of $\frac{\partial \tilde{p}}{\partial t}$ (resp., $\frac{\partial \tilde{p}}{\partial x}$) by 0 in $(\tau,T)\times\mathbb{R}_+$. We define $\alpha \in L^2((\tau,T),V')$ by

$$\langle \alpha, v\rangle = \int_{\{x<\gamma(t)\}} DJ(u)v + \frac{1}{2}\int_\Gamma \eta x^2 \frac{\partial \tilde{p}}{\partial x}|_\Gamma v|_\Gamma n_\Gamma ds,$$

where $n_\Gamma(t,\gamma(t)) = \frac{1}{\sqrt{1+(\gamma'(t))^2}}$. Note that $n_\Gamma ds = dt$. We have

$$\begin{array}{rcl} \frac{\partial p}{\partial t} + \frac{\partial^2}{\partial x^2}(\frac{\eta x^2}{2}p) - \frac{\partial}{\partial x}(rxp) - rp - \alpha &=& -DJ(u), \quad t\in(0,T),\ x>0, \\ p(T) &=& 0, \\ p &=& 0, \quad 0<t<T,\ x<\gamma(t), \end{array}$$

or in other words, for all $w \in L^2(0,T;V)\cap C^0([0,T]\times\mathbb{R}_+)$, with $\frac{\partial w}{\partial t} \in L^2(0,T;V')$ and $w(t)=0$ for $t<t_0$,

$$\begin{array}{l} \left\langle \dfrac{\partial w}{\partial t} + A_\eta(t)w, p \right\rangle + \langle DJ(u(\eta)) - \alpha, w\rangle = 0, \\ p = 0 \quad \text{a.e. in } \{u(\eta)=u_\circ\}, \\ \langle \alpha, w\rangle = 0 \quad \text{if } w \text{ satisfies } w=0 \text{ in } \{u(\eta)=u_\circ\}. \end{array} \tag{5.30}$$

THEOREM 5.4. *If $\eta \in \mathcal{H}$ such that the regularity Assumption 4 on $u(\eta)$ and on the free boundary is fulfilled, then the functional $\xi \mapsto J(u(\xi))$ is differentiable in $\mathcal{H}$ at η, and with p given by* (5.30)*, its differential is*

$$\xi \mapsto \int \frac{x^2}{2}\xi \frac{\partial u(\eta)}{\partial x}\frac{\partial p}{\partial x} + \left(x\xi + \frac{x^2}{2}\frac{\partial \xi}{\partial x}\right)\frac{\partial u(\eta)}{\partial x}p. \tag{5.31}$$

Proof. Let $\delta\eta$ be a small variation of η in Y, such that $\min \eta+\delta\eta > 0$. We note that $\delta u = u(\eta+\delta\eta)-u(\eta)$, $\delta\mu = \mu(\eta+\delta\eta)-\mu(\eta)$, and $\delta J = J(u(\eta+\delta\eta))-J(u(\eta))$. We know from Proposition 4.1 that $\|\delta u\|_{L^2(0,T;V)} + \|\delta u\|_{L^\infty(0,T;L^2(\mathbb{R}_+))} \le C\|\delta\eta\|_{L^\infty((0,T)\times\mathbb{R}_+)}$, so $\delta J = \langle DJ(u(\eta)), \delta u\rangle + o(\delta\eta)$, where the notation $o(\delta\eta)$ is used for a function of $\delta\eta$ such that

$$\lim_{\|\delta\eta\|_{L^\infty((0,T)\times\mathbb{R}_+)} + \|x\frac{\partial\delta\eta}{\partial x}\|_{L^\infty((0,T)\times\mathbb{R}_+)}\to 0} \frac{|o(\delta\eta)|}{\|\delta\eta\|_{L^\infty((0,T)\times\mathbb{R}_+)} + \|x\frac{\partial\delta\eta}{\partial x}\|_{L^\infty((0,T)\times\mathbb{R}_+)}} = 0.$$

Subtracting the equation satisfied by $(u(\eta),\mu(\eta))$ from the one satisfied by $(u(\eta)+\delta u, \mu(\eta)+\delta\mu)$, we obtain that

$$\begin{array}{l} \left\langle \dfrac{\partial \delta u}{\partial t} + A_{\eta+\delta\eta}(t)\delta u, w \right\rangle \\ \qquad + \displaystyle\int \frac{x^2}{2}\delta\eta \frac{\partial u(\eta)}{\partial x}\frac{\partial w}{\partial x} + \left(x\delta\eta + \frac{x^2}{2}\frac{\partial\delta\eta}{\partial x}\right)\frac{\partial u(\eta)}{\partial x}w - \langle \delta\mu, w\rangle = 0, \end{array} \tag{5.32}$$

because A_η is linear with respect to η. From estimate (4.1), we deduce from (5.32) that

$$\left\langle \frac{\partial \delta u}{\partial t} + A_\eta(t)\delta u, w \right\rangle + \int \frac{x^2}{2}\delta\eta \frac{\partial u(\eta)}{\partial x}\frac{\partial w}{\partial x} + \left(x\delta\eta + \frac{x^2}{2}\frac{\partial \delta\eta}{\partial x}\right)\frac{\partial u(\eta)}{\partial x} w - \langle \delta\mu, w\rangle = o(\delta\eta). \tag{5.33}$$

We have also

$$\langle \mu, \delta u\rangle + \langle \delta\mu, u(\eta) - u_\circ\rangle + \langle \delta\mu, \delta u\rangle = 0, \tag{5.34}$$

and estimate (4.2) indicates that $\langle \delta\mu, \delta u\rangle \leq C\|\delta\eta\|^2_{L^\infty((0,T)\times\mathbb{R}_+)}$.

Since $\delta\eta = 0$ for $t < t_0$, we have that $\delta u = 0$ for $t < t_0$. We deduce from (5.30) that

$$\begin{array}{l}\langle DJ(u(\eta)), \delta u\rangle = \\ \displaystyle\int \frac{x^2}{2}\delta\eta\frac{\partial u(\eta)}{\partial x}\frac{\partial p}{\partial x} + \left(x\delta\eta + \frac{x^2}{2}\frac{\partial\delta\eta}{\partial x}\right)\frac{\partial u(\eta)}{\partial x}p - \langle\delta\mu, p\rangle + \langle\alpha, \delta u\rangle + o(\delta\eta).\end{array} \tag{5.35}$$

We aim at proving that the terms $\langle\alpha, \delta u\rangle$ and $\langle\delta\mu, p\rangle$ are $o(\delta\eta)$.

Let us start with the term $\langle\alpha, \delta u\rangle$. The measure α can be written as the sum of two terms:

(1) A measure $\alpha_a = 1_{\{u(\eta)=u_\circ\}}DJ(u(\eta))$ which is absolutely continuous with respect to the Lebesgue measure, and supported in the region $\{u(\eta) = u_\circ\}$. From Assumption 3, we see that

$$\alpha_a = 1_{\{u(\eta)=u_\circ\}}2(u_\circ - u_g). \tag{5.36}$$

We know that $\langle\alpha_a, \delta u\rangle = 2\int 1_{\{u(\eta)=u_\circ\}}(u_\circ - u_g)(u(\eta+\delta\eta) - u_\circ)$. Therefore from (4.2),

$$\begin{array}{l}|\langle\alpha_a, \delta u\rangle| \leq 2(K + \|u_g\|_{L^\infty((0,T)\times\mathbb{R}_+)})\displaystyle\int 1_{\{u(\eta)=u_\circ\}}(u(\eta+\delta\eta) - u_\circ) \\ \leq C\|\eta\|^2_{L^\infty((0,T)\times\mathbb{R}_+)}.\end{array}$$

(2) A singular measure $\alpha_s \in L^2(0,T;V')$ supported by Γ:

$$\langle\alpha_s, \delta u\rangle = \int_{t=t_0}^T h(t)\gamma(t)\delta u(t,\gamma(t))dt, \qquad h(t) = \eta(t,\gamma(t))\gamma(t)\frac{\partial\tilde{p}}{\partial x}(t,\gamma(t)). \tag{5.37}$$

So

$$\begin{array}{rcl}|\langle\alpha_s, \delta u\rangle| = & \left|\displaystyle\int_{t=t_0}^T h(t)\gamma(t)\delta u(t,\gamma(t))dt\right| \\ = & \left|\displaystyle\int_{t=t_0}^T h(t)1_{\delta\gamma(t)<0}\int_{\gamma(t)+\delta\gamma(t)}^{\gamma(t)}\frac{\partial}{\partial x}(x\delta u)dx\right| \\ \leq & \|h\sqrt{(\delta\gamma)_-}\|_{L^2(t_0,T)}\left(\displaystyle\int_{t=t_0}^T 1_{\delta\gamma(t)<0}\int_{\gamma(t)+\delta\gamma(t)}^{\gamma(t)}\left(\frac{\partial}{\partial x}(x\delta u)\right)^2 dx\right)^{\frac{1}{2}}.\end{array}$$

But from Proposition 4.2, $(\delta\gamma)_-$ converges to 0 in $L^1(t_0,T)$. Consider a subsequence $(\delta\eta_n,\delta\gamma_n)$ such that $\|h\sqrt{(\delta\gamma_n)_-}\|_{L^2(t_0,T)} \to \ell$. It is possible to extract a second subsequence still called $(\delta\eta_n,\delta\gamma_n)$ such that $(\delta\gamma_n)_- \to 0$ a.e., but we know that $0 \le (\delta\gamma)_- \le K$, so Lebesgue's theorem tells us that $\ell = 0$. Therefore, we have that $\lim_{\delta\eta\to 0} \|h\sqrt{(\delta\gamma_n)_-}\|_{L^2(t_0,T)} = 0$. From this and (4.1), we conclude that $\langle \alpha_s, \delta u\rangle = o(\delta\eta)$.

We have proved that $\langle \alpha, \delta u\rangle = o(\delta\eta)$. There remains to study the term $\langle \delta\mu, p\rangle$:

$$\begin{aligned}
|\langle \delta\mu, p\rangle| &= rK \left| \int_{t=t_0}^{T} 1_{\delta\gamma(t)>0} \int_{\gamma(t)}^{\gamma(t)+\delta\gamma(t)} p \right| \\
&\le \int_{t=t_0}^{T} 1_{\delta\gamma(t)>0} \sqrt{\delta\gamma(t)} \sqrt{\int_{\gamma(t)}^{\gamma(t)+\delta\gamma(t)} p^2} \\
&\le \int_{t=t_0}^{T} 1_{\delta\gamma(t)>0} (\delta\gamma(t))^{\frac{3}{2}} \sqrt{\int_{\gamma(t)}^{\gamma(t)+\delta\gamma(t)} \left(\frac{\partial p}{\partial x}\right)^2},
\end{aligned}$$

and we deduce from Proposition 4.4 and from the argument based on Lebesgue's theorem, which is detailed above, that $\langle \delta\mu, p\rangle = o(\delta\eta)$.

We have proved Theorem 5.4. □

6. The calibration problem.

6.1. Toward the calibration problem. A first step toward the calibration problem is to consider problem (5.3), where $\mathcal{H}$ and J_R are defined in section 5.1, and where J is now

$$J(u) = (u(T,x_{ob}) - \bar{u})^2, \tag{6.1}$$

where x_{ob} and $\bar{u}$ are positive numbers. Of course, for any penalty parameter ϵ, it is possible to consider problem (5.6). For this new choice of J, a result similar to Lemma 5.1 holds.

LEMMA 6.1. *Let $(\epsilon_n)_n$ be a sequence of penalty parameters such that $\epsilon_n \to 0$ as $n \to \infty$, and let $\eta^*_{\epsilon_n}, u^*_{\epsilon_n}$ be a solution of the problem* (5.6). *Consider a subsequence such that $\eta^*_{\epsilon_n} \to \eta^*$ weakly in Y and strongly in $L^\infty(0,T,W^{1,\infty}(\mathbb{R}_+))$, and $u^*_{\epsilon_n} \to u^*$ weakly in $L^2(0,T;V)$. We have for any x_1, x_2, $0 < x_1 < x_2 < +\infty$,*

$$\lim_{n\to\infty} \|u^*_{\epsilon_n} - u^*\|_{L^\infty((0,T)\times(x_1,x_2))} = 0. \tag{6.2}$$

Moreover, η^*, u^* is a solution to (5.3) *and, calling $\mu^* = rK1_{\{u^*=u_\circ\}}$,*

$$\lim_{n\to\infty} \|rK1_{\{x<K\}}\mathcal{V}_{\epsilon^*_n}(u^*_{\epsilon_n} - u_\circ) - \mu^*\|_{L^p((0,T)\times\mathbb{R}_+)} = 0 \quad \forall p,\ 1 \le p < +\infty, \tag{6.3}$$

and

$$\lim_{n\to\infty} \left\| x\frac{\partial u^*_{\epsilon_n}}{\partial x} - x\frac{\partial u^*}{\partial x} \right\|_{L^\infty(0,T;L^2(\mathbb{R}_+))} = 0. \tag{6.4}$$

Proof. It can be shown exactly as in the proof of Lemma 5.1 that u^* satisfies (2.1) with $\eta = \eta^*$. Moreover, from Proposition 4.3,

$$\lim_{n\to\infty} \|u(\eta_{\epsilon^*_n}) - u^*\|_{L^\infty((0,T)\times(x_1,x_2))} = 0.$$

Combining this with the fact that $u(\eta^*_{\epsilon_n}) \leq u^*_{\epsilon_n} \leq u(\eta^*_{\epsilon_n}) + \epsilon_n$ yields that $u^*_{\epsilon_n}$ converges uniformly to u^* in $[0,T] \times [x_1, x_2]$. This proves that $\lim_{n\to\infty} J(u^*_{\epsilon_n}) = J(u^*)$ and we conclude as in the proof of Lemma 5.1 that (η^*, u^*) is a solution to (5.3). We prove (6.3) and (6.4) as in the proof of Lemma 5.1. □

Remark 10. Let $(\eta^*_{\epsilon_n}, u^*_{\epsilon_n})$ be a subsequence converging to (η^*, u^*) as in Lemma 6.1. It is clear from the continuity of u^* and from (6.2) that if $u^*(T, x_{ob}) > u_\circ(x_{ob})$, then there exists a positive real number a and an integer N such that for $n > N$, $u^*_{\epsilon_n}(t,x) > u_\circ(x) + \epsilon_n$ for all (t,x) with $|x - x_{ob}| \leq a$ and $t > T - a$. Let $(\eta^*_{\epsilon_n}, u^*_{\epsilon_n})$ be a subsequence converging to (η^*, u^*) as in Lemma 6.1. Assume that $u^*(T, x_{ob}) > u_\circ(x_{ob})$ and let a and N be as in Remark 10. It is possible to derive necessary optimality conditions for problem (5.6) for $n > N$. We drop the index n to simplify the notation.

LEMMA 6.2. *Assume that there exits a positive number a such that $u^*_\epsilon(t,x) > u_\circ(x) + \epsilon$ for all (t,x) with $|x - x_{ob}| \leq a$ and $T - t \leq a$. There exists a unique $p^*_\epsilon \in L^2((0,T) \times \mathbb{R}_+)$ such that, for all $v \in Z$ (the space Z is defined in* (5.12)*),*

$$\int_0^T \int_{\mathbb{R}_+} \left(\frac{\partial v}{\partial t} - \frac{\eta^*_\epsilon x^2}{2} \frac{\partial^2 v}{\partial x^2} - rx \frac{\partial v}{\partial x} + rv - rK 1_{\{x<K\}} \mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ) v \right) p^*_\epsilon \tag{6.5}$$
$$= 2(u^*_\epsilon(T, x_{ob}) - \bar{u}) v((T, x_{ob})),$$

and $\|p^*_\epsilon\|_{L^2((0,T)\times\mathbb{R}_+)}$ is bounded by a constant independent of ϵ in the subsequence.

Moreover, we have, in the sense of distributions in the domain $t < T$, $x > 0$,

$$\frac{\partial p^*_\epsilon}{\partial t} + \frac{\partial^2}{\partial x^2} \left(\frac{\eta^*_\epsilon x^2}{2} p^*_\epsilon \right) - \frac{\partial}{\partial x}(rx p^*_\epsilon) - r p^*_\epsilon + rK 1_{\{x<K\}} \mathcal{V}'_\epsilon(u^*_\epsilon - u_\circ) p^*_\epsilon = 0, \tag{6.6}$$

and for a smooth function ϕ taking the value 1 for $|x - x_{ob}| \geq \frac{a}{2}$, $T - t \geq \frac{a}{2}$ and vanishing in a neighborhood of (T, x_{ob}), we have that $\phi p^*_\epsilon \in L^2(0,T;V) \cap \mathcal{C}^0([0,T]; L^2(\mathbb{R}_+))$, with norms bounded independently of ϵ.

Proof. We call b the bilinear form on $L^2((0,T) \times \mathbb{R}_+) \times Z$ defined by

$$b(q,v) = \int_0^T \int_{\mathbb{R}_+} \left(\frac{\partial v}{\partial t} - \frac{\eta^*_\epsilon x^2}{2} \frac{\partial^2 v}{\partial x^2} - rx \frac{\partial v}{\partial x} + rv \right) q.$$

It is clear that b is continuous and that there exists a positive constant c, independent of η^*_ϵ in the class defined by Assumption 1, such that

$$\inf_{q \in L^2((0,T)\times\mathbb{R}_+)} \sup_{v \in Z} \frac{b(q,v)}{\|q\|_{L^2((0,T)\times\mathbb{R}_+)} \|v\|_Z} \geq c.$$

To prove this inf-sup condition, take $v \in L^2(0,T;V) \cap H^1(0,T;L^2(R_+))$ as the weak solution of

$$\frac{\partial v}{\partial t} - \frac{\eta^*_\epsilon x^2}{2} \frac{\partial^2 v}{\partial x^2} - rx \frac{\partial v}{\partial x} + rv = q, \quad t > 0, \qquad v(0,\cdot) = 0.$$

It can be proved that $x \frac{\partial v}{\partial x} \in L^2(0,T;V)$ and that $\|v\|_{L^2(0,T;V)}$ and $\|x \frac{\partial v}{\partial x}\|_{L^2(0,T;V)}$ are bounded by $C\|q\|_{L^2(0,T;L^2(R_+))}$ with C independent of η^*_ϵ in the class defined by Assumption 1. Therefore $\|\frac{x^2}{2} \frac{\partial^2 v}{\partial x^2}\|_{L^2(0,T;L^2(R_+))} \leq C\|q\|_{L^2(0,T;L^2(R_+))}$ and from

the equation satisfied by v, $\|\frac{\partial v}{\partial t}\|_{L^2(0,T;L^2(R_+))} \le C\|q\|_{L^2(0,T;L^2(R_+))}$. Hence, $\|v\|_Z \le C\|q\|_{L^2(0,T;L^2(R_+))}$, and the inf-sup condition above is proved. In other words, calling B the continuous linear operator from $L^2(0,T;L^2(\mathbb{R}_+)$ to the dual of Z defined by $\langle Bp, v\rangle = b(p,v)$, the range of B is closed. Also, it is clear that B^T is injective, so B is surjective. Using again the inf-sup condition, we have that B is an isomorphism from $L^2((0,T)\times\mathbb{R}_+)$ onto the dual of Z, and its inverse is continuous with a norm independent of η_ϵ^*.

Therefore, there exists a unique $g_\epsilon \in L^2(0,T;L^2(\mathbb{R}_+))$ such that for all $v \in Z$,

$$b(g_\epsilon, v) = 2(u_\epsilon^*(T,x_{ob}) - \bar{u})v((T,x_{ob}))$$

and $\|g_\epsilon\|_{L^2(0,T;L^2(R_+))}$ is bounded independently of η_ϵ^*. Furthermore, for any smooth function ϕ vanishing near (T,x_{ob}), $\phi g_\epsilon \in L^2(0,T;V)$ with a norm bounded independently of ϵ.

Call now $\tilde{p}_\epsilon \in L^2(0,T;V)\cap \mathcal{C}^0([0,T];L^2(\mathbb{R}_+))$ the weak solution of the following problem:

(6.7)
$$\begin{aligned}
&\frac{\partial \tilde{p}_\epsilon}{\partial t} + \frac{\partial^2}{\partial x^2}\left(\frac{\eta_\epsilon^* x^2}{2}\tilde{p}_\epsilon\right) - \frac{\partial}{\partial x}(rx\tilde{p}_\epsilon) - r\tilde{p}_\epsilon + rK1_{\{x<K\}}\mathcal{V}_\epsilon'(u_\epsilon^* - u_\circ)\tilde{p}_\epsilon \\
&\qquad = -rK1_{\{x<K\}}\mathcal{V}_\epsilon'(u_\epsilon^* - u_\circ)g_\epsilon = -rK1_{\{x<K\}}\mathcal{V}_\epsilon'(u_\epsilon^* - u_\circ)\phi g_\epsilon, \quad t<T,\ x>0, \\
&\tilde{p}_\epsilon(T) = 0,
\end{aligned}$$

where ϕ is the smooth function introduced in the statement of Lemma 6.2. By looking at the equation satisfied by $\tilde{p}_\epsilon + \phi g_\epsilon$, we see that $\|\tilde{p}_\epsilon + \phi g_\epsilon\|_{L^2(0,T;V)}$ is bounded independently of ϵ. It is clear that $p_\epsilon^* = \tilde{p}_\epsilon + g_\epsilon$ satisfies (6.5). By using now the bilinear form

$$\tilde{b}(q,v) = \int_0^T \int_{\mathbb{R}_+} \left(\frac{\partial v}{\partial t} - \frac{\eta_\epsilon^* x^2}{2}\frac{\partial^2 v}{\partial x^2} - rx\frac{\partial v}{\partial x} + rv - rK1_{\{x<K\}}\mathcal{V}_\epsilon'(u_\epsilon^* - u_\circ)v\right) q,$$

and repeating the same argument as for b (using the monotonicity properties of $\mathcal{V}_\epsilon$), the solution of (6.5) is clearly unique.

We have also proved that $\|p_\epsilon\|_{L^2((0,T)\times\mathbb{R}_+)}$ and $\|\phi p_\epsilon\|_{L^2(0,T;V)}$ are bounded independently of ϵ. □

The optimality conditions for problem (5.6) are the following:

1. u_ϵ^* satisfies (2.10) for $\eta = \eta_\epsilon^*$.
2. There exists an adjoint state $p_\epsilon^* \in L^2((0,T)\times\mathbb{R}_+)$ solution of the problem (6.5).
3. For all $\eta \in \mathcal{H}$,

$$\langle DJ_R(\eta_\epsilon^*), \eta - \eta_\epsilon^*\rangle + \int_0^T \int_{\mathbb{R}_+} \frac{x^2}{2}(\eta - \eta_\epsilon^*)p_\epsilon^* \frac{\partial^2 u_\epsilon^*}{\partial x^2}\, dt\, dx \ge 0. \tag{6.8}$$

The next step is to pass to the limit when $\epsilon \to 0$.

THEOREM 6.3. *Let ϵ_n be a sequence of penalty parameters going to zero, and let $(\eta_{\epsilon_n}^*, u_{\epsilon_n}^*)$ be a sequence of solutions to* (5.6) *converging to (η^*, u^*) as in Lemma* 6.1. *Assume that there exits a positive number a such that $u_{\epsilon_n}^*(t,x) > u_\circ(x) + \epsilon_n$ for all*

(t,x) with $|x-x_{ob}|\leq a$ and $T-t\leq a$. There exists a subsequence denoted n_k such that $p^*_{\epsilon_{n_k}}$ converges weakly to p^* in $L^2(0,T,L^2(\mathbb{R}_+))$ and $\phi p^*_{\epsilon_{n_k}}$ converges weakly to ϕp^* in $L^2(0,T;V)$, where ϕ is given in Lemma 6.2. Moreover, there exists a Radon measure α^* such that for all $v\in Z$ (the space Z is defined in (5.12)),

$$\int_0^T\int_{\mathbb{R}_+}\left(\frac{\partial v}{\partial t}-\frac{\eta^* x^2}{2}\frac{\partial^2 v}{\partial x^2}-rx\frac{\partial v}{\partial x}+rv\right)p^*+\langle\alpha^*,v\rangle = 2(u^*(T,x_{ob})-\bar u)v((T,x_{ob})). \tag{6.9}$$

The function p^* satisfies

$$\frac{\partial p^*}{\partial t}+\frac{\partial^2}{\partial x^2}\left(\frac{\eta^* x^2}{2}p^*\right)-\frac{\partial}{\partial x}(rxp^*)-rp^*-\alpha^*=0 \tag{6.10}$$

in the sense of distributions. Furthermore we have

$$\mu^*|p^*|=0, \tag{6.11}$$

and, for any $x_1>0$ and any function $\psi\in C^0([0,T]\times\mathbb{R}_+)$ such that $\psi(t,x)=0$ if $x\leq x_1$,

$$\langle\alpha^*,|u^*-u_\circ|\psi\rangle=0. \tag{6.12}$$

Finally, for any $\eta\in\mathcal{H}$,

$$\langle DJ_R(\eta^*),\eta-\eta^*\rangle+\int_0^T\int_{\mathbb{R}_+}\frac{x^2}{2}(\eta-\eta^*)p^*\frac{\partial^2 u^*}{\partial x^2}\,dt\,dx\geq 0. \tag{6.13}$$

Proof. We follow essentially the proof of Theorem 5.2, but we have to take into account the singular behavior of p^*_ϵ at (T,x_{ob}) which necessitates additional technicalities. For ϕ introduced in Lemma 6.2, we consider the parabolic PDE satisfied by p^*_ϵ and we take as a test function $\rho_\delta(p^*_\epsilon)\chi(x)\phi(t,x)$, where ρ is defined in (5.30) and where χ is a smooth cut-off function $\chi:\mathbb{R}_+\to[0,1]$, taking the value 1 for $x\leq\bar K$ and 0 for $x\geq 2\bar K$. Calling $G_\delta(p)=\int_0^p\rho_\delta(q)dq$, we obtain that there exists a constant $C_{\bar K}$ independent of δ and ϵ such that

$$\int_t^T\int_{\mathbb{R}_+}\left(\frac{\eta^*_\epsilon x^2}{2}\rho'_\delta(p^*_\epsilon)\left(\frac{\partial p^*_\epsilon}{\partial x}\right)^2\phi(s,x)\chi(x)-rK\mathcal{V}'_\epsilon(u^*_\epsilon-u_\circ)p^*_\epsilon\rho_\delta(p^*_\epsilon)\right)dsdx + \int_{\mathbb{R}_+}G_\delta(p^*_\epsilon(t,x))\phi(t,x)\chi(x)dx\leq C_{\bar K}.$$

From this, we infer that there exists a positive constant C such that

$$rK\int_0^T\int_0^K|\mathcal{V}'_\epsilon(u^*_\epsilon-u_\circ)p^*_\epsilon|dtdx\leq C,$$

and it is possible to extract a subsequence ϵ_{n_k} such that $p^*_{\epsilon_{n_k}}\to p^*$ weakly in $L^2((0,T)\times\mathbb{R}_+)$, $\phi p^*_{\epsilon_{n_k}}\to p^*$ weakly in $L^2(0,T;V)$, and $-rK1_{\{x\leq K\}}\mathcal{V}'_{\epsilon_{n_k}}(u^*_{\epsilon_{n_k}}-u_\circ)p^*_{\epsilon_{n_k}}\to\alpha^*$ weakly* in $(L^\infty((0,T)\times\mathbb{R}_+))^*$. Equation (6.10) is satisfied in the sense of distributions and (6.9) is obtained as well by passing to the limit. The proofs of (6.11) and (6.12) follow the same line as in the proof of Theorem 5.2.

There remains to prove (6.13).

For that we write $p_{\epsilon^*}\frac{\partial^2}{\partial x^2}u_{\epsilon^*}$ as $\phi p_{\epsilon^*}\frac{\partial^2}{\partial x^2}u_{\epsilon^*}+p_{\epsilon^*}(1-\phi)\frac{\partial^2}{\partial x^2}u_{\epsilon^*}$ and we use the fact that for a subsequence,

- $\phi p_{\epsilon^*}\to\phi p^*$ in $L^2(0,T;V)$ weakly and $p^*_\epsilon\to p^*$ in $L^2((0,T)\times\mathbb{R}_+)$ weakly,
- $x\frac{\partial u^*_\epsilon}{\partial x}\to x\frac{\partial u^*}{\partial x}$ in $L^\infty(0,T;L^2(\mathbb{R}_+))$,
- $x^2(1-\phi)\frac{\partial^2 u^*_\epsilon}{\partial x^2}\to x^2(1-\phi)\frac{\partial^2 u^*}{\partial x^2}$ in $L^2(0,T;L^2(\mathbb{R}_+))$,
- $\eta^*_\epsilon\to\eta^*$ in $L^\infty([0,T];W^{1,\infty}(\mathbb{R}_+))$.

The decomposition of $p_{\epsilon^*}\frac{\partial^2}{\partial x^2}u_{\epsilon^*}$ and these convergence properties enable us to pass to the limit in $\int_0^T\int_{\mathbb{R}_+}\frac{x^2}{2}(\eta-\eta^*_\epsilon)p^*_\epsilon\frac{\partial^2 u^*_\epsilon}{\partial x^2}\,dt\,dx$ and to obtain (6.13). ∎

6.2. The calibration problem. The calibration problem consists of finding η from the observations of

- the price x_o of the underlying asset today,
- the prices $(\bar u_i)_{i\in I}$ of a family of American puts with different maturities and different strikes $(T_i,K_i)_{i\in I}$.

We call $T=\max_{i\in I}T_i$. For $\mathcal{H}$ and J_R given in section 5.1, we consider the least square problem: find $\eta\in\mathcal{H}$ minimizing

$$J(\eta)+J_R(\eta),\qquad J(\eta)=\sum_{i\in I}|u_i(T_i,x_\circ)-\bar u_i|^2, \tag{6.14}$$

where

(6.15)
$$\left.\begin{aligned}
&\frac{\partial u_i}{\partial t}-\frac{\eta(T_i-t,x)x^2}{2}\frac{\partial^2 u_i}{\partial x^2}-rx\frac{\partial u_i}{\partial x}+ru_i\ge 0,\\
&u_i\ge(K_i-x)_+\\
&\left(\frac{\partial u_i}{\partial t}-\frac{\eta(T_i-t,x)x^2}{2}\frac{\partial^2 u_i}{\partial x^2}-rx\frac{\partial u_i}{\partial x}+ru_i\right)(u_i-(K_i-x)_+)=0,
\end{aligned}\right\}\quad \begin{array}{l}t<T_i,\\ x>0,\end{array}$$
$$u_i(t=0,x)=(K_i-x)_+,\quad x>0,$$

Note that we have changed the time variable so that the expiration date of all the options becomes 0.

THEOREM 6.4. *Assuming that η_* is a minimum of* (6.14), (6.15) *obtained as the limit of a sequence of minimizers for the corresponding penalized problems as the penalty parameter goes to* 0, *and such that for all $i\in I$, (T_i,x_i) lies in the zone where $u_i^*>(K_i-x)_+$ (where u_i^* is the solution of* (6.15) *with $\eta=\eta^*$), we have the following necessary condition of optimality: for $i\in I$, there exists p_i in $L^2(0,T_i,L^2(\mathbb{R}_+))$ and a Radon measure α_i such that for all $v\in Z_i$,*

$$\begin{aligned}&\int_0^{T_i}\int_{\mathbb{R}_+}\left(\frac{\partial v}{\partial t}-\frac{\eta^*(T_i-t)x^2}{2}\frac{\partial^2 v}{\partial x^2}-rx\frac{\partial v}{\partial x}+rv\right)p_i+\langle\alpha_i,v\rangle\\ &\quad=2(u_i^*(T_i,x_i)-\bar u_i)v((T_i,x_i)),\end{aligned} \tag{6.16}$$

with

$$1_{\{u_i^*=(K_i-x)_+\}}|p_i|=0, \tag{6.17}$$

and, for any $x_1 > 0$ *and any function* $\psi \in C^0([0,T_i]\times\mathbb{R}_+)$ *such that* $\psi(t,x)=0$ *if* $x \le x_1$,

$$\langle \alpha_i, |u_i^* - (K_i - x)_+|\psi\rangle = 0. \tag{6.18}$$

For any $\eta \in \mathcal{H}$,

$$\langle DJ_R(\eta^*), \eta-\eta^*\rangle + \sum_{i\in I}\int_0^{T_i}\int_{\mathbb{R}_+} \frac{x^2}{2}(\eta-\eta^*)(T_i-t,x)p_i\frac{\partial^2 u_i^*}{\partial x^2}\,dt\,dx \ge 0. \tag{6.19}$$

REFERENCES

[1] Y. Achdou, *A numerical procedure for the calibration of the volatility with American options*, Appl. Math. Finance, accepted.

[2] Y. Achdou and O. Pironneau, *Volatility smile by multilevel least square*, Int. J. Theor. Appl. Finance, 5 (2002), pp. 619–643.

[3] M. Avellaneda, M. Friedman, C. Holmes, and D. Samperi, *Calibrating volatility surfaces via relative entropy minimization*, Appl. Math. Finance, 4 (1997), pp. 37–64.

[4] M. Bergounioux and K. Kunisch, *On the structure of Lagrange multipliers for state-constrained optimal control problems*, Systems Control Lett., 48 (2003), pp. 169–176.

[5] M. Bergounioux and F. Mignot, *Optimal control of obstacle problems: Existence of Lagrange multipliers*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 45–70 (electronic).

[6] A. Bove, B. Franchi, and E. Obrecht, *Straightening of a noncylindrical region and evolution equations*, Rend. Sem. Mat. Univ. Padova, 71 (1984), pp. 209–216.

[7] R. M. Brown, W. Hu, and G. M. Lieberman, *Weak solutions of parabolic equations in non-cylindrical domains*, Proc. Amer. Math. Soc., 125 (1997), pp. 1785–1792.

[8] T. F. Coleman, Y. Li, and A. Verma, *Reconstructing the unknown local volatility function*, J. Comput. Finance, 2 (1999), pp. 77–100.

[9] R. Cont and P. Tankov, *Financial Modelling with Jump Processes*, Chapman & Hall/CRC Fin. Math. Ser., Chapman & Hall/CRC, Boca Raton, FL, 2004.

[10] B. Dupire, *Pricing and hedging with smiles*, in Mathematics of Derivative Securities (Cambridge, 1995), Publ. Newton Inst. 15, Cambridge University Press, Cambridge, UK, 1997, pp. 103–111.

[11] J.-P. Fouque, G. Papanicolaou, and K. R. Sircar, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, Cambridge, UK, 2000.

[12] A. Friedman, *Regularity theorems for variational inequalities in unbounded domains and applications to stopping time problems*, Arch. Rational Mech. Anal., 52 (1973), pp. 134–160.

[13] A. Friedman and W. Shen, *A variational inequality approach to financial valuation of retirement benefits based on salary*, Finance Stoch., 6 (2002), pp. 273–302.

[14] S. Heston, *A closed-form solution for options with stochastic volatility with applications to bond and currency options*, Oxford University Press, Oxford, UK, Review of Financial Studies, 6 (1993), pp. 327–343.

[15] M. Hintermüller, *Inverse coefficient problems for variational inequalities: Optimality conditions and numerical realization*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 129–152.

[16] K. Ito and K. Kunisch, *Optimal control of elliptic variational inequalities*, Appl. Math. Optim., 41 (2000), pp. 343–364.

[17] N. Jackson, E. Süli, and S. Howison, *Computation of deterministic volatility surfaces*, J. Comput. Finance, 2 (1998/1999), pp. 5–32.

[18] P. Jaillet, D. Lamberton, and B. Lapeyre, *Variational inequalities and the pricing of American options*, Acta Appl. Math., 21 (1990), pp. 263–289.

[19] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Pure Appl. Math. 88, Academic Press, London, 1980.

[20] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Uralćeva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1967.

[21] R. Lagnado and S. Osher, *Reconciling differences*, Risk, 10 (1997), pp. 79–83.

[22] R. Lagnado and S. Osher, *A technique for calibrating derivative security pricing models: Numerical solution of an inverse problem*, J. Comput. Finance, 1 (1997), pp. 13–25.

[23] D. Lamberton and B. Lapeyre, *Introduction au calcul stochastique appliqué à la finance*, Ellipses, Paris, 1997.

[24] D. Lamberton, *Critical price for an American option near maturity*, in Seminar on Stochastic Analysis, Random Fields and Applications (Ascona, 1993), Progr. Probab. 36, Birkhäuser, Basel, Switzerland, 1995, pp. 353–358.
[25] J.-L. Lions, *Sur les problèmes mixtes pour certains systèmes paraboliques dans des ouverts non cylindriques*, Ann. Inst. Fourier (Grenoble), 7 (1957), pp. 143–182.
[26] J.-L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod, Paris, 1969.
[27] J.-L. Lions and E. Magenes, *Problèmes aux Limites Non Homogènes et Applications*, Vols. 1 and 2, Dunod, Paris, 1968.
[28] F. Mignot and J.-P. Puel, *Contrôle optimal d'un système gouverné par une inéquation variationnelle parabolique*, C. R. Acad. Sci. Paris Sér. I Math., 298 (1984), pp. 277–280.
[29] O. A. Oleĭnik, *Linear equations of second order with non-negative characteristic form*, Amer. Math. Soc. Transl. Ser. 2, 65 (1967), pp. 167–200.
[30] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag, New York, 1984 (corrected reprint of the 1967 original).
[31] P. Willmott, J. Dewynne, and J. Howison, *Option Pricing: Mathematical Models and Computations*, Oxford Financial Press, Oxford, UK, 1993.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1616–1639

CONTROLS INSENSITIZING THE OBSERVATION OF A QUASI-GEOSTROPHIC OCEAN MODEL*

ENRIQUE FERNÁNDEZ-CARA†, GALINA C. GARCIA‡, AND AXEL OSSES§

Abstract. We consider a linear quasi-geostrophic ocean model with partially known initial conditions. We search for controls that make the observation locally insensitive to the perturbations of the initial data. Their existence is equivalent to the null controllability property for an associated cascade Stokes-like system. Thanks to the presence of the Coriolis term, we are able to prove the existence of such controls. Our strategy is the following. First, we prove a unique continuation property for the adjoint of the state system that leads to approximate controllability; then, under certain assumptions, an observability inequality is established for the adjoint. The proof is inspired by the arguments leading to the unique continuation property. This inequality leads to the desired null controllability result.

Key words. insensitizing controls, Carleman inequalities, unique continuation, null controllability, ocean model

AMS subject classifications. 93B05, 35B37, 35B60, 35Q30

DOI. 10.1137/S0363012903433607

1. Introduction and main results.

1.1. Incomplete initial data ocean model. Let Ω be a nonempty open bounded and connected subset of $\mathbb{R}^2$, with boundary Γ of class $\mathcal{C}^2$ and outwards unit normal vector $\nu = \nu(x)$. Let ω be a nonempty open subset of Ω, $T > 0$, $Q = \Omega \times (0,T)$, and $\Sigma = \Gamma \times (0,T)$. In this paper, we will consider a linear quasi-geostrophic ocean model [1, 15, 16] described by the following equations:

$$\begin{cases} u_t - A\Delta u + \gamma u + (f_0 + \beta x_2)\,\mathrm{k} \wedge u + \dfrac{1}{\rho_0}\nabla p = \mathcal{T} + h1_\omega & \text{in } Q, \\ \operatorname{div} u = 0 & \text{in } Q, \\ u = 0 & \text{on } \Sigma, \\ u(0) = u_0 + \tau\widehat{u}_0 & \text{in } \Omega, \end{cases} \tag{1.1}$$

where $u(x,t)$ and $p(x,t)$, respectively, denote the velocity and the pressure of the fluid at $(x,t) = (x_1, x_2, t) \in \mathbb{R}^2 \times \mathbb{R}_+$. In this model, A represents the horizontal *eddy viscosity* coefficient, γ is the bottom *friction* coefficient, ρ_0 is the fluid density, and $(f_0 + \beta x_2)\mathrm{k} \wedge u$ is the Coriolis term, with $\mathrm{k} \wedge u = (-u_2, u_1)$. In the right-hand side, 1_ω denotes the characteristic function of ω and $\mathcal{T}$ is a given source in $L^2(Q)^2$. The term

*Received by the editors August 22, 2003; accepted for publication (in revised form) July 1, 2004; published electronically March 11, 2005.

http://www.siam.org/journals/sicon/43-5/43360.html

†Departamento de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, Aptdo. 1160, 41080 Sevilla, Spain (cara@numer.us.es). This author's work was partially supported by D.G.E.S. (Spain) grants BFM2000-1317 and BFM2003-06446.

‡Facultad de Ingeniería, Universidad Católica de la Santísima Concepción, Casilla 297, Concepción, Chile (galina@ucsc.cl). This author's work was supported by FONDAP in Applied Mathematics, CONICYT Ph.D. grants, and CONICYT-INRIA cooperation agreements (Chile).

§Departamento de Ingeniería Matemática, Universidad de Chile, Casilla 170/3 Correo 3, Santiago, Chile, and Centro de Modelamiento Matemático, UMI 2807/Universidad de Chile-CNRS, Santiago, Chile (axosses@dim.uchile.cl). This author's work was partially supported by FONDAP in Applied Mathematics, FONDECYT-CONICYT 1030808-7030059, and ECOS-CONICYT C01E02 grants (Chile).

$\tau\widehat{u}_0$, where $\tau \in \mathbb{R}$, represents a small unknown perturbation of the initial velocity field u_0, and $h = h(x,t)$ is a control function to be determined.

Notice that the Coriolis force is represented by a zero order coupling term in the equations. It introduces a different behavior of the system depending on the direction in space. To simplify the presentation of the results, we will assume that $A = 1$, $\gamma = 1$, $f_0 = 1$, $\beta = 1$, and $\rho_0 = 1$.

We introduce the following spaces, which are usual in the analysis of Stokes systems:

$$H = \{v \in L^2(\Omega)^2 : \ \operatorname{div} v = 0 \text{ in } \Omega, \ v \cdot \nu = 0 \text{ on } \Gamma\},$$
$$V = \{v \in H_0^1(\Omega)^2 : \ \operatorname{div} v = 0 \text{ in } \Omega\}, \qquad W = H^2(\Omega)^2 \cap V.$$

Recall that

$$W \hookrightarrow V \hookrightarrow H \equiv H' \hookrightarrow V' \hookrightarrow W',$$

where the embeddings are dense and compact.

For any given $u_0, \tau\widehat{u}_0 \in H$ with $\|\widehat{u}_0\|_{0,\Omega} = 1$, any $\mathcal{T} \in L^2(Q)^2$, and any $h \in L^2(\omega \times (0,T))^2$, the linear system (1.1) possesses a unique solution (u,p), with $u \in L^2(0,T;V) \cap H^1(0,T;V')$ and $p \in W^{-1,\infty}(0,T;L^2(\Omega))$. ($p$ is unique up to an additive distribution only depending on t.) This is easily proved by adapting the arguments of [17] to the presence of a skew-symmetric Coriolis term in the equations. Notice that if we had $u_0 + \tau\widehat{u}_0 \in V$, then the couple (u,p) would satisfy $u \in L^2(0,T;W) \cap H^1(0,T;H)$ and $p \in L^2(0,T;H^1(\Omega))$.

We will be concerned with the search of controls such that the velocity measurements over an observation set are either insensitive or almost insensitive to small variations of the initial conditions. To do this, we will use *insensitizing control theory*.

1.2. Insensitizing controls and controllability. Let $\mathcal{O}$ be an open nonempty subset of Ω and let us introduce the following functional, defined on the family of solutions to (1.1):

$$\Phi(u) = \frac{1}{2}\int_0^T \int_{\mathcal{O}} |u(x,t)|^2 \, dx \, dt. \tag{1.2}$$

The notion of *insensitizing controls* was introduced by Lions [13]. In the context of (1.1)–(1.2), it reads as follows.

DEFINITION 1.1. *We say that the control $h \in L^2(\omega \times (0,T))^2$ is Φ insensitizing if*

$$\frac{d}{d\tau}\Phi(u)\Big|_{\tau=0} = 0 \qquad \forall \widehat{u}_0 \in H \quad \text{with} \quad \|\widehat{u}_0\|_{0,\Omega} = 1. \tag{1.3}$$

On the other hand, we say that $h \in L^2(\omega \times (0,T))^2$ is Φ ε-insensitizing if

$$\left|\frac{d}{d\tau}\Phi(u)\Big|_{\tau=0}\right| \le \varepsilon \qquad \forall \widehat{u}_0 \in H \quad \text{with} \quad \|\widehat{u}_0\|_{0,\Omega} = 1. \tag{1.4}$$

Of course, in (1.3) and in (1.4) u is, together with p, the solution to (1.1).

The Φ insensitizing (resp., Φ ε-insensitizing) controls h must be interpreted as those leading to an observation $\Phi(u)$ that is locally independent (resp., almost independent) at the initial perturbation $\tau\widehat{u}_0$. The existence of such controls is a pertinent

question, since it is realistic to assume that the true initial conditions for (1.1) are unknown. In fact, as noticed in [13], it would be more convenient to search for Ψ insensitizing (or Ψ ε-insensitizing) controls, where

$$\Psi(u) = \frac{1}{2}\int_0^T \int_{\mathcal{O}} |\operatorname{curl} u(x,t)|^2 \, dx\, dt,$$

but this is beyond the scope of this article and will be the subject of future work.

It is easy to characterize the insensitivity (resp., ε-insensitivity) property in terms of exact null controllability (resp., approximate controllability) of a related cascade system. Indeed, let $(\bar{u}, \bar{p})$ and (q, r) be the solutions of the following systems:

$$\begin{cases} \bar{u}_t - \Delta \bar{u} + \bar{u} + (1 + x_2)\, k \wedge \bar{u} + \nabla \bar{p} = \mathcal{T} + h 1_\omega & \text{in } Q, \\ \operatorname{div} \bar{u} = 0 & \text{in } Q, \\ \bar{u} = 0 & \text{on } \Sigma, \\ \bar{u}(0) = u_0 & \text{in } \Omega, \end{cases} \tag{1.5}$$

$$\begin{cases} -q_t - \Delta q + q - (1 + x_2)\, k \wedge q + \nabla \pi = \bar{u} 1_{\mathcal{O}} & \text{in } Q, \\ \operatorname{div} q = 0 & \text{in } Q, \\ q = 0 & \text{on } \Sigma, \\ q(T) = 0 & \text{in } \Omega. \end{cases} \tag{1.6}$$

Then the control h is Φ insensitizing (resp., Φ ε-insensitizing) if and only if

$$q(0) = 0 \qquad (\text{resp., } \ \|q(0)\|_{0,\Omega} \leq \varepsilon). \tag{1.7}$$

Indeed, in view of (1.2), condition (1.3) is equivalent to

$$\int_0^T \int_{\mathcal{O}} \bar{u} \cdot u_\tau \, dx\, dt = 0 \quad \left(\text{resp., (1.4) is equivalent to } \left| \int_0^T \int_{\mathcal{O}} \bar{u} \cdot u_\tau \, dx\, dt \right| \leq \varepsilon \right),$$

where $\bar{u}$ is the solution of (1.5) and u_τ is the solution of (1.1) differentiated with respect to τ. Using the definition of (q, π) and integrating by parts, we obtain

$$\int_\Omega q(0) \cdot \widehat{u}_0 \, dx = 0 \quad \left(\text{resp., } \left| \int_\Omega q(0) \cdot \widehat{u}_0 \, dx \right| \leq \varepsilon \right) \quad \forall \widehat{u}_0 \in H \ \text{ with } \|\widehat{u}_0\|_{0,\Omega} = 1.$$

This is equivalent to (1.7). See [18] for more detail.

Notice that since $\bar{u} \in L^2(0,T;V)$, we also have $q \in L^2(0,T;W) \cap H^1(0,T;H)$ and $\pi \in L^2(0,T;H^1(\Omega))$.

We are thus in the presence of a null controllability problem (resp., an approximate controllability problem) for a cascade system, where the control h is not acting directly in the system satisfied by q (the function we want to drive to zero after a time interval of length T) but indirectly, through $\bar{u}1_{\mathcal{O}}$.

1.3. Main results. There have been several recent results concerning the existence of insensitizing and ε-insensitizing controls for parabolic problems.

Thus, in [2] the existence of ε-insensitizing controls for linear heat equations with partially known initial and boundary conditions was established. The same was also obtained for semilinear heat equations with globally Lipschitz-continuous nonlinearities. Since then, it has been proved in [18] that insensitizing controls exist for the same equations completed with zero initial data, under suitable assumptions

on the source term. In [3], the authors extended these results to other more general (slightly superlinear) nonlinearities.

In this paper, we deal with the insensitizing and ε-insensitizing problems for the case of the Stokes-type equations (1.1). Our results were sketched in [7]. These are the first insensitivity results in the literature for equations of this type, as far as we know.

We will assume that the following geometrical hypothesis is satisfied, as in the previous references:

$$\omega \cap \mathcal{O} \neq \emptyset. \tag{1.8}$$

Our main results are the following.

THEOREM 1.2. *Let $T > 0$ and assume that (1.8) is satisfied. Then, for each $\varepsilon > 0$ there exists a control $h \in L^2(\omega \times (0,T))^2$ which is Φ ε-insensitizing.*

THEOREM 1.3. *Under the assumptions of Theorem 1.2, if we also have $u_0 = 0$ and*

$$\int_0^T \int_\Omega \exp\left(Mt^{-4}\right) \mathcal{T}^2 \, dx \, dt < +\infty \tag{1.9}$$

for an appropriate constant M depending on Ω, ω, $\mathcal{O}$, and T, then there exists a control $h \in L^2(\omega \times (0,T))^2$ which is Φ insensitizing.

It was proved in [18] for the linear heat equation that, in general, we cannot expect the existence of insensitizing controls for nonvanishing initial data in $L^2(\Omega)$ when $\Omega \setminus \overline{\omega} \neq \emptyset$. The proof of this result is based on a counterexample for which the appropriate *observability inequality* fails when the initial data belong to $L^2(\Omega)$. Similar arguments could be used for Stokes systems. In view of this, it is reasonable to impose in Theorem 1.3 that $u_0 = 0$.

This paper is organized as follows. In section 2, we prove Theorem 1.2, where we obtain a unique continuation result for an adjoint cascade system thanks to the presence of the Coriolis term. In section 3, we prove Theorem 1.3. In this section, we show that insensitizing controls do exist if an appropriate observability inequality holds. We deduce this observability inequality in section 3.2 by means of an appropriate global Carleman inequality for the same adjoint cascade system. The proof of this global Carleman inequality is given in section 3.1 and follows a chain of estimates based on the steps of the unique continuation proof. At the end of this section and to be self-contained, we give the proof of a standard global Carleman estimate for Stokes-like systems that is needed in section 3.1. Finally, in section 4, we summarize the key points of this article in some final remarks.

2. Proof of Theorem 1.2. We can assume without loss of generality that $\mathcal{T} = 0$ and $u_0 = 0$ in (1.5)–(1.6). It is well known that the existence of ε-insensitizing controls for (1.5)–(1.6) is equivalent to a *unique continuation property* of the associate adjoint system

$$\begin{cases} \phi_t - \Delta\phi + \phi + (1 + x_2)\, k \wedge \phi + \nabla\theta = 0 \quad \text{in } Q, \\ \operatorname{div} \phi = 0 \quad \text{in } Q, \\ \phi = 0 \quad \text{on } \Sigma, \\ \phi(0) = \phi_0 \quad \text{in } \Omega, \end{cases} \tag{2.1}$$

$$\begin{cases} -z_t - \Delta z + z - (1 + x_2)\, k \wedge z + \nabla r = \phi 1_{\mathcal{O}} \quad \text{in } Q, \\ \operatorname{div} z = 0 \quad \text{in } Q, \\ z = 0 \quad \text{on } \Sigma, \\ z(T) = 0 \quad \text{in } \Omega \end{cases} \tag{2.2}$$

for a given $\phi_0 \in H$. This coupled system possesses a unique solution (ϕ, θ), (z, r), with at least $\phi, z \in L^2(0,T;V) \cap H^1(0,T;V')$ and $\theta, r \in W^{-1,\infty}(0,T;L^2(\Omega))$. (Again, θ and r are unique up to a distribution depending only on t.)

Using (1.5)–(1.6) and (2.1)–(2.2) the following duality identity is easily deduced:

$$\int_0^T \int_\omega h \cdot z \, dx \, dt = \int_\Omega q(0) \cdot \phi_0 \, dx \qquad \forall h \in L^2(\omega \times (0,T))^2.$$

It is clear from this last identity that the set $\{ q(0) : h \in L^2(\omega \times (0,T))^2 \}$ is dense in H if the following unique continuation result holds.

LEMMA 2.1. *Assume* (1.8). *Let* (ϕ, θ), (z, r) *be the solution to* (2.1)–(2.2) *with* $\phi_0 \in H$. *Then, if* $z = 0$ *in* $\omega \times (0,T)$, *we necessarily have* $z \equiv \phi \equiv 0$ *and* $\nabla r \equiv \nabla \theta \equiv 0$ *in* Q.

Proof. This is a direct consequence of a more general unique continuation result. To state this result precisely, let $\widetilde{\omega} = \omega \cap \mathcal{O} \neq \emptyset$ and let us set

$$C_1(\widetilde{\omega}) = \{(x_1, x_2) \in \Omega \,:\, \exists x_1^0 \ \text{ s.t. } (x_1^0, x_2) \in \widetilde{\omega}\}, \quad \Sigma_1(\widetilde{\omega}) = (\Gamma \cap \overline{C}_1) \times (0,T). \tag{2.3}$$

($C_1(\widetilde{\omega})$ is the horizontal component of $\widetilde{\omega}$.) We will prove that if $\phi = (\phi_1, \phi_2)$ is together with θ, z, and r a solution of

$$\begin{cases} \phi_t - \Delta\phi + \phi + (1 + x_2)\, k \wedge \phi + \nabla\theta = 0 \quad \text{in } Q, \\ \operatorname{div} \phi = 0 \quad \text{in } Q, \\ \phi_1 = 0 \quad \text{on } \Sigma_1, \end{cases} \tag{2.4}$$

$$\begin{cases} -z_t - \Delta z + z - (1 + x_2)\, k \wedge z + \nabla r = \phi 1_{\widetilde{\omega}} \quad \text{in } Q, \\ \operatorname{div} z = 0 \quad \text{in } Q, \end{cases}$$

and $z = 0$ in $\widetilde{\omega} \times (0,T)$, then $\phi \equiv 0$.

To prove this assertion, we divide the proof into two steps. Without loss of generality, we can assume that $\widetilde{\omega}$ is connected; otherwise we would replace $\widetilde{\omega}$ by one of its connected components.

In a first step, we deduce from the fact that $z = 0$ in $\widetilde{\omega} \times (0,T)$ that $\phi_2 = 0$ and ϕ_1 is constant if they are restricted to $\widetilde{\omega} \times (0,T)$. Thus, since $z = 0$ in $\widetilde{\omega} \times (0,T)$ we notice that $\operatorname{curl} \phi = 0$ in $\widetilde{\omega} \times (0,T)$ by applying the curl operator in the equation of z in (2.4). Using this fact, if we now apply the curl operator to the first equation in (2.4), thanks to the presence of the Coriolis term we obtain that

$$\operatorname{curl}\left((1 + x_2)\, k \wedge \phi\right) = \phi_2 + \operatorname{div} \phi = \phi_2 = 0$$

in $\widetilde{\omega} \times (0,T)$. Now, since $\operatorname{div} \phi = 0$ and $\operatorname{curl} \phi = 0$ in $\widetilde{\omega} \times (0,T)$, we have $\nabla \phi_1 = 0$ in $\widetilde{\omega} \times (0,T)$. Therefore ϕ_1 is constant in $\widetilde{\omega} \times (0,T)$ and we certainly obtain $\phi = (\text{Const.}, 0)$ in $\widetilde{\omega} \times (0,T)$.

In a second step, let us introduce

$$\psi = \frac{\partial \phi}{\partial x_1}, \quad \pi = \frac{\partial \theta}{\partial x_1} \tag{2.5}$$

and the coefficient matrix:

$$a = \begin{bmatrix} 1 & -(1 + x_2) \\ (1 + x_2) & 1 \end{bmatrix} \in L^\infty_{loc}(Q). \tag{2.6}$$

Then we have

$$\begin{cases} \psi_t - \Delta\psi + a\psi + \nabla\pi = 0 \quad \text{in } Q, \\ \operatorname{div}\psi = 0 \quad \text{in } Q, \end{cases} \tag{2.7}$$

$$(\psi, \pi) \in L^2_{loc}(Q)^2 \times \mathcal{D}'(Q) \tag{2.8}$$

with $\psi = 0$ in $\widetilde{\omega} \times (0,T)$. Here, we use a sharp uniqueness property for the Stokes system (2.7) proved in [5] that says that, under the regularity determined by (2.6) and (2.8), one has $\psi \equiv 0$ in Q. Now, from (2.5) we obtain $\partial\phi_i/\partial x_1 \equiv 0$ for $i = 1, 2$ in Q. Since $\operatorname{div}\phi = 0$ in Q we also have $\nabla\phi_2 = 0$ in Q and, from the fact that $\phi_2 = 0$ in $\widetilde{\omega} \times (0,T)$, we deduce that $\phi_2 \equiv 0$ in Q (recall that Ω is connected).

On the other hand, since $\partial\phi_1/\partial x_1 = 0$ in Q and $\phi_1 = 0$ on Σ_1 , we see that $\phi_1 = 0$ in $C_1 \times (0,T)$. Finally we have $\phi = (\phi_1, \phi_2) = 0$ in $C_1 \times (0,T)$, which is an open subset of Q. We can conclude that $\phi \equiv 0$ in Q using again the uniqueness property of [5]. (We can also use here the weaker result proved in [4].) □

Remark 1. The method used in the second part of the proof of Lemma 2.1 leads to the following uniqueness property in any dimension n. Let (ϕ, θ) be the solution of

$$\begin{cases} \phi_t - \Delta\phi + a\,\phi + \nabla\theta = 0 \quad \text{in } Q, \\ \operatorname{div}\phi = 0 \quad \text{in } Q, \\ \phi = 0 \quad \text{on } \Sigma_1(\omega), \end{cases} \tag{2.9}$$

where $Q = \Omega \times (0,T)$, Ω is a nonempty open bounded connected subset of $\mathbb{R}^n$, ω is an open nonempty subset of Ω, Σ_1 and C_1 are as defined in (2.3), and $a \in L^\infty(Q)$. If a is a function independent of x_1 in Q and ϕ is independent of x_1 in $\omega \times (0,T)$, then ϕ vanishes in Q. Indeed, let us introduce $\psi = \partial\phi/\partial x_1$, $\pi = \partial\theta/\partial x_1$, which satisfy a Stokes problem similar to (2.9), and this problem does not involve ϕ explicitly since a is independent of x_1. Now, from the uniqueness property in [5], $\psi \equiv 0$ in Q. Consequently $\partial\phi/\partial x_1 = 0$ in Q and $\phi = 0$ on Σ_1, so we have $\phi = 0$ in $C_1 \times (0,T)$. Using the unique continuation property in [5] once again, we obtain that $\phi \equiv 0$ in Q.

Remark 2. The previous remark shows that the Coriolis term plays a crucial role only in the first part of the proof of Lemma 2.1. In fact, the presence of the Coriolis term allows us to prove that $\operatorname{curl}\phi = 0$ in $\widetilde{\omega} \times (0,T)$ implies that the second component of ϕ vanishes in $\widetilde{\omega} \times (0,T)$. This will also be important in the deduction of the Carleman inequality later.

Remark 3. In the proof of the previous lemma, it is not possible to use the results of [4] concerning uniqueness properties of the Stokes system when one of the components of ϕ vanishes in $\widetilde{\omega} \times (0,T)$. This is because the results in [4] require that the coefficient a, introduced in (2.6), satisfy $a_{12} = 0$.

3. Proof of Theorem 1.3. The proof of the existence of insensitizing controls for (1.1), i.e., the exact null controllability for (1.5)–(1.6), relies on the following observability result for the cascade adjoint system (2.1)–(2.2).

PROPOSITION 3.1. *Assume that $\omega \cap \mathcal{O} \neq \emptyset$. There exist positive constants M and K, depending only on Ω, ω, $\mathcal{O}$, and T, such that the inequality*

$$\int_0^T \int_\Omega \exp\left(-Mt^{-4}\right)|z|^2\,dx\,dt \leq K \int_0^T \int_\omega |z|^2\,dx\,dt \tag{3.1}$$

holds for every solution of (2.1)–(2.2) *with $\phi_0 \in H$.*

The proof of this result is based on a global Carleman inequality (see Theorem 3.3), as will be seen in section 3.2. This Carleman inequality will be proved in section 3.1.

Let us now give the proof of Theorem 1.3 using Proposition 3.1. Thus, let us assume that (1.8) is satisfied, $u_0 = 0$, and (1.9) holds with M being the constant furnished by Proposition 3.1.

The approximate control h of minimal norm in $L^2(\omega \times (0,T))^2$ corresponding to $u_0 = 0$, a source term $\mathcal{T}$ satisfying (1.9), and tolerance $\varepsilon > 0$ can be obtained by minimizing in $L^2(\Omega)^2$ the following convex functional [6, 14]:

$$J_\varepsilon(\phi_0) = \frac{1}{2}\int_0^T\int_\omega |z|^2\,dx\,dt + \int_0^T\int_\Omega \mathcal{T}\cdot z\,dx\,dt + \varepsilon\|\phi_0\|_{0,\Omega}. \tag{3.2}$$

Thus, the minimum of J_ε is attained at some $\widehat{\phi}_{0\varepsilon} \in L^2(\Omega)^2$. We denote by $(\widehat{\phi}_\varepsilon, \widehat{\theta}_\varepsilon)$, $(\widehat{z}_\varepsilon, \widehat{r}_\varepsilon)$ the corresponding solution to (2.1)–(2.2) with $\phi_0 = \widehat{\phi}_{0\varepsilon}$; then the control function defined as

$$h_\varepsilon = \widehat{z}_\varepsilon 1_\omega \tag{3.3}$$

is such that the associated solution $(\bar{u}_\varepsilon, \bar{p}_\varepsilon)$, $(q_\varepsilon, \pi_\varepsilon)$ to (1.5)–(1.6) with $u_0 = 0$ satisfies $\|q_\varepsilon(0)\|_{0,\Omega} \le \varepsilon$.

It is not difficult to see that

$$\liminf_{\|\phi_0\|_{0,\Omega}\to\infty} \frac{J_\varepsilon(\phi_0)}{\|\phi_0\|_{0,\Omega}} \ge \varepsilon.$$

The proof of this inequality is classical; see [6]. It is implied by the unique continuation property for the cascade adjoint system that we presented above (see Lemma 2.1).

Furthermore, the following optimality condition must be satisfied at $\widehat{\phi}_{0\varepsilon}$:

$$\int_0^T\int_\omega |\widehat{z}_\varepsilon|^2\,dx\,dt + \int_0^T\int_\Omega \mathcal{T}\cdot\widehat{z}_\varepsilon\,dx\,dt + \varepsilon\|\widehat{\phi}_{0\varepsilon}\|_{0,\Omega} = 0. \tag{3.4}$$

By replacing (3.3) in (3.4), introducing the weight $e^{Mt^{-4}}$, and using (3.1) and Young's inequality, we easily deduce that

$$\int_0^T\int_\omega |h_\varepsilon|^2\,dx\,dt \le K^2\int_0^T\int_\Omega \exp(Mt^{-4})|\mathcal{T}|^2\,dx\,dt.$$

Since $\{h_\varepsilon\}$ is uniformly bounded in $L^2(\omega\times(0,T))^2$, then up to a subsequence, still denoted $\{h_\varepsilon\}$, we have

$$\begin{aligned} &h_\varepsilon \rightharpoonup h \quad \text{weakly in} \quad L^2(\omega\times(0,T))^2,\\ &\bar{u}_\varepsilon \to \bar{u} \quad \text{strongly in} \quad L^2(Q)^2, \quad \text{and}\\ &q_\varepsilon \to q \quad \text{strongly in} \quad L^2(Q)^2, \end{aligned}$$

as $\varepsilon \to 0$. Of course, we have denoted here by $(\bar{u}_\varepsilon, \bar{p}_\varepsilon)$, $(q_\varepsilon, \pi_\varepsilon)$ and $(\bar{u}, \bar{p})$, (q, π) the solutions to (1.5)–(1.6) associated with h_ε and h, respectively. Notice that $\|q_\varepsilon(0)\|_{0,\Omega} \le \varepsilon$ and consequently we have $q(0) = 0$. This ends the proof of Theorem 1.3.

3.1. A global Carleman estimate. The goal of this section is to present an estimate of the Carleman kind for the solutions to the adjoint cascade system (2.1)–(2.2). As mentioned above, this estimate will be crucial for the proof of Proposition 3.1.

Let us first introduce an open ball B_0 such that $B_0 \subset\subset \omega \cap \mathcal{O}$ and an auxiliary function $\eta_0 \in \mathcal{C}^2(\overline{\Omega})$ satisfying

$$\eta_0(x) > 0 \ \ \forall x \in \Omega, \quad \eta_0 = 0 \ \text{ on } \partial\Omega, \quad |\nabla \eta_0(x)| > 0 \ \ \forall x \in \overline{\Omega \setminus B_0}\,. \tag{3.5}$$

The existence of such a function is proved in [9].

Let us also introduce the weight functions

$$\alpha(x,t) = \frac{e^{2\lambda\|\eta_0\|_\infty} - e^{\lambda\eta_0}}{t^4(T-t)^4}, \quad \widehat{\alpha}(t) = \min_{\overline{\Omega}} \alpha(x,t), \quad \alpha^*(t) = \max_{\overline{\Omega}} \alpha(x,t),$$
$$\varphi(x,t) = \frac{e^{\lambda\eta_0}}{t^4(T-t)^4}, \qquad \widehat{\varphi}(t) = \max_{\overline{\Omega}} \varphi(x,t), \quad \varphi^*(t) = \min_{\overline{\Omega}} \varphi(x,t).$$

The following property of the functions α^* and $\widehat{\alpha}$ will be needed later.

LEMMA 3.2. *For any $a > 1$ there exists $\lambda_a > 0$ such that*

$$a\,\widehat{\alpha}(t) > \alpha^*(t) \quad \forall \lambda > \lambda_a\,, \quad \forall t \in (0,T).$$

Proof. The proof is elementary. It suffices to notice that we have $a(e^{2x} - e^x) > e^{2x} - 1$ if $a > 1$ and x is sufficiently large. ∎

The main result in this section is the following.

THEOREM 3.3. *Assume that $\omega \cap \mathcal{O} \neq \emptyset$ and let the functions α, φ, $\widehat{\alpha}$, and $\widehat{\varphi}$ be as above. For each $\widehat{\gamma} \in (0,1)$, there exist constants $\widehat{s}$, $\widehat{\lambda}$, and $\widehat{C}$ depending on Ω, ω, $\mathcal{O}$, T, and $\widehat{\gamma}$ such that one has*

$$\begin{aligned}
&\int_0^T\!\!\int_\Omega e^{-2s\alpha}\left(\frac{1}{s\varphi}(|z_t|^2 + |\Delta z|^2) + s\lambda^2\varphi|\nabla z|^2 + s^3\lambda^4\varphi^3|z|^2\right) dx\,dt \\
&\quad + \int_0^T\!\!\int_\Omega e^{-2s\alpha}\left(\frac{1}{s\varphi}(|\phi_t|^2 + |\Delta\phi|^2) + s\lambda^2\varphi|\nabla\phi|^2 + s^3\lambda^4\varphi^3|\phi|^2\right) dx\,dt \\
&\quad \le \widehat{C}\int_0^T\!\!\int_\omega e^{-(1+\widehat{\gamma})s\widehat{\alpha}} s^{63}\lambda^{32}\widehat{\varphi}^{67}|z|^2\,dx\,dt
\end{aligned} \tag{3.6}$$

for any $s > \widehat{s}$ and $\lambda > \widehat{\lambda}$ and for every solution (ϕ,θ), (z,r) to (2.1)–(2.2) associated with initial data $\phi_0 \in H$.

The proof will be divided in several steps and will be given in the following subsections. First, we will apply a global Carleman estimate for the Stokes system to (2.1) and (2.2). This will lead to the estimate (3.10). Then, to deduce (3.6), we will have to estimate the integral in the right-hand side of (3.10) containing ϕ in terms of z. To this end, we will follow the steps of the proof of Lemma 2.1 in reverse order.

3.1.1. Step 1: A first direct Carleman estimate. Let $I(s,\lambda;v)$ stand for the quantity

$$I(s,\lambda;v) = \int_0^T\!\!\int_\Omega e^{-2s\alpha}\left(\frac{1}{s\varphi}(|v_t|^2 + |\Delta v|^2) + s\lambda^2\varphi|\nabla v|^2 + s^3\lambda^4\varphi^3|v|^2\right) dx\,dt \tag{3.7}$$

for any positive s and λ and any sufficiently regular function $v = v(x,t)$. We then have the following.

LEMMA 3.4. *For each $\gamma_1 \in (0,1)$ there exist positive constants s_1, λ_1, and C_1, depending on Ω, ω, $\mathcal{O}$, T, and γ_1, with the following properties:*

$$
\begin{aligned}
I(s,\lambda;z) \le C_1 \Bigg\{ & \int_0^T \int_{B_0} e^{-(1+\gamma_1)s\widehat{\alpha}} s^7 \lambda^4 \widehat{\varphi}^{15/2} |z|^2 \, dx\, dt \\
& + \int_0^T \int_{B_0} e^{-2s\widehat{\alpha}} (s\lambda\widehat{\varphi})^2 |\phi|^2 \, dx\, dt \\
& + \int_0^T \int_{\mathcal{O}} e^{-2s\alpha} \left((s\varphi)^{1/2} |\phi|^2 + \frac{1}{s^3\varphi^{7/2}} |\phi_t|^2 \right) dx\, dt \Bigg\}
\end{aligned} \tag{3.8}
$$

and

$$
I(s,\lambda;\phi) \le C_1 \int_0^T \int_{B_0} e^{-(1+\gamma_1)s\widehat{\alpha}} s^7 \lambda^4 \widehat{\varphi}^{15/2} |\phi|^2 \, dx\, dt \tag{3.9}
$$

for any $s > s_1$ and $\lambda > \lambda_1$ and for every solution of (2.1)–(2.2) *with $\phi_0 \in H$.*

The proof of Lemma 3.4 is similar to the proof of other recent global Carleman inequalities for the Stokes system. The main ideas are due to Imanuvilov [10, 11]; also see [8] for other related results. The proof is presented in the appendix.

Let us fix $\widehat{\gamma}$, with $0 < \widehat{\gamma} < 1$. We are now going to deduce several estimates that hold for "sufficiently large s and λ." By this we mean that they are satisfied for any $s > \bar{s}$ and any $\lambda > \bar{\lambda}$, where $\bar{s}$ and $\bar{\lambda}$ are (large) positive constants depending only on Ω, ω, $\mathcal{O}$, T, and $\widehat{\gamma}$.

In what follows, C denotes a generic constant, not necessarily the same at each occurrence, depending on Ω, ω, $\mathcal{O}$, T, and (possibly) $\widehat{\gamma}$.

Let γ_1 be given in $(\widehat{\gamma}, 1)$. In view of Lemma 3.4 applied to γ_1, we get

$$
I(s,\lambda;z) + I(s,\lambda;\phi) \le C \int_0^T \int_{B_0} e^{-(1+\gamma_1)s\widehat{\alpha}} s^7 \lambda^4 \widehat{\varphi}^{15/2} (|z|^2 + |\phi|^2) \, dx\, dt \tag{3.10}
$$

for s and λ large enough.

Indeed, the last two integrals in (3.8) can be absorbed by the left-hand side of $I(s,\lambda;\phi)$, since

$$
Cs^{-3}\varphi^{-7/2} \le \frac{1}{2}(s\varphi)^{-1} \quad \text{and} \quad C(s\varphi)^{1/2} \le \frac{1}{2} s^3 \varphi^3
$$

for sufficiently large s.

3.1.2. Step 2: An estimate of ϕ in terms of curl ϕ. To simplify the notation, let us set $a = 7$ and $b = 15/2$. Then

$$
I(s,\lambda;z) + I(s,\lambda;\phi) \le C \int_0^T \int_{B_0} e^{-(1+\gamma_1)s\widehat{\alpha}} s^a \lambda^4 \widehat{\varphi}^b (|z|^2 + |\phi|^2) \, dx\, dt. \tag{3.11}
$$

We will denote by $B_1, B_2, \ldots$ a sequence of balls centered at the same point as B_0 and satisfying

$$
B_0 \subset\subset B_1 \subset\subset \cdots \subset\subset \omega \cap \mathcal{O}.
$$

It is not a restriction to assume that their common center is the origin. This will be supposed in what follows for simplicity. We will consider some functions $\xi_i \in \mathcal{C}_0^\infty(B_i)$ satisfying

$$\begin{gathered} 0 \le \xi_i \le 1, \quad \xi_i(x) = 1 \text{ in } B_{i-1}\,, \\ \xi_i^{-1/2}\nabla\xi_i \in L^\infty(\Omega), \quad \xi_i^{-1/2}\Delta\xi_i \in L^\infty(\Omega). \end{gathered} \tag{3.12}$$

(See [18] for a justification of the existence of these ξ_i.)

Since $\operatorname{div}\phi = 0$, $\phi = 0$ on Σ, and Ω is connected, we can introduce the stream function ψ satisfying

$$\phi = \operatorname{curl}\psi \equiv \left(\frac{\partial\psi}{\partial x_2}, -\frac{\partial\psi}{\partial x_1}\right),$$

with $\psi = 0$ on one connected component of Σ and $\frac{\partial\psi}{\partial n} = 0$ on Σ.

Let us set $\rho_1(t) = e^{-(1+\gamma_1)s\widehat{\alpha}}s^a\lambda^4\widehat{\varphi}^b$. Then we have

$$\int_0^T\int_{B_0}\rho_1|\phi|^2\,dx\,dt \le \int_0^T\int_{B_1}\rho_1\xi_1|\nabla\psi|^2\,dx\,dt.$$

We will now give an estimate of the last integral in terms of $|\operatorname{curl}\phi|^2$. To this end, let us introduce the vorticity w, given by

$$w = \operatorname{curl}\phi = \frac{\partial\phi_2}{\partial x_1} - \frac{\partial\phi_1}{\partial x_2}.$$

Applying the curl operator to (2.1), we obtain

$$\begin{cases} w_t - \Delta w + w - \dfrac{\partial\psi}{\partial x_1} = 0 \quad \text{in } Q, \\ \Delta\psi + w = 0 \quad \text{in } Q. \end{cases} \tag{3.13}$$

To estimate $|\nabla\psi|^2$, we multiply by $\rho_1\xi_1\psi$ the second equation of (3.13). Then, we integrate by parts with respect to the space variable x and we get

$$\int_0^T\int_{B_1}\rho_1\xi_1|\nabla\psi|^2\,dx\,dt = \int_0^T\int_{B_1}\rho_1\xi_1\psi w\,dx\,dt + \frac{1}{2}\int_0^T\int_{B_1}\rho_1(\Delta\xi_1)|\psi|^2\,dx\,dt. \tag{3.14}$$

Notice that using $I(s,\lambda;\phi)$, we can get upper bounds for $|\psi|^2$, $|\nabla\psi|^2$, and $|\psi_t|^2$. Indeed, from the definition of α^*, φ^*, and $\widehat{\varphi}$, we have

$$\begin{aligned} I(s,\lambda;\phi) &\ge \int_0^T\int_\Omega e^{-2s\alpha}\left(\frac{1}{s\varphi}|\nabla\psi_t|^2 + s^3\lambda^4\varphi^3|\nabla\psi|^2\right)dx\,dt \\ &\ge \int_0^T\int_\Omega e^{-2s\alpha^*}\left(\frac{1}{s\widehat{\varphi}}|\nabla\psi_t|^2 + s^3\lambda^4(\varphi^*)^3|\nabla\psi|^2\right)dx\,dt \\ &\ge C\int_0^T\int_\Omega e^{-2s\alpha^*}\left(\frac{1}{s\widehat{\varphi}}|\psi_t|^2 + s^3\lambda^4(\varphi^*)^3\left(|\psi|^2 + |\nabla\psi|^2\right)\right)dx\,dt. \end{aligned} \tag{3.15}$$

Here we have used the fact that $\psi = 0$ on one of the connected components of Σ to apply Poincaré's inequality.

With this information, we will be able to absorb the first integral in (3.14). Indeed, after using Young's inequality, we can estimate this term as follows:

$$\int_0^T\int_{B_1}\rho_1\xi_1\psi w\,dx\,dt \le \delta\int_0^T\int_\Omega e^{-2s\alpha^*}s^3\lambda^4(\varphi^*)^3|\psi|^2\,dx\,dt + C_\delta\int_0^T\int_{B_1} e^{-2(1+\gamma_1)s\widehat{\alpha}+2s\alpha^*}s^{2a-3}\lambda^4\widehat{\varphi}^{2b-3}|w|^2\,dx\,dt. \tag{3.16}$$

Now, if we introduce γ_2 with $0<\gamma_2<2\gamma_1-1$, then $(1+2\gamma_1-\gamma_2)/2>1$ and, from Lemma 3.2, we see that $(1+2\gamma_1-\gamma_2)\widehat{\alpha}/2>\alpha^*$ for λ sufficiently large. Consequently, it can be assumed that

$$-2(1+\gamma_1)\widehat{\alpha}+2\alpha^*<-(1+\gamma_2)\widehat{\alpha}$$

and we can replace $e^{-2(1+\gamma_1)s\widehat{\alpha}+2s\alpha^*}$ by $e^{-(1+\gamma_2)s\widehat{\alpha}}$ in the last integral in (3.16):

$$\int_0^T\int_{B_1}\rho_1\xi_1\psi w\,dx\,dt \le \delta\int_0^T\int_\Omega e^{-2s\alpha^*}s^3\lambda^4(\varphi^*)^3|\psi|^2\,dx\,dt + C_\delta\int_0^T\int_{B_1} e^{-(1+\gamma_2)s\widehat{\alpha}}s^{2a-3}\lambda^4\widehat{\varphi}^{2b-3}|w|^2\,dx\,dt. \tag{3.17}$$

Notice that if we had chosen γ_1 sufficiently close to 1 before, then we would still have the possibility of choosing γ_2 satisfying $\widehat{\gamma}<\gamma_2<2\gamma_1-1$.

On the other hand, by choosing δ sufficiently small, we can absorb the first term in the right-hand side of (3.17) with $I(s,\lambda;\phi)$.

It remains in this step to estimate the last integral in (3.14). Assume that ξ_1 has been constructed as before but also satisfying

$$\xi_1(x)=\begin{cases}1 & \text{in } |x|<r_0,\\ \widehat{\Psi}\left(\frac{|x|-r_0}{r_1-a-r_0}\right) & \text{in } r_0\le|x|\le r_1-a,\\ 0 & \text{in } |x|>r_1-a,\end{cases}$$

where r_i denotes the radius of B_i, a is small enough, and $\widehat{\Psi}$ is a function satisfying $\widehat{\Psi}\in\mathcal{C}^\infty([0,1])$,

$$\widehat{\Psi}(0)=1,\quad \widehat{\Psi}(1)=0,\quad \text{and}\quad \widehat{\Psi}^{(n)}(0)=\widehat{\Psi}^{(n)}(1)=0\quad\forall n\ge1.$$

Let us set

$$\eta(x)=\int_{\bar{x}_1}^{x_1}\Delta\xi_1(y_1,x_2)\,dy_1$$

where for each $x=(x_1,x_2)\in\bar{B}_1$ we take $\bar{x}_1<x_1$ and $(\bar{x}_1,x_2)\in\partial B_1$. Notice that $\frac{\partial\eta}{\partial x_1}=\Delta\xi_1$. It is also easy to see that Supp $\eta\subset\bar{B}_1(0;r_1-a)$. And now, using the first equation in (3.13), we observe that

$$\begin{aligned}\frac{1}{2}\int_0^T\int_{B_1}\rho_1(\Delta\xi_1)|\psi|^2\,dx\,dt &= \frac{1}{2}\int_0^T\int_{B_1}\rho_1\frac{\partial\eta}{\partial x_1}|\psi|^2\,dx\,dt\\ &= -\int_0^T\int_{B_1}\rho_1\eta\psi\frac{\partial\psi}{\partial x_1}\,dx\,dt\\ &= -\int_0^T\int_{B_1}\rho_1\eta\psi(w_t-\Delta w+w)\,dx\,dt.\end{aligned} \tag{3.18}$$

Remark 4. Notice that we used the term $\frac{\partial \psi}{\partial x_1}$ in equation (3.13) to estimate $|\psi|^2$ over B_1. The term comes from Coriolis force and it is absent in the Stokes system.

We will now estimate this last integral in the right-hand side of (3.18). Concerning the product $\rho_1 \eta \psi w_t$, we can integrate by parts with respect to time in $B_1 \times (0,T)$ and then apply Young's inequality to deduce that

$$\begin{aligned}\int_0^T\!\!\int_{B_1} \rho_1 \eta \psi w_t \, dx\, dt &= -\int_0^T\!\!\int_{B_1} (\rho_1 \eta \psi_t w + \rho_1' \eta \psi w)\, dx\, dt \\ &\le \delta \int_0^T\!\!\int_\Omega e^{-2s\alpha^*} \left(\frac{1}{s\widehat{\varphi}} |\psi_t|^2 + s^3 \lambda^4 (\varphi^*)^3 |\psi|^2 \right) dx\, dt \\ &\quad + C_\delta \int_0^T\!\!\int_{B_1} e^{-(1+\gamma_2)s\widehat{\alpha}} (s^{2a+1}\lambda^8 \widehat{\varphi}^{2b+1} + s^{2a-1}\lambda^4 \widehat{\varphi}^{2b-1/2}) |w|^2 \, dx\, dt \end{aligned} \tag{3.19}$$

for sufficiently large s and λ.

To obtain this inequality, we first used that

$$|\rho_1'| = \left| (e^{-(1+\gamma_1)s\widehat{\alpha}} s^a \lambda^4 \widehat{\varphi}^b)_t \right| \le C e^{-(1+\gamma_1)s\widehat{\alpha}} s^{a+1} \lambda^4 \widehat{\varphi}^{b+5/4}.$$

Then, we noticed that

$$\begin{aligned}&\int_0^T\!\!\int_{B_1} \rho_1' \eta \psi w \, dx\, dt \\ &\quad \le \delta \int_0^T\!\!\int_\Omega e^{-2s\alpha^*} s^3 \lambda^4 (\varphi^*)^3 |\psi|^2 \, dx\, dt \\ &\quad + C_\delta \int_0^T\!\!\int_{B_1} e^{-2(1+\gamma_1)s\widehat{\alpha} + 2s\alpha^*} s^{2a-1} \lambda^4 \widehat{\varphi}^{2b+5/2} (\varphi^*)^{-3} |w|^2 \, dx\, dt, \end{aligned}$$

and, finally, we took s and λ large enough to have

$$e^{-2(1+\gamma_1)s\widehat{\alpha} + 2s\alpha^*} \widehat{\varphi}^{2b+5/2} (\varphi^*)^{-3} \le e^{-(1+\gamma_2)s\widehat{\alpha}} \widehat{\varphi}^{2b-1/2}.$$

We can simplify the estimate (3.19) by using the inequality

$$s^{2a-1} \widehat{\varphi}^{2b-1/2} \le C s^{2a+1} \widehat{\varphi}^{2b+1},$$

which must hold for large s. Thus, we obtain

$$\begin{aligned}&\int_0^T\!\!\int_{B_1} \rho_1 \eta \psi w_t \, dx\, dt \\ &\quad \le \delta \int_0^T\!\!\int_\Omega e^{-2s\alpha^*} \left(\frac{1}{s\widehat{\varphi}} |\psi_t|^2 + s^3 \lambda^4 (\varphi^*)^3 |\psi|^2 \right) dx\, dt \\ &\quad + C_\delta \int_0^T\!\!\int_{B_1} e^{-(1+\gamma_2)s\widehat{\alpha}} s^{2a+1} \lambda^8 \widehat{\varphi}^{2b+1} |w|^2 \, dx\, dt. \end{aligned} \tag{3.20}$$

Notice that the first integral in the right-hand side of (3.20) also appears in (3.15) and can be absorbed later by choosing δ small enough.

Let us now consider the term $\rho_1 \eta \psi (\Delta w)$ in the last integral of (3.18). Let us integrate by parts with respect to the space variable x, let us use the identity $\Delta \psi = w$,

and let us apply Young's inequality. Arguing as before, we obtain

$$\begin{aligned}\int_0^T\int_{B_1}\rho_1\eta\psi(\Delta w)\,dx\,dt &= \int_0^T\int_{B_1}\rho_1\left((\Delta\eta)\psi w+2\nabla\eta\cdot\nabla\psi\, w+\eta|w|^2\right)\,dx\,dt\\ &\le \delta\int_0^T\int_\Omega e^{-2s\alpha^*}s^3\lambda^4(\varphi^*)^3(|\psi|^2+|\nabla\psi|^2)\,dx\,dt\\ &\quad + C_\delta\int_0^T\int_{B_1}e^{-(1+\gamma_2)s\widehat{\alpha}}s^{2a-3}\lambda^4\widehat{\varphi}^{\,2b-3}|w|^2\,dx\,dt \qquad (3.21)\end{aligned}$$

for any sufficiently large s and λ.

Finally, arguing in a similar way, we can also estimate the last term $\rho_1\eta\psi w$ in (3.18):

$$\begin{aligned}\int_0^T\int_{B_1}\rho_1\eta\psi w\,dx\,dt &\le \delta\int_0^T\int_{B_1}e^{-2s\alpha^*}s^3\lambda^4(\varphi^*)^3|\psi|^2\,dx\,dt\\ &\quad + C_\delta\int_0^T\int_{B_1}e^{-(1+\gamma_2)s\widehat{\alpha}}s^{2a-3}\lambda^4\widehat{\varphi}^{\,2b-3}|w|^2\,dx\,dt. \qquad (3.22)\end{aligned}$$

From (3.18) and (3.20)–(3.22), we find that

$$\begin{aligned}&\frac12\int_0^T\int_{B_1}\rho_1(\Delta\xi_1)|\psi|^2\,dx\,dt\\ &\qquad\le 3\delta\int_0^T\int_\Omega e^{-2s\alpha^*}\left(\frac{1}{s\widehat{\varphi}}|\psi_t|^2+s^3\lambda^4(\varphi^*)^3(|\psi|^2+|\nabla\psi|^2)\right)\,dx\,dt\\ &\qquad\quad + C_\delta\int_0^T\int_{B_1}\rho_2|w|^2\,dx\,dt, \qquad (3.23)\end{aligned}$$

where

$$\rho_2(t)=e^{-(1+\gamma_2)s\widehat{\alpha}}s^{2a+1}\lambda^8\widehat{\varphi}^{\,2b+1}.$$

Replacing the estimates (3.16) and (3.23) in (3.10), with $\delta>0$ sufficiently small, we obtain

$$I(s,\lambda;z)+I(s,\lambda;\phi)\le C\left\{\int_0^T\int_{B_0}\rho_1|z|^2\,dx\,dt+\int_0^T\int_{B_1}\rho_2|\operatorname{curl}\phi|^2\,dx\,dt\right\}. \qquad (3.24)$$

3.1.3. Step 3: An estimate of $\operatorname{curl}\phi$ in terms of z. Let us apply the curl operator to (2.2). For $\zeta=\operatorname{curl} z$, we obtain the following:

$$-\zeta_t-\Delta\zeta+\zeta-z_2=w1_{\mathcal{O}}\quad\text{in }\mathcal{O}\times(0,T).$$

Recall that $\xi_2\in\mathcal{C}_0^\infty(B_2)$ satisfies (3.12) and $B_1\subset\subset B_2\subset\subset\omega\cap\mathcal{O}$. After multiplying the above equation by $\rho_2\xi_2 w$, integrating by parts in Q, and using (3.13), it follows that

$$\begin{aligned}\int_0^T\int_{B_2}\rho_2\xi_2|w|^2\,dx\,dt &= -\int_0^T\int_{B_2}\rho_2\xi_2\phi_2\zeta\,dx\,dt+\int_0^T\int_{B_2}\rho_2'\xi_2 w\zeta\,dx\,dt\\ &\quad -\int_0^T\int_{B_2}\rho_2((\Delta\xi_2)w\zeta+2(\nabla\xi_2\cdot\nabla w)\zeta+\xi_2 wz_2)\,dx\,dt. \qquad (3.25)\end{aligned}$$

As before, we choose γ_3 satisfying $0 < \gamma_3 < 2\gamma_2 - 1$. Then, for sufficiently large λ we have $(1 + 2\gamma_2 - \gamma_3)\widehat{\alpha}/2 > \alpha^*$ and, consequently,

$$-2(1+\gamma_2)\widehat{\alpha} + 2s\alpha < -2(1+\gamma_2)\widehat{\alpha} + 2s\alpha^* < -(1+\gamma_3)\widehat{\alpha}.$$

Notice once more that if γ_1 is sufficiently close to 1, then we can choose γ_3 in $(\widehat{\gamma}, 1)$.

Now, proceeding as in the previous step, we see that

$$\left|\int_0^T\int_{B_2} \rho_2\xi_2\phi_2\zeta\,dx\,dt\right| \le \delta\int_0^T\int_\Omega e^{-2s\alpha}s^3\lambda^4\varphi^3|\phi_2|^2\,dx\,dt + C_\delta\int_0^T\int_{B_2}\rho_3\frac{1}{s^4\lambda^4\widehat{\varphi}^4}\xi_2^2|\zeta|^2\,dx\,dt$$

for any small $\delta > 0$ (to be fixed later). Here, ρ_3 stands for the function

$$\rho_3(t) = e^{-(1+\gamma_3)s\widehat{\alpha}(t)}s^{4a+3}\lambda^{16}\widehat{\varphi}^{4b+3}(t).$$

We also have

$$\int_0^T\int_{B_2}\rho_2'\xi_2 w\zeta\,dx\,dt \le \delta\int_0^T\int_{B_2}\rho_2\xi_2|w|^2\,dx\,dt + C_\delta\int_0^T\int_{B_2}\rho_2 s^2\widehat{\varphi}^{5/2}\xi_2|\zeta|^2\,dx\,dt.$$

Furthermore, after separating the terms in the last integral in (3.25), we find that

$$\left|\int_0^T\int_{B_2}\rho_2(\Delta\xi_2)w\zeta\,dx\,dt\right| \le \delta\int_0^T\int_{B_2}\rho_2\xi_2|w|^2\,dx\,dt + C_\delta\int_0^T\int_{B_2}\rho_2\frac{|\Delta\xi_2|^2}{\xi_2}|\zeta|^2\,dx\,dt$$

and

$$\left|\int_0^T\int_{B_2}\rho_2(\nabla\xi_2\cdot\nabla w)\,\zeta\,dx\,dt\right| \le \delta\int_0^T\int_\Omega e^{-2s\alpha}\frac{1}{s\varphi}|\Delta\phi|^2\,dx\,dt + C_\delta\int_0^T\int_{B_2}\rho_3|\nabla\xi_2|^2|\zeta|^2\,dx\,dt.$$

In this last estimate we have used that $|\nabla w|^2 = |\Delta\phi|^2$. Finally,

$$\left|\int_0^T\int_{B_2}\rho_2\xi_2 w z_2\,dx\,dt\right| \le \delta\int_0^T\int_{B_2}\rho_2\xi_2|w|^2\,dx\,dt + C_\delta\int_0^T\int_{B_2}\rho_2\xi_2|z_2|^2\,dx\,dt.$$

In view of (3.25) and all these inequalities, we obtain

$$\begin{aligned}\int_0^T\int_{B_2}\rho_2\xi_2|\mathrm{curl}\,\phi|^2\,dx\,dt &\le \frac{\delta}{1-3\delta}\int_0^T\int_\Omega e^{-2s\alpha}\left(\frac{1}{s\varphi}|\Delta\phi|^2 + s^3\lambda^4\varphi^3|\phi_2|^2\right)dx\,dt\\ &\quad + C_\delta\int_0^T\int_{B_2}\left(\rho_2|z_2|^2 + \rho_3\bar{\xi}_2|\mathrm{curl}\,z|^2\right)dx\,dt \qquad (3.26)\end{aligned}$$

for some $\bar{\xi}_2 \in \mathcal{C}_0^\infty(B_2)$.

It remains to estimate the previous integral of $\rho_3\bar{\xi}_2|\mathrm{curl}\, z|^2$. Arguing as above, we see that

$$\begin{aligned}\int_0^T\int_{B_2}\rho_3\bar{\xi}_2|\mathrm{curl}\, z|^2\,dx\,dt &\le 2\int_0^T\int_{B_2}\rho_3\bar{\xi}_2|\nabla z|^2\,dx\,dt\\ &= -2\int_0^T\int_{B_2}\rho_3(\nabla\bar{\xi}_2\cdot\nabla z+\bar{\xi}_2(\Delta z))z\,dx\,dt\\ &\le \delta\int_0^T\int_\Omega e^{-2s\alpha}\left(s\lambda^2\varphi|\nabla z|^2+\frac{1}{s\varphi}|\Delta z|^2\right)dx\,dt\\ &\quad + C_\delta\int_0^T\int_{B_2}\rho_4|z|^2\,dx\,dt, \end{aligned}\tag{3.27}$$

where

$$\rho_4(t)=e^{-(1+\gamma_4)s\widehat{\alpha}}s^{8a+7}\lambda^{32}\widehat{\varphi}^{8b+7}$$

for some γ_4 satisfying $0<\gamma_4<2\gamma_3-1$. For the reasons stated above, it is clear that γ_4 can be assumed to satisfy $\widehat{\gamma}<\gamma_4<1$.

Choosing $\delta>0$ small enough and replacing the estimates (3.26) and (3.27) in (3.24), we obtain

$$I(s,\lambda;z)+I(s,\lambda;\phi)\le C\int_0^T\int_\omega\rho_4|z|^2\,dx\,dt\tag{3.28}$$

for all large s and λ. Taking into account the definition of ρ_4, that $\gamma_4>\widehat{\gamma}$, $a=7$, and $b=15/2$, we see that (3.6) holds.

This ends the proof of Theorem 3.3.

3.2. Proof of Proposition 3.1. Let us now give the proof of the observability inequality (3.1) for solutions of system (2.1)–(2.2), which relies on the above result. First, we observe that from the continuous dependence of the solution of (2.2), we have

$$\int_{T/2}^T\int_\Omega|z|^2\,dx\,dt\le C\int_{T/2}^T\int_{\mathcal{O}}|\phi|^2\,dx\,dt,\tag{3.29}$$

and from classical energy estimates for system (2.1), using the fact that $(k\wedge\phi)\cdot\phi=0$, we obtain an energy decreasing property:

$$\|\phi(t+T/4)\|_{0,\Omega}^2\le C\|\phi(t)\|_{0,\Omega}^2\qquad\forall t\in(T/4,3T/4)\,.$$

If we integrate the last inequality over the time interval $(T/4,3T/4)$ and change the integral variable $t\to t+T/4$ on the left-hand integral, we can easily deduce that

$$\int_{T/2}^T\int_\Omega|\phi|^2\,dx\,dt\le C\int_{T/4}^{3T/4}\int_\Omega|\phi|^2\,dx\,dt,\tag{3.30}$$

where C is independent of ϕ.

In what follows, we will fix s, λ, and $\widehat{\gamma}$ as in Theorem 3.3, depending on Ω, ω, $\mathcal{O}$, and T, such that (3.6) is satisfied.

Let us first prove that there exist positive constants M, C_1 such that

$$\int_0^{T/2}\int_\Omega \exp(-Mt^{-4})|z|^2\,dx\,dt + \int_{T/2}^{T}\int_\Omega |\phi|^2\,dx\,dt \le C_1\int_0^T\int_\omega |z|^2\,dx\,dt. \tag{3.31}$$

For this estimate, let us first notice that for some constants M and C, one has

$$e^{-2s\alpha(x,t)}\varphi(x,t)^3 \ge Ce^{-Mt^{-4}} \quad \forall (x,t)\in \overline{\Omega}\times(0,T/2);$$

this is easy to see, in view of the definitions of α and φ. Now, using (3.6), we get

$$\begin{aligned}\int_0^{T/2}\int_\Omega \exp(-Mt^{-4})|z|^2\,dx\,dt &\le \frac{1}{s^3\lambda^4}\,I(s,\lambda;z)\\ &\le CK\int_0^T\int_\omega e^{-(1+\widehat{\gamma})s\widehat{\alpha}}s^{63}\widehat{\varphi}^{67}|z|^2\,dx\,dt,\end{aligned}$$

and since the weight $e^{-(1+\widehat{\gamma})s\widehat{\alpha}}\widehat{\varphi}^{67}$ is bounded, we can estimate the first term in the left-hand side of (3.31). On the other hand, to obtain an estimate of ϕ in terms of z, let us recall the inequality (3.6). Since $e^{-2s\alpha}t^{-12}(T-t)^{-12}$ is bounded from below far from $t=0$ and $t=T$, in view of (3.30), we have

$$\begin{aligned}\int_{T/2}^T\int_\Omega |\phi|^2\,dx\,dt &\le \int_{T/4}^{3T/4}\int_\Omega |\phi|^2\,dx\,dt\\ &\le C\int_{T/4}^{3T/4}\int_\Omega e^{-2s\alpha}s^3\lambda^4\varphi^3|\phi|^2\,dx\,dt \le C\int_0^T\int_\omega e^{-(1+\widehat{\gamma})s\widehat{\alpha}}\widehat{\varphi}^{67}|z|^2\,dx\,dt.\end{aligned}$$

As before, from the fact that $e^{-(1+\widehat{\gamma})s\widehat{\alpha}}\widehat{\varphi}^{67}$ is bounded, we are able to estimate the second term in the left-hand side of (3.31).

Finally, the desired observability inequality (3.1) is obtained using the energy estimate (3.29) and (3.31):

$$\begin{aligned}\int_0^T\int_\Omega \exp(-Mt^{-4})|z|^2\,dx\,dt &\le \int_0^{T/2}\int_\Omega \exp(-Mt^{-4})|z|^2\,dx\,dt + \int_{T/2}^T\int_\Omega |z|^2\,dx\,dt\\ &\le C\left(\int_0^{T/2}\int_\Omega \exp(-Mt^{-4})|z|^2\,dx\,dt + \int_{T/2}^T\int_\Omega |\phi|^2\,dx\,dt\right)\\ &\le C\int_0^T\int_\omega |z|^2\,dx\,dt.\end{aligned}$$

Appendix. Proof of Lemma 3.4. Let us recall that this proof is given for the sake of completeness, but it is essentially an adaptation to our framework of the arguments presented in [8] and [11]. Let us consider the system

$$\begin{cases} -z_t - \Delta z + z - (1+x_2)\,\mathrm{k}\wedge z + \nabla r = \phi 1_{\mathcal{O}} \quad \text{in } Q,\\ \operatorname{div} z = 0 \quad \text{in } Q,\\ z = 0 \quad \text{on } \Sigma,\\ z(T) = 0 \quad \text{in } \Omega,\end{cases} \tag{3.32}$$

where $\phi \in L^2(0,T;W)\cap H^1(0,T;H)$.

Recall that B_0 is an open ball satisfying $B_0 \subset\subset \omega \cap \mathcal{O}$ and the auxiliary function η_0 satisfies $\eta_0 \in \mathcal{C}^2(\overline{\Omega})$,

$$\eta_0(x) > 0 \ \ \forall x \in \Omega, \quad \eta_0 = 0 \ \text{ on } \partial\Omega, \quad |\nabla \eta_0(x)| > 0 \ \ \forall x \in \overline{\Omega \setminus B_0}\,.$$

We will need an additional open ball $B_{00} \subset\subset B_0$, such that we still have

$$|\nabla \eta_0(x)| > 0 \ \ \forall x \in \overline{\Omega \setminus B_{00}}\,.$$

We will divide the proof of Lemma 3.4 into several steps.

Step 1. Following [8], we apply some well-known Carleman estimates for the heat equation to (3.32). Thus, there exist constants s_0, λ_0, and $C > 0$ depending on Ω, ω, and T such that for every $\lambda > \lambda_0$ and $s > s_0$, the following estimate holds:

$$\begin{aligned} I(s,\lambda;z) \le C\Bigg\{ &\int_0^T \int_{B_{00}} e^{-2s\alpha} s^3\lambda^4\varphi^3 |z|^2 \,dx\,dt \\ &+ \int_0^T \int_\Omega e^{-2s\alpha} \left(|\nabla r|^2 + |(1+x_2)k \wedge z|^2 + |\phi 1_{\mathcal{O}}|^2 \right) dx\,dt \Bigg\}. \end{aligned} \tag{3.33}$$

Recall that the definitions of $I(s,\lambda;z)$ and the weights α and φ are given in section 3.1.

Of course, we can choose s large enough to absorb the previous term $|(1+x_2)k \wedge z|^2$ with the left-hand side (3.33). We then have

$$\begin{aligned} I(s,\lambda;z) \le C\Bigg\{ &\int_0^T \int_{B_{00}} e^{-2s\alpha} s^3\lambda^4\varphi^3 |z|^2 \,dx\,dt \\ &+ \int_0^T \int_\Omega e^{-2s\alpha} |\nabla r|^2 \,dx\,dt + \int_0^T \int_{\mathcal{O}} e^{-2s\alpha} |\phi|^2 \,dx\,dt \Bigg\} \end{aligned} \tag{3.34}$$

for any $\lambda > \lambda_0$ and any $s > s_{01}$.

Step 2. To estimate the pressure gradient ∇r in (3.34), we first apply the divergence operator to (3.32), i.e., we write

$$\Delta r(t) = \operatorname{div}((1+x_2)\mathrm{k} \wedge z)(t) \quad \text{in } \ \Omega, \qquad t \in (0,T), \tag{3.35}$$

and then we use the following result by Imanuvilov and Puel [12], which is satisfied by weak solutions to second order elliptic equations.

LEMMA 3.5. *Let us set $\beta(x) = e^{\lambda\eta_0(x)}$ and let $v \in H^1(\Omega)$ be a solution of*

$$\Delta v = \operatorname{div} h \quad \text{in} \quad \Omega, \tag{3.36}$$

where $h \in L^2(\Omega)^2$. Then there exist positive constants τ_2, λ_{01}, and C such that

$$\begin{aligned} \int_\Omega e^{2\tau\beta} |\nabla v|^2 \,dx \le C\Bigg\{ &\tau \int_\Omega e^{2\tau\beta} \beta |h|^2 \,dx + \tau^{1/2} e^{2\tau} \|g\|^2_{1/2,\partial\Omega} \\ &+ \tau^2\lambda^2 \int_{B_{00}} e^{2\tau\beta} \beta^2 |v|^2 \,dx + \int_{B_{00}} e^{2\tau\beta} |\nabla v|^2 \,dx \Bigg\} \end{aligned} \tag{3.37}$$

for any $\tau > \tau_2$ and any $\lambda > \lambda_{01}$, where $g = v|_{\partial\Omega}$. ∎

In particular, we have the following for $r(t)$ and $g(t) = r(t)|_{\partial\Omega}$:

$$\int_\Omega e^{2\tau\beta}|\nabla r(t)|^2\,dx \le C\Big\{\tau\int_\Omega e^{2\tau\beta}\beta|(1+x_2)k\wedge z(t)|^2\,dx + \tau^{1/2}e^{2\tau}\|g(t)\|^2_{1/2,\partial\Omega} + \tau^2\lambda^2\int_{B_{00}} e^{2\tau\beta}\beta^2|r(t)|^2\,dx + \int_{B_{00}} e^{2\tau\beta}|\nabla r(t)|^2\,dx\Big\}. \tag{3.38}$$

To estimate the last integral in (3.38), let us introduce an open set B_{01} such that $B_{00}\subset\subset B_{01}\subset\subset B_0$ and a function $\xi_{01}\in\mathcal{C}_0^2(B_{01})$ such that

$$0\le\xi_{01}\le 1 \quad\text{and}\quad \xi_{01}=1 \text{ in } B_{00}\,.$$

Integrating by parts, it follows from (3.35) that

$$\begin{aligned}\int_{B_{00}} e^{2\tau\beta}|\nabla r(t)|^2\,dx &\le \int_{B_{01}} e^{2\tau\beta}\xi_{01}|\nabla r(t)|^2\,dx\\ &= -\int_{B_{01}} e^{2\tau\beta}\xi_{01}\,\mathrm{div}((1+x_2)\mathrm{k}\wedge z)(t)r(t)\,dx\\ &\quad -\frac{1}{2}\int_{B_{01}} e^{2\tau\beta}\nabla\xi_{01}\cdot\nabla|r(t)|^2\,dx - \int_{B_{01}}\xi_{01}\nabla e^{2\tau\beta}\cdot\nabla|r(t)|^2\,dx.\end{aligned}$$

Integrating again by parts, applying Young's inequality, and taking into account that $|\Delta(e^{2\tau\beta}\xi_{01})|\le C\tau^2\lambda^2\beta^2e^{2\tau\beta}$ for some positive constant C, after some straightforward computations we deduce that

$$\int_{B_{00}} e^{2\tau\beta}|\nabla r(t)|^2\,dx \le C\Big\{\tau^2\lambda^2\int_{B_{01}} e^{2\tau\beta}\beta^2|r(t)|^2\,dx + \int_{B_{01}} e^{2\tau\beta}|z(t)|^2\,dx\Big\}.$$

Replacing this inequality in (3.38), we obtain the following for each $t\in(0,T)$:

$$\int_\Omega e^{2\tau\beta}|\nabla r(t)|^2\,dx \le C\Big\{\tau\int_\Omega e^{2\tau\beta}\beta|z(t)|^2\,dx + \tau^{1/2}e^{2\tau}\|g(t)\|^2_{1/2,\partial\Omega} + \tau^2\lambda^2\int_{B_{01}} e^{2\tau\beta}\beta^2|r(t)|^2\,dx\Big\}.$$

Now, let us put $\tau = s/(t^4(T-t)^4)$ and let us choose $s > s_{02} = \max(s_{01}\,, \tau_2(T/2)^8)$. Then $\tau > \tau_2$. Let us multiply by $\exp(-2s\exp(2\lambda\|\eta_0\|_\infty)/(t^4(T-t)^4))$ the previous inequality and let us integrate with respect to t in $(0,T)$. This leads to the estimate

$$\int_0^T\int_\Omega e^{-2s\alpha}|\nabla r|^2\,dx\,dt \le C\Big\{\int_0^T\int_\Omega e^{-2s\alpha}s\varphi|z|^2\,dx\,dt + \int_0^T e^{-2s\alpha^*}(s\varphi^*)^{1/2}\|g(t)\|^2_{1/2,\partial\Omega}\,dt + \int_0^T\int_{\omega_1} e^{-2s\alpha}(s\lambda\varphi)^2|r|^2\,dx\,dt\Big\}, \tag{3.39}$$

where α^* and φ^* were introduced in section 3.1.

The first term in the right-hand side of (3.39) can be absorbed by the left-hand side $I(s,\lambda;z)$ in (3.33) for s large enough. Hence, we obtain

$$I(s,\lambda;z) \le C\left\{\int_0^T\int_{\omega_0} e^{-2s\alpha}s^3\lambda^4\varphi^3|z|^2\,dx\,dt + \int_0^T e^{-2s\alpha^*}(s\varphi^*)^{1/2}\|g(t)\|^2_{1/2,\partial\Omega}\,dt\right.$$
$$\left. + \int_0^T\int_{\omega_1} e^{-2s\alpha}(s\lambda\varphi)^2|r|^2\,dx\,dt + \int_0^T\int_{\mathcal{O}} e^{-2s\alpha}|\phi|^2\,dx\,dt\right\} \tag{3.40}$$

for any $\lambda > \lambda_{01}$ and any $s > s_{03}$.

Step 3. Step 3 estimates the norm of the trace of the pressure on the boundary. To this end, we introduce three new functions:

$$\chi(t) = e^{-s\alpha^*(t)}(s\varphi^*(t))^{1/4}, \quad \tilde z = \chi(t)z, \qquad \tilde r = \chi(t)r.$$

From (3.32), we see that $(\tilde z,\tilde r)$ satisfies

$$\begin{cases} -\tilde z_t - \Delta\tilde z + \tilde z + \nabla\tilde r = -\chi' z + \chi(1+x_2)\,\mathrm{k}\wedge z + \chi\phi 1_{\mathcal{O}} & \text{in } Q,\\ \operatorname{div}\tilde z = 0 & \text{in } Q,\\ \tilde z = 0 & \text{on } \Sigma,\\ \tilde z(T) = 0 & \text{in } \Omega. \end{cases}$$

Using the continuity of the trace operator and standard a priori estimates for the pressure, we deduce that

$$\begin{aligned}\int_0^T \|\tilde r(t)\|^2_{1/2,\partial\Omega}\,dt &\le \int_0^T \|\tilde r(t)\|^2_{1,\Omega}\,dt\\ &\le C\left\{\int_0^T\int_\Omega e^{-2s\alpha^*}s^{5/2}(\varphi^*)^3|z|^2\,dx\,dt\right.\\ &\qquad \left. + \int_0^T\int_{\mathcal{O}} e^{-2s\alpha^*}(s\varphi^*)^{1/2}|\phi|^2\,dx\,dt\right\}.\end{aligned}$$

We have used here that $|\chi'(t)|^2 \le Ce^{-2s\alpha^*}s^{5/2}(\varphi^*(t))^3$ for all $t\in(0,T)$. We thus obtain a new estimate from (3.40):

$$I(s,\lambda;z) \le C\left\{\int_0^T\int_{B_{00}} e^{-2s\alpha}s^3\lambda^4\varphi^3|z|^2\,dx\,dt + \int_0^T\int_{B_{01}} e^{-2s\alpha}(s\lambda\varphi)^2|r|^2\,dx\,dt\right.$$
$$\left. + \int_0^T\int_{\mathcal{O}} e^{-2s\alpha}(s\varphi)^{1/2}|\phi|^2\,dx\,dt\right\} \tag{3.41}$$

for any $\lambda > \lambda_{01}$ and any $s > s_{04}$.

Step 4. It remains to estimate the local term in the right-hand side of (3.41) containing $|r|^2$ in terms of z and ϕ.

Assume that the pressure r has been normalized in such a way that

$$\int_{B_{01}} r(t)\,dx = 0 \quad \forall t\in(0,T).$$

Then there exists $C>0$ such that

$$\int_{B_{01}} |r(t)|^2\,dx \le C\int_{B_{01}} |\nabla r(t)|^2\,dx \quad \forall t\in(0,T)$$

and also

$$\int_0^T\int_{B_{01}} e^{-2s\alpha}(s\lambda\varphi)^2|r|^2\,dx\,dt \le C\int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2|\nabla r|^2\,dx\,dt,$$

where the functions $\widehat{\alpha}=\widehat{\alpha}(t)$ and $\widehat{\varphi}=\widehat{\varphi}(t)$ were introduced in section 3.1.

From (3.32), we see that

$$\int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2|\nabla r|^2\,dx\,dt \le C\Bigg\{\int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2(|z|^2+|\phi|^2)\,dx\,dt + \int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2(|z_t|^2+|\Delta z|^2)\,dx\,dt\Bigg\}.$$

Therefore, in view of (3.41), we obtain

$$\begin{aligned} I(s,\lambda;z) \le C\Bigg\{&\int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}\left(s^3\lambda^4\widehat{\varphi}^3|z|^2+(s\lambda\widehat{\varphi})^2|\phi|^2\right)\,dx\,dt\\ &+\int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2(|z_t|^2+|\Delta z|^2)\,dx\,dt\\ &+\int_0^T\int_{\mathcal{O}} e^{-2s\alpha}(s\varphi)^{1/2}|\phi|^2\,dx\,dt\Bigg\}. \end{aligned} \tag{3.42}$$

Step 5. The rest of the proof deals with the estimates of the local integrals containing $|\Delta z|^2$ and $|z_t|^2$. First, we will be concerned with $|\Delta z|^2$.

Let us introduce a function $\xi_0\in\mathcal{C}_0^4(B_0)$ such that

$$0\le\xi_0\le 1 \quad\text{and}\quad \xi_0=1 \text{ in } B_{01}\,.$$

Let us set $\widehat{z}(x,t)=e^{-s\widehat{\alpha}}\widehat{\varphi}\,\xi_0\Delta z(T-t)$. We want to estimate the norm $\|\widehat{z}\|_{L^2(B_{01}\times(0,T))^2}$. Following the arguments in [8] (see Step 4), we can deduce that

$$\begin{aligned} &\int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2|\Delta z|^2\,dx\,dt = \int_0^T\int_{B_{01}} s^2\lambda^2|\widehat{z}|^2\,dx\,dt\\ &\le C\left(\int_0^T\int_{B_0} e^{-2s\widehat{\alpha}}s^4\lambda^2\widehat{\varphi}^{9/2}|z|^2\,dx\,dt+\int_0^T\int_{B_0} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2|\phi|^2\,dx\,dt\right). \end{aligned} \tag{3.43}$$

Thus, from (3.42) we have

$$\begin{aligned} &I(s,\lambda;z)\\ &\le C\Bigg(\int_0^T\int_{B_0} e^{-2s\widehat{\alpha}}s^4\lambda^4\widehat{\varphi}^{9/2}|z|^2\,dx\,dt+\int_0^T\int_{B_0} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2|\phi|^2\,dx\,dt\\ &\qquad+\int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2|z_t|^2\,dx\,dt+\int_0^T\int_{\mathcal{O}} e^{-2s\alpha}(s\varphi)^{1/2}|\phi|^2\,dx\,dt\Bigg). \end{aligned} \tag{3.44}$$

Step 6. Now we want to estimate $|z_t|^2$. Due to the regularity properties of ϕ, we can use here a more straightforward argument than in [8], where the right-hand side belongs only to $L^2(Q)^2$.

First, notice that

$$\int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2|z_t|^2\,dx\,dt \le \delta\int_0^T\int_{B_{01}} e^{-2s\alpha}\frac{1}{s\varphi}|z_t|^2\,dx\,dt + \delta\int_0^T\int_{B_{01}} e^{-2s\alpha^*}\frac{1}{s^3(\varphi^*)^{7/2}}|z_{tt}|^2\,dx\,dt + C_\delta\int_0^T\int_{B_{01}} e^{-4s\alpha^*+2s\alpha^*}s^7\lambda^4\widehat{\varphi}^{15/2}|z|^2\,dx\,dt.$$

This is easily obtained by integrating by parts in time. We will later choose $\delta > 0$ small enough.

We have the following auxiliary result.

LEMMA 3.6. *Let (z, r) be the solution of* (3.32). *Then the following estimate holds:*

$$\int_0^T\int_\Omega e^{-2s\alpha^*}\frac{1}{s^3(\varphi^*)^{7/2}}|z_{tt}|^2\,dx\,dt \le C\left(I(s,\lambda;z) + \int_0^T\int_{\mathcal{O}} e^{-2s\alpha^*}\left(\frac{1}{s\varphi^*}|\phi|^2 + \frac{1}{s^3(\varphi^*)^{7/2}}|\phi_t|^2\right)dx\,dt\right). \tag{3.45}$$

Proof. Multiply (3.32) by $e^{-2s\alpha^*}s^{-2}(\varphi^*)^{-9/4}z_{tt}$ and integrate in Q. Noticing that

$$\left|(e^{-2s\alpha^*}(\varphi^*)^{-9/4})_t\right| \le Ce^{-2s\alpha^*}s(\varphi^*)^{-1},$$

after some computations we deduce that

$$\begin{aligned}\int_0^T\int_\Omega e^{-2s\alpha^*}\frac{1}{s^2(\varphi^*)^{9/4}}|\nabla z_t|^2\,dx\,dt &\le C\left\{\int_0^T\int_\Omega e^{-2s\alpha^*}\frac{1}{s\varphi^*}(|z|^2+|z_t|^2)\,dx\,dt\right.\\ &\quad\left.+\int_0^T\int_\Omega e^{-2s\alpha^*}s(\varphi^*)^{1/4}|\nabla z|^2\,dx\,dt + \int_0^T\int_{\mathcal{O}} e^{-2s\alpha^*}\frac{1}{s\varphi^*}|\phi|^2\,dx\,dt\right\}\\ &\quad+\frac{1}{2}\int_0^T\int_\Omega e^{-2s\alpha^*}\frac{1}{s^3(\varphi^*)^{7/2}}|z_{tt}|^2\,dx\,dt.\end{aligned} \tag{3.46}$$

On the other hand, if we compute the time derivative of (3.32) and then we multiply the result by $e^{-2s\alpha^*}s^{-3}(\varphi^*)^{-7/2}z_{tt}$, we find that

$$\begin{aligned}\int_0^T\int_\Omega e^{-2s\alpha^*}\frac{1}{s^3(\varphi^*)^{7/2}}|z_{tt}|^2\,dx\,dt &\le \int_0^T\int_\Omega e^{-2s\alpha^*}\frac{1}{s^2(\varphi^*)^{9/4}}|\nabla z_t|^2\,dx\,dt\\ &\quad+C\left(\int_0^T\int_\Omega e^{-2s\alpha^*}\frac{1}{s\varphi^*}|z_t|^2\,dx\,dt + \int_0^T\int_{\mathcal{O}} e^{-2s\alpha^*}\frac{1}{s^3(\varphi^*)^{7/2}}|\phi_t|^2\,dx\,dt\right).\end{aligned} \tag{3.47}$$

From (3.46) and (3.47), we see that (3.45) holds. ∎

In view of this lemma, we have

$$\begin{aligned}
&\int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2|z_t|^2\,dx\,dt\\
&\quad\le C\delta\left(I(s,\lambda;z)+\int_0^T\int_{\mathcal{O}} e^{-2s\alpha^*}\left(\frac{1}{s\varphi^*}|\phi|^2+\frac{1}{s^3(\varphi^*)^{7/2}}|\phi_t|^2\right)dx\,dt\right)\\
&\quad+C_\delta\int_0^T\int_{B_{01}} e^{-4s\alpha^*+2s\alpha^*}s^7\lambda^4\widehat{\varphi}^{15/2}|z|^2\,dx\,dt.
\end{aligned}$$

If we assume that $\gamma_1<1$, then $(3-\gamma_1)/2>1$, and from Lemma 3.2 we deduce that $(3-\gamma_1)\widehat{\alpha}/2>\alpha^*$ for sufficiently large λ, say, $\lambda>\lambda_{02}$. Consequently, $-4\widehat{\alpha}+2\alpha^*<-(1+\gamma_1)\widehat{\alpha}$ and

$$\begin{aligned}
&\int_0^T\int_{B_{01}} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2|z_t|^2\,dx\,dt\\
&\quad\le C\delta\left(I(s,\lambda;z)+\int_0^T\int_{\mathcal{O}} e^{-2s\alpha^*}\left(\frac{1}{s\varphi^*}|\phi|^2+\frac{1}{s^3(\varphi^*)^{7/2}}|\phi_t|^2\right)dx\,dt\right)\\
&\quad+C_\delta\int_0^T\int_{B_{01}} e^{-(1+\gamma_1)s\widehat{\alpha}}s^7\lambda^4\widehat{\varphi}^{15/2}|z|^2\,dx\,dt
\end{aligned}$$

for any $\lambda>\lambda_{02}$ and any $s>s_{04}$.

From (3.44) and this estimate, choosing $\delta>0$ small enough, we find

$$\begin{aligned}
&I(s,\lambda;z)\\
&\quad\le C\left\{\int_0^T\int_{B_0} e^{-(1+\gamma_1)s\widehat{\alpha}}s^7\lambda^4\widehat{\varphi}^{15/2}|z|^2\,dx\,dt+\int_0^T\int_{B_0} e^{-2s\widehat{\alpha}}(s\lambda\widehat{\varphi})^2|\phi|^2\,dx\,dt\right.\\
&\qquad\left.+\int_0^T\int_{\mathcal{O}} e^{-2s\alpha}\left((s\varphi)^{1/2}|\phi|^2+\frac{1}{s^3(\varphi^*)^{7/2}}|\phi_t|^2\right)dx\,dt\right\}
\end{aligned}$$

for all $\lambda>\lambda_{02}$ and $s>s_{04}$.

Obviously, this yields (3.8). The proof of (3.9) is very similar and in fact much simpler, since the left-hand side of (2.1) is zero.

Thus, we have proved Lemma 3.4 for $\lambda_1=\lambda_{02}$ and $s_1=s_{04}$ (two parameters depending on Ω, ω, $\mathcal{O}$, and T).

4. Some final remarks. The geometrical hypothesis $\omega\cap\mathcal{O}\neq\emptyset$ is required to prove the existence of both ε-insensitizing and insensitizing controls. In the first case, this assumption is used to prove a unique continuation property (Lemma 2.1). In the case of insensitizing controls, it is used to prove an observability inequality. The problem is completely open when $\omega\cap\mathcal{O}=\emptyset$ (see [18]).

The existence of insensitizing controls is guaranteed by the null controllability property of a cascade system of quasi-geostrophic equations (1.5)–(1.6). In this case, the control acts indirectly on one variable through the other one. Of course, this controllability property is stronger than the null controllability of a single quasi-geostrophic system.

To prove the null controllability property of the cascade system, there are two main difficulties.

First is the need for a unique continuation result for the adjoint system (2.2)–(2.1). The presence of the Coriolis term permits us to relate the second component of the velocity and its associated vorticity, and this is a key point in the proof of uniqueness (see Remark 2).

The second problem is the need for an observability inequality for the adjoint. This inequality comes from an appropriate (global) Carleman estimate. The main idea is to estimate ϕ in terms of $\operatorname{curl}\phi$ in a ball contained in $\omega \cap \mathcal{O}$ in the right-hand side of (3.10). This is possible again due to the presence of the Coriolis term (in fact, our method does not work in the case of the usual Stokes equations). We rewrite the system using the stream function and the vorticity and we see that the Coriolis term leads to an expression of the horizontal derivative of the stream function in terms of the vorticity. In this way, we are able to avoid estimates of pressure terms, which are in general very hard to deduce (see the appendix in section 3). Moreover, the weight in the right-hand side of (3.10) is larger than the weight in the left-hand side. Accordingly, the terms in the right cannot be absorbed directly as in the case of the heat equation (see [18]) and this fact requires some additional work.

Acknowledgments. The authors wish to thank J.-P. Puel of the Laboratoire de Mathématiques Appliquées, Université Versailles/Saint Quentin-en-Yvelines (France) and S. Guerrero of the Departamento de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla (Spain), for very useful discussions.

REFERENCES

[1] R. Bermejo and P. G. Del Sastre, *Numerical studies of the long-term dynamics of the 2D Navier-Stokes equations applied to ocean circulation*, in XVII CEDYA: Congress on Differential Equations and Applications, Universidad de Salamanca, 2001, pp. 15–34.

[2] O. Bodart and C. Fabre, *Controls insensitizing the norm of the solution of a semilinear heat equation*, J. Math. Anal. Appl., 195 (1995), pp. 658–683.

[3] O. Bodart, M. González-Burgos, and R. Pérez-García, *Insensitizing controls for a semilinear heat equation with a superlinear nonlinearity*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 677–682.

[4] C. Fabre, *Uniqueness results for Stokes equations and their consequences in linear and nonlinear control problems*, ESAIM Control Optim. Calc. Var., 1 (1996), pp. 267–302.

[5] C. Fabre and G. Lebeau, *Régularité et unicité pour le problème de Stokes*, Comm. Partial Differential Equations, 27 (2002), pp. 437–475.

[6] C. Fabre, J.-P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[7] E. Fernández-Cara, G. C. García, and A. Osses, *Insensitizing controls for a large-scale ocean circulation model*, C. R. Math. Acad. Sci. Paris, 337 (2003), pp. 265–270.

[8] E. Fernández-Cara, S. Guerrero, O. Y. Imanuvilov, and J.-P. Puel, *Local exact controllability of the Navier-Stokes system*, J. Math. Pures Appl. (9), 83 (2004), pp. 1501–1542.

[9] A. Fursikov and O. Y. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Research Institute of Mathematics, Seoul National University, Korea, 1996.

[10] O. Y. Imanuvilov, *On exact controllability of Navier-Stokes equations*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 97–131.

[11] O. Y. Imanuvilov, *Remarks on exact controllability for the Navier-Stokes equations*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 39–72.

[12] O. Y. Imanuvilov and J.-P. Puel, *Global Carleman estimates for weak solutions of elliptic non homogeneous Dirichlet problems*, Int. Math. Res. Not., 16 (2003), pp. 883–913.

[13] J. L. Lions, *Quelques notions dans l'analyse et le contrôle de systèmes à données incomplètes*, in Proceedings of the XIth Congress on Differential Equations and Applications/First Congress on Applied Mathematics, University of Málaga, 1990, pp. 43–54.

[14] J. L. Lions, *Remarks on approximate controllability*, J. Anal. Math., 59 (1992), pp. 103–116.

[15] G. I. Marchuk, V. I. Agoshkov, and V. P. Shutyaev, *Adjoint Equations and Perturbation Algorithms in Nonlinear Problems*, CRC Press, Boca Raton, FL, 1996.
[16] P. G. Myers and A. J. Weaver, *A diagnostic barotropic finite-element ocean circulation model*, J. Atmos. Ocean Tech., 12 (1995), pp. 511–526.
[17] R. Temam, *Navier-Stokes Equations*, 2nd ed., North-Holland, Amsterdam, 1984.
[18] L. de Teresa, *Insensitizing control for a semilinear heat equation*, Comm. Partial Differential Equations, 25 (2000), pp. 39–72.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1640–1675

ADAPTIVE WAVELET METHODS FOR LINEAR-QUADRATIC ELLIPTIC CONTROL PROBLEMS: CONVERGENCE RATES*

WOLFGANG DAHMEN† AND ANGELA KUNOTH‡

Abstract. We propose an adaptive algorithm based on wavelets for the fast numerical solution of control problems governed by elliptic boundary value problems with distributed or Neumann boundary control. A quadratic cost functional that may involve fractional Sobolev norms of the state and the control is to be minimized subject to linear constraints in weak form. Placing the problem into the framework of (biorthogonal) wavelets allows us to formulate the functional and the constraints equivalently in terms of ℓ_2-norms of wavelet expansion coefficients and constraints in the form of an ℓ_2 automorphism. The resulting first order necessary conditions are then derived as a (still infinite) system in ℓ_2. Applying the machinery developed in [A. Cohen, W. Dahmen, and R. DeVore, *Math. Comp.*, 70 (2001), pp. 27–75; A. Cohen, W. Dahmen, and R. DeVore, *Found. Comput. Math.*, 2 (2002), pp. 203–245], we propose an adaptive method which can be interpreted as an inexact gradient descent method, where in each iteration step the primal and the adjoint system need to be solved up to a prescribed accuracy. Convergence of the adaptive algorithm is proved. In addition, we show that the adaptive algorithm is asymptotically optimal, that is, the convergence rate achieved for computing the solution up to a desired target tolerance is asymptotically the same as the wavelet-best N-term approximation of the solution, and the total computational work is proportional to the number of computational unknowns.

Key words. optimal control, elliptic boundary value problem, wavelets, infinite ℓ_2-system, preconditioning, adaptive refinements, inexact iterations, convergence, convergence rates, optimal complexity

AMS subject classifications. 65K10, 65N99, 93B40

DOI. 10.1137/S0363012902419199

1. Introduction.

A new type of adaptive wavelet method for the numerical solution of a wide class of variational problems has been developed and analyzed in a series of papers [CDD1, CDD2, CDD3]. These methods have been shown to exhibit asymptotically computational complexity in the following sense. If the solution can be approximated (using ideal complete information) by N terms from the underlying wavelet basis with accuracy $\mathcal{O}(N^{-s})$ (in the energy norm), then the scheme recovers for a certain range of decay rates s, depending on the wavelet basis, the solution with any desired target accuracy ε at a computational expense that stays proportional to $\varepsilon^{-1/s}$, uniformly in ε, and matches in this sense the optimal work/accuracy rate of best *N-term approximation*.

Moreover, the underlying analysis has lead to a new algorithmic paradigm that can be summarized as follows:

(i) Establish *well-posedness* of the underlying variational problem, which is to identify a Hilbert space (energy space) for which the operator induced by the

*Received by the editors December 8, 2002; accepted for publication (in revised form) June 9, 2004; published electronically March 11, 2005. This work was supported in part by the Deutsche Forschungsgemeinschaft SFB 401, RWTH Aachen, SFB 611, Universität Bonn, and the European Community's Human Potential Programme under contract HPRN-CT-2002-00286 (Breaking Complexity).

http://www.siam.org/journals/sicon/43-5/41919.html

†Institut für Geometrie und Praktische Mathematik, RWTH Aachen, 52056 Aachen, Germany (dahmen@igpm.rwth-aachen.de, www.igpm.rwth-aachen.de/dahmen).

‡Institut für Angewandte Mathematik, Universität Bonn, Wegelerstr. 6, 53115 Bonn, Germany (kunoth@iam.uni-bonn.de, www.iam.uni-bonn.de/∼kunoth).

variational problem is boundedly invertible as a mapping from this Hilbert space onto its dual.

(ii) Transform the original problem into an equivalent one that is now well posed in the Euclidean metric ℓ_2. This is usually done by finding a *wavelet basis* that is a Riesz basis for the energy space.

(iii) Exploit (ii) so as to devise an iterative scheme for the (still infinite dimensional) transformed problem on ℓ_2 that has a fixed error reduction per step.

(iv) Perform the ideal iteration from (iii) approximately by *adaptively* applying the involved operators in wavelet coordinates within suitable dynamically updated accuracy tolerances.

The objective for this paper is to explore the use of such concepts in the context of *optimal control problems* with PDE constraints. We are primarily motivated by the following two aspects. By their very nature, such control problems tend to have a rather demanding computational complexity, so that the use of schemes that minimize computational complexity is very tempting. Second, since the above paradigm tries to stay with the infinite dimensional well-posed problem as long as possible, it turns out to inherit the stability of the infinite dimensional problem in the following sense. Compatibility conditions on finite dimensional trial spaces that may arise in coupled problems, such as in the form of the LBB condition for saddle point problems, do not arise in the adaptive context; see [CDD2, DDU]. Moreover, the fact that suitable scalings of one and the same wavelet basis form Riesz bases for a whole range of Sobolev spaces allows one to treat in a convenient way (at least up to equivalence) a variety of such norms in the objective functional whose realization (of any equivalent version) poses severe difficulties in conventional settings. We view the fact that the use of such norms (even when realized only up to equivalence) offers greater flexibility in balancing data and regularization as an additional *modeling* tool. We shall further comment on this point below.

To bring out the basic mechanisms, we deliberately confine the discussion to rather simple types of control problems with linear constraints, including Dirichlet and Neumann problems with distributed or Neumann boundary controls. The setting will be described in section 2, along with some examples that will guide the subsequent developments. While the above paradigm was developed mainly for PDEs or singular integral equations, the first issue will be to formulate the optimal control problem in a way that allows us to branch into the above road map. This will be done in section 3, which provides the background for (ii). These formulations also shed some light on principal preferences concerning the formulation of reduced minimization problems through elimination of the state variable. Moreover, whether the penalty term in the cost functional is needed for regularization becomes obvious; see sections 3.4 and 3.5. These issues might be less apparent when working from the outset with a finite dimensional discretized problem.

To understand the potential but also the limitations of the concepts developed in this paper, it is important to distinguish two slightly different problem scenarios. In the first, a specific cost functional is imposed by the (physical) problem background, and it is mandatory to compute an optimal solution with respect to this special objective function. The quadratic forms appearing in this case induce typically (weighted) L_2-norms or first order Sobolev norms (with possibly varying diffusion coefficients). We shall refer to this case in the following as the *mandatory case*.

On the other hand, in particular, when the cost functional involves a regularization term, the purpose of formulating an optimal control problem is often to accom-

plish a possibly good compromise between fitting the data under the given constraints and a necessary regularization. One usually has only estimates for the regularization parameter and, in principle, one obtains different optimal solutions when varying this parameter slightly. In such a case there is possibly no *unique* optimization problem associated with the underlying task, but the formulation of the objective functional becomes part of the modeling process; see, e.g., [BBDM]. An interesting way to enrich this modeling process, beyond varying a single scale weight, is to incorporate mechanisms that allow one to affect contributions with different characteristic length scales in a different way when searching for a good compromise between data fit and regularization. This could help, for instance, to exploit best a possibly higher regularity of the observed data to suppress unwanted undulations in the final solution. Moreover, imposing more regularity on the control might significantly improve its practical executability in applications. While this is not possible by a global weight (or weight function implanted in a single norm), smoothness norms of Sobolev or Besov type would serve that purpose. (See, e.g., [CDLL] for application of such norms in image compression or, e.g., [CK] in the context of scattered data fitting.) Such norms, especially when the smoothness order is not an integer, typically do not have any canonical, physically motivated representer. Using extensions combined with Fourier transforms, or interpolation between integer order cases, or intrinsic norms defined by double integrals of difference quotients, or factor norms when dealing with traces, one obtains different variants whose common feature is the above-mentioned weighting of different length scales. Lacking an a priori physical reason for employing any specific version of such a norm, it is natural to choose one that supports the numerical treatment best. We shall refer to this scenario (typically associated with noninteger smoothness norms) when no physically mandatory optimal solution can be identified as the *ambiguous case*.

Unfortunately, not much is known about this latter scenario, primarily because in conventional discretization settings none of the above-mentioned equivalent norms is easily (or is not at all) realized in practice. In the ambiguous case, when in principle any *representer* can be used, wavelet concepts offer a promising alternative that renders such mechanisms feasible. In fact, norm equivalences based on simple scalings suggest themselves and, in particular, support (iii) and (iv). Of course, preferring one equivalent norm in the objective functional over another one will change the solution, although the constraints remain unchanged. However, in the case of compatible data, the same solution is obtained for all equivalent representers of the chosen norm. Moreover, it will be seen that the well-posedness of the corresponding variational problem remains invariant.

We present here for the first time to our knowledge rigorous convergence and complexity estimates for adaptive methods for optimal control problems that treat in full generality the ambiguous case and also, for certain naturally arising objective functionals, the mandatory case.

Therefore, the general control problem described initially in section 2.1 is to be viewed only as a *reference model*, where in the ambiguous case the involved norms are specified only up to equivalence. To facilitate a possibly unified treatment of both scenarios, we employ the concept of *Riesz operators* when formulating the problem in wavelet coordinates. This enables us to retain formally the same cost functional when dealing with the mandatory case. In favor of a unified treatment and to keep technicalities at a manageable level, we present a detailed description and analysis of the adaptive concept first in the ambiguous case. We sketch at the end the modifications

needed to address the mandatory case for certain naturally arising norms.

In section 4 we briefly collect some relevant facts from [CDD2, CDD3] that will later be used for (iv). One major task in the present context is the formulation of a convergent (ideal) iteration (iii) and a way that makes (iv) feasible. This is the objective of section 5. Having started out from a rather general setting we will have by then narrowed, step by step, requirements on the computational ingredients that will imply optimal complexity at the end and guide the construction of the scheme. The complexity analysis in section 6 will finally allow us to identify specific evaluation schemes that will be seen to render our adaptive solver for the optimal control problems under consideration to have optimal work/accuracy rates in the above sense. It is perhaps worth noting that the analysis brings out some distinctions between the inherent computational complexity of problems with distributed versus Neumann boundary control. In section 7 we address some special cases regarding simplifications of the scheme. Moreover, we indicate for natural objective functionals ways of treating the mandatory case. Finally, in section 8 some concluding remarks are made.

Throughout the paper, we use the following notational conventions, unless specific constants have to be identified. The relation $a \sim b$ stands for $a \lesssim b$ and $a \gtrsim b$, where the latter relation means that b can be estimated from above by a constant multiple of a independent of all parameters on which a or b may depend.

2. Problems in optimal control. We shall be concerned with the following abstract class of problems in optimal control that will serve as a first simple model for studying adaptive solution concepts in such a context. Several specifications will guide the subsequent analysis.

2.1. Abstract linear-quadratic control problems. Let Y and Q denote the *state* and the *control space*, respectively, which are assumed to be (closed subspaces of) Hilbert spaces, with topological duals Y', Q' and associated dual forms $\langle\cdot,\cdot\rangle_{Y'\times Y}$, $\langle\cdot,\cdot\rangle_{Q'\times Q}$. When there is no risk of confusion we write briefly $\langle\cdot,\cdot\rangle$.

In many applications the *states* y are measured in a weaker norm corresponding here to a Hilbert space Z hosting the observed data y_*. Specifically, suppose that $T : Y \to R$ is a continuous linear operator from Y onto its range R

$$\|Tv\|_R \lesssim \|v\|_Y, \quad v \in Y, \tag{2.1}$$

and that R is continuously embedded in Z. We shall always identify norms by using the respective space as a subscript. In contrast, the regularity imposed on the *control* u, represented here by a Hilbert space U, is often higher than that required in a natural variational formulation. Thus, in summary we shall assume the validity of the continuous embeddings

$$\|w\|_Z \lesssim \|w\|_R, \quad w \in R, \qquad \|v\|_Q \lesssim \|v\|_U, \quad v \in U. \tag{2.2}$$

Our objective is to minimize quadratic functionals of the form

$$J(y,u) = \frac{1}{2}\|Ty - y_*\|_Z^2 + \frac{\omega}{2}\|u\|_U^2, \tag{2.3}$$

subject to linear constraints, that will be described next. We shall assume that $a(\cdot,\cdot) : Y \times Y \to \mathbb{R}$ is a bilinear continuous Y-*elliptic* form, i.e.,

$$a(v,v) \sim \|v\|_Y^2, \qquad v \in Y. \tag{2.4}$$

It will sometimes be convenient to refer to the linear operator $A : Y \to Y'$ defined by $\langle Aw, v\rangle = a(w, v)$, $w, v \in Y$.

The last ingredient is a linear continuous operator $E : Q \to Y'$, describing an action on the control.

The abstract *linear-quadratic control problem* can now be formulated as follows.

(CP). *For given observations $y_* \in Z$, a right-hand side $f \in Y'$, and a weight parameter $\omega > 0$, minimize the quadratic functional* (2.3) *over $(y, u) \in Y \times Q$ subject to the linear constraints*

$$a(y, v) = \langle f + Eu, v\rangle, \quad v \in Y. \tag{2.5}$$

Remark 2.1. Of course, when the observed data are *compatible* in the sense that $y_* \equiv TA^{-1}f$, (CP) has the trivial solution $u \equiv 0$ yielding $J(y, u) \equiv 0$.

2.2. Some examples. In all the following, $\Omega \subset \mathbb{R}^d$ denotes a bounded Lipschitz domain. The choice $Z = U = L_2(\Omega)$ in the functional (2.3) is classical (see [Li]), perhaps partly due to the difficulty of evaluating the norms that could be termed *natural* (such as fractional trace norms) with regard to the underlying variational formulation, namely, the norms $\|\cdot\|_Y, \|\cdot\|_Q$. Here we explicitly allow for use also of natural norms for observing the state y, unless, for statistical reasons, measurements are meaningful only in weaker norms, such as L_2. It will be shown below that Sobolev or even Besov norms on Ω or (part of) its boundary $\partial\Omega$ for a certain range of regularity scales can be dealt with by our approach.

Although the problems with *distributed control* are perhaps of a rather academic nature, they serve as good illustrations for the essential mechanisms.

2.2.1. Dirichlet problem with distributed control. In our first example we consider such a distributed control problem with the following identification of the above ingredients:

$$a(v, w) := \int_\Omega \nabla v \cdot \nabla w \, dx, \quad Y = H_0^1(\Omega),\ Q = H^{-1}(\Omega) = Y'. \tag{2.6}$$

This gives rise to constraints whose strong form is given by the standard second order Dirichlet problem with distributed control,

$$\begin{aligned} -\Delta y &= f + u && \text{in } \Omega, \\ y &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{2.7}$$

Admissible choices for Z, U, satisfying (2.2), are then

$$Z := H_{00}^{1-\gamma}(\Omega), \quad U = H^{\beta-1}(\Omega) := (H_{00}^{1-\beta}(\Omega))', \quad 0 \le \gamma, \beta \le 1, \tag{2.8}$$

where $H_{00}^\gamma(\Omega)$ consists of those elements in $H^\gamma(\Omega)$ whose trivial extension by zero belongs to $H^\gamma(\mathbb{R}^d)$. Thus, for $\gamma > 0$ states are measured in a weaker norm, while for $\beta > 0$ additional smoothness is imposed on the control when compared with the natural norms. In particular, the classical case $U = Z = L_2(\Omega)$ is covered. In all these cases the operators T, E are the canonical injections $T = I$, $E = I$, which, for the regularity scales in (2.8), are indeed bounded.

2.2.2. Neumann problem with distributed control. Choosing

$$a(v,w) := \int_\Omega (\nabla v \cdot \nabla w + vw)dx, \quad Y := H^1(\Omega), \quad Q = (H^1(\Omega))' = Y', \tag{2.9}$$

(2.4) holds. Denoting by ι the trace operator to $\partial\Omega$, mapping functions in $Y = H^1(\Omega)$ to $H^{1/2}(\partial\Omega)$, we consider next the constraint

$$a(y,v) = \langle \tilde{f}, v\rangle + \int_{\partial\Omega} g(\iota v)\, ds + \langle u, v\rangle \qquad \text{for all } v \in Y \tag{2.10}$$

and for given $\tilde{f} \in Y'$, $g \in H^{-1/2}(\partial\Omega)$. Its strong form is the second order nonhomogeneous Neumann problem with distributed control

$$\begin{aligned} -\Delta y + y &= \tilde{f} + u && \text{in } \Omega, \\ \frac{\partial y}{\partial n} &= g && \text{on } \partial\Omega, \end{aligned} \tag{2.11}$$

where $\frac{\partial}{\partial n}$ is the normal derivative in the direction of the outward normal. The constraints (2.10) can be formulated as an operator equation

$$Ay = f + u, \tag{2.12}$$

where the data f are defined by $\langle f, v\rangle := \langle \tilde{f}, v\rangle + \int_{\partial\Omega} g(\iota v)\, ds$ and A is boundedly invertible from Y to Y'.

In analogy to (2.8) we can take here

$$Z = H^{1-\gamma}(\Omega), \quad U = (H^{1-\beta}(\Omega))', \quad 0 \le \gamma, \beta \le 1. \tag{2.13}$$

Again $T = I$ and $E = I$ are then the canonical injections.

One can also prescribe as observations boundary conditions of Dirichlet type y_* on $\partial\Omega$, in which case the range of T is $R = H^{1/2}(\partial\Omega)$, which will be referred to as the natural observation space. In this case $Z = H^{1/2-\gamma}(\partial\Omega)$ are admissible observation spaces for $\gamma \ge 0$. Here $T : H^1(\Omega) \to H^{1/2}(\partial\Omega)$ coincides with the trace operator. Hence, the optimal control problem is to steer the states toward Dirichlet boundary conditions, while the constraints (2.11) involve Neumann boundary conditions.

2.2.3. Neumann problem with Neumann boundary control. Now let the boundary $\partial\Omega$ be decomposed into two parts, $\overline{\partial\Omega} = \overline{\Gamma_N} \cup \overline{\Gamma_c}$, where Γ_c has nonvanishing $d-1$ dimensional measure. For $a(\cdot,\cdot)$ from (2.9), consider the constraint

$$a(y,v) = \langle \tilde{f}, v\rangle + \int_{\Gamma_c} g(\iota v)\, ds + \int_{\Gamma_c} u(\iota v)\, ds \qquad \text{for all } v \in Y := H^1(\Omega) \tag{2.14}$$

and given $\tilde{f} \in Y'$, $g \in (H^{1/2}(\Gamma_c))'$, whose strong form is the second order Neumann problem

$$\begin{aligned} -\Delta y + y &= \tilde{f} && \text{in } \Omega, \\ \frac{\partial y}{\partial n} &= \begin{cases} 0 & \text{on } \Gamma_N, \\ g + u & \text{on } \Gamma_c. \end{cases} \end{aligned} \tag{2.15}$$

To identify the remaining ingredients, note first that for the right-hand side of (2.14) to be well defined, the control must belong to $Q = (H^{1/2}(\Gamma_c))'$. Thus, the operator E is the adjoint of the trace operator ι to the *control boundary* Γ_c, defined as

$$\langle Eq, w\rangle_{(H^1(\Omega))'\times H^1(\Omega)} := \int_{\Gamma_c} q(\iota w)\, ds. \tag{2.16}$$

That is, $E : (H^{1/2}(\Gamma_c))' \to (H^1(\Omega))'$ is an extension operator to Ω. Thus, the formulation of the constraint as an operator equation reads in this case

$$Ay = f + Eu. \tag{2.17}$$

As in the previous cases, one could choose Z to be a space defined on Ω. A more frequent practical situation is to approximate prescribed conditions for the state on some part of the boundary.

To this end, denote by $\Gamma_\circ \subseteq \partial\Omega$ an *observation boundary* (again with strictly positive measure) and by $T : H^1(\Omega) \to H^{1/2}(\Gamma_\circ) =: R$ the trace operator to this part of the boundary. Then admissible choices for Z are $H^{1/2-\gamma}(\Gamma_\circ)$, $\gamma \geq 0$. For the control, we have $Q = (H^{1/2}(\Gamma_c))'$ so that $U = L_2(\Gamma_c)$ would require the optimal control to be somewhat smoother. For these choices, the functional (2.3) is of the form

$$J(y,u) = \frac{1}{2}\|Ty - y_*\|^2_{H^{1/2}(\Gamma_\circ)} + \frac{\omega}{2}\|u\|^2_{L_2(\Gamma_c)}. \tag{2.18}$$

Again we could take $Z = H^{1/2-\gamma}(\Gamma_\circ)$ for $0 \leq \gamma$ instead. For the choice $Z = L_2(\Gamma_\circ)$ and $U = L_2(\Gamma_c)$, the functional (2.3) with constraints (2.14) was treated in [BKR] by employing an adaptive finite element solver. The case $\Gamma_\circ = \Gamma_c = \partial\Omega$ and $Z = U = L_2(\partial\Omega)$ is classical [Li].

Remark 2.2. Note that the choice of a specific representer of the broken trace norms (that arise above as *natural norms*) and of any other smoothness norm of noninteger order (that enable scale-dependent fitting criteria) is ambiguous. Whether one uses, for instance, a factor norm or an intrinsic norm based on a parametrization of the boundary manifold in the case of trace norms is neither essential nor physically predetermined.

Remark 2.3. For linear-quadratic elliptic problems with Dirichlet boundary controls the constraints are usually formulated as saddle point problems (see, e.g., [K2]), which no longer satisfy the ellipticity condition (2.4). The techniques developed below can also be extended to this situation; see [CDD2, K3]. However, to make the basic mechanisms as transparent as possible, we confine the present discussion to the case of elliptic constraints.

3. Reformulation of (CP). The standard approach to control problems like (CP) would be to derive the necessary conditions for optimality in terms of an adjoint equation in the functional analytic setting (see, e.g., [Li]) and to discretize the resulting conditions by choosing suitable *finite dimensional* trial spaces. Here we will deviate from such a procedure in several ways. The first step is to transform the original problem (CP) into an *equivalent* (hence still infinite dimensional) problem, which is now formulated entirely in ℓ_2. The use of appropriate Riesz operators allows us to leave the original reference objective functional unchanged and to treat the mandatory case. However, in the ambiguous case associated with smoothness norms of noninteger order, these norms are only given up to equivalence (see Remark 2.2)

and consequently the Riesz maps are unspecified. Furthermore, their numerical realization would generally be practically infeasible, except when using one specific norm equivalence induced by wavelet expansions. Therefore we shall specify the objective functionals by employing norm equivalences induced by scaled wavelet expansions. This particular reformulation will be seen to offer the following advantages:

- All the previous special cases take on a *unified format*. All norms (including those with negative order or fractional trace norms) are represented by ℓ_2-norms. In particular, the more flexible options of balancing the observation part against the regularization part in the cost functional with the aid of different regularity requirements are retained and become computationally feasible.
- There is no need for inverting ill-conditioned linear systems.
- It provides the foundation for adaptive solution strategies.
- Aside from complexity issues, such adaptive strategies have stabilizing effects in cases where discretizations usually have to obey compatibility constraints, such as the LBB-condition for saddle point problems.

The transformation hinges on the availability of appropriate *wavelet bases*, which are described next.

3.1. Wavelet coordinates. In the following we shall assume that for each Hilbert space $H \in \{Y, Z, R, Q, U\}$ we have a collection of functions

$$\Psi_H = \{\psi_{H,\lambda} : \lambda \in \mathbb{I}_H\} \subset H \tag{3.1}$$

—a *wavelet basis*—with the following properties at our disposal. $\mathbb{I}_H$ is an infinite index set whose elements λ encode different features such as *scale* $|\lambda|$ and spatial location $k = k(\lambda)$. In the simplest case of wavelets on the real line one has $\psi_{H,\lambda} = 2^{j/2}\psi(2^j \cdot -k)$, $j, k \in \mathbb{Z}$, normalized in L_2. Thus λ represents (j, k) and $|\lambda| = j$. We dispense at this point with further technical details about the actual construction of such wavelet bases but collect only those properties that are relevant in the present context.

Locality (L). The functions $\psi_{H,\lambda}$ are local, and the widths of their support are decreasing with growing discretization level $|\lambda|$,

$$\operatorname{diam}(\operatorname{supp}\psi_{H,\lambda}) \sim 2^{-|\lambda|}. \tag{3.2}$$

Cancellation property (CanP). There exists an integer $\tilde{m} = \tilde{m}_H$ such that

$$\langle v, \psi_{H,\lambda}\rangle \lesssim 2^{-|\lambda|(d/2+\tilde{m})}|v|_{W^{\tilde{m}}_\infty(\operatorname{supp}\,\psi_{H,\lambda})}, \tag{3.3}$$

where d is the dimension of the underlying domain or manifold. Thus, integrating against a wavelet has the effect of taking an $\tilde{m}$th order difference which annihilates the smooth part of v. In fact, this is typically realized (for wavelets defined on Euclidean domains) by constructing Ψ_H in such a way that it possesses a *dual* or *biorthogonal* basis $\tilde{\Psi}_H \subset H'$ such that the multiresolution spaces $\tilde{S}_j := \operatorname{span}\{\tilde{\psi}_{H,\lambda} : |\lambda| < j\}$ contain all polynomials of order $\tilde{m}$. Here *dual basis* means that $\langle \psi_{H,\lambda}, \tilde{\psi}_{H,\nu}\rangle = \delta_{\lambda,\nu}$, $\lambda, \nu \in \mathbb{I}_H$. Here and in what follows the tilde is to express that the collection is a dual basis to a primal one for the space identified by the subscript. The role of dual bases will be addressed again below.

This cancellation property entails quasi-sparse representations of a wide class of operators.

Riesz basis property (R). This is perhaps the most crucial requirement. Every $v \in H$ has a unique expansion in terms of Ψ_H,

$$v = \sum_{\lambda \in \mathbb{I}_H} v_\lambda \, \psi_{H,\lambda} =: \mathbf{v}^T \, \Psi_H, \quad \mathbf{v} := (v_\lambda)_{\lambda \in \mathbb{I}_H}, \tag{3.4}$$

and its expansion coefficients satisfy the following *norm equivalence*: There exist finite positive constants c_H, C_H such that

$$c_H \|\mathbf{v}\|_{\ell_2(\mathbb{I}_H)} \le \|\mathbf{v}^T \Psi_H\|_H \le C_H \|\mathbf{v}\|_{\ell_2(\mathbb{I}_H)}, \quad \mathbf{v} \in \ell_2(\mathbb{I}_H). \tag{3.5}$$

Thus, wavelet expansions induce isomorphisms between certain function and sequence spaces.

By duality arguments one can show that (3.5) is equivalent to the existence of a biorthogonal collection

$$\tilde{\Psi}_H := \{\tilde{\psi}_{H,\lambda} : \lambda \in \mathbb{I}_H\} \subset H', \quad \langle \psi_{H,\lambda}, \tilde{\psi}_{H,\mu} \rangle = \delta_{\lambda,\mu}, \qquad \lambda, \mu \in \mathbb{I}_H, \tag{3.6}$$

which is a Riesz basis in H', i.e.,

$$C_H^{-1} \|\tilde{\mathbf{v}}\|_{\ell_2(\mathbb{I}_H)} \le \|\tilde{\mathbf{v}}^T \tilde{\Psi}_H\|_{H'} \le c_H^{-1} \|\tilde{\mathbf{v}}\|_{\ell_2(\mathbb{I}_H)} \tag{3.7}$$

holds for any $\tilde{v} = \tilde{\mathbf{v}}^T \, \tilde{\Psi}_H \in H'$; see, e.g., [D1, D3, D4, K1].

We shall need a little more information about how bases with the above properties are constructed. In all our examples the Hilbert space $H \in \{Y, Q, R, Z, U\}$ is actually (a closed subspace of) a Sobolev space $H^s = H^s(G)$ or its dual (possibly determined by homogeneous boundary conditions), where G is either the domain Ω or (part of) its boundary. The basis Ψ_H for H is then typically obtained from an *anchor* basis $\Psi = \{\psi_\lambda : \lambda \in \mathbb{I} = \mathbb{I}_H\}$, which is a Riesz basis for $L_2(G)$, i.e., $\|\psi_\lambda\|_{L_2(G)} \sim 1$, whose dual basis $\tilde{\Psi}$ is therefore also a Riesz basis for $L_2(G)$. In fact, Ψ and $\tilde{\Psi}$ are constructed in such a way that rescaled versions of *both bases* $\Psi, \tilde{\Psi}$ form Riesz bases for a whole range of Sobolev (sub-)spaces H^s, for $0 < s < \gamma, \tilde{\gamma}$, respectively. From this fact one derives then that for each $s \in (-\tilde{\gamma}, \gamma)$ the collection

$$\Psi_s := \{2^{-s|\lambda|} \psi_\lambda : \lambda \in \mathbb{I}\} =: \mathbf{D}^{-s} \Psi \tag{3.8}$$

is a Riesz basis for H^s (with the above interpretation of H^s as a dual when s is negative) [D1]; i.e., there exist positive constants c_s, C_s such that

$$c_s \, \|\mathbf{v}\|_{\ell_2(\mathbb{I})} \le \|\mathbf{v}^T \Psi_s\|_{H^s} \le C_s \, \|\mathbf{v}\|_{\ell_2(\mathbb{I})}, \quad \mathbf{v} \in \ell_2(\mathbb{I}), \tag{3.9}$$

holds for each $s \in (-\tilde{\gamma}, \gamma)$. Analogous relations hold for $\tilde{\Psi}$ with reversed roles of γ and $\tilde{\gamma}$. We shall make use of the following consequence of this fact. For $t \in (-\tilde{\gamma}, \gamma)$ the mapping

$$D^t : v = \mathbf{v}^T \Psi \mapsto (\mathbf{D}^t \mathbf{v})^T \Psi = \mathbf{v}^T \mathbf{D}^t \Psi = \sum_{\lambda \in \mathbb{I}} v_\lambda \, 2^{t|\lambda|} \psi_\lambda \tag{3.10}$$

acts as a shift operator between Sobolev scales, i.e.,

$$\|D^t v\|_{H^s} \sim \|v\|_{H^{s+t}} \sim \|\mathbf{D}^{s+t} \mathbf{v}\|_{\ell_2(\mathbb{I})}, \quad \text{provided that} \quad s, \; s+t \in (-\tilde{\gamma}, \gamma). \tag{3.11}$$

Concrete constructions of wavelet bases with the above properties for parameters $\gamma, \tilde{\gamma}$ ranging in most cases up to $3/2$ on bounded Euclidean domains and also on closed

piecewise parametrically defined manifolds can be found in [CTU, CM, DKU, DS1, DS2, DSt]. Note that in the above examples the relevant Sobolev regularity indices range then between -1 and 1 so that these bases allow us to exploit relations like (3.11) when the metrics in the spaces Z and U differ from the natural norms in the way indicated above. Thus we shall henceforth assume the validity of the above properties (L), (CanP), and (R) in appropriate ranges, as detailed in the next section.

It should be noted, however, that such norm equivalences are by no means restricted to simple Sobolev spaces, as indicated by the following example.

Remark 3.1. Let H be the Hilbert space endowed with the norm $\|v\|_H^2 := \langle \nabla v, a\nabla v\rangle + \langle wv, v\rangle =: a(v,v)$, where a is a uniformly positive definite but possibly spatially varying diffusion matrix and w is a nonnegative spatially varying weight function. Starting with a basis Ψ for $L_2(\Omega)$, say, the matrix $\mathbf{D}^{-s}$ in (3.8) (for $s=1$) should then be replaced by $\mathbf{D}^{-1}$, where $\mathbf{D}$ now has the diagonal entries $a(\psi_\lambda,\psi_\lambda)^{1/2}$. This automatically incorporates spatially varying weights or diffusion terms in the normalization of the wavelet basis.

In what follows, it will be convenient to make systematic use of the following shorthand notation that already has been used to some extent above. We will view Ψ both as in (3.1) as a *collection* of functions as well as a (possibly infinite) (column) *vector* containing all functions always assembled in some fixed order. For a countable collection of functions Θ and some single function σ, the quantities $\langle\Theta,\sigma\rangle$ and $\langle\sigma,\Theta\rangle$ are to be understood as the column, respectively, row, vector with entries $\langle\theta,\sigma\rangle$, respectively, $\langle\sigma,\theta\rangle$, $\theta\in\Theta$. For two collections Θ,Σ, the term $\langle\Theta,\Sigma\rangle$ is then a (possibly infinite) matrix with entries $(\langle\theta,\sigma\rangle)_{\theta\in\Theta,\ \sigma\in\Sigma}$ for which $\langle\Theta,\Sigma\rangle = \langle\Sigma,\Theta\rangle^T$. This also implies for a (possibly infinite) matrix $\mathbf{C}$ that $\langle\mathbf{C}\Theta,\Sigma\rangle = \mathbf{C}\langle\Theta,\Sigma\rangle$ and $\langle\Theta,\mathbf{C}\Sigma\rangle = \langle\Theta,\Sigma\rangle\mathbf{C}^T$. In this notation, the expansion coefficients in (3.4) and (3.7) can explicitly be expressed as $\mathbf{v}^T = \langle v,\tilde\Psi\rangle$ and $\tilde{\mathbf{v}} = \langle\Psi,\tilde v\rangle$. Furthermore, the *biorthogonality* or *duality conditions* (3.6) can be reexpressed as $\langle\Psi,\tilde\Psi\rangle = \mathbf{I}$ with the infinite identity matrix.

3.2. Wavelet representation of operators. The last important ingredient concerns *wavelet representations* of operators. Suppose that Ψ_H,Ψ_M are Riesz bases for Hilbert spaces H,M. As before, we shall always denote by $\tilde\Psi_H,\tilde\Psi_M$ the respective dual bases, i.e., $\langle\Psi_V,\tilde\Psi_V\rangle = \mathbf{I}$, $V\in\{H,M\}$, where $\langle\cdot,\cdot\rangle$ is the corresponding duality. Suppose $L: H\to M$ is a linear operator. Any image $Lv\in M$ is naturally expanded with respect to Ψ_M as $Lv = \langle Lv,\tilde\Psi_M\rangle\Psi_M$. Inserting the expansion $v=\mathbf{v}^T\Psi_H$ with respect to Ψ_H yields

$$Lv = \mathbf{v}^T\langle L\Psi_H,\tilde\Psi_M\rangle\Psi_M = (\langle\tilde\Psi_M, L\Psi_H\rangle\mathbf{v})^T\Psi_M. \tag{3.12}$$

Since we shall make frequent use of this operation, we record the relevant facts in the following remark.

Remark 3.2. The wavelet representation of $L: H\to M$ (with respect to the bases $\Psi_H,\tilde\Psi_M$ of H, M', respectively) is given by

$$\mathbf{L} := \langle\tilde\Psi_M, L\Psi_H\rangle, \qquad Lv = (\mathbf{L}\mathbf{v})^T\Psi_M. \tag{3.13}$$

Thus, the expansion coefficients of Lv (in the basis that spans the range space of L) are obtained by applying the *infinite* matrix $\mathbf{L} = \langle\tilde\Psi_M, L\Psi_H\rangle$ to the coefficient vector of v. Moreover, boundedness of L implies boundedness of $\mathbf{L}$ in ℓ_2, i.e.,

$$\|Lv\|_M \lesssim \|v\|_H, \quad v\in H, \quad \text{implies} \quad \|\mathbf{L}\| := \sup_{\|\mathbf{v}\|_{\ell_2(\mathbb{I}_H)}\le 1}\|\mathbf{L}\mathbf{v}\|_{\ell_2(\mathbb{I}_M)} \lesssim 1. \tag{3.14}$$

Proof. The first part of the assertion was established by the preceding observations. For (3.14), one infers from (3.5) and (3.13) that

$$\|\mathbf{Lv}\|_{\ell_2(\mathbb{I}_M)} \sim \|(\mathbf{Lv})^T\Psi_M\|_M = \|Lv\|_M \lesssim \|v\|_H \sim \|\mathbf{v}\|_{\ell_2(\mathbb{I}_H)},$$

which confirms the claim. □

We shall employ *Riesz operators* $R_H : H \to H'$ defined by

$$(v,w)_H = \langle v, R_H w\rangle, \quad v,w \in H, \tag{3.15}$$

where $(\cdot,\cdot)_H$ denotes the inner product in H. Thus, R_H maps H boundedly invertibly onto its dual H'. By Remark 3.2, we therefore have (with the role of M being played now by H')

$$\mathbf{R}_H := \langle \Psi_H, R_H\Psi_H\rangle = (\Psi_H,\Psi_H)_H \tag{3.16}$$

and

$$\|\mathbf{R}_H\|,\ \|\mathbf{R}_H^{-1}\| \lesssim 1. \tag{3.17}$$

For general surveys on the application of wavelets to operator equations, we refer to [Co, D2, D3].

3.3. Equivalent control problems in ℓ_2. Now we are in the position to transform the abstract control problem (CP) into wavelet coordinates. We begin with the constraints (2.5). Following the above recipe (3.12), we expand y in Ψ_Y and u in a wavelet basis Ψ_U for $U \subset Q$; see (2.2). As explained in section 3.1, the basis Ψ_U is conveniently obtained as a scaled version of a basis for Q, i.e.,

$$\mathbf{D}_Q\Psi_U = \Psi_Q, \tag{3.18}$$

where the diagonal elements of $\mathbf{D}_Q$ are nondecreasing. Hence, expanding $u = \mathbf{u}^T\Psi_U$ in terms of Ψ_Q, gives

$$u = \mathbf{u}^T\Psi_U = (\mathbf{D}_Q^{-1}\mathbf{u})^T\Psi_Q. \tag{3.19}$$

The operators A and E, in turn, are bounded operators from Y, respectively, Q, to Y'. By Remark 3.2, their wavelet representations and that of $f \in Y'$ are therefore given by

$$\mathbf{A} := a(\Psi_Y,\Psi_Y), \quad \mathbf{E} := \langle \Psi_Y, E\Psi_Q\rangle, \quad \boldsymbol{f} := \langle \Psi_Y, f\rangle. \tag{3.20}$$

Since $\mathbf{E}$ acts on coordinates relative to Ψ_Q we can express Eu in (2.5) as $Eu = (\mathbf{E}\mathbf{D}_Q^{-1}\mathbf{u})^T\tilde{\Psi}_Y$. Thus, the constraints (2.5) take on the form

$$\mathbf{A}\mathbf{y} = \boldsymbol{f} + \mathbf{E}\mathbf{D}_Q^{-1}\mathbf{u}. \tag{3.21}$$

To simplify the notation we shall suppress in the following the subscripts $\ell_2(\mathbb{I})$ and write briefly $\|\cdot\| := \|\cdot\|_{\ell_2(\mathbb{I})}$ because we shall be dealing with only Euclidean norms, and the ranges of indices $\mathbb{I}$ will always be clear from the context.

It is well known that the ellipticity (2.4) and the Riesz basis property (R) (3.5) (for $H = Y$) imply the following fact; see, e.g., [D2].

Remark 3.3. The matrix $\mathbf{A}$ is a boundedly invertible mapping of $\ell_2(\mathbb{I}_Y)$ onto itself, i.e., $\mathbf{A}$ is onto $\ell_2(\mathbb{I}_Y)$ and there exist finite positive constants $c_{\mathbf{A}}, C_{\mathbf{A}}$ such that

$$c_{\mathbf{A}}\|\mathbf{v}\| \le \|\mathbf{A}\mathbf{v}\| \le C_{\mathbf{A}}\|\mathbf{v}\|, \qquad \mathbf{v} \in \ell_2(\mathbb{I}_Y). \tag{3.22}$$

Similarly, since the operator T is bounded as a mapping from Y to R, it is natural to represent it with respect to a pair of bases $\Psi_Y, \tilde{\Psi}_R$, i.e.,

$$\mathbf{T} = \langle \tilde{\Psi}_R, T\Psi_Y \rangle. \tag{3.23}$$

We infer from Remark 3.2 that there exist finite positive constants $C_{\mathbf{T}}, C_{\mathbf{E}}$ such that

$$\|\mathbf{T}\mathbf{v}\| \le C_{\mathbf{T}}\|\mathbf{v}\|, \qquad \|\mathbf{E}\mathbf{v}\| \le C_{\mathbf{E}}\|\mathbf{v}\| \tag{3.24}$$

for $\mathbf{v} \in \ell_2(\mathbb{I}_Y)$ and $\mathbf{v} \in \ell_2(\mathbb{I}_Q)$, respectively.

In general, R is a continuously embedded subspace of Z, and the dual pair of bases $\Psi_R, \tilde{\Psi}_R$ are scaled versions of a corresponding pair $\Psi_Z, \tilde{\Psi}_Z$; i.e., there is a diagonal matrix $\mathbf{D}_Z$ with nondecreasing diagonal entries such that

$$\Psi_Z = \mathbf{D}_Z\Psi_R, \quad \tilde{\Psi}_Z = \mathbf{D}_Z^{-1}\tilde{\Psi}_R. \tag{3.25}$$

We shall make frequent use of the fact that

$$\|\mathbf{D}_Q^{-1}\|, \ \|\mathbf{D}_Z^{-1}\| \le 1, \tag{3.26}$$

where we assume without loss of generality that all entries on the respective diagonals are bounded by one.

The observed data have to be expanded in Ψ_Z. By (2.2), T is also continuous as an operator from Y to Z. Thus the representations of T with respect to $\Psi_Y, \tilde{\Psi}_Z$ and of y_* are given by

$$\mathbf{D}_Z^{-1}\mathbf{T} = \langle \tilde{\Psi}_Z, T\Psi_Y \rangle, \qquad \mathbf{y}_* := \mathbf{D}_Z\langle \tilde{\Psi}_Z, y_* \rangle. \tag{3.27}$$

We are now in a position to express the functional $J(y,u)$ in (2.3) in terms of wavelet coordinates. On account of (3.15) and (3.16), we have the exact representation

$$J(y,u) = \frac{1}{2}\|\mathbf{R}_Z^{1/2}\mathbf{D}_Z^{-1}(\mathbf{T}\mathbf{y} - \mathbf{y}_*)\|^2 + \frac{\omega}{2}\|\mathbf{R}_U^{1/2}\mathbf{u}\|^2. \tag{3.28}$$

A few comments on possible interrelations on the various involved bases are in order. The spaces Z and U may be defined over domains different from Ω, such as traces as explained at the end of section 2.2.2 and in section 2.2.3. In these cases T is a trace operator and E an extension. Here we cannot simply interrelate Ψ_Z and Ψ_Y since traces of wavelets generally are not wavelets. Thus, the bases for Z and U have to be provided independently.

This is different, however, in the following special case.

Special case (SC). Suppose that the space Z belongs to the Sobolev scale over Ω; see (2.8), (2.13). In this case T is the identity and $R = Y$. Hence, (3.25) becomes

$$\Psi_Z = \mathbf{D}_Z\Psi_Y, \qquad \tilde{\Psi}_Z = \mathbf{D}_Z^{-1}\tilde{\Psi}_Y, \tag{3.29}$$

which clearly forms a dual pair for Z. As a mapping into Z, T is the canonical injection so that, according to (3.27),

$$\mathbf{D}_Z^{-1}\mathbf{T} = \mathbf{D}_Z^{-1}, \quad \text{i.e.,} \quad \mathbf{T} = \mathbf{I}. \tag{3.30}$$

If Z does not coincide with Y, Z induces a weaker topology than Y so that the entries of the diagonal matrix $\mathbf{D}_Z$ increase in scale. For instance, for $Y = H^t$, $Z = H^{t-\gamma}$, $0 \le \gamma$, one has $(\mathbf{D}_Z)_{\lambda,\lambda} \sim 2^{\gamma|\lambda|}$. Recall that in this case the mapping E is also just the identity ((2.7), (2.12)), i.e., $Q = Y'$ and Ψ_Q should span the range of A. So we may as well take

$$\Psi_Q := \tilde{\Psi}_Y, \quad \mathbf{E} = \mathbf{I}, \tag{3.31}$$

in this case.

In summary, we have shown that the solution to the abstract control problem (CP) is equivalent to finding the coefficient arrays $\mathbf{y}, \mathbf{u}$ of $y \in Y, u \in U$, that solve the following *abstract control problem in wavelet coordinates.*

(CPW). *Given data* $\mathbf{y}_* \in \ell_2(\mathbb{I}_Z)$, $\boldsymbol{f} \in \ell_2(\mathbb{I}_Y)$ *and a weight parameter* $\omega > 0$, *minimize the quadratic functional*

$$\widetilde{\mathbf{J}}(\mathbf{y}, \mathbf{u}) := \frac{1}{2}\|\mathbf{R}_Z^{1/2}\mathbf{D}_Z^{-1}(\mathbf{T}\mathbf{y} - \mathbf{y}_*)\|^2 + \frac{\omega}{2}\|\mathbf{R}_U^{1/2}\mathbf{u}\|^2 \quad (= J(y,u)) \tag{3.32}$$

over $(\mathbf{y}, \mathbf{u}) \in \ell_2(\mathbb{I}_Y) \times \ell_2(\mathbb{I}_Q)$ *subject to the linear constraints* (3.21)

$$\mathbf{A}\mathbf{y} = \boldsymbol{f} + \mathbf{E}\mathbf{D}_Q^{-1}\mathbf{u}.$$

3.4. Further auxiliary formulations. Next we shall derive equivalent formulations of (CPW), on which the subsequent numerical treatment will be based. By standard arguments (see, e.g., [Li]), the unique minimum for (CPW) is obtained by solving the first order necessary conditions for $\widetilde{\mathbf{J}}$ which can, e.g., be derived by first eliminating $\mathbf{y}$ in (3.32). In view of (3.22), we can invert (3.21) to obtain $\mathbf{y} = \mathbf{A}^{-1}\boldsymbol{f} + \mathbf{A}^{-1}\mathbf{E}\mathbf{D}_Q^{-1}\mathbf{u}$. Substituting this into (3.32) yields a functional that depends only on $\mathbf{u}$,

$$\widetilde{\mathbf{J}}(\mathbf{u}) \;=\; \frac{1}{2}\|\mathbf{R}_Z^{1/2}\mathbf{D}_Z^{-1}\left(\mathbf{T}\mathbf{A}^{-1}\mathbf{E}\mathbf{D}_Q^{-1}\mathbf{u} - (\mathbf{y}_* - \mathbf{T}\mathbf{A}^{-1}\boldsymbol{f})\right)\|^2 + \frac{\omega}{2}\|\mathbf{R}_U^{1/2}\mathbf{u}\|^2. \tag{3.33}$$

Abbreviating

$$\mathbf{Z} := \mathbf{D}_Z^{-1}\mathbf{T}\mathbf{A}^{-1}\mathbf{E}\mathbf{D}_Q^{-1}, \quad \mathbf{G} := \mathbf{D}_Z^{-1}(\mathbf{y}_* - \mathbf{T}\mathbf{A}^{-1}\boldsymbol{f}), \tag{3.34}$$

$\widetilde{\mathbf{J}}$ is of the form

$$\widetilde{\mathbf{J}}(\mathbf{u}) \;=\; \frac{1}{2}\|\mathbf{R}_Z^{1/2}(\mathbf{Z}\mathbf{u} - \mathbf{G})\|^2 + \frac{\omega}{2}\|\mathbf{R}_U^{1/2}\mathbf{u}\|^2. \tag{3.35}$$

This is a standard least squares functional whose minimizer is characterized by the normal equations that were in the present format derived in [K2].

PROPOSITION 3.4. *The functional* $\widetilde{\mathbf{J}}$ *is twice differentiable on* $\ell_2(\mathbb{I}_Q)$ *with first and second variation given by*

$$\delta\widetilde{\mathbf{J}}(\mathbf{u}) = (\mathbf{Z}^T\mathbf{R}_Z\mathbf{Z} + \omega\mathbf{R}_U)\mathbf{u} - \mathbf{Z}^T\mathbf{R}_Z\mathbf{G}, \qquad \delta^2\widetilde{\mathbf{J}}(\mathbf{u}) = \mathbf{Z}^T\mathbf{R}_Z\mathbf{Z} + \omega\mathbf{R}_U. \tag{3.36}$$

Thus, $\widetilde{\mathbf{J}}$ *is strictly convex so that a unique minimizer exists.*

Proof. By (3.24), (3.22), (3.17), and the fact that the scaling matrices $\mathbf{D}_Z, \mathbf{D}_Q$ have nondecreasing diagonal entries (3.26), we conclude that

$$\widetilde{\mathbf{Q}} := \mathbf{Z}^T\mathbf{R}_Z\mathbf{Z} + \omega\mathbf{R}_U \tag{3.37}$$

is bounded on $\ell_2(\mathbb{I}_Q)$. Moreover, $\widetilde{\mathbf{Q}}$ is positive definite since $\mathbf{Z}^T\mathbf{R}_Z\mathbf{Z}$ is at least positive semidefinite and, by (3.17), we have $\|\mathbf{R}_U\mathbf{v}\| \sim \|\mathbf{v}\|$. Hence, for any fixed $\omega > 0$ there exist finite positive constants $c_{\widetilde{\mathbf{Q}}}, C_{\widetilde{\mathbf{Q}}}$ such that

$$c_{\widetilde{\mathbf{Q}}}\|\mathbf{v}\| \le \|\widetilde{\mathbf{Q}}\mathbf{v}\| \le C_{\widetilde{\mathbf{Q}}}\|\mathbf{v}\|, \qquad \mathbf{v} \in \ell_2(\mathbb{I}_Q). \tag{3.38}$$

This confirms the claim. ∎

In summary, the solution of (CPW) is uniquely determined by solving $\delta\widetilde{\mathbf{J}}(\mathbf{u}) = 0$, i.e., the system

$$\widetilde{\mathbf{Q}}\mathbf{u} = \widetilde{\mathbf{g}} \qquad \text{where} \qquad \widetilde{\mathbf{g}} := \mathbf{Z}^T\mathbf{R}_Z\mathbf{G}. \tag{3.39}$$

The above argument reveals also the following facts concerning the necessity of regularization.

Remark 3.5. The operator $\widetilde{\mathbf{Q}}$ is well conditioned in the sense of (3.38), uniformly in $\omega \ge 0$, if and only if $\mathbf{Z}$ is a topological automorphism on ℓ_2. Obviously this prohibits the appearance of nontrivial scalings and requires $\mathbf{E}, \mathbf{T}$ to be boundedly invertible and hence excludes trace operators.

Specifically, in the situation (SC), i.e., $\mathbf{E} = \mathbf{T} = \mathbf{I}$, $\widetilde{\mathbf{Q}}$ takes on the form

$$\widetilde{\mathbf{Q}} = \mathbf{D}_Q^{-1}\mathbf{A}^{-T}\mathbf{D}_Z^{-1}\mathbf{R}_Z\mathbf{D}_Z^{-1}\mathbf{A}^{-1}\mathbf{D}_Q^{-1} + \omega\mathbf{R}_U. \tag{3.40}$$

Note that when employing natural norms, i.e., $\mathbf{D}_Q = \mathbf{D}_Z = \mathbf{I}$,

$$\widetilde{\mathbf{Q}} = \mathbf{A}^{-T}\mathbf{R}_Z\mathbf{A}^{-1} + \omega\mathbf{R}_U$$

obviously satisfies (3.38) also for $\omega = 0$, which means that in this case regularization is not necessary. Moreover, (3.39) is equivalent to

$$(\mathbf{I} + \omega\mathbf{D}_Q\mathbf{A}\mathbf{D}_Z\mathbf{R}_Z^{-1}\mathbf{D}_Z\mathbf{A}^T\mathbf{D}_Q\mathbf{R}_U)\mathbf{u} = \mathbf{D}_Q(\mathbf{A}\mathbf{y}_* - \boldsymbol{f}). \tag{3.41}$$

Proof. The first part, especially (3.40), follows from substituting the definition of $\mathbf{Z}$. (3.41) is obtained by multiplying (3.39) with $\mathbf{Z}^{-1}\mathbf{R}_Z^{-1}\mathbf{Z}^{-T}$ from the left. ∎

The system (3.39) would offer a natural link to the setting in [CDD1] because it is symmetric positive definite and, thus, invites the application of gradient iterations which could then be carried out approximately by adaptive applications of $\widetilde{\mathbf{Q}}$. However, even when approximate evaluation schemes for infinite matrices like $\mathbf{A}$ were available, one immediately encounters some severe obstructions in applying $\widetilde{\mathbf{Q}}$ and evaluating the right-hand side, due to the appearance of (a) the Riesz operators and (b) the inverses involved in the definition of $\mathbf{Z}$ and $\mathbf{G}$.

To this end, the formulation (3.41) looks tempting since it no longer involves the inverse of $\mathbf{A}$ but requires only inverting a Riesz map. However, it is not very helpful unless in (SC) one employs only the natural norms $Z = Y, U = Q$ since otherwise the matrices $\mathbf{D}_Z, \mathbf{D}_Q$ are not bounded in ℓ_2. In this latter very special case $\mathbf{D}_Z = \mathbf{I}$, $\mathbf{D}_Q = \mathbf{I}$ so that (3.41) becomes

$$(\mathbf{I} + \omega\mathbf{A}\mathbf{R}_Z^{-1}\mathbf{A}^T\mathbf{R}_U)\mathbf{u} = \mathbf{A}\mathbf{y}_* - \boldsymbol{f}. \tag{3.42}$$

In the example discussed in section 2.2.1 we have $Y = Z = H_0^1(\Omega)$, $U = Q = H^{-1}(\Omega)$. When $\|\cdot\|_Z$ is the first order Sobolev seminorm, $\mathbf{R}_Z$ is the wavelet representation $\boldsymbol{\Delta} = \langle\nabla\Psi_Y, \nabla\Psi_Y\rangle$ of the Laplacian, while $\mathbf{R}_U = \boldsymbol{\Delta}^{-1}$. Thus (3.42) provides

$$(\mathbf{I} + \omega\mathbf{A}\boldsymbol{\Delta}^{-1}\mathbf{A}^T\boldsymbol{\Delta}^{-1})\mathbf{u} = \mathbf{A}\mathbf{y}_* - \boldsymbol{f}. \tag{3.43}$$

We shall discuss the advantage of this formulation later in more detail but must concede that, for the reasons mentioned above, it is not appropriate for more general situations because, in contrast to (3.39), (3.41) is generally ill conditioned on the infinite dimensional level.

To facilitate nevertheless the application of $\widetilde{\mathbf{Q}}$ for the above general scope of problems we shall employ further auxiliary reformulations of (CPW) to be derived next.

3.5. The Euler equations. We define the Lagrangian and introduce as an additional variable the Lagrange parameter $\mathbf{p} \in \ell_2(\mathbb{I}_Y)$,

$$\mathrm{Lagr}(\mathbf{y},\mathbf{p},\mathbf{u}) := \widetilde{\mathbf{J}}(\mathbf{y},\mathbf{u}) + \langle \mathbf{p}, \mathbf{A}\mathbf{y} - \boldsymbol{f} - \mathbf{E}\mathbf{D}_Q^{-1}\mathbf{u}\rangle, \tag{3.44}$$

where $\widetilde{\mathbf{J}}(\mathbf{y},\mathbf{u})$ has been defined in (3.32). Straightforward calculations yield the first order Euler–Lagrange equations whose solution also yields the minimizer of (3.32); see, e.g., [Z] or [K2].

Remark 3.6. The solution $\mathbf{u}$ of the system (3.39) is a component of the solution $(\mathbf{y},\mathbf{p},\mathbf{u})$ of the weakly coupled system of Euler equations in wavelet coordinates

$$\begin{aligned} \mathbf{A}\mathbf{y} &= \boldsymbol{f} + \mathbf{E}\mathbf{D}_Q^{-1}\mathbf{u}, \\ \mathbf{A}^T\mathbf{p} &= -\mathbf{T}^T\mathbf{D}_Z^{-1}\mathbf{R}_Z\mathbf{D}_Z^{-1}\left(\mathbf{T}\mathbf{y} - \mathbf{y}_*\right), \qquad \text{(EE)} \\ \omega\mathbf{R}_U\mathbf{u} &= \mathbf{D}_Q^{-1}\mathbf{E}^T\mathbf{p}. \end{aligned} \tag{3.45}$$

(3.46)

The first equation of (EE) is often denoted as the *state* or *primal equation*, while the second equation is called the *costate* or *adjoint equation*.

The system (EE) can, of course, be reformulated as a saddle point problem

$$\begin{pmatrix} \omega\mathbf{R}_U & \mathbf{0} & -\mathbf{D}_Q^{-1}\mathbf{E}^T \\ \mathbf{0} & \mathbf{T}^T\mathbf{D}_Z^{-1}\mathbf{R}_Z\mathbf{D}_Z^{-1}\mathbf{T} & \mathbf{A}^T \\ -\mathbf{E}\mathbf{D}_Q^{-1} & \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{T}^T\mathbf{D}_Z^{-1}\mathbf{R}_Z\mathbf{D}_Z^{-1}\mathbf{y}_* \\ \boldsymbol{f} \end{pmatrix}. \tag{3.47}$$

In particular, in the case (SC) when using natural norms in (2.3), i.e., $\mathbf{D}_Q = \mathbf{D}_Z = \mathbf{I}$, we have

$$\begin{pmatrix} \omega\mathbf{R}_U & \mathbf{0} & -\mathbf{I} \\ \mathbf{0} & \mathbf{R}_Z & \mathbf{A}^T \\ -\mathbf{I} & \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{R}_Z\mathbf{y}_* \\ \boldsymbol{f} \end{pmatrix}. \tag{3.48}$$

Due to the appearance of the isomorphisms $\mathbf{R}_U, \mathbf{R}_Z$ on the diagonal of the upper left two-by-two block, this system satisfies the *inf-sup* condition and, therefore, defines a boundedly invertible mapping on $\ell_2(\mathbb{I}_Q) \times \ell_2(\mathbb{I}_Y)^2$. Thus, one can immediately apply the results from [DDU] on adaptive Uzawa iterations for well-posed saddle point problems. Corresponding optimal complexity estimates apply whenever the matrix $\mathbf{A}$ is compressible, which is the case in all the above examples.

In general, however, when $\mathbf{T}$ is a trace operator, the block $\mathbf{T}^T\mathbf{D}_Z^{-1}\mathbf{R}_Z\mathbf{D}_Z^{-1}\mathbf{T}$ has a nontrivial kernel. To apply the Uzawa strategy, one first must use an augmented Lagrangian formulation so as to have a well-defined Schur complement. This can be done along the lines described in [DDU].

Here we prefer the formally somewhat different approach based on the system (EE). This approach also applies, in principle, to constraints in the form of a saddle point system, as pointed out in Remark 2.3; see [K3].

Our strategy will be to solve (3.39) with the aid of a (perturbed) descent scheme which requires approximating at each stage the *residual* $\widetilde{\mathbf{g}} - \widetilde{\mathbf{Q}}\mathbf{u}^k$. Such an approximation will be based on the following observation, namely, that the residual of (3.39) is just the residual of the third equation in (EE), sometimes referred to as the *design equation*.

LEMMA 3.7. *For any* $\mathbf{v} \in \ell_2(\mathbb{I}_Q)$, *one has the representation*

$$\widetilde{\mathbf{Q}}\mathbf{v} - \widetilde{\mathbf{g}} = \omega \mathbf{R}_U \mathbf{v} - \mathbf{D}_Q^{-1}\mathbf{E}^T\mathbf{p}, \tag{3.49}$$

where $\mathbf{p}$ *is the solution of the first two equations in* (EE). *Thus, for any given* $\mathbf{v}$, *the sequence* $\mathbf{p}$ *is determined by solving*

$$\mathbf{A}^T\mathbf{p} = -\mathbf{T}^T\mathbf{D}_Z^{-1}\mathbf{R}_Z\mathbf{D}_Z^{-1}\left(\mathbf{T}\mathbf{y} - \mathbf{y}_*\right), \quad \textit{where} \quad \mathbf{A}\mathbf{y} = \boldsymbol{f} + \mathbf{E}\mathbf{D}_Q^{-1}\mathbf{v}. \tag{3.50}$$

Proof. By definition we infer from (3.34), (3.37), and (3.39) that

$$\widetilde{\mathbf{Q}}\mathbf{v} - \widetilde{\mathbf{g}} = \omega \mathbf{R}_U \mathbf{v} + \mathbf{Z}^T\mathbf{R}_Z(\mathbf{Z}\mathbf{v} - \mathbf{D}_Z^{-1}(\mathbf{y}_* - \mathbf{T}\mathbf{A}^{-1}\boldsymbol{f})).$$

The second term on the right-hand side of this equality can be written as

$$\begin{aligned}\mathbf{Z}^T\mathbf{R}_Z\mathbf{D}_Z^{-1}\mathbf{T}\mathbf{A}^{-1}\mathbf{E}\mathbf{D}_Q^{-1}\mathbf{v} - \mathbf{Z}^T\mathbf{R}_Z\mathbf{D}_Z^{-1}(\mathbf{y}_* - \mathbf{T}\mathbf{A}^{-1}\boldsymbol{f}) \\ = \mathbf{Z}^T\mathbf{R}_Z\mathbf{D}_Z^{-1}\mathbf{T}\mathbf{A}^{-1}(\mathbf{E}\mathbf{D}_Q^{-1}\mathbf{v} + \boldsymbol{f}) - \mathbf{Z}^T\mathbf{R}_Z\mathbf{D}_Z^{-1}\mathbf{y}_*.\end{aligned}$$

Thus, taking $\mathbf{y}$ as the solution of the second equation in (3.50), this reduces to

$$-\mathbf{Z}^T\mathbf{R}_Z\mathbf{D}_Z^{-1}(\mathbf{y}_* - \mathbf{T}\mathbf{y}) = -\mathbf{D}_Q^{-1}\mathbf{E}^T\mathbf{A}^{-T}\mathbf{T}^T\mathbf{D}_Z^{-1}\mathbf{R}_Z\mathbf{D}_Z^{-1}(\mathbf{y}_* - \mathbf{T}\mathbf{y}) = -\mathbf{D}_Q^{-1}\mathbf{E}^T\mathbf{p},$$

where we have used the first equation in (3.50). This finishes the proof. ∎

Note that one immediately infers from (3.24) and (3.26) that one still has

$$\|\mathbf{D}_Z^{-1}\mathbf{T}\mathbf{v}\| \le C_{\mathbf{T}}\|\mathbf{v}\|, \quad \|\mathbf{E}\mathbf{D}_Q^{-1}\mathbf{v}\| \le C_{\mathbf{E}}\|\mathbf{v}\|. \tag{3.51}$$

3.6. Special cost functionals: A scaling model. We have succeeded so far in realizing steps (i) and (ii) of the paradigm described in the introduction, namely, to reformulate (CP) as an equivalent well-posed problem over ℓ_2. The realization of step (iii) will be based on the residual representation (3.49). Lemma 3.7 suggests approximating the residual with the aid of approximate solutions of the equations (3.50). This, in turn, will be done iteratively as well. Aside from approximately applying the infinite matrices $\mathbf{A}, \mathbf{A}^T$, this also requires the application of the Riesz matrices $\mathbf{R}_Z, \mathbf{R}_U$.

Remark 3.8. When $Z, U \in \{H^1, L_2\}$, the matrices $\mathbf{R}_H$ are for $H = Z, U$ of the form $\langle \Psi_H, \Psi_H\rangle$, $\langle \nabla\Psi_H, \nabla\Psi_H\rangle$ or $c_1\langle \Psi_H, \Psi_H\rangle + c_2\langle \nabla\Psi_H, \nabla\Psi_H\rangle$; see Remark 3.1. In this case, $\mathbf{R}_H$ can be shown to be compressible in the sense of [CDD1]; see also section 5.1 concerning the application of $\mathbf{A}, \mathbf{A}^T$. Thus, the efficient approximate application of such matrices can be carried out with the aid of the schemes developed in [BCDU, CDD1]. Similarly, when $U = H^{-1}$ we have $\mathbf{R}_U = \boldsymbol{\Delta}^{-1}$, in which case the design equation (3.46) becomes

$$\omega\mathbf{u} = \boldsymbol{\Delta}\mathbf{D}_Q^{-1}\mathbf{p}, \tag{3.52}$$

which corresponds to the application of the Laplacian; see section 2.2.1. We shall discuss the consequences of these facts in more detail later.

As explained in Remark 2.2, in the ambiguous case associated with Sobolev spaces Z, U of noninteger order the corresponding Riesz operators $\mathbf{R}_Z, \mathbf{R}_U$ are not specified.

In conventional discretization settings, the evaluation of neither representer of such a norm is practically feasible. In the present framework, however, we can resort to one specific representer which supports the numerical treatment best. As explained earlier, the rational, of course, is that changing the regularity scales in the different parts of the cost functional has a much more subtle effect than changing the (single scale) weight ω and that this effect is realized by any equivalent norm; see [BK] for numerical explorations of this issue. Formally, this amounts to expressing Sobolev norms as sequence norms, based on (3.9). In the present framework this means to replace the norm automorphisms $\mathbf{R}_Z, \mathbf{R}_U$ (see (3.17)) simply by the identity $\mathbf{I}$ in the definition of $\widetilde{\mathbf{J}}$ in (3.32). This leads us to consider the following *specified abstract control problem in wavelet coordinates.*

(SCPW). *Given data* $\mathbf{y}_* \in \ell_2(\mathbb{I}_Z)$, $\boldsymbol{f} \in \ell_2(\mathbb{I}_Y)$, *and a weight parameter* $\omega > 0$, *minimize the quadratic functional*

$$\mathbf{J}(\mathbf{y}, \mathbf{u}) := \frac{1}{2}\|\mathbf{D}_Z^{-1}(\mathbf{T}\mathbf{y} - \mathbf{y}_*)\|^2 + \frac{\omega}{2}\|\mathbf{u}\|^2 \tag{3.53}$$

over $(\mathbf{y}, \mathbf{u}) \in \ell_2(\mathbb{I}_Y) \times \ell_2(\mathbb{I}_Q)$ *subject to the linear constraints* (3.21)

$$\mathbf{A}\mathbf{y} = \boldsymbol{f} + \mathbf{E}\mathbf{D}_Q^{-1}\mathbf{u}.$$

Remark 3.9. Problem (SCPW) is related to the reference problem (CPW) (respectively, (CP)) as follows. The constraints remain unchanged as (3.21) is the exact wavelet representation of (2.5). Furthermore, there exist finite positive constants c_J, C_J such that for any $y = \mathbf{y}^T\Psi_Y \in Y$, and any $u = \mathbf{u}^T\Psi_U \in U$, one has

$$c_J\, \mathbf{J}(\mathbf{y}, \mathbf{u}) \le J(y, u) \le C_J\, \mathbf{J}(\mathbf{y}, \mathbf{u}). \tag{3.54}$$

Moreover, in the case of *compatible data* $y_* = TA^{-1}f$, the respective minimizers coincide.

Proof. The fact that $\|\mathbf{D}_Z^{-1}(\mathbf{y} - \mathbf{y}_*)\| \sim \|\mathbf{R}_Z^{1/2}\mathbf{D}_Z^{-1}(\mathbf{y} - \mathbf{y}_*)\| = \|Ty - y_*\|_Z$ follows immediately from (3.17). This proves (3.54), which, in turn, implies the rest of the assertion. ∎

In case of incompatible data, the solution of (SCPW) may certainly depend on the specific equivalent norm although the constraints remain unchanged. Thus, in the mandatory case the formulation (SCPW) may not yield a solution that could be termed optimal. On the other hand, in the ambiguous case (see the scenario described in Remark 2.2) (SCPW) may very well be expected to capture the essential features of the original extremal problem reflecting the mutual effect of different regularity scales in the cost functional. Specifically, realizing more regular controls should support their practical executability. In this case, we shall refer to $\mathbf{J}(\cdot, \cdot)$ as a *representer* of $J(\cdot, \cdot)$ from (2.3).

The situation that a specific norm in the cost functional is essential arises most likely for the norms described in Remark 3.1. As pointed out in Remark 3.8, the corresponding Riesz operators are then numerically accessible. In favor of a unified and clearer treatment we shall focus first entirely on (SCPW), however. Later we shall indicate the necessary modifications for dealing with the mandatory situation (CPW) when the involved Riesz operators are accessible in the sense of Remark 3.8.

For the representer defined in (3.53) the reduced cost functional takes on the form

$$\mathbf{J}(\mathbf{u}) := \frac{1}{2}\|\mathbf{Z}\mathbf{u}-\mathbf{G}\|^2+\frac{\omega}{2}\|\mathbf{u}\|^2, \tag{3.55}$$

whose minimization is equivalent to solving

$$\mathbf{Q}\mathbf{u}=\mathbf{g}, \quad \text{where} \quad \mathbf{Q}:=\mathbf{Z}^T\mathbf{Z}+\omega\mathbf{I},\ \mathbf{g}:=\mathbf{Z}^T\mathbf{G}, \tag{3.56}$$

where $\mathbf{G},\mathbf{Z}$ are defined in (3.34). By the same reasoning as before we still have

$$c_{\mathbf{Q}}\|\mathbf{v}\|\le\|\mathbf{Q}\mathbf{v}\|\le C_{\mathbf{Q}}\|\mathbf{v}\|, \qquad \mathbf{v}\in\ell_2(\mathbb{I}_Q), \tag{3.57}$$

with finite positive constants $c_{\mathbf{Q}},C_{\mathbf{Q}}$ that may depend on ω under the same circumstances described in Remark 3.5. Thus, in principle, the well-posedness of the optimal control problem is not affected by exchanging equivalent norms.

In view of (3.57), there exists a fixed positive parameter α such that the *descent iteration*

$$\mathbf{u}^{k+1}=\mathbf{u}^k+\alpha(\mathbf{g}-\mathbf{Q}\mathbf{u}^k) \tag{3.58}$$

reduces the error in each step by at least a factor $\rho<1$, i.e.,

$$\|\mathbf{u}-\mathbf{u}^{k+1}\|\le\rho\|\mathbf{u}-\mathbf{u}^k\|, \quad k=0,1,2,\dots, \tag{3.59}$$

where $\mathbf{u}$ is the exact solution of (3.56). Our ultimate goal is to carry out this iteration approximately with dynamically updated accuracy tolerances.

The application of $\mathbf{Q}$ will be facilitated by the following obvious modification of Remark 3.6. The solution $\mathbf{u}$ of the system (3.56) results from the following Euler equations in wavelet coordinates:

$$\begin{aligned}
&\mathbf{A}\mathbf{y}=\boldsymbol{f}+\mathbf{E}\mathbf{D}_Q^{-1}\mathbf{u},\\
&\mathbf{A}^T\mathbf{p}=-\mathbf{T}^T\mathbf{D}_Z^{-2}\left(\mathbf{T}\mathbf{y}-\mathbf{y}_*\right), &\text{(EE}')\quad (3.60)\\
&\omega\mathbf{u}=\mathbf{D}_Q^{-1}\mathbf{E}^T\mathbf{p}. &(3.61)
\end{aligned}$$

Likewise, according to Lemma 3.7, one has for any $\mathbf{v}\in\ell_2(\mathbb{I}_Q)$, the representation

$$\mathbf{Q}\mathbf{v}-\mathbf{g}=\omega\mathbf{v}-\mathbf{D}_Q^{-1}\mathbf{E}^T\mathbf{p}, \tag{3.62}$$

where $\mathbf{p}$ is the solution of the first two equations in (EE′). Thus, for any given $\mathbf{v}$, the sequence $\mathbf{p}$ is given by solving

$$\mathbf{A}^T\mathbf{p}=-\mathbf{T}^T\mathbf{D}_Z^{-2}\left(\mathbf{T}\mathbf{y}-\mathbf{y}_*\right), \quad \text{where} \quad \mathbf{A}\mathbf{y}=\boldsymbol{f}+\mathbf{E}\mathbf{D}_Q^{-1}\mathbf{v}. \tag{3.63}$$

4. Basic concepts. In this section we collect the main conceptual tools from [CDD1, CDD2, CDD3] that will be needed to treat, in the general situation, first (3.56) for the numerical solution of (SCPW). The core issue will be the application of $\mathbf{Q}$ and the evaluation of the right-hand-side $\mathbf{g}$ (which will be easier than that of $\widetilde{\mathbf{Q}}$ and $\widetilde{\mathbf{g}}$).

4.1. Perturbed iterations. The basic strategy applies, in principle, to any system of the form

$$\mathbf{Mq} = \mathbf{z}, \tag{4.1}$$

where $\mathbf{M} : \ell_2 \to \ell_2$ is a (possibly) infinite matrix satisfying

$$c_{\mathbf{M}}\|\mathbf{v}\| \le \|\mathbf{Mv}\| \le C_{\mathbf{M}}\|\mathbf{v}\|, \quad \mathbf{v} \in \ell_2, \tag{4.2}$$

for some finite positive constants $c_{\mathbf{M}}, C_{\mathbf{M}}$, as well as

$$\rho := \|\mathbf{I} - \alpha\mathbf{M}\| < 1 \tag{4.3}$$

for some positive number α. Clearly, due to the positive definiteness of $\mathbf{Q}$ and by (3.57), $\mathbf{M} = \mathbf{Q}$ falls into this category.

Given (4.3), the gradient descent iteration

$$\mathbf{q}^{k+1} = \mathbf{q}^k + \alpha(\mathbf{z} - \mathbf{Mq}^k), \quad k = 0, 1, 2, \dots, \tag{4.4}$$

will converge with a fixed error reduction rate $\rho < 1$ per step. Of course, this iteration cannot be carried out exactly because $\mathbf{M}$ is an infinite matrix and the data $\mathbf{z}$ could be an infinite array. However, one can hope that perturbed iterations with dynamical accuracy tolerances, that are suitably updated in the course of the iteration, will still converge. Thus, we need a routine with the following property.

RES $[\eta, \mathbf{M}, \mathbf{z}, \mathbf{v}] \to \mathbf{r}_\eta$ determines for a given tolerance $\eta > 0$ a finitely supported sequence $\mathbf{r}_\eta$ satisfying

$$\|\mathbf{z} - \mathbf{Mv} - \mathbf{r}_\eta\| \le \eta. \tag{4.5}$$

There is a further ingredient whose role is at this stage not apparent yet but which will eventually play a crucial role in controlling the complexity of the scheme.

COARSE $[\eta, \mathbf{w}] \to \mathbf{w}_\eta$ determines for any finitely supported input vector $\mathbf{w}$ a vector $\mathbf{w}_\eta$ with smallest possible support such that

$$\|\mathbf{w} - \mathbf{w}_\eta\| \le \eta. \tag{4.6}$$

The precise description of COARSE can be found in [CDD1]. The idea is to sort the entries of $\mathbf{w}$ by size and to subtract squares of their moduli starting from the smallest one until the sum reaches η^2. The sorting usually introduces a logarithmic term of the size of $\mathbf{w}$. A quasi-sorting based on binary binning can be shown to avoid the logarithmic term at the expense of the resulting support size being at most a fixed constant of the minimal size; see [B]. This will suffice for the subsequent analysis so that it is justified to suppress logarithmic terms in what follows.

Let us suppose for the moment that the routine RES is already at our disposal. We shall first devise the precise form of a perturbed iteration that converges in the following sense. For every target accuracy ε it produces after finitely many steps a finitely supported approximate solution with accuracy ε.

Following [CDD2], to arrive at the right interplay between the routines RES and COARSE, we need the following control parameter. Given (an estimate of) the reduction rate ρ and the step size parameter α from (4.3), fix any constant $a > 1$ and let

$$K := \min\{\ell \in \mathbb{N} : \rho^{\ell-1}(\alpha\ell + \rho) \le \tfrac{1}{2(1+a)}\}. \tag{4.7}$$

(Here the upper bound $(2+2a)^{-1}$ stems from the analysis in [CDD3] and will be used again below.) Denoting in the following always by $\mathbf{q}$ the exact solution of (4.1), a perturbed version of (4.4) can now be formulated as follows.

SOLVE $[\varepsilon, \mathbf{M}, \mathbf{z}, \overline{\mathbf{q}}^0, \varepsilon_0] \to \overline{\mathbf{q}}_\varepsilon$.

(i) Fix a target accuracy $\varepsilon > 0$. Given an initial guess $\overline{\mathbf{q}}^0$ along with an error bound $\|\mathbf{q} - \overline{\mathbf{q}}^0\| \le \varepsilon_0$, set $j = 0$.

(ii) If $\varepsilon_j \le \varepsilon$, stop and set $\overline{\mathbf{q}}_\varepsilon := \overline{\mathbf{q}}^j$. Otherwise set $\mathbf{v}^0 := \overline{\mathbf{q}}^j$.

(ii.1) For $k = 0, \dots, K-1$ compute RES $[\rho^k \varepsilon_j, \mathbf{M}, \mathbf{z}, \mathbf{v}^k] \to \mathbf{r}^k$ and

$$\mathbf{v}^{k+1} := \mathbf{v}^k + \alpha \mathbf{r}^k. \tag{4.8}$$

(ii.2) Apply COARSE $[\frac{a\varepsilon_j}{2(1+a)}, \mathbf{v}^K] \to \overline{\mathbf{q}}^{j+1}$; set $\varepsilon_{j+1} := \frac{1}{2}\varepsilon_j$, $j+1 \to j$ and go to (ii).

In the case that no particular initial guess is known, step (i) is replaced by the default

(i)′ Fix a target accuracy $\varepsilon > 0$. Set $j = 0$ and

$$\overline{\mathbf{q}}^0 = \mathbf{0}, \quad \varepsilon_0 := c_{\mathbf{M}}^{-1} \|\mathbf{z}\|. \tag{4.9}$$

In this case we use the short notation SOLVE $[\varepsilon, \mathbf{M}, \mathbf{z}] \to \overline{\mathbf{q}}_\varepsilon$.

The choice of the interior tolerance $\rho^k \varepsilon_j$ in step (ii.1) yields the following estimate from [CDD2] regarding the final iterate $\mathbf{v}^K$ resulting from step (ii.1). Inserting the exact iterate of (4.4) with initial value $\overline{\mathbf{w}}^j$ denoted by $\overline{\mathbf{v}}^K(\overline{\mathbf{w}}^j)$, we get

$$\begin{aligned} \|\mathbf{v}^K - \mathbf{q}\| &\le \|\mathbf{v}^K - \overline{\mathbf{v}}^K(\overline{\mathbf{w}}^j)\| + \|\overline{\mathbf{v}}^K(\overline{\mathbf{w}}^j) - \mathbf{q}\| \\ &\le \alpha K \rho^{K-1} \varepsilon_j + \rho^K \|\overline{\mathbf{w}}^j - \mathbf{q}\| \le (\alpha K + \rho) \rho^{K-1} \varepsilon_j. \end{aligned} \tag{4.10}$$

Employing the choice of K in (4.7), this yields

$$\|\mathbf{v}^K - \mathbf{q}\| \le \frac{\varepsilon_j}{2(1+a)}. \tag{4.11}$$

The particular form of the constants for the interior estimates that can be seen in (4.10) will be employed later in section 5.

Straightforward perturbation arguments reveal the following result; see [CDD2, CDD3].

PROPOSITION 4.1. *The iterates* $\overline{\mathbf{q}}^j$ *generated by* SOLVE $[\varepsilon, \mathbf{M}, \mathbf{z}]$ *satisfy*

$$\|\mathbf{q} - \overline{\mathbf{q}}^j\| \le \varepsilon_j, \qquad j \in \mathbb{N}_0, \tag{4.12}$$

where $\varepsilon_j = 2^{-j}\varepsilon_0$.

Of course, the estimates for α rely on the constants in the norm equivalences (3.5) and in the relation (4.2). Thus there may be only a poor estimate for ρ, which, in turn, gives rise to an overly pessimistic choice for the number K defined in (4.7) of perturbed iterations in each block (ii) of SOLVE prior to a coarsening step. Therefore, we recall from [CDD3] that step (ii) can be terminated based on monitoring the approximate residuals as follows. By (4.2), we have

$$\|\mathbf{q} - \mathbf{v}\| \le c_{\mathbf{M}}^{-1} \|\mathbf{z} - \mathbf{M}\mathbf{v}\|. \tag{4.13}$$

Choose any $\bar{\rho} < 1$ and define $\bar{K}$ by (4.7) with respect to $\bar{\rho}$. Replacing ρ by $\bar{\rho}$ in the definition of the tolerances in step (ii) of SOLVE would take $M := \max\{K, \bar{K}\}$ steps to ensure that in the $(j+1)$st call of (ii) $\|\mathbf{q} - \mathbf{v}^M\| \le \varepsilon_j/10$. Now suppose that the

ρ is expected to be a too-pessimistic estimate of the true reduction rate. Choosing, e.g., $\bar{\rho} := 1/2$ and setting $\eta_k := 2^{-k}\varepsilon_j = \bar{\rho}^{-k}\varepsilon_j$ as tolerances in the $(j+1)$st call of (ii), we infer from (4.13) and (4.5) that

$$\|\mathbf{q} - \mathbf{v}^k\| \le c_{\mathbf{M}}^{-1}\|\mathbf{z} - \mathbf{M}\mathbf{v}^k\| \le c_{\mathbf{M}}^{-1}\left(\eta_k + \|\mathbf{r}^k\|\right) =: \delta_k,$$

where $\mathbf{r}^k$ is the approximate residual produced in step (ii.1) of SOLVE. By the previous remarks, we can terminate the iteration in step (ii) of SOLVE when either $k = K-1$ or the *computable a posteriori bound* δ_k satisfies $\delta_k \le \varepsilon_j/2(1+a)$, which might happen much earlier than predicted by (4.7). In fact, a refined argument reveals that a weaker error reduction in (4.11) suffices; see [CDD3]. Of course, the constant $c_{\mathbf{M}}$ is usually also only estimated. However, a poor estimate enters the above a posteriori termination criterion in a less severe way than a poor estimate for ρ. Nevertheless, to keep the exposition as simple as possible, we confine the subsequent discussion to the above version of SOLVE, bearing in mind that variants of the above sort are automatically covered by the complexity analysis.

4.2. Complexity analysis. Of course, the main issues are the actual realization of the routine RES and to understand its complexity. The realization will depend on the concrete application, which here will be the control problem (SCPW). Here we outline first a suitable framework for the complexity analysis. Striving for schemes that are in some sense *optimal*, the meaning of optimality must first be clarified.

We say that the scheme SOLVE has an *optimal work/accuracy rate* s if the following is true: whenever the error of best N-term approximation

$$\sigma_N(\mathbf{q}) := \|\mathbf{q} - \mathbf{q}_N\| := \min_{\#\operatorname{supp}\mathbf{v}\le N}\|\mathbf{q} - \mathbf{v}\| \tag{4.14}$$

decays like $\mathcal{O}(N^{-s})$, then the solution $\overline{\mathbf{q}}_\varepsilon$ is produced by SOLVE at an expense that also stays proportional to $\varepsilon^{-1/s}$ and in that sense matches the best N-term rate. Clearly this implies that $\#\operatorname{supp}\overline{\mathbf{q}}_\varepsilon$ also stays proportional to $\varepsilon^{-1/s}$. Thus, our benchmark is "best N-term approximation," which is obviously the best one can hope for.

Clearly this best N-term approximation $\mathbf{q}_N$ of $\mathbf{q}$ is given by taking the N largest terms in modulus from $\mathbf{q}$. When $\mathbf{q}$ is the (unknown) solution of (4.1), this knowledge is certainly not available. Nevertheless, the formulation of appropriate complexity criteria will be based on a characterization of those sequences $\mathbf{v}$ for which the best N-term approximation error decays at a particular rate (*Lorentz spaces*). Following [CDD1], consider sequences that are *sparse* in the sense that for any given threshold $0 < \eta \le 1$, say, the number of terms exceeding that threshold is controlled by some function of this threshold. Specifically, set for some $0 < \tau < 2$

$$\ell_\tau^w := \{\mathbf{v} \in \ell_2 :\ \#\{\lambda \in \mathbb{I} : |v_\lambda| > \eta\} \le C_{\mathbf{v}}\,\eta^{-\tau} \text{ for all } 0 < \eta \le 1\}, \tag{4.15}$$

i.e., the set ℓ_τ^w consists of all those sequences $\mathbf{v} \in \ell_2$ for which there exists a constant $C_{\mathbf{v}}$ such that for all $0 < \eta \le 1$ the number of terms v_λ whose moduli exceed the threshold η is bounded by $C_{\mathbf{v}}\eta^{-\tau}$. Note that this determines a strict subspace of ℓ_2 only when $\tau < 2$, and the sequence is sparser the smaller τ is. Denote for a given $\mathbf{v} \in \ell_\tau^w$ by $C_{\mathbf{v}}$ the smallest constant for which (4.15) holds. Then one has

$$|\mathbf{v}|_{\ell_\tau^w} := \sup_{n\in\mathbb{N}} n^{1/\tau} v_n^* = C_{\mathbf{v}}^{1/\tau}, \tag{4.16}$$

where $\mathbf{v}^* = (v_n^*)_{n\in\mathbb{N}}$ is a nondecreasing rearrangement of $\mathbf{v}$. The quantity

$$\|\mathbf{v}\|_{\ell_\tau^w} := \|\mathbf{v}\| + |\mathbf{v}|_{\ell_\tau^w} \tag{4.17}$$

can be shown to be a quasi-norm for ℓ_τ^w [CDD1]. Furthermore, because of the continuous embeddings

$$\ell_\tau \hookrightarrow \ell_\tau^w \hookrightarrow \ell_{\tau+\varepsilon} \hookrightarrow \ell_2 \qquad \text{for } \tau < \tau + \varepsilon < 2, \tag{4.18}$$

ℓ_τ^w is very close to ℓ_τ, which justifies calling it *weak* ℓ_τ. Now we can recall the following result from [CDD1], which relates the sequences in ℓ_τ^w to best N-term approximation.

PROPOSITION 4.2. *Let positive real numbers s and τ be related by*

$$\frac{1}{\tau} = s + \frac{1}{2}. \tag{4.19}$$

Then a sequence $\mathbf{v}$ belongs to ℓ_τ^w if and only if

$$\|\mathbf{v} - \mathbf{v}_N\| \lesssim N^{-s} \qquad \text{and} \qquad \sigma_N(\mathbf{v}) \lesssim N^{-s}\,\|\mathbf{v}\|_{\ell_\tau^w}, \tag{4.20}$$

where as before $\mathbf{v}_N$ denotes a best N-term approximation of $\mathbf{v}$.

Depending on the space H which is characterized by the wavelet basis Ψ_H, the fact that an array of wavelet coefficients satisfies $\mathbf{v} \in \ell_\tau$ is typically equivalent to the fact that the expansion $\mathbf{v}^T\Psi_H$ belongs to a certain Besov space which describes a much weaker regularity measure than a Sobolev space of corresponding order. In view of (4.18), Proposition 4.2 therefore expresses how much loss of regularity can be *compensated* by best N-term approximation, i.e., by judiciously placing the degrees of freedom in a nonlinear way to retain a certain optimal order of error decay. We shall return to this issue later.

As will be seen in Theorem 4.3, a key criterion for a scheme SOLVE to exhibit an optimal work/accuracy rate can be formulated through the following property of the respective residual approximation.

τ^*-sparsity. *The routine* RES *is called τ^*-sparse for some $0 < \tau^* < 2$ if the following holds: whenever the solution $\mathbf{q}$ of (4.1) belongs to ℓ_τ^w for some $\tau^* < \tau < 2$, then for any $\mathbf{v}$ with finite support the output $\mathbf{r}_\eta$ of* RES $[\eta, \mathbf{M}, \mathbf{z}, \mathbf{v}]$ *satisfies*

(i)
$$\begin{aligned} \|\mathbf{r}_\eta\|_{\ell_\tau^w} &\lesssim \max\{\|\mathbf{v}\|_{\ell_\tau^w}, \|\mathbf{q}\|_{\ell_\tau^w}\}, \\ \#\operatorname{supp}\mathbf{r}_\eta &\lesssim \eta^{-1/s}\max\{\|\mathbf{v}\|_{\ell_\tau^w}^{1/s}, \|\mathbf{q}\|_{\ell_\tau^w}^{1/s}\}, \end{aligned} \tag{4.21}$$

where s and τ are related as before by (4.19);

(ii) *the number of floating point operations needed to compute $\mathbf{r}_\eta$ stays proportional to $\#\operatorname{supp}\mathbf{r}_\eta$.*

Furthermore, the constants in (i), (ii) *depend only on τ as $\tau \to \tau^*$.*

In this context we shall always make the following tacit assumption. Given data are always to be considered completely accessible. In practical terms this may mean that depending on some final target accuracy (in view of (3.38)) sufficiently many of the corresponding coefficients of explicitly given data are determined in a preprocessing step and then ordered by size, so that COARSE can be applied to a finitely supported array. For notational simplicity we shall not distinguish between the ideal exact data and such an approximation.

The following result can then be extracted from the analysis in [CDD2] (see also [CDD3] for nonlinear problems) and was used in [DUV].

THEOREM 4.3. *If* RES *is τ^*-sparse and if the exact solution $\mathbf{q}$ of* (4.1) *belongs to ℓ_τ^w for some $\tau > \tau^*$, then for every $\varepsilon > 0$ algorithm* SOLVE $[\varepsilon, \mathbf{M}, \mathbf{z}]$ *produces after*

finitely many steps an output $\overline{\mathbf{q}}_\varepsilon$ *(which, according to Proposition* 4.1*, always satisfies* $\|\mathbf{q}-\overline{\mathbf{q}}_\varepsilon\|<\varepsilon$*) with the following properties: for* s *and* τ *related by* (4.19)*, one has*

$$\#\operatorname{supp}\overline{\mathbf{q}}_\varepsilon \lesssim \varepsilon^{-1/s}\|\mathbf{q}\|_{\ell^w_\tau}^{1/s}, \qquad \|\overline{\mathbf{q}}_\varepsilon\|_{\ell^w_\tau} \lesssim \|\mathbf{q}\|_{\ell^w_\tau}, \tag{4.22}$$

and the number of floating point operations needed to compute $\overline{\mathbf{q}}_\varepsilon$ *remains proportional to* $\#\operatorname{supp}\overline{\mathbf{q}}_\varepsilon$.

Thus, τ^*-sparsity of the routine RES implies asymptotically optimal work/accuracy rates of the scheme SOLVE for a certain range of decay rates given by τ^*. We stress that the algorithm itself does *not* require any a priori knowledge about the solution such as its actual best N-term approximation rate. Theorem 4.3 also shows that controlling the ℓ^w_τ-norm of the quantities generated in the course of the computation is crucial. This finally explains the role of COARSE in step (ii.2) of SOLVE through the following result; see, e.g., [CDD3].

LEMMA 4.4 (coarsening lemma). *Let* $\mathbf{v}\in\ell^w_\tau$ *and let* $\mathbf{w}$ *be any finitely supported approximation such that* $\|\mathbf{v}-\mathbf{w}\|\le\eta$*. Then, for any fixed* $a>1$*, the output* $\mathbf{w}_\eta$ *of* COARSE$\,[a\eta,\mathbf{w}]$ *satisfies*

$$\#\operatorname{supp}\mathbf{w}_\eta \lesssim \|\mathbf{v}\|_{\ell^w_\tau}^{1/\tau}\,\eta^{-1/s}, \quad \|\mathbf{v}-\mathbf{w}_\eta\|\le(1+a)\eta, \qquad \|\mathbf{w}_\eta\|_{\ell^w_\tau}\lesssim\|\mathbf{v}\|_{\ell^w_\tau}, \tag{4.23}$$

where the constants in the first and third estimate depend on a *when* $a\to 1$.

Thus, knowing an error bound for a given finitely supported approximation $\mathbf{w}$, a certain coarsening using only knowledge about $\mathbf{w}$, produces a new approximation to (the possibly unknown) $\mathbf{v}$ which gives rise to a slightly larger error but realizes up to a uniform constant the optimal relation between support and accuracy. In the scheme SOLVE this means that by the coarsening step, the ℓ^w_τ-norms of the iterates $\mathbf{v}^K$ are controlled. Recall from (4.11) that the choice of the constant K in (4.7), which controls the number of iterations in step (ii.1), guarantees that in the $(j+1)$st outer iteration of SOLVE the iterate $\mathbf{v}^K$ satisfies $\|\mathbf{q}-\mathbf{v}^K\|\le\frac{1}{2(1+a)}\varepsilon_j$. The threshold $\frac{a\varepsilon_j}{2(1+a)}$ in step (ii.2) ensures, on account of (4.23), that the error after coarsening is still bounded by $\frac{1}{2}\varepsilon_j$. At the same time, if $\mathbf{q}\in\ell^w_\tau$, then $\|\overline{\mathbf{q}}^j\|_{\ell^w_\tau}$ remains bounded and $\#\operatorname{supp}\overline{\mathbf{q}}^j$ increases at most like $\varepsilon_j^{-1/s}$, which is the best possible N-term rate for sequences in ℓ^w_τ. Thus to ensure an overall optimal work/accuracy rate, one has to show that the ℓ^w_τ-norms of the intermediate iterates $\mathbf{v}^k$ in step (ii.1) of SOLVE cannot grow too much, which is indeed guaranteed by τ^*-sparsity.

The proportionally constants in the above complexity estimates can, in principle, be assessed or at least estimated in any concrete case. They will depend on the Riesz constants of the employed wavelet bases, on the ellipticity and continuity constants in the operator equation (which determine the constants in (4.2) and (4.3)), on the choice of the constant a in the coarsening step, and on the concrete realization of the approximate application of matrices like $\mathbf{A}$; see section 5.1.

The remainder of this paper is now devoted to the construction and analysis of a concrete realization of SOLVE—termed $\text{SOLVE}_{\text{SCPW}}$—for the problem (SCPW) such that the corresponding routine RES_{SCPW} is τ^*-sparse.

5. The scheme $\text{SOLVE}_{\text{SCPW}}$.

Since $\mathbf{Q}$ from (3.56) still involves several inverses of matrices it is not so clear how to realize a residual approximation in an economical way—recall obstruction (b) in section 3.4.

5.1. Realization of the routine RES_{SCPW}.

The realization of the routine RES_{SCPW} for the problem (3.56) will be based on the residual representation (3.62);

see also Lemma 3.7. However, this requires solving the two auxiliary systems (3.21), (3.60) in (EE′). Since the residual has to be only approximated, these systems will have to be solved only approximately. These approximate solutions, in turn, will be provided again by calls of the scheme SOLVE but this time with respect to suitable residual schemes tailored to the systems in (EE′). In all our examples the matrix $\mathbf{A}$ appearing in (EE′) is symmetric positive definite and the choice of wavelet bases ensures the validity of (3.22). Hence, (4.2) and (4.3) hold for $\mathbf{M} = \mathbf{A}$ and $\mathbf{M} = \mathbf{A}^T$ so that the scheme SOLVE can indeed be invoked. We hasten to mention, however, that the symmetry and positive definiteness of $\mathbf{A}$ is not essential. As long as (3.22) holds, which means that the operator equation induced by the constraints is well posed (which is still the case, e.g., for many saddle point problems), we can multiply the systems in (EE′) by $\mathbf{A}^T$, respectively, $\mathbf{A}$, to arrive at a least squares formulation with $\mathbf{M} = \mathbf{A}^T\mathbf{A}$ or $\mathbf{M} = \mathbf{A}\mathbf{A}^T$, still satisfying (4.2) but now yielding symmetric positive definite systems to ensure (4.3). However, to keep the exposition as simple as possible, we confine the following discussion to the case that $\mathbf{A}$ already satisfies (in addition to (3.22)) (4.3).

Note also that although we conceptually use the fact that a gradient descent for the reduced problem (3.56) reduces the error for $\mathbf{u}$ in each step by a fixed amount, the use of (EE′) for the evaluation of the residuals will generate as byproducts approximate solutions to the full Euler–Lagrange system, i.e., we shall obtain approximations to the exact solution triple $(\mathbf{y}, \mathbf{p}, \mathbf{u})$ of (EE′).

Under this hypothesis, we have to formulate next the ingredients for suitable versions $\text{SOLVE}_{\text{PRM}}$ and $\text{SOLVE}_{\text{ADJ}}$ of SOLVE for the systems in (EE′). Specifically, this requires identifying residual schemes RES_{PRM} and RES_{ADJ} for the systems $\text{SOLVE}_{\text{PRM}}$ and $\text{SOLVE}_{\text{ADJ}}$. The main task in both cases is to apply scaling matrices and operators $\mathbf{A}, \mathbf{T}, \mathbf{E}$ along with their transposes. Of course, the application of a scaling matrix can be done exactly and hence is easily realized. To ease notation we may therefore combine conceptually the application of $\mathbf{T}$ followed by $\mathbf{D}_Z^{-1}$ in a single operator application, which will be reflected by the notation

$$\mathbf{T}_Z := \mathbf{D}_Z^{-1}\mathbf{T}, \quad \mathbf{y}_Z := \mathbf{D}_Z^{-1}\mathbf{y}_*. \tag{5.1}$$

Again we assume for the moment that routines for the application of these operators are available, i.e., that for any $\mathbf{L} \in \{\mathbf{A}, \mathbf{A}^T, \mathbf{T}_Z, \mathbf{T}_Z^T, \mathbf{E}, \mathbf{E}^T\}$ we have a scheme with the following property at our disposal. We shall later discuss their concrete realization.

$\text{APPLY}[\eta, \mathbf{L}, \mathbf{v}] \to \mathbf{w}_\eta$ determines for any finitely supported input vector $\mathbf{v}$ and any tolerance $\eta > 0$ a finitely supported output $\mathbf{w}_\eta$ which satisfies

$$\|\mathbf{L}\mathbf{v} - \mathbf{w}_\eta\| \le \eta. \tag{5.2}$$

The scheme $\text{SOLVE}_{\text{PRM}}$ for the first system in (EE′) is now defined by

$$\text{SOLVE}_{\text{PRM}}[\eta, \mathbf{A}, \mathbf{E}, \mathbf{D}_Q, \boldsymbol{f}, \mathbf{v}, \overline{\mathbf{y}}^0, \varepsilon_0] := \text{SOLVE}[\eta, \mathbf{A}, \boldsymbol{f} + \mathbf{E}\mathbf{D}_Q^{-1}\mathbf{v}, \overline{\mathbf{y}}^0, \varepsilon_0], \tag{5.3}$$

where $\overline{\mathbf{y}}^0$ is an initial guess for the solution $\mathbf{y}$ of $\mathbf{A}\mathbf{y} = \boldsymbol{f} + \mathbf{E}\mathbf{D}_Q^{-1}\mathbf{v}$ with accuracy ε_0 and where the scheme RES for step (ii) in SOLVE is in this case realized by a new routine RES_{PRM}, which is defined as follows.

$\text{RES}_{\text{PRM}}[\eta, \mathbf{A}, \mathbf{E}, \mathbf{D}_Q, \boldsymbol{f}, \mathbf{v}, \overline{\mathbf{y}}] \to \mathbf{r}_\eta$ determines for any positive tolerance η a given finitely supported $\mathbf{v}$ and any finitely supported input $\overline{\mathbf{y}}$ a finitely supported approximate residual $\mathbf{r}_\eta$ satisfying (4.5), that is,

$$\|\boldsymbol{f} + \mathbf{E}\mathbf{D}_Q^{-1}\mathbf{v} - \mathbf{A}\overline{\mathbf{y}} - \mathbf{r}_\eta\| \le \eta, \tag{5.4}$$

as follows:

(i) $\text{APPLY}[\frac{1}{3}\eta, \mathbf{A}, \overline{\mathbf{y}}] \to \mathbf{w}_\eta$;
(ii) $\text{COARSE}[\frac{1}{3}\eta, \boldsymbol{f}] \to \boldsymbol{f}_\eta$;
$\text{APPLY}[\frac{1}{3}\eta, \mathbf{E}, \mathbf{D}_Q^{-1}\mathbf{v}] \to \mathbf{z}_\eta$;
(iii) set $\mathbf{r}_\eta := \boldsymbol{f}_\eta + \mathbf{z}_\eta - \mathbf{w}_\eta$.

In fact, noting that $\boldsymbol{f} + \mathbf{E}\mathbf{D}_Q^{-1}\mathbf{v} - \mathbf{A}\overline{\mathbf{y}} - \mathbf{r}_\eta = (\boldsymbol{f} - \boldsymbol{f}_\eta) + (\mathbf{E}\mathbf{D}_Q^{-1}\mathbf{v} - \mathbf{z}_\eta) + (\mathbf{w}_\eta - \mathbf{A}\overline{\mathbf{y}})$, by triangle inequality (5.4) is an immediate consequence of the choice of the tolerances in steps (i)–(iii) of RES_{PRM}.

Similarly, we need a version of SOLVE for the approximate solution of the second system in (EE′), which with the notation (5.1) now reads $\mathbf{A}^T\mathbf{p} = -\mathbf{T}_Z^T(\mathbf{T}_Z\mathbf{y} - \mathbf{y}_Z)$. This depends therefore on $\mathbf{y}_Z = \mathbf{D}_Z^{-1}\mathbf{y}_*$, an approximate solution $\overline{\mathbf{y}}$ of the primal system, and possibly on some initial guess $\overline{\mathbf{p}}^0$ with accuracy ε_0. Specifically, we set here

$$\text{SOLVE}_{\text{ADJ}}[\eta, \mathbf{A}, \mathbf{T}_Z, \mathbf{y}_Z, \overline{\mathbf{y}}, \overline{\mathbf{p}}^0, \varepsilon_0] := \text{SOLVE}[\eta, \mathbf{A}^T, -\mathbf{T}_Z^T(\mathbf{T}_Z\overline{\mathbf{y}} - \mathbf{y}_Z), \overline{\mathbf{p}}^0, \varepsilon_0]. \tag{5.5}$$

As usual we will assume that the data $\boldsymbol{f}, \mathbf{y}_Z$ are approximated in a preprocessing step with sufficient accuracy (depending on the final target accuracy for solving (3.56)) by finite arrays whose entries are ordered by size and hence can be treated by COARSE.

Again we have to identify a suitable residual approximation scheme RES_{ADJ} for step (ii) of this version of SOLVE where the main issue is the approximate evaluation of the right-hand side. The routine RES_{ADJ} is defined as follows.

$\text{RES}_{\text{ADJ}}[\eta, \mathbf{A}, \mathbf{T}_Z, \mathbf{y}_Z, \overline{\mathbf{y}}, \overline{\mathbf{p}}] \to \mathbf{r}_\eta$ determines for any positive tolerance η, given finitely supported data $\overline{\mathbf{y}}, \mathbf{y}_Z$ and any finitely supported input $\overline{\mathbf{p}}$ an approximate residual $\mathbf{r}_\eta$ satisfying (4.5), i.e.,

$$\| - \mathbf{T}_Z^T(\mathbf{T}_Z\overline{\mathbf{y}} - \mathbf{y}_Z) - \mathbf{A}^T\overline{\mathbf{p}} - \mathbf{r}_\eta\| \le \eta, \tag{5.6}$$

as follows:

(i) $\text{APPLY}[\frac{1}{3}\eta, \mathbf{A}^T, \overline{\mathbf{p}}] \to \mathbf{w}_\eta$;
(ii) $\text{APPLY}[\frac{1}{6\,C_{\mathbf{T}}}\eta, \mathbf{T}_Z, \overline{\mathbf{y}}] \to \mathbf{z}_\eta$ with $C_{\mathbf{T}}$ from (3.51);
$\text{COARSE}[\frac{1}{6\,C_{\mathbf{T}}}\eta, \mathbf{y}_Z] \to (\mathbf{y}_Z)_\eta$; set $\mathbf{d}_\eta := (\mathbf{y}_Z)_\eta - \mathbf{z}_\eta$;
$\text{APPLY}[\frac{1}{3}\eta, \mathbf{T}_Z^T, \mathbf{d}_\eta] \to \mathbf{v}_\eta$;
(iii) set $\mathbf{r}_\eta := \mathbf{v}_\eta - \mathbf{w}_\eta$.

To confirm the validity of (5.6), note that by steps (i)–(iii) of RES_{ADJ}

$$\begin{aligned} - \mathbf{T}_Z^T(\mathbf{T}_Z\overline{\mathbf{y}} - \mathbf{y}_Z) - \mathbf{A}^T\overline{\mathbf{p}} - \mathbf{r}_\eta \\ = -\mathbf{T}_Z^T\left((\mathbf{T}_Z\overline{\mathbf{y}} - \mathbf{y}_Z) - \mathbf{d}_\eta\right) + (\mathbf{T}_Z^T\mathbf{d}_\eta - \mathbf{v}_\eta) + (\mathbf{w}_\eta - \mathbf{A}^T\overline{\mathbf{p}}), \end{aligned}$$

so that (5.6) follows, in view of (3.51) and the tolerances above, by triangle inequality.

Recall that the exact solution $\mathbf{u}$ of (3.56) is the third component of the solution triple $(\mathbf{y}, \mathbf{p}, \mathbf{u})$ of the Euler–Lagrange system (EE′). We shall consistently use this notation for the exact solutions of the respective systems.

We are now in a position to define the residual scheme for the version of SOLVE applied to (3.56). We shall refer to this specification as $\text{SOLVE}_{\text{SCPW}}$. Likewise the corresponding residual scheme is denoted by RES_{SCPW}. We shall use the constants from (3.22) and (3.51). Since the scheme is based on (3.62) (see also Lemma 3.7), it will therefore involve several parameters stemming from the auxiliary systems (EE′).

$\text{RES}_{\text{SCPW}}[\eta, \mathbf{Q}, \mathbf{g}, \tilde{\mathbf{y}}, \delta_y, \tilde{\mathbf{p}}, \delta_p, \mathbf{v}, \delta_v] \to (\mathbf{r}_\eta, \tilde{\mathbf{y}}, \delta_y, \tilde{\mathbf{p}}, \delta_p)$ determines for any approximate solution triple $(\tilde{\mathbf{y}}, \tilde{\mathbf{p}}, \mathbf{v})$ of the system (EE′) satisfying

$$\|\mathbf{y} - \tilde{\mathbf{y}}\| \le \delta_y, \quad \|\mathbf{p} - \tilde{\mathbf{p}}\| \le \delta_p, \quad \|\mathbf{u} - \mathbf{v}\| \le \delta_v, \tag{5.7}$$

an approximate residual $\mathbf{r}_\eta$ such that

$$\|\mathbf{g} - \mathbf{Q}\mathbf{v} - \mathbf{r}_\eta\| \le \eta. \tag{5.8}$$

Moreover, the initial approximations $\tilde{\mathbf{y}}, \tilde{\mathbf{p}}$ are overwritten by new approximations $\tilde{\mathbf{y}}, \tilde{\mathbf{p}}$ satisfying (5.7) with new bounds δ_y and δ_p defined in (5.10), as follows:

(i) $\textsc{Solve}_{\textsc{prm}}[\frac{c_{\mathbf{A}}\eta}{3C_{\mathbf{E}}C_{\mathbf{T}}^2}, \mathbf{A}, \boldsymbol{f}, \mathbf{v}, \tilde{\mathbf{y}}, \delta_y] \to \mathbf{y}_\eta$;
(ii) $\textsc{Solve}_{\textsc{adj}}[\frac{\eta}{3C_{\mathbf{E}}}, \mathbf{A}, \mathbf{T}_Z, \mathbf{y}_Z, \mathbf{y}_\eta, \tilde{\mathbf{p}}, \delta_p] \to \mathbf{p}_\eta$;
(iii) $\textsc{Apply}[\frac{\eta}{3}, \mathbf{D}_Q^{-1}\mathbf{E}^T, \mathbf{p}_\eta] \to \mathbf{q}_\eta$;
(iv) set $\mathbf{r}_\eta := \mathbf{q}_\eta - \omega\mathbf{v}$;
(v) set

$$\xi_y := \frac{C_{\mathbf{E}}}{c_{\mathbf{A}}}\,\delta_v + \frac{c_{\mathbf{A}}}{3C_{\mathbf{E}}C_{\mathbf{T}}^2}\,\eta, \quad \xi_p := \frac{C_{\mathbf{T}}^2 C_{\mathbf{E}}}{c_{\mathbf{A}}^2}\,\delta_v + \frac{2}{3C_{\mathbf{E}}}\,\eta, \tag{5.9}$$

and replace $\tilde{\mathbf{y}}, \delta_y$ as well as $\tilde{\mathbf{p}}, \delta_p$ by the new values

$$\begin{aligned} \tilde{\mathbf{y}} &:= \textsc{Coarse}[a\xi_y, \mathbf{y}_\eta], \qquad \delta_y := (1+a)\,\xi_y, \\ \tilde{\mathbf{p}} &:= \textsc{Coarse}[a\xi_p, \mathbf{p}_\eta], \qquad \delta_p := (1+a)\,\xi_p. \end{aligned} \tag{5.10}$$

((5.9) already indicates the conditions on the tolerance η and the accuracy bound δ_v under which the new error bounds in (5.10) are actually tighter. The precise relation between η and δ_v in the context of $\textsc{Solve}_{\textsc{scpw}}$ will emerge from the complexity analysis in section 6; see (6.2).) Let us confirm the claimed estimates (5.8) and (5.10). To this end, let for any given input $\mathbf{v}$ the exact solution to the first system in (EE′) be denoted by $\mathbf{y}_\mathbf{v}$. Moreover, let $\mathbf{p}_\mathbf{v}$ be the exact solution of the second system in (EE′) with $\mathbf{y}$ replaced by $\mathbf{y}_\mathbf{v}$. Finally, let $\hat{\mathbf{p}}$ be the exact solution of the second system but with $\mathbf{y}$ replaced by the approximate solution $\mathbf{y}_\eta$ of the first equation in (EE′). By step (iv) in $\textsc{Res}_{\textsc{scpw}}$ and (3.49), we have

$$\mathbf{g} - \mathbf{Q}\mathbf{v} - \mathbf{r}_\eta = \mathbf{D}_Q^{-1}\mathbf{E}^T\mathbf{p}_\mathbf{v} - \mathbf{q}_\eta = \mathbf{D}_Q^{-1}\mathbf{E}^T(\mathbf{p}_\mathbf{v} - \mathbf{p}_\eta) + \mathbf{D}_Q^{-1}\mathbf{E}^T\mathbf{p}_\eta - \mathbf{q}_\eta.$$

Hence it follows that

$$\|\mathbf{g} - \mathbf{Q}\mathbf{v} - \mathbf{r}_\eta\| \le \frac{\eta}{3} + C_{\mathbf{E}}\|\mathbf{p}_\eta - \mathbf{p}_\mathbf{v}\|. \tag{5.11}$$

To estimate the second term, note that

$$\mathbf{p}_\mathbf{v} - \hat{\mathbf{p}} = \mathbf{A}^{-T}\mathbf{T}_Z^T\mathbf{T}_Z(\mathbf{y}_\mathbf{v} - \mathbf{y}_\eta),$$

and therefore, by (3.22), (3.51), and step (i),

$$\|\mathbf{p}_\mathbf{v} - \hat{\mathbf{p}}\| \le c_{\mathbf{A}}^{-1}C_{\mathbf{T}}^2\|\mathbf{y}_\mathbf{v} - \mathbf{y}_\eta\| \le \frac{\eta}{3C_{\mathbf{E}}}. \tag{5.12}$$

Thus, by step (ii) and (5.12), $\|\mathbf{p}_\eta - \mathbf{p}_\mathbf{v}\| \le \frac{2\eta}{3C_{\mathbf{E}}}$, which together with (5.11) confirms (5.8).

Adhering to the above notational conventions, the first system in (EE′) yields $\mathbf{y} - \mathbf{y}_\mathbf{v} = \mathbf{A}^{-1}\mathbf{E}\mathbf{D}_Q^{-1}(\mathbf{w} - \mathbf{v})$ so that by (5.7), (3.22), and (3.51),

$$\|\mathbf{y} - \mathbf{y}_\eta\| \le \|\mathbf{y} - \mathbf{y}_\mathbf{v}\| + \|\mathbf{y}_\mathbf{v} - \mathbf{y}_\eta\| \le \frac{C_{\mathbf{E}}}{c_{\mathbf{A}}}\,\delta_v + \frac{c_{\mathbf{A}}}{3C_{\mathbf{E}}C_{\mathbf{T}}^2}\,\eta, \tag{5.13}$$

which is the value of ξ_y in step (v). Likewise we infer from the second system in (EE′) that

$$\mathbf{p}-\mathbf{p}_\eta=\mathbf{p}-\hat{\mathbf{p}}+\hat{\mathbf{p}}-\mathbf{p}_\eta=\mathbf{A}^{-T}\mathbf{T}_Z^T\mathbf{T}_Z(\mathbf{y}-\mathbf{y}_\eta)+\hat{\mathbf{p}}-\mathbf{p}_\eta.$$

Hence, by (3.22), (3.24), and step (ii), we obtain

$$\|\mathbf{p}-\mathbf{p}_\eta\|\le\frac{C_{\mathbf{T}}^2}{c_{\mathbf{A}}}\,\xi_y+\frac{\eta}{3C_{\mathbf{E}}}=\frac{C_{\mathbf{T}}^2C_{\mathbf{E}}}{c_{\mathbf{A}}^2}\,\delta_v+\frac{2}{3C_{\mathbf{E}}}\,\eta, \tag{5.14}$$

which is the value of ξ_p in step (v). The estimates (5.7) with the new bounds defined in (5.10) are now an immediate consequence of the coarsening step in (v) and the triangle inequality. This concludes the confirmation of all estimates stated in RES_{SCPW}.

It remains to initialize the scheme $\text{SOLVE}_{\text{SCPW}}$. Again we assume that $\boldsymbol{f}$ and $\mathbf{y}_Z$ are given and fully accessible. Choosing $\overline{\mathbf{u}}^0\equiv\mathbf{0}$ we infer from (3.38), (3.34), and (3.39) that

$$\begin{aligned}\|\overline{\mathbf{u}}^0-\mathbf{u}\|&\le c_{\mathbf{Q}}^{-1}\|\mathbf{Q}\overline{\mathbf{w}}^0-\mathbf{g}\|=c_{\mathbf{Q}}^{-1}\|\mathbf{g}\|\\&=c_{\mathbf{Q}}^{-1}\|\mathbf{D}_Q^{-1}\mathbf{E}^T\mathbf{A}^{-T}\mathbf{T}_Z^T(\mathbf{y}_Z-\mathbf{T}_Z\mathbf{A}^{-1}\boldsymbol{f})\|\\&\le\frac{C_{\mathbf{E}}C_{\mathbf{T}}}{c_{\mathbf{A}}}\left(\|\mathbf{y}_Z\|+\frac{C_{\mathbf{T}}}{c_{\mathbf{A}}}\|\boldsymbol{f}\|\right)\\&=:\varepsilon_0.\end{aligned} \tag{5.15}$$

Moreover, for $\tilde{\mathbf{y}}^0:=\mathbf{0}$ one has

$$\begin{aligned}\|\mathbf{y}-\tilde{\mathbf{y}}^0\|=\|\mathbf{A}^{-1}(\boldsymbol{f}+\mathbf{E}\mathbf{D}_Q^{-1}\mathbf{u})\|&\le c_{\mathbf{A}}^{-1}\left(\|\boldsymbol{f}\|+C_{\mathbf{E}}c_{\mathbf{Q}}^{-1}\|\mathbf{g}\|\right)\\&\le c_{\mathbf{A}}^{-1}\left(\|\boldsymbol{f}\|+C_{\mathbf{E}}\varepsilon_0\right)=:\delta_{y,0}.\end{aligned} \tag{5.16}$$

Similarly, for $\tilde{\mathbf{p}}^0:=\mathbf{0}$ we obtain

$$\|\tilde{\mathbf{p}}^0-\mathbf{p}\|=\|\mathbf{A}^{-T}\mathbf{T}_Z^T(\mathbf{T}_Z\mathbf{y}-\mathbf{y}_Z)\|\le c_{\mathbf{A}}^{-1}(C_{\mathbf{T}}^2\delta_{y,0}+C_{\mathbf{T}}\|\mathbf{y}_Z\|)=:\delta_{p,0}. \tag{5.17}$$

The scheme $\text{SOLVE}_{\text{SCPW}}$ takes now the following form with the error reduction factor $\rho=\rho(\mathbf{Q})$ from (3.59) and K given by (4.7) with α from (3.58).

$\text{SOLVE}_{\text{SCPW}}[\varepsilon,\mathbf{Q},\mathbf{g}]\to(\overline{\mathbf{u}}_\varepsilon,\overline{\mathbf{y}}_\varepsilon,\overline{\mathbf{p}}_\varepsilon)$.

(i) Let $\overline{\mathbf{q}}^0:=\mathbf{0}$ and let ε_0 be given by (5.15). Moreover, let $\tilde{\mathbf{y}}:=\mathbf{0}$, $\tilde{\mathbf{p}}:=\mathbf{0}$ and set $j=0$. Finally, let $\delta_y:=\delta_{y,0},\delta_p:=\delta_{p,0}$ be defined by (5.16), (5.17), respectively.

(ii) If $\varepsilon_j\le\varepsilon$, stop and set $\overline{\mathbf{u}}_\varepsilon:=\overline{\mathbf{u}}^j$, $\overline{\mathbf{y}}_\varepsilon=\tilde{\mathbf{y}}$, $\overline{\mathbf{p}}_\varepsilon=\tilde{\mathbf{p}}$.
Otherwise set $\mathbf{v}^0:=\overline{\mathbf{u}}^j$.

(ii.1) For $k=0,\dots,K-1$, compute
$\text{RES}_{\text{SCPW}}[\rho^k\varepsilon_j,\mathbf{Q},\mathbf{g},\tilde{\mathbf{y}},\delta_y,\tilde{\mathbf{p}},\delta_p,\mathbf{v}^k,\delta_k]\to(\mathbf{r}^k,\tilde{\mathbf{y}},\delta_y,\tilde{\mathbf{p}},\delta_p)$,
where $\delta_0:=\varepsilon_j$ and $\delta_k:=\rho^{k-1}(\alpha k+\rho)\varepsilon_j$;
set

$$\mathbf{v}^{k+1}:=\mathbf{v}^k+\alpha\mathbf{r}^k. \tag{5.18}$$

(ii.2) Apply $\text{COARSE}[\frac{a\varepsilon_j}{2(1+a)},\mathbf{v}^K]\to\overline{\mathbf{u}}^{j+1}$; set $\varepsilon_{j+1}:=\frac{1}{2}\varepsilon_j$, $j+1\to j$ and go to (ii).

(The particular choice of the interior tolerance δ_k in step (ii.1) is based on the estimate (4.10).) Since when overwriting $\tilde{\mathbf{y}}$, $\tilde{\mathbf{p}}$ at the last stage before termination of $\text{SOLVE}_{\text{SCPW}}$ one has $\delta_k \le \varepsilon$, $\eta \le \varepsilon$, the following fact is an immediate consequence of (5.10).

Remark 5.1. The outputs $\overline{\mathbf{y}}_\varepsilon$ and $\overline{\mathbf{p}}_\varepsilon$ produced by $\text{SOLVE}_{\text{SCPW}}$ in addition to $\overline{\mathbf{u}}_\varepsilon$ are approximations to the exact solutions $\mathbf{y}, \mathbf{p}$ of (EE) satisfying

$$\|\mathbf{y} - \overline{\mathbf{y}}_\varepsilon\| \le (1+a)\varepsilon \left(\frac{C_{\mathbf{E}}}{c_{\mathbf{A}}} + \frac{c_{\mathbf{A}}}{3C_{\mathbf{E}}C_{\mathbf{T}}^2} \right), \tag{5.19}$$

$$\|\mathbf{p} - \overline{\mathbf{p}}_\varepsilon\| \le (1+a)\varepsilon \left(\frac{C_{\mathbf{T}}^2 C_{\mathbf{E}}}{c_{\mathbf{A}}^2} + \frac{2}{3C_{\mathbf{E}}} \right). \tag{5.20}$$

6. The complexity of $\text{SOLVE}_{\text{SCPW}}$. In view of the definition of RES_{PRM} and RES_{ADJ} entering RES_{SCPW}, the scheme $\text{SOLVE}_{\text{SCPW}}$ ultimately hinges on the availability of suitable schemes APPLY for the operators $\mathbf{L} \in \{\mathbf{A}, \mathbf{A}^T, \mathbf{T}_Z, \mathbf{T}_Z^T, \mathbf{E}, \mathbf{E}^T\}$. We shall adhere to our strategy of narrowing down step by step the requirements on our algorithmic ingredients, and we first identify conditions on the APPLY schemes that ensure τ^*-sparsity of RES_{SCPW} as formulated in section 4.2. It is no surprise that the key requirement is that the approximate application of each of these operators has a work/accuracy rate that is within some range comparable to best N-term approximation. Precisely, we say that $\text{APPLY}[\cdot, \mathbf{L}, \cdot]$ is τ^*-*efficient* for some $0 < \tau^* < 2$ if for any finitely supported $\mathbf{v} \in \ell_\tau^w$, for $0 < \tau^* < \tau < 2$, the output $\mathbf{w}_\eta$ of $\text{APPLY}[\eta, \mathbf{L}, \mathbf{v}]$ satisfies

$$\|\mathbf{w}_\eta\|_{\ell_\tau^w} \lesssim \|\mathbf{v}\|_{\ell_\tau^w}, \quad \#\text{supp}\, \mathbf{w}_\eta \lesssim \eta^{-1/s} \|\mathbf{v}\|_{\ell_\tau^w}^{1/s} \quad \eta \to 0, \tag{6.1}$$

where the constants depend only on τ as $\tau \to \tau^*$ and where s is related to τ by (4.19). Moreover, the number of floating point operations needed to compute $\mathbf{w}_\eta$ is to remain proportional to $\#\text{supp}\, \mathbf{w}_\eta$.

One should note that the existence of a τ^*-efficient scheme for an operator $\mathbf{L}$ has the following important implication that follows immediately from Proposition 4.2.

Remark 6.1. If one can find a τ^*-efficient scheme for $\mathbf{L}$, then $\mathbf{L}$ is bounded on ℓ_τ^w for every $\tau > \tau^*$.

Proof. For convenience, the proof from [CDD1] is included here. For $\mathbf{v} \in \ell_\tau^w$ and $\eta > 0$ there exists a $\tilde{\mathbf{v}}$ with $\|\mathbf{v} - \tilde{\mathbf{v}}\| \le \eta/(2\|\mathbf{L}\|)$ and $\#\text{supp}\, \tilde{\mathbf{v}} \lesssim \eta^{-1/s} \|\mathbf{v}\|_{\ell_\tau^w}^{1/s}$. Now by definition of $\mathbf{w}_\eta = \text{APPLY}\,[\eta/2, \mathbf{L}, \tilde{\mathbf{v}}]$ and τ^*-efficiency of $\mathbf{L}$ (6.1), one has for $\tau > \tau^*$, the estimate $\|\mathbf{L}\tilde{\mathbf{v}} - \mathbf{w}_\eta\| \le \eta/2$ while $\#\text{supp}\, \mathbf{w}_\eta \lesssim \eta^{-1/s} \|\tilde{\mathbf{v}}\|_{\ell_\tau^w}^{1/s} \le \eta^{-1/s} \|\mathbf{v}\|_{\ell_\tau^w}^{1/s}$. Since $\|\mathbf{L}\mathbf{v} - \mathbf{w}_\eta\| \le \eta$, we have identified a vector $\mathbf{w}_\eta$ with support $\lesssim \eta^{-1/s}$ that approximates $\mathbf{L}\mathbf{v}$ within accuracy η. Hence we can invoke Proposition 4.2 to conclude that $\|\mathbf{L}\mathbf{v}\|_{\ell_\tau^w} \lesssim \|\mathbf{v}\|_{\ell_\tau^w}$ as claimed. ∎

PROPOSITION 6.2. *If the* APPLY *schemes in* RES_{PRM} *and* RES_{ADJ} *are* τ^**-efficient for some* $\tau^* < 2$*, then* RES_{SCPW} *is* τ^**-sparse whenever there exists a constant* C *such that*

$$C\eta \ge \delta_v, \tag{6.2}$$

$$\max\{\|\tilde{\mathbf{y}}\|_{\ell_\tau^w}, \|\tilde{\mathbf{p}}\|_{\ell_\tau^w}, \|\mathbf{v}\|_{\ell_\tau^w}\} \le C\left(\|\mathbf{y}\|_{\ell_\tau^w} + \|\mathbf{p}\|_{\ell_\tau^w} + \|\mathbf{u}\|_{\ell_\tau^w}\right), \tag{6.3}$$

where $\mathbf{v}$ *is the current finitely supported input and where* $\tilde{\mathbf{y}}, \tilde{\mathbf{p}}$ *are the initial guesses for the exact solution components* $(\mathbf{y}, \mathbf{p})$.

Proof. Since $\text{SOLVE}_{\text{SCPW}}$ actually determines an approximation to the full triple $(\mathbf{y}, \mathbf{p}, \mathbf{u})$, the notion of τ^*-sparseness of RES_{SCPW} refers to properties of the whole triple. Thus, we have to assume that each of the solution components belongs to ℓ_τ^w for some $\tau > \tau^*$. By Remark 6.1 and our hypothesis on τ^*-efficiency, each $\mathbf{L} \in \{\mathbf{A}, \mathbf{A}^T, \mathbf{T}_Z, \mathbf{T}_Z^T, \mathbf{E}\mathbf{D}_Q^{-1}, \mathbf{D}_Q^{-1}\mathbf{E}^T\}$ is bounded on ℓ_τ^w for $\tau > \tau^*$. Thus, for the first system in (EE′) this implies

$$\|\boldsymbol{f}\|_{\ell_\tau^w} \lesssim \|\mathbf{y}\|_{\ell_\tau^w} + \|\mathbf{u}\|_{\ell_\tau^w}. \tag{6.4}$$

Likewise we have

$$\|\mathbf{T}_Z^T \mathbf{y}_Z\|_{\ell_\tau^w} \lesssim \|\mathbf{p}\|_{\ell_\tau^w} + \|\mathbf{y}\|_{\ell_\tau^w}. \tag{6.5}$$

Now, by the assumption (6.2), the quotients δ_v/η, δ_y/η, and δ_p/η are bounded. Therefore, according to step (i) in RES_{SCPW}, the scheme $\text{SOLVE}_{\text{PRM}}$ will invoke only a uniformly bounded finite number of iteration blocks (ii) with corresponding residual approximations RES_{PRM}. From the τ^*-efficiency of $\mathbf{A}$ and $\mathbf{E}\mathbf{D}_Q^{-1}$ and Remark 6.1, we infer that

$$\|\mathbf{y}_\eta\|_{\ell_\tau^w} \lesssim \|\boldsymbol{f}\|_{\ell_\tau^w} + \|\mathbf{v}\|_{\ell_\tau^w} + \|\tilde{\mathbf{y}}\|_{\ell_\tau^w} \lesssim \|\mathbf{y}\|_{\ell_\tau^w} + \|\mathbf{u}\|_{\ell_\tau^w} + \|\mathbf{v}\|_{\ell_\tau^w} + \|\tilde{\mathbf{y}}\|_{\ell_\tau^w}, \tag{6.6}$$

where we used (6.4). Likewise one concludes that the output $\mathbf{p}_\eta$ of step (ii) of RES_{SCPW} satisfies

$$\begin{aligned} \|\mathbf{p}_\eta\|_{\ell_\tau^w} &\lesssim \|\tilde{\mathbf{p}}\|_{\ell_\tau^w} + \|\mathbf{T}_Z^T \mathbf{y}_Z\|_{\ell_\tau^w} + \|\tilde{\mathbf{y}}\|_{\ell_\tau^w} \\ &\lesssim \|\tilde{\mathbf{p}}\|_{\ell_\tau^w} + \|\tilde{\mathbf{y}}\|_{\ell_\tau^w} + \|\mathbf{p}\|_{\ell_\tau^w} + \|\mathbf{y}\|_{\ell_\tau^w}, \end{aligned} \tag{6.7}$$

where we used (6.5) in the last step. (4.21) follows now from (6.6), (6.7), and (6.3). The second part of (4.21) and (ii) of the τ^*-sparseness of RES_{SCPW} can be concluded from τ^*-efficiency of the APPLY schemes in RES_{PRM} and RES_{ADJ}. This confirms the claim. ∎

THEOREM 6.3. *Assume that the* APPLY *schemes appearing in* RES_{PRM} *and* RES_{ADJ} *are* τ^**-efficient for some* $\tau^* < 2$ *and that the components of the solution* $(\mathbf{y}, \mathbf{p}, \mathbf{u})$ *of* (EE) *all belong to the respective space* ℓ_τ^w *for some* $\tau > \tau^*$. *Then the approximate solutions* $\mathbf{y}_\varepsilon, \mathbf{p}_\varepsilon, \mathbf{u}_\varepsilon$, *produced by* $\text{SOLVE}_{\text{SCPW}}$ *for any target accuracy* ε, *satisfy*

$$\|\mathbf{y}_\varepsilon\|_{\ell_\tau^w} + \|\mathbf{p}_\varepsilon\|_{\ell_\tau^w} + \|\mathbf{u}_\varepsilon\|_{\ell_\tau^w} \lesssim \|\mathbf{y}\|_{\ell_\tau^w} + \|\mathbf{p}\|_{\ell_\tau^w} + \|\mathbf{u}\|_{\ell_\tau^w} \tag{6.8}$$

and

$$(\#\operatorname{supp} \mathbf{y}_\varepsilon) + (\#\operatorname{supp} \mathbf{p}_\varepsilon) + (\#\operatorname{supp} \mathbf{u}_\varepsilon) \lesssim \left(\|\mathbf{y}\|_{\ell_\tau^w}^{1/s} + \|\mathbf{p}\|_{\ell_\tau^w}^{1/s} + \|\mathbf{u}\|_{\ell_\tau^w}^{1/s}\right) \varepsilon^{-1/s}, \tag{6.9}$$

where the constants depend only on τ *when* τ *approaches* τ^*. *Moreover, the number of floating point operations required during the execution of* $\text{SOLVE}_{\text{SCPW}}$ *remains proportional to the right-hand side of* (6.9).

Proof. According to Theorem 4.3, it remains to show that at each stage when RES_{SCPW} is called in step (ii.1) of $\text{SOLVE}_{\text{SCPW}}$, the hypotheses (6.2) and (6.3) in Proposition 6.2 are satisfied for some fixed constant C. The claim follows then from Theorem 4.3.

The validity of (6.2) is a consequence of the bounds (5.10) for the initial guesses, the values of η and δ_k in the kth perturbed iteration of the $(j+1)$st call of step (ii.1) of

$\text{SOLVE}_{\text{SCPW}}$, and the initialization bounds (5.15), (5.16), and (5.17). By the coarsening Lemma 4.4 and the coarsening in step (v) of RES_{SCPW}, we know that

$$\|\tilde{\mathbf{y}}\|_{\ell_\tau^w} \lesssim \|\mathbf{y}\|_{\ell_\tau^w}, \quad \|\tilde{\mathbf{p}}\|_{\ell_\tau^w} \lesssim \|\mathbf{p}\|_{\ell_\tau^w}. \tag{6.10}$$

Moreover, since in the $(j+1)$st call of step (ii) in $\text{SOLVE}_{\text{SCPW}}$ $\mathbf{v}^K$ satisfies $\|\mathbf{u}-\mathbf{v}^K\| \le \varepsilon_j/2(1+a)$ (see [CDD3] or (4.10)), we conclude from step (ii.2) in $\text{SOLVE}_{\text{SCPW}}$ and Lemma 4.4 that

$$\|\overline{\mathbf{u}}^j\|_{\ell_\tau^w} \lesssim \|\mathbf{u}\|_{\ell_\tau^w}, \quad j \in \mathbb{N}_0. \tag{6.11}$$

Combining (6.10) and (6.11) confirms the validity of (6.3). ∎

Thus the practical realization of $\text{SOLVE}_{\text{SCPW}}$ providing optimal work/accuracy rates for a possibly large range of decay rates of the error of best N-term approximation hinges on the availability of τ^*-efficient APPLY schemes with possibly small τ^* for the involved operators.

Distributed control. In this regard we discuss first the example in section 2.2.1 for *natural norms*, i.e., $Z = H_0^1(\Omega)$ and $U = Y' = Q = H^{-1}(\Omega)$. In this case, one has $\mathbf{E} = \mathbf{T} = \mathbf{D}_Z = \mathbf{D}_Q = \mathbf{I}$ and $\mathbf{A} = \mathbf{A}^T$. Since the identity mapping is τ^*-efficient for any $\tau^* < 2$, we only have to discuss the τ^*-efficiency of $\mathbf{A}$ defined by (3.20). The fact that one can indeed devise efficient schemes for the approximate application of wavelet representations of a wide class of operators, including differential operators, is a consequence of the cancellation properties (3.3) of wavelets together with the norm equivalences (3.5) for the relevant function spaces. In fact, such representations turn out to be *quasi-sparse* in the following sense. Recall that a matrix $\mathbf{A}$ is called *s^*-compressible* if for any $0 < s < s^*$ there exists a matrix $\mathbf{A}_j$ with at most $\le \alpha_j 2^j$ nonzero entries per row and column such that

$$\|\mathbf{A} - \mathbf{A}_j\| \le \alpha_j 2^{-sj}, \qquad j \in \mathbb{N}_0, \tag{6.12}$$

where $\{\alpha_j\}_{j\in\mathbb{N}_0}$ is any summable sequence.

Denote for a finitely supported vector $\mathbf{v}$ its best 2^j-approximations (given by the 2^j largest wavelet coefficients) by $\mathbf{v}_{[j]} := \mathbf{v}_{2^j}$. Following [CDD1], the expansion

$$\mathbf{w}_j := \mathbf{A}_j\mathbf{v}_{[0]} + \mathbf{A}_{j-1}(\mathbf{v}_{[1]} - \mathbf{v}_{[0]}) + \cdots + \mathbf{A}_0(\mathbf{v}_{[j]} - \mathbf{v}_{[j-1]}) \tag{6.13}$$

approximates $\mathbf{A}\mathbf{v}$. In fact, combining the a priori knowledge (6.12) with the a posteriori information $\|\mathbf{v}_{[k]} - \mathbf{v}_{[k-1]}\|$, one can see that for any finitely supported input $\mathbf{v}$ the error $\|\mathbf{A}\mathbf{v} - \mathbf{w}_j\|$ tends to zero when j grows. Thus, given a tolerance $\eta > 0$, one chooses the smallest j so that the bound for $\|\mathbf{A}\mathbf{v} - \mathbf{w}_j\|$ is less than or equal to η. This leads to a concrete scheme with the following properties.

$\text{APPLY}\,[\eta, \mathbf{A}, \mathbf{v}] \to \mathbf{w}_\eta$ computes for a given tolerance $\eta > 0$ a finitely supported sequence $\mathbf{w}_\eta$ satisfying

$$\|\mathbf{A}\mathbf{v} - \mathbf{w}_\eta\| \le \eta. \tag{6.14}$$

A detailed description and analysis of this routine can be found in [CDD1]. Its implementation has been discussed in [BCDU]. The following essential complexity estimate is taken from [CDD1].

THEOREM 6.4. *If $\mathbf{A}$ is s^*-compressible, then $\mathbf{A}$ is bounded on ℓ_τ^w for $s < s^*$, where τ and s are related by* (4.19), $\frac{1}{\tau} = s + \frac{1}{2}$. *Moreover, for a finitely supported vector $\mathbf{v}$ the output $\mathbf{w}_\eta$ of* $\text{APPLY}\,[\eta, \mathbf{A}, \mathbf{v}]$ *satisfies*

$$\|\mathbf{w}_\eta\|_{\ell_\tau^w} \lesssim \|\mathbf{v}\|_{\ell_\tau^w}, \qquad \#\operatorname{supp}\mathbf{w}_\eta,\ \#\text{flops} \lesssim \eta^{-1/s}\|\mathbf{v}\|_{\ell_\tau^w}^{1/s}. \tag{6.15}$$

Thus, the above scheme APPLY is τ^*-efficient for $\tau^* = (s^* + 1/2)^{-1}$ whenever $\mathbf{A}$ is s^*-compressible. It is known that s^* is larger the higher the regularity and the order of cancellation properties of the wavelets are for all the differential operators considered in section 2. Bounds for s^* in terms of these quantities for families of spline wavelets can be found, e.g., in [BCDU]. Hence, Theorem 6.3 ensures asymptotically optimal complexity bounds in the range $\tau > \tau^*$, i.e., the scheme $\text{SOLVE}_{\text{SCPW}}$ recovers rates of the error of best N-term approximation of order N^{-s} for $s < s^*$.

Now consider the same example but with a strictly larger space $Z \supset Y$ and a strictly smaller space $U \subset Q = Y'$. While one still has $\mathbf{E} = \mathbf{T} = \mathbf{I}$, the matrices $\mathbf{D}_Z, \mathbf{D}_Q$ are nontrivial scalings of the form $\mathbf{D}_Z = \mathbf{D}^\gamma$, i.e., $(\mathbf{D}_Z)_{\lambda,\nu} = 2^{\gamma|\lambda|}\delta_{\lambda,\nu}$, and $\mathbf{D}_Q = \mathbf{D}^\beta$ for some positive numbers $\gamma, \beta > 0$. The system (EE′) then takes the form

$$\begin{aligned} \mathbf{A}\mathbf{y} &= \boldsymbol{f} + \mathbf{D}^{-\beta}\mathbf{u}, \\ \mathbf{A}^T\mathbf{p} &= -\mathbf{D}^{-2\gamma}(\mathbf{y} - \mathbf{y}_*), \\ \omega\mathbf{u} &= \mathbf{D}^{-\beta}\mathbf{p}. \end{aligned} \tag{6.16}$$

First, the scaling smoothes the right-hand sides of the first two equations and favors their sparsity. Since it is important to understand the effect of such scalings in the present context, in particular in connection with the application of the operators $\mathbf{T}, \mathbf{E}$, we shall next make this point more precise. It is perhaps instructive to relate this first briefly to regularity issues. To this end, suppose that Ψ_H is a Riesz basis for a Sobolev space $H = H^t$ with respect to a bounded domain in $\mathbb{R}^d$. Besov spaces $B_p^{t+\gamma}(L_p)$, measuring smoothness in L_p, are then known, by the Sobolev embedding theorem, to be embedded in H^t whenever $p^{-1} \le \gamma/d + 1/2$. The smaller p is, the larger is $B_p^{t+\gamma}(L_p)$, which admits more singularities compared with the space $H^{t+\gamma}$ of the same smoothness level but measured in L_2. For a certain range of smoothness indices γ, depending on the regularity of the wavelets, an equivalent seminorm for $B_p^{t+\gamma}(L_p)$ can be given as a weighted sequence norm of wavelet coefficients (see, e.g., [D2, DeV]),

$$|v|^p_{B_p^{t+\gamma}(L_p)} := \sum_{\lambda\in\mathbb{I}} 2^{|\lambda| p d\left(\frac{\gamma}{d} + \frac{1}{2} - \frac{1}{p}\right)} |v_\lambda|^p. \tag{6.17}$$

Note that the scaling for t is already incorporated in the normalization of the wavelet basis. Specifically, when $\tau^{-1} = \gamma/d + 1/2$, the Besov norm $\|\cdot\|_{B_\tau^{t+\gamma}(L_\tau)}$ is equivalent to the ℓ_τ-norm of the wavelet coefficients. Therefore, the membership to such a Besov space on the Sobolev embedding line is, in view of (4.18), *almost* equivalent to saying that the sequence of wavelet coefficients belongs to the corresponding weak ℓ_τ-space and hence has a best N-term approximation rate of $N^{-\gamma/d}$; see Proposition 4.2. This means that nonlinear approximation, like best N-term approximation, compensates for the loss of regularity between $H^{t+\gamma}$ and $B_\tau^{t+\gamma}(L_\tau)$ in that the same approximation rate (in terms of degrees of freedoms) is preserved for $B_\tau^{t+\gamma}(L_\tau)$, that would be achievable in the smaller space $H^{t+\gamma}$ with the aid of uniform mesh refinements. It immediately follows now from (6.17) that

$$v = \mathbf{v}^T\Psi_H \in B_\tau^{t+\gamma}(L_\tau) \quad \text{implies} \quad w = (\mathbf{D}^{-\beta}\mathbf{v})^T\Psi_H \in B_\tau^{t+\gamma+\beta}(L_\tau) \subset B_{\tau'}^{t+\gamma}(L_{\tau'}), \tag{6.18}$$

where τ' is the critical index for the higher smoothness level $\gamma + \beta$.

$$\frac{1}{\tau'} = \frac{\gamma+\beta}{d} + \frac{1}{2}. \tag{6.19}$$

Note that τ' and τ are related by $\frac{1}{\tau'} - \frac{1}{\tau} = \frac{\beta}{d}$. The next observation says that scaling by $\mathbf{D}^{-\beta}$ effects exactly the same shift also between *weak* ℓ_τ-spaces that *precisely* characterize best N-term approximation whose order therefore grows, on account of Proposition 4.2 by β/d.

PROPOSITION 6.5. *One has that*

$$\mathbf{p} \in \ell^w_\tau \quad \textit{implies} \quad \mathbf{D}^{-\beta}\mathbf{p} \in \ell^w_{\tau'}, \quad \textit{where} \quad \frac{1}{\tau'} := \frac{1}{\tau} + \frac{\beta}{d}. \tag{6.20}$$

Moreover, this result is sharp in the sense that for no $\tau'' < \tau'$ there holds $\mathbf{D}^{-\beta}\mathbf{p} \in \ell^w_{\tau''}$ for all $\mathbf{p} \in \ell^w_\tau$.

Proof. Let $C > 1$ be a fixed constant that later will be chosen at our convenience. Let $\mathcal{P}$ be the class of those $\mathbf{p} \in \ell^w_\tau$ with $\|\mathbf{p}\|_{\ell^w_\tau} \le C^{1/\tau}$ and let $\tilde{\mathbf{p}} := \mathbf{D}^{-\beta}\mathbf{p}$. Consider the set

$$\tilde{\Lambda}_{(J)} := \{\lambda : |\tilde{p}_\lambda| \ge \eta_J\}, \qquad \text{where } \eta_J := 2^{-J\tau/\tau'}.$$

In view of the definition of $\ell^w_{\tau'}$ in (4.15), we have to show that the cardinality of $\tilde{\Lambda}_{(J)}$ increases at most like $2^{J\tau}$. (Standard arguments imply then that $\#\{\lambda : |\tilde{p}_\lambda| \ge \eta\} \lesssim \eta^{-\tilde{\tau}}$ for any $\eta \le 1$.) To this end, we determine first which of the sets

$$\Lambda_j := \{\lambda : 2^{-j} \le |p_\lambda| < 2^{-j+1}\}$$

is always fully contained in $\tilde{\Lambda}_{(J)}$. We know from [CDD1] that $\#\Lambda_j \le C2^{j\tau}$. Without loss of generality we may assume that $\#\Lambda_j = C2^{j\tau}$ to cover the largest possible sets. Since the entries of $\tilde{\mathbf{p}}$ arise from those of $\mathbf{p}$ by scaling with the weights $2^{-\beta|\lambda|} \le 1$, the smaller the levels $|\lambda|$ in these weights, the better the chance for $\lambda \in \Lambda_j$ to belong to $\tilde{\Lambda}_{(J)}$. Thus, to ensure that for any $\mathbf{p} \in \mathcal{P}$ the set Λ_j is contained in $\tilde{\Lambda}_{(J)}$, we must be able to find $C2^{j\tau}$ indices λ with possibly small $|\lambda|$ such that $2^{-\beta|\lambda|}2^{-j} \ge \eta_J$. This count clearly involves the spatial dimension d since $c2^{jd}$ indices λ of level $|\lambda| = j$ can occur. Here the constant c depends on the spatial dimension of the functions whose wavelet coefficients are considered. Thus, the smallest possible maximum level L_j of these indices is therefore determined by $c2^{dL_j} = C2^{j\tau}$, i.e., $L_j = j\tau/d + \left(\log_2 \frac{C}{c}\right)/d$. Assume for convenience that $C/c \ge 1$. Then, for $\lambda \in \Lambda_j$ we conclude

$$|\tilde{p}_\lambda| \ge 2^{-\beta|\lambda|}2^{-j} \ge 2^{-(\beta L_j + j)} = 2^{-j\left(\frac{d+\beta\tau}{d}\right)}(C/c)^{\beta/d} \ge 2^{-j\tau/\tau'}. \tag{6.21}$$

Thus, $\Lambda_j \subset \tilde{\Lambda}_{(J)}$ for $j \le J$. On the other hand, since for $\lambda \in \Lambda_j$ one also has $|\tilde{p}_\lambda| \le 2^{-\beta|\lambda|-j+1}$, not all the indices in Λ_j can always be contained in $\tilde{\Lambda}_{(J)}$ for $j \ge J$ and any choice of $\mathbf{p} \in \mathcal{P}$. To determine the maximum number of $\lambda \in \Lambda_{J+k}$ for $k \in \mathbb{N}$ that can belong to $\tilde{\Lambda}_{(J)}$ for any $\mathbf{p} \in \mathcal{P}$, we must have in view of (6.21) $|\lambda| \le \ell_k$, where $\beta\ell_k + J + k \le J\tau/\tau'$. Using that $(\tau/\tau') - 1 = \beta\tau/d$, straightforward calculations yield $\ell_k \le \frac{J\tau}{d} - \frac{k}{\beta}$. Thus, we can assign at most $2^{d\ell_k}$ scaling weights to Λ_j to keep $|\tilde{p}_\lambda| \ge \eta_J$ for that many $\lambda \in \Lambda_{J+k}$. Moreover, the set Λ_{J+k} is disjoint from $\tilde{\Lambda}_{(J)}$ whenever $2^{-(J+k)+1} \le \eta_J$, which is the case when $k \ge 1 + J\tau\beta/d$. Hence, we have

$$\sum_{k\in\mathbb{N}} \#(\Lambda_{J+k} \cap \tilde{\Lambda}_{(J)}) \le \sum_{k=1}^{J\tau\beta/d} 2^{d\ell_k} = \sum_{k=1}^{J\tau\beta/d} 2^{\frac{d}{\beta}\left(\frac{J\tau\beta}{d} - k\right)} \lesssim 2^{J\tau}. \tag{6.22}$$

Since $\mathbb{I} = \bigcup_{j\ge 0} \Lambda_j$ and $\sum_{j\le J} \#\Lambda_j \lesssim 2^{J\tau}$, we conclude from (6.22) that $\#\tilde{\Lambda}_{(J)} \lesssim 2^{J\tau} = \eta^{-\tau'}$ for $\eta = 2^{-J\tau/\tau'}$. This confirms (6.20).

To verify the rest of the assertion, consider $\mathbf{p}$, whose decreasing rearrangement is given by $p_n^* = n^{-1/\tau}$ while $\tilde{p}_n^* = 2^{-\beta(j+1)} n^{-1/\tau}$ for $2^{dj} < n \le 2^{d(j+1)}$. Then

$$\begin{aligned}\sigma_{2^{dj}}(\tilde{\mathbf{p}})^2 &= \sum_{n>2^{dj}} (\tilde{p}_n^*)^2 \gtrsim \sum_{j=J}^{\infty} \sum_{2^{jd}<n\le 2^{d(j+1)}} n^{-2/\tau} 2^{-2\beta(j+1)} \\ &\gtrsim \sum_{j=J}^{\infty} 2^{-2dj\left(\frac{1}{\tau}+\frac{\beta}{d}-\frac{1}{2}\right)} \gtrsim \left(2^{-dJs'}\right)^2,\end{aligned}$$

where $s' = \frac{1}{\tau'} - \frac{1}{2}$. Thus, by Proposition 4.2, $\mathbf{p} \notin \ell_{\tau''}^w$ for any $\tau'' < \tau'$, which finishes the proof. □

Therefore, whatever the sparsity class of the adjoint variable $\mathbf{p}$, the third equation in (6.16) says, in view of Proposition 6.5, that the control $\mathbf{u}$ is even sparser. Thus, although the control $\mathbf{u}$ may be accurately recovered with relatively few degrees of freedom, the overall solution complexity in the above case is bounded from below by the (perhaps) less sparse auxiliary variable $\mathbf{p}$.

As a possible remedy one might think of introducing the variable $\tilde{\mathbf{p}} = \mathbf{D}^{-\beta}\mathbf{p}$ (to replace $\mathbf{p}$ by a sparser variable) and rewrite the second system in (6.16) as

$$\tilde{\mathbf{A}}^T \tilde{\mathbf{p}} = \mathbf{D}^{-\beta}\mathbf{D}^{-2\gamma}(\mathbf{y}_* - \mathbf{y}), \quad \tilde{\mathbf{A}} := \mathbf{D}^{\beta}\mathbf{A}\mathbf{D}^{-\beta}.$$

To apply our complexity analysis we could assume now that $\tilde{\mathbf{p}}$ has a certain sparsity which would then naturally be the same as the sparsity of $\mathbf{u}$. But although the matrix $\tilde{\mathbf{A}}$ has still the same spectrum as $\mathbf{A}$ and hence is an ℓ_2-automorphism for which the gradient iteration would still converge, it is presumably less compressible. In fact, the unsymmetric scaling means that we only have an estimate of the form

$$|(\tilde{\mathbf{A}})_{\lambda,\nu}| \lesssim 2^{\beta(|\nu|-|\lambda|)} 2^{-\sigma||\lambda|-|\nu||},$$

where σ results from the regularity of the wavelets and determines the original compressibility of $\mathbf{A}$.

So, in summary, in the case of a distributed control the solution complexity is determined not by the sparseness of the control but by that of the remaining variables in (EE′).

Boundary control. The situation is different for the example from section 2.2.3. Recall that in this case $Y = H^1(\Omega)$, $Q = (H^{1/2}(\Gamma_c))'$ so that $E : (H^{1/2}(\Gamma_c))' \to (H^1(\Omega))'$ is the extension operator defined by (2.16). Choosing $Z = H^{1-\gamma}(\Omega)$ for $0 \le \gamma \le 1$ as the observation space, T is the canonical injection and

$$\mathbf{T} = \mathbf{I}, \qquad \mathbf{T}_Z = \mathbf{D}^{\gamma} = \mathbf{D}_Z. \tag{6.23}$$

Choosing bases $\Psi_Q \subset Q = (H^{1/2}(\Gamma_c))'$, $\tilde{\Psi}_Q \subset Q' = H^{1/2}(\Gamma_c)$, and $\Psi_Y \subset Y = H^1(\Omega)$, $\tilde{\Psi}_Y \subset (H^1(\Omega))'$, (2.16), and (3.20) say that $\mathbf{E}$ is given by

$$\mathbf{E} = \langle \iota\Psi_Y, \tilde{\Psi}_Q \rangle, \quad \mathbf{E}^T = \langle \tilde{\Psi}_Q, \iota\Psi_Y \rangle. \tag{6.24}$$

Thus, the entries of $\mathbf{E}$ are inner products of traces of wavelets on Ω with wavelets on the control boundary Γ_c. Choosing $U = L_2(\Gamma_c)$, the Euler–Lagrange system (EE′) now reads

$$\begin{aligned}\mathbf{A}\mathbf{y} &= \boldsymbol{f} + \mathbf{E}\mathbf{D}^{-1/2}\mathbf{u}, \\ \mathbf{A}^T\mathbf{p} &= -\mathbf{D}^{-2\gamma}(\mathbf{y} - \mathbf{y}_*), \\ \omega\mathbf{u} &= \mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{p}.\end{aligned} \tag{6.25}$$

Recall that the sparsity of solutions of the first two systems in (6.25) is exploited by the compressibility of $\mathbf{A}$ up to the limiting index s^* which depends only on the cancellation properties and the regularity of Ψ_Y and not on the particular differential operator. In contrast, as shown in [M], $\mathbf{E}$ is τ^*-efficient only for $\tau^* \geq 1$, i.e., $\mathbf{E}^T$ is not bounded on ℓ_τ^w for $\tau < 1$. In other words, $\mathbf{E}^T$ is at most s^*-compressible for $s^* = 1/2$. This is because traces of wavelets are in general no longer wavelets, so this factor does not have any cancellation properties that help keep the entries of $\mathbf{E}$ small. Thus, in this case, even when $\mathbf{p}$ is highly sparse in the sense that $\mathbf{p}$ belongs to ℓ_τ^w for τ much smaller than 1, the application of $\mathbf{E}^T$ in the third equation of (6.25) reduces that sparsity when computing the control $\mathbf{u}$. However, the scaling $\mathbf{D}_Q^{-1} = \mathbf{D}^{-1/2}$ raises the order of compressibility by Proposition 6.5 to $s^* = 1$. This also can be seen directly because it enhances the decay of the entries of $\mathbf{E}^T$ along each row. Without the attenuation caused by the scaling, the latter decay is weak due to the lack of vanishing moments of the traces of domain wavelets.

7. Special cases and alternatives. We address in this section some special issues concerning alternative solution strategies as well as the mandatory case when retaining the exact reference objective functional (CP), respectively, (CPW). Several cases arise.

(I) Recall from Remark 3.5 that in the special situation (SC) with natural norms, (3.56) is equivalent to the equation

$$\mathbf{M}\mathbf{u} := (\mathbf{I} + \omega \mathbf{A}\mathbf{A}^T)\mathbf{u} = \mathbf{A}\mathbf{y}_* - \boldsymbol{f}. \tag{7.1}$$

A twofold application of the scheme APPLY to $\mathbf{A}$, $\mathbf{A}^T$ immediately yields a τ^*-efficient evaluation scheme for $\mathbf{M}$. Thus we can apply here the scheme SOLVE introduced in section 4.1 directly without any inner iterations. Moreover, the solution complexity is now completely governed by the compressibility of $\mathbf{A}$ and by the sparsity of the control $\mathbf{u}$.

In this case one can therefore still efficiently solve the exact reference problem (CPW) that involves the Riesz maps. This is illustrated best by the example (3.43). Using appropriate wavelet bases, $\mathbf{A}, \boldsymbol{\Delta}$ are ℓ_2-automorphisms, so that the application of SOLVE to $\mathbf{M}\mathbf{u} = \mathbf{A}\mathbf{y}_* - \boldsymbol{f}$ with $\mathbf{M} := (\mathbf{I} + \omega \mathbf{A}\boldsymbol{\Delta}^{-1}\mathbf{A}^T\boldsymbol{\Delta}^{-1})$ is feasible. The application of $\mathbf{M}$ can then be realized in a way similar to $\text{SOLVE}_{\text{SCPW}}$ through suitable inner iterations. An advantage is that now only systems involving the Laplacian have to be solved instead of possibly more complicated operator equations with variable coefficients.

(II) So far $\text{SOLVE}_{\text{SCPW}}$ has been confined to the problem (3.56), which is only a representer of (CP) in the ambiguous case. However, at least for the cases listed in Remark 3.8 (where the necessity for realizing a specific norm is more conceivable), we can extend $\text{SOLVE}_{\text{SCPW}}$ so as to apply it to the original reference formulation (CPW). In this case, one could solve (3.39) based on the system (EE) and Lemma 3.7. In view of (3.45) and (3.46) in the system (EE), this now requires the additional application of $\mathbf{R}_Z$ and $\mathbf{R}_U$ in the scheme $\text{SOLVE}_{\text{ADJ}}$. This is feasible for the cases listed in Remark 3.8. The scheme APPLY can then be applied with optimal complexity. This requires an evaluation of the Riesz operators within suitable dynamically updated tolerances which have to be taken into proper account. Since the reasoning is completely analogous to the previous case, we omit the details.

Nevertheless, the question remains how to extend the scheme $\text{SOLVE}_{\text{SCPW}}$ to scenarios where a specific cost functional should be kept although the norms do not fall

into the category described in Remark 3.1 and therefore usually do not have canonical representers. One could then strive for representers of (CP) that approximate the specific reference formulation in a quantitatively better way. This could be done by using better approximations to $\mathbf{R}_Z, \mathbf{R}_U$ which are still efficiently applicable. These applications have to be incorporated in $\text{SOLVE}_{\text{SCPW}}$ as explained above. In a nonadaptive context the effect of such strategies is investigated in [BK], where $\mathbf{R}_Z, \mathbf{R}_U$ are replaced by scaled L_2-mass matrices of the respective wavelet bases.

8. Concluding remarks. We have developed a class of fully adaptive schemes for the solution of optimal control problems with elliptic boundary problems as constraints. The approach is based on a gradient descent iteration for the corresponding full infinite dimensional variational problem in wavelet coordinates. The various possible ways of formulating this infinite dimensional variational problem through elimination of the state variable reveal preferences with regard to well-posedness that would perhaps not be so apparent when working from the beginning with a finite dimensional discretization. The numerical realization of the adaptive solution concepts relies on the adaptive application of the involved operators within stage-dependent dynamically updated tolerances. The complexity of such schemes is shown to hinge on the properties of these application routines. Concrete realizations of such schemes are exhibited in several simple cases. This sheds some light on the different inherent complexity properties of distributed versus boundary control problems also in connection with different choices of norms in the objective functional. We refer to [BK] for first numerical experiments with algorithms of the above form with uniform refinements where the influence of different norms as a modeling tool is explored.

We have not considered here so far the quantitative role of the regularization parameter ω in (2.3), (6.16), or (6.25). Its variation affects all scales simultaneously, while the diagonal scalings representing different smoothness norms treat high and low frequencies differently. The concepts presented here unfold their full potential when the problem context allows one to modify the cost functional within classes of equivalent norms. We have indicated the limitations when this is not the case, as well as possible remedies. This issue is also addressed in the experiments in [BK]. It turns out that when no regularization is needed, the scheme is robust when ω tends to zero and produces the correct solution. Corresponding numerical experiments for the above adaptive version will be presented and discussed in a forthcoming paper.

Acknowledgments. We are indebted to the referees for valuable comments and suggestions that helped us improve the presentation of the material. A comment in one of the reports motivated part of the material presented in section 7.

REFERENCES

[B] A. Barinka, *Fast Evaluation Tools for Adaptive Wavelet Schemes*, Ph.D. dissertation, RWTH Aachen, 2004.

[BCDU] A. Barinka, T. Barsch, Ph. Charton, A. Cohen, S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet schemes for elliptic problems—Implementation and numerical experiments*, SIAM J. Sci. Comput., 23 (2001), pp. 910–939.

[BKR] R. Becker, H. Kapp, and R. Rannacher, *Adaptive finite element methods for optimal control of partial differential equations: Basic concept*, SIAM J. Control Optim., 39 (2000), pp. 113–132.

[BBDM] T. Binder, L. Blank, W. Dahmen, and W. Marquardt, *Multiscale concepts for moving horizon optimization*, in Online Optimization for Large Scale Systems, M. Grötschel, S.O. Krumke, and J. Rambau, eds., Springer, Berlin, 2001, pp. 341–362.

[BK] C. Burstedde and A. Kunoth, *Wavelet methods for linear-quadratic control problems*, Preprint 127, SFB 611, Universität Bonn, December 2003, submitted for publication.

[CTU] C. Canuto, A. Tabacco, and K. Urban, *The wavelet element method, part* I: *Construction and analysis*, Appl. Comput. Harm. Anal., 6 (1999), pp. 1–52.

[CK] D. Castaño and A. Kunoth, *Multilevel regularization of wavelet based fitting of scattered data—Some experiments*, Numer. Algorithms, to appear.

[CDLL] A. Chambolle, R. DeVore, N. Lee, and B. Lucier, *Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage*, IEEE Trans. Image Proc., 7 (1998), pp. 319–333.

[Co] A. Cohen, *Numerical Analysis of Wavelet Methods*, Handbook of Numerical Analysis II, vol. 8, P.G. Ciarlet and J.L. Lions, eds., Elsevier Science, New York, 1998.

[CDD1] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations—Convergence rates*, Math. Comp., 70 (2001), pp. 27–75.

[CDD2] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods* II*—Beyond the elliptic case*, Found. Comput. Math., 2 (2002), pp. 203–245.

[CDD3] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet schemes for nonlinear variational problems*, SIAM J. Numer. Anal., 41 (2003), pp. 1785–1823.

[CM] A. Cohen and R. Masson, *Wavelet adaptive methods for second order elliptic problems, boundary conditions and domain decomposition*, Numer. Math., 86 (2000), pp. 193–238.

[DDU] S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet methods for saddle point problems—optimal convergence rates*, SIAM J. Numer. Anal., 40 (2002), pp. 1230–1262.

[D1] W. Dahmen, *Stability of multiscale transformations*, J. Fourier. Anal. Appl., 2 (1996), pp. 341–361.

[D2] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer. 6, Cambridge University Press, Cambridge, UK, 1997, pp. 55–228.

[D3] W. Dahmen, *Wavelet methods for PDEs—Some recent developments*, J. Comput. Appl. Math., 128 (2001), pp. 133–185.

[D4] W. Dahmen, *Multiscale and wavelet methods for operator equations*, in Multiscale Problems and Methods in Numerical Simulation, C. Canuto, ed., Lecture Notes in Math. 1825, Springer, Heidelberg, 2003, pp. 31–96.

[DKS] W. Dahmen, A. Kunoth, and R. Schneider, *Wavelet least square methods for boundary value problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1985–2013.

[DKU] W. Dahmen, A. Kunoth, and K. Urban, *Biorthogonal spline-wavelets on the interval—Stability and moment conditions*, Appl. Comput. Harm. Anal., 6 (1999), pp. 132–196.

[DS1] W. Dahmen and R. Schneider, *Composite wavelet bases for operator equations*, Math. Comp., 68 (1999), pp. 1533–1567.

[DS2] W. Dahmen and R. Schneider, *Wavelets on manifolds* I: *Construction and domain decomposition*, SIAM J. Math. Anal., 31 (1999), pp. 184–230.

[DSt] W. Dahmen and R. Stevenson, *Element-by-element construction of wavelets satisfying stability and moment conditions*, SIAM J. Numer. Anal., 37 (1999), pp. 319–352.

[DUV] W. Dahmen, K. Urban, and J. Vorloeper, *Adaptive wavelet methods—Basic concepts and applications to the Stokes problem*, in Wavelet Analysis, D.-X. Zhou, ed., World Scientific, Englewood Cliffs, NJ, 2002, pp. 39–80.

[DeV] R. DeVore, *Nonlinear Approximation*, Acta Numer. 7, Cambridge University Press, Cambridge, UK, 1998, pp. 51–150.

[K1] A. Kunoth, *Wavelet methods—Boundary value problems and control problems*, in Advances in Numerical Mathematics, Teubner, Leipzig, Stuttgart, 2001.

[K2] A. Kunoth, *Fast iterative solution of saddle point problems in optimal control based on wavelets*, Comput. Optim. Appl., 22 (2002), pp. 225–259.

[K3] A. Kunoth, *Adaptive wavelet schemes for an elliptic control problem with Dirichlet boundary control*, Numer. Algorithms, to appear.

[Li] J.L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, 1971.

[M] M. Mommer, *Fictitious Domain—Lagrange Multiplier Approach: Smoothness Analysis*, IGPM preprint, 2003.

[Z] E. Zeidler, *Nonlinear Functional Analysis and its Applications;* III: *Variational Methods and Optimization*, Springer, Berlin, 1985.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1676–1713

SENSITIVITY ANALYSIS USING ITÔ–MALLIAVIN CALCULUS AND MARTINGALES, AND APPLICATION TO STOCHASTIC OPTIMAL CONTROL*

EMMANUEL GOBET† AND RÉMI MUNOS†

Abstract. We consider a multidimensional diffusion process $(X_t^\alpha)_{0\le t\le T}$ whose dynamics depends on a parameter α. Our first purpose is to write as an expectation the sensitivity $\nabla_\alpha J(\alpha)$ for the expected cost $J(\alpha) = \mathbb{E}(f(X_T^\alpha))$, in order to evaluate it using Monte Carlo simulations. This issue arises, for example, from stochastic control problems (where the controller is parameterized, which reduces the control problem to a parametric optimization one) or from model misspecifications in finance. Previous evaluations of $\nabla_\alpha J(\alpha)$ using simulations were limited to smooth cost functions f or to diffusion coefficients not depending on α (see Yang and Kushner, *SIAM J. Control Optim.*, 29 (1991), pp. 1216–1249). In this paper, we cover the general case, deriving three new approaches to evaluate $\nabla_\alpha J(\alpha)$, which we call the *Malliavin calculus approach*, the *adjoint approach*, and the *martingale approach*. To accomplish this, we leverage Itô calculus, Malliavin calculus, and martingale arguments. In the second part of this work, we provide discretization procedures to simulate the relevant random variables; then we analyze their respective errors. This analysis proves that the discretization error is essentially linear with respect to the time step. This result, which was already known in some specific situations, appears to be true in this much wider context. Finally, we provide numerical experiments in random mechanics and finance and compare the different methods in terms of variance, complexity, computational time, and time discretization error.

Key words. sensitivity analysis, parameterized control, Malliavin calculus, weak approximation

AMS subject classifications. 90C31, 93E20, 60H30

DOI. 10.1137/S0363012902419059

1. Introduction.

We consider a d-dimensional stochastic differential equation (SDE) defined by

$$X_t = x + \int_0^t b(s, X_s, \alpha)\, ds + \sum_{j=1}^{q} \int_0^t \sigma_j(s, X_s, \alpha)\, dW_s^j, \tag{1.1}$$

where α is a parameter (taking values in $\mathcal{A} \subset \mathbb{R}^m$) and $(W_t)_{0\le t\le T}$ is a standard Brownian motion in $\mathbb{R}^q$ on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0\le t\le T}, \mathbb{P})$, with the usual assumptions on the filtration $(\mathcal{F}_t)_{0\le t\le T}$.

We first aim at evaluating the sensitivity w.r.t. α of the expected cost

$$J(\alpha) = \mathbb{E}\left(f(X_T)\right), \tag{1.2}$$

for a given terminal cost f and for a fixed time T. The sensitivity of more general functionals including instantaneous costs like $\mathbb{E}\big(\int_0^T g(t, X_t)dt + f(X_T)\big) = \int_0^T \mathbb{E}(g(t, X_t))dt + \mathbb{E}(f(X_T))$ will follow by discretizing the integral and applying the sensitivity estimator for each time.

This evaluation is a typical issue raised in various applications. A first example is the analysis of the impact on the expected cost $J(\alpha)$ of a misspecification of a stochastic model (defined by a SDE with coefficients $\bar{b}(t, x)$ and $(\bar{\sigma}_j(t, x))_{1\le j\le q}$). The issue

*Received by the editors December 3, 2002; accepted for publication (in revised form) July 5, 2004; published electronically March 11, 2005.
http://www.siam.org/journals/sicon/43-5/41905.html
†Centre de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau Cedex, France (emmanuel.gobet@polytechnique.fr, remi.munos@polytechnique.fr).

may be formulated by setting $b(t,x,\alpha) = \bar{b}(t,x) + \sum_{i=1}^{m} \alpha_i \phi_i(t,x)$ (and analogously for $(\sigma_j(t,x,\alpha))_{1\le j\le q}$), then computing the sensitivities at the point $\alpha = 0$. In finance, misspecifications in option pricing procedures usually concern the diffusion coefficients $(\bar{\sigma}_j(t,x))_{1\le j\le q}$ (the *volatility* of the assets). There are also some connections with the so-called model risk problem (see Cvitanić and Karatzas [CK99]).

Stochastic control is another field requiring sensitivity analysis. For instance, if the controlled SDE is defined by $dX_t = \bar{b}(t,X_t,u_t)\ dt + \sum_{j=1}^{q} \bar{\sigma}_j(t,X_t,u_t) dW_t^j$, the problem is to find the maximal value of $\mathbb{E}(f(X_T))$ among the admissible policies $(u_t)_{0\le t\le T}$. In low dimensions (say 1 or 2), numerical methods based on the dynamic programming principle can be successfully implemented (see Kushner and Dupuis [KD01] for some references), but they become inefficient in higher dimensions. Alternatively, one can use *policy search* algorithms (see [BB01] and references therein). It consists in seeking a *good* policy in a feedback form using a parametric representation, that is, $u_t = u(t,X_t,\alpha)$: in that case, one puts $b(t,x,\alpha) = \bar{b}(t,x,u(t,x,\alpha))$ and $\sigma_j(t,x,\alpha) = \bar{\sigma}_j(t,x,u(t,x,\alpha))$. The policy function $u(t,x,\alpha)$ can be parameterized through a linear approximation (a linear combination of basis functions) or through a nonlinear one (e.g., with neural networks, see Rumelhart and McClelland [RM86] or Haykin [Hay94] for general references). Then, one might use a standard parametric optimization procedure such as the stochastic gradient method or other stochastic approximation algorithms (see Polyak [Pol87]; Benveniste, Metivier, and Priouret [BMP90]; Kushner and Yin [KY97]), which require sensitivity estimations of $J(\alpha)$ w.r.t. α, such as $\nabla_\alpha J(\alpha)$. This gradient is the quantity we will focus on in this paper.

Since the setting is a priori multidimensional, we propose a Monte Carlo approach for the numerical computations. The evaluation of $J(\alpha)$ is standard and has been widely studied. For an introduction to numerical approximations of SDEs, we refer the reader to Kloeden and Platen [KP95], for instance. To our knowledge, there are three different approaches to compute $\nabla_\alpha J(\alpha)$ in our context:

1. The *resampling method* (see Glasserman and Yao [GY92], L'Ecuyer and Perron [LP94] for instance), which consists in computing different values of $J(\alpha)$ for some close values of the parameter α and then forming some appropriate differences to approximate the derivatives. However, not only is it costly when the dimension of the parameter α is large, but it also provides biased estimators.
2. The *pathwise method* (proposed in our context by Yang and Kushner [YK91]), which consists in putting the gradient inside the expectation, involving ∇f and $\nabla_\alpha X_T$. Then, $\nabla_\alpha J(\alpha)$ is expressed as an expectation (see Proposition 1.1 below) and Monte Carlo methods can be used. One limitation of this method is that the cost function f has to be smooth.
3. The so-called *likelihood method* or *score method* (introduced by Glynn [Gly86, Gly87], Reiman and Weiss [RW86]; see also Broadie and Glasserman [BG96] for applications to the computation of Greeks in finance), in which the gradient is rewritten as $\mathbb{E}(f(X_T)H)$ for some random variable H. There is no uniqueness in this representation, since we can add to H any random variables orthogonal to X_T. Unlike the pathwise method, this method is not limited to smooth cost functions. Usually, H is equal to $\nabla_\alpha(\log(p(\alpha,X_T)))$, where $p(\alpha,.)$ is the density w.r.t. the Lebesgue measure of the law of X_T. This has some strong limitations in our context since this quantity is generally unknown. However, Yang and Kushner [YK91] provide explicit weights H, under the restrictions that α concerns only b (and not σ_j) and that the diffusion coefficient is elliptic, using the Girsanov theorem (see Proposition 2.6).

A **first purpose** of this work is to handle more general situations where both coefficients defining the SDE (1.1) depend on α. To address this issue, we provide three new approaches to express the sensitivity of $J(\alpha)$ with respect to α.

1. The first one is an extension of the likelihood approach method to the case of diffusion coefficients depending on α. It uses a direct integration-by-parts formula of the Malliavin calculus. This idea has been used recently in a financial context in the paper by Fournié et al. [FLL$^+$99] to compute option prices' sensitivities. These techniques have also been used efficiently by the first author to derive asymptotic properties of statistical procedures when we estimate parameters defining a SDE (see [Gob01b, Gob02]). Actually, our true contribution concerns essentially a situation where ellipticity is replaced by a weaker (but standard) nondegeneracy condition, which addresses random mechanics problems or portfolio optimization problems in finance.
2. The second approach is rather different from previous methods. Indeed, we initially focus on the adjoint point of view (see Bensoussan [Ben88] or Peng [Pen90]) to finally derive new formulae, involving again some integration-by-parts formula, but written in a simple way (using only Itô's calculus). In stochastic control problems, adjoint processes are related to backward SDEs (see Yong and Zhou [YZ99], e.g.), and their simulation is an extremely difficult and costly task. Here, we circumvent this difficulty since we only need to express them as explicit conditional expectations, which is feasible.
3. The third approach follows from martingale arguments applied to the expected cost and leads to an original representation, which appears to be surprisingly simple.

To compare these new methods with the previous ones, we will measure in section 5, on the one hand, the variance of the random variables involved in the resulting formulae for $\nabla_\alpha J(\alpha)$, and on the other hand, the computational time. Surprisingly, the three methods that we propose behave similarly in terms of variance, but the most efficient in terms of computational time is certainly the *martingale approach* (see Tables 5.1, 5.2, 5.3, 5.4, and 5.5).

Another element of comparison is the influence of the time step h, which is used to approximately simulate the random variables. The analysis of these discretization errors is the second significant part of this work. The relevant random variables are essentially written as the product of the cost function $f(X_T)$ by a random variable H, and simulations are based on Euler schemes. Although H has a complex form, we first propose an approximation algorithm and then we analyze the induced error w.r.t. the time step h. This part of the paper is original: previous results in the literature concern the approximation of $\mathbb{E}(f(X_T))$ (see Bally and Talay [BT96a]) or more generally of some smooth functionals of X (see Kohatsu-Higa and Pettersson [KHP00], [KHP02]). Here, regarding the techniques, we improve estimates given in [KHP00] since we do not need to add a small perturbation to the processes. Our multidimensional framework also raises extra difficulties compared to [KHP02], and we develop specific localization techniques that are interesting for themselves.

Outline of the paper. In the following, we make some assumptions and define the notations which will be used throughout the paper. We also recall the *pathwise* approach in Proposition 1.1. In section 2, after giving some standard facts on the Malliavin calculus, we develop our three approaches to computing the sensitivity of $J(\alpha)$ w.r.t. α: these are the so-called *Malliavin calculus* approach (Propositions 2.5 and 2.8), the *adjoint* approach (Theorem 2.11), and the *martingale* approach (Theorem 2.12). In section 3, we provide simulation procedures to compute $\nabla_\alpha J(\alpha)$ by

the usual Monte Carlo approach using the methods developed before and analyze the influence of the time step h used for Euler-type schemes. A significant part of the paper covers these analyses which have never been developed before in the literature. The approximation results are stated in Theorems 3.1, 3.2, 3.4, and 3.5, while their proofs are postponed to section 4. Finally, numerical experiments in section 5 illustrate the developed methods: we compare the computational time, the complexity, the variance, and the time discretization error of the estimators on many examples borrowed from finance and control.

Assumptions. In our applications, the parameter is a priori multidimensional, but since in the following we will look at sensitivities w.r.t. α coordinatewise, it is not a restriction to assume that the parameter space $\mathcal{A}$ is a subset of $\mathbb{R}$ ($m = 1$).

The process defined in (1.1) depends on the parameter α, but we deliberately omit this dependence in the notation. Furthermore, the initial condition $X_0 = x$ is fixed throughout the paper. We note σ_j, the jth column vector of σ.

To study the sensitivity of J (defined in (1.2)) w.r.t. α, we may assume that coefficients are smooth enough: in what follows, k is an integer greater than 2.

ASSUMPTION (R_k). *The functions b and σ are of class C^1 w.r.t. the variables t, x, α, and for some $\eta > 0$, the following Hölder continuity condition holds:*

$$\sup_{(t,x,\alpha,\alpha')\in[0,T]\times\mathbb{R}^d\times\mathcal{A}\times\mathcal{A}} \frac{|g(t,x,\alpha) - g(t,x,\alpha')|}{|\alpha - \alpha'|^\eta} < \infty$$

for $g = \partial_\alpha b$ and $g = \partial_\alpha \sigma$. Furthermore, for any $\alpha \in \mathcal{A}$, the functions $b(\cdot,\cdot,\alpha)$, $\sigma(\cdot,\cdot,\alpha)$, $\partial_\alpha b(\cdot,\cdot,\alpha)$, and $\partial_\alpha\sigma(\cdot,\cdot,\alpha)$ are of class $C^{\lfloor k/2\rfloor,k}$ w.r.t. (t,x); the functions $\partial_\alpha b$ and $\partial_\alpha\sigma$ are uniformly bounded in (t,x,α), and the derivatives of b, σ, $\partial_\alpha b$, and $\partial_\alpha\sigma$ w.r.t. (t,x) are uniformly bounded as well.

Note that b and σ may be unbounded. We do not assert that the assumption above is the weakest possible, but it is sufficient for our purpose. At several places, the diffusion coefficient will be required to be uniformly elliptic, in the following sense.

ASSUMPTION (E). *σ is a squared matrix ($q = d$) such that the matrix $\sigma\sigma^*$ satisfies a uniform ellipticity condition:*

$$\forall (t,x) \in [0,T]\times\mathbb{R}^d, \qquad [\sigma\sigma^*](t,x,\alpha) \geq \mu_{\min}\, \mathrm{I}_d$$

for a real number $\mu_{\min} > 0$.

Notation.

- *Sensitivity estimators.* To clarify the connection between our methods and the estimators H which are derived, we will write $H_T^{Path.}$ for the pathwise approach (Proposition 1.1), $H_T^{Mall.Ell.}$ (resp., $H_T^{Mall.Gen.}$) for the Malliavin calculus approach in the elliptic case (resp., in the general case) (Propositions 2.5 and 2.8), $H_T^{b,Adj.}$ and $H_T^{\sigma,Adj.}$ for the adjoint approach (Theorem 2.11), and $H_T^{Mart.}$ for the martingale approach (Theorem 2.12). The subscript T refers to the time in the expected cost (1.2). Their approximations using some discretization procedure with N time steps will be denoted $H_T^{Path.,N}$, $H_T^{Mall.Ell.,N}$, and so on.
- *Differentiation.* As usual, derivatives w.r.t. α will be simply denoted with a dot, for instance, $\partial_\alpha J = \dot{J}$. If no ambiguity is possible, we will omit to write explicitly the parameter α in b, $\sigma_j \cdots$. We adopt the following usual convention on the gradients: if $\psi : \mathbb{R}^{p_2} \mapsto \mathbb{R}^{p_1}$ is a differentiable function, its gradient $\nabla_x\psi(x) = (\partial_{x_1}\psi(x), \ldots, \partial_{x_{p_2}}\psi(x))$ takes values in $\mathbb{R}^{p_1}\otimes\mathbb{R}^{p_2}$. At many places, $\nabla_x\psi(x)$ will simply be denoted $\psi'(x)$.

- *Linear algebra.* The rth column of a matrix A will be denoted A_r (or $A_{r,t}$ if A is a time dependent matrix), and the rth element of a vector a will be denoted a_r (or $a_{r,t}$ if a is a time dependent vector). A^* stands for the transpose of A. For a matrix A, the matrix obtained by keeping only the last r rows (resp., the last r columns) will be denoted $\Pi_r^R(A)$ (resp., $\Pi_r^C(A)$). For $i \in \{1, \dots, d\}$, we set $e^i = (0 \cdots 0\ 1\ 0 \cdots 0)^*$, where 1 is the ith coordinate.
- *Constants.* We will keep the same notation $K(T)$ for all finite, nonnegative, and nondecreasing functions: they do not depend on x, the function f, or further discretization steps h, but they may depend on the coefficients $b(\cdot)$ and $\sigma(\cdot)$. The generic notation $K(x,T)$ stands for any function bounded by $K(T)(1+|x|^Q)$ for $Q \geq 0$.

When a function $g(s,x,\alpha)$ is evaluated at $x = X_s^\alpha$, we may sometimes use the short notation g_s if no ambiguity is possible. For instance, (1.1) may be written as $X_t = x + \int_0^t b_s ds + \sum_{j=1}^q \int_0^t \sigma_{j,s} dW_s^j$.

Other processes related to $(X_t)_{0\leq t\leq T}$. To the diffusion X under (R_2), we may associate its flow, i.e., the Jacobian matrix $Y_t := \nabla_x X_t$, the inverse of its flow $Z_t = {Y_t}^{-1}$, and the pathwise derivative of X_t w.r.t. α, which we denote $\dot{X}_t$ (see Kunita [Kun84]). These processes solve

$$Y_t = \mathrm{I}_d + \int_0^t b'_s\, Y_s\, ds + \sum_{j=1}^q \int_0^t \sigma'_{j,s}\, Y_s\, dW_s^j, \tag{1.3}$$

$$Z_t = \mathrm{I}_d - \int_0^t Z_s (b'_s - \sum_{j=1}^q (\sigma'_{j,s})^2)\, ds - \sum_{j=1}^q \int_0^t Z_s \sigma'_{j,s}\, dW_s^j, \tag{1.4}$$

$$\dot{X}_t = \int_0^t \left(\dot{b}_s + b'_s\, \dot{X}_s\right) ds + \sum_{j=1}^q \int_0^t \left(\dot{\sigma}_{j,s} + \sigma'_{j,s}\, \dot{X}_s\right) dW_s^j. \tag{1.5}$$

Actually, since the process $(\dot{X}_t)_{0\leq t\leq T}$ satisfies a linear equation, it can also simply be written using Y_t and Z_t (apply Theorem 56 from p. 271 of Protter [Pro90]):

$$\dot{X}_t = Y_t \int_0^t Z_s \left[\left(\dot{b}_s - \sum_{j=1}^q \sigma'_{j,s} \dot{\sigma}_{j,s} \right) ds + \sum_{j=1}^q \dot{\sigma}_{j,s} dW_s^j \right]. \tag{1.6}$$

If f is continuously differentiable with an appropriate growth condition (in order to apply the Lebesgue differentiation theorem), one immediately obtains the following result (see also Yang and Kushner [YK91]), which we call the *pathwise approach*.

PROPOSITION 1.1. *Assume* (R_2). *One has* $\dot{J}(\alpha) = \mathbb{E}\big(H_T^{Path.}\big)$ *with*

$$H_T^{Path.} = f'(X_T)\dot{X}_T.$$

Hence, the gradient can still be written as an expectation, which is crucial for a Monte Carlo evaluation. One purpose of the paper is to extend this result to the case of nondifferentiable functions, by essentially writing $\dot{J}(\alpha) = \mathbb{E}\left(f(X_T)H\right)$ for some random variable H.

In what follows, we will make two types of assumption on f.

ASSUMPTION (H). *f is a bounded measurable function.*

Actually, the above boundedness property of f is not important, since in what follows, we essentially use the fact that the random variable $f(X_T)$ belongs to any $\mathbf{L}^p$. However, this assumption simplifies the analysis.

ASSUMPTION (H′). *f is a bounded measurable function and satisfies the following continuity estimate for $p_0 > 1$:*

$$\int_0^T \frac{\|f(X_T) - f(X_t)\|_{\mathbf{L}^{p_0}}}{T-t}\, dt < +\infty.$$

This $\mathbf{L}^p$-smoothness assumption of $f(X_T) - f(X_t)$ is obviously satisfied for uniformly Hölder functions with exponent β, but also for some nonsmooth functions, such as the indicator function of a domain.

PROPOSITION 1.2. *Let D be a domain of $\mathbb{R}^d$: suppose that either it has a compact and smooth boundary (say, of class C^2; see* [GT77]*), or it is a convex polyhedron ($D = \cap_{i=1}^I D_i$, where $(D_i)_{1\le i\le I}$ are half-spaces). Assume* (E), (R$_2$), *and bounded coefficients b and σ. Then, the function $f = \mathbf{1}_D$ satisfies the assumption* (H′) *(for any $p_0 > 1$).*

Proof. Since $\|f(X_T) - f(X_t)\|_{\mathbf{L}^p}^p \le \mathbb{E}|\mathbf{1}_D(X_T) - \mathbf{1}_D(X_t)| \le \mathbb{P}(X_T \in D, X_t \notin D) + \mathbb{P}(X_T \notin D, X_t \in D)$, we only need to prove that $\mathbb{P}(X_T \in D, X_t \notin D) \le K(T)(T-t)^\beta$ with $\beta > 0$. Now, recall the standard exponential inequality $\mathbb{P}(\|X_u - x\| \ge \delta) \le K(T)\exp(-c\frac{\delta^2}{u})$ (with $c > 0$) available for $u \in]0, T]$ and $\delta \ge 0$ (see, e.g., Lemma 4.1 in [Gob00]). Combining this with the Markov property, it follows that $\mathbb{P}(X_T \in D, X_t \notin D) \le K(T)\mathbb{E}(\mathbf{1}_{X_t \notin D}\exp(-c\frac{d^2(X_t, D^c)}{(T-t)}))$. Then, a direct estimation of the above expectation using in particular a Gaussian upper bound for the density of the law of X_t (see Friedman [Fri64]) yields easily the required estimate with $\beta = \frac{1}{2}$ (see Lemma 2.8 in [Gob01a] for details). □

2. Sensitivity formulae. In this section, we present three different approaches to evaluate $\dot{J}(\alpha)$. Before this, we introduce the Malliavin calculus material necessary to our computations.

2.1. Some basic results on the Malliavin calculus. The reader may refer to Nualart [Nua95] (section 2.2 for the case of diffusion processes) for a detailed exposition of this section.

Put $\mathcal{H} = \mathbf{L}^2([0,T], \mathbb{R}^q)$: we will consider elements of $\mathcal{H}$ written as a row vector. For $h(.) \in \mathcal{H}$, denote by $W(h)$ the Wiener stochastic integral $\int_0^T h(t)\, dW_t$.

Let $\mathcal{S}$ denote the class of random variables of the form $F = f(W(h_1), \dots, W(h_N))$, where f is a C^∞-function with derivatives having a polynomial growth, $(h_1, \dots, h_N) \in \mathcal{H}^N$ and $N \ge 1$. For $F \in \mathcal{S}$, we define $\mathcal{D}F = (\mathcal{D}_t F := (\mathcal{D}_t^1 F, \dots, \mathcal{D}_t^q F))_{t\in[0,T]}$, its derivative, as the $\mathcal{H}$-valued random variable given by $\mathcal{D}_t F = \sum_{i=1}^N \partial_{x_i} f(W(h_1), \dots, W(h_N))\ \ h_i(t)$. The operator $\mathcal{D}$ is closable as an operator from $\mathbf{L}^p(\Omega)$ to $\mathbf{L}^p(\Omega, \mathcal{H})$, for any $p \ge 1$. Its domain is denoted by $\mathbb{D}^{1,p}$ w.r.t. the norm $\|F\|_{1,p} = [\mathbb{E}|F|^p + \mathbb{E}(\|\mathcal{D}F\|_{\mathcal{H}}^p)]^{1/p}$. We can define the iteration of the operator $\mathcal{D}$ in such a way that for a smooth random variable F, the derivative $\mathcal{D}^k F$ is a random variable with values on $\mathcal{H}^{\otimes k}$. As in the case $k = 1$, the operator $\mathcal{D}^k$ is closable from $S \subset \mathbf{L}^p(\Omega)$ into $\mathbf{L}^p(\Omega; \mathcal{H}^{\otimes k})$, $p \ge 1$. If we define the norm $\|F\|_{k,p} = [\mathbb{E}|F|^p + \sum_{j=1}^k \mathbb{E}(\|\mathcal{D}^j F\|_{\mathcal{H}^{\otimes j}}^p)]^{1/p}$, we denote its domain by $\mathbb{D}^{k,p}$. Finally, set $\mathbb{D}^{k,\infty} = \cap_{p\ge1}\mathbb{D}^{k,p}$ and $\mathbb{D}^\infty = \cap_{k,p\ge1}\mathbb{D}^{k,p}$. One has the following chain rule property.

PROPOSITION 2.1. *Fix $p \ge 1$. For $f \in C_b^1(\mathbb{R}^d, \mathbb{R})$ and $F = (F_1, \dots, F_d)^*$ a random vector whose components belong to $\mathbb{D}^{1,p}$, $f(F) \in \mathbb{D}^{1,p}$ and for $t \ge 0$, one has*

$\mathcal{D}_t(f(F)) = f'(F)\mathcal{D}_t F$, *with the notation*

$$\mathcal{D}_t F = \begin{pmatrix} \mathcal{D}_t F_1 \\ \vdots \\ \mathcal{D}_t F_d \end{pmatrix} \in \mathbb{R}^d \otimes \mathbb{R}^q.$$

We now introduce δ, the Skorohod integral, defined as the adjoint operator of $\mathcal{D}$.

DEFINITION 2.2. *δ is a linear operator on $\mathbf{L}^2([0,T]\times\Omega,\mathbb{R}^q)$ with values in $\mathbf{L}^2(\Omega)$ such that*

1. *the domain of δ (denoted by $\mathrm{Dom}(\delta)$) is the set of processes $u \in \mathbf{L}^2([0,T]\times\Omega,\mathbb{R}^q)$ such that $\left|\mathbb{E}\big(\int_0^T \mathcal{D}_t F \cdot u_t\, dt\big)\right| \le c(u)\,\|F\|_{\mathbf{L}^2}$ for any $F \in \mathbb{D}^{1,2}$.*
2. *if u belongs to $\mathrm{Dom}(\delta)$, then $\delta(u)$ is the element of $\mathbf{L}^2(\Omega)$ characterized by the integration-by-parts formula*

$$\forall F \in \mathbb{D}^{1,2}, \qquad \mathbb{E}\big(F\ \delta(u)\big) = \mathbb{E}\left(\int_0^T \mathcal{D}_t F \cdot u_t\ dt\right). \tag{2.1}$$

In the following proposition, we outline a few properties of the Skorohod integral.

PROPOSITION 2.3.

1. *The space of weakly differentiable $\mathcal{H}$-valued variables $\mathbb{D}^{1,2}(\mathcal{H})$ belongs to $\mathrm{Dom}(\delta)$.*
2. *If u is an adapted process belonging to $\mathbf{L}^2([0,T]\times\Omega,\mathbb{R}^q)$, then the Skorohod integral and the Itô integral coincide: $\delta(u) = \int_0^T u_t\, dW_t$.*
3. *If F belongs to $\mathbb{D}^{1,2}$, then for any $u \in \mathrm{Dom}(\delta)$ such that $\mathbb{E}(F^2 \int_0^T \|u_t\|^2\, dt) < +\infty$, one has*

$$\delta(F\ u) = F\ \delta(u) - \int_0^T \mathcal{D}_t F \cdot u_t\ dt, \tag{2.2}$$

whenever the right-hand side above belongs to $\mathbf{L}^2(\Omega)$.

Concerning the solution of SDEs, it is well known that under (R_k) $(k \ge 2)$ for any $t \ge 0$, the random variable X_t (resp., Y_t, Z_t, and $\dot{X}_t$) belongs to $\mathbb{D}^{k,\infty}$ (resp., $\mathbb{D}^{k-1,\infty}$). Furthermore, one has the following estimates: $\mathbb{E}\left(\sup_{0\le t\le T} \|\mathcal{D}_{r_1,\dots,r_{k'}} U_t\|^p\right) \le K(T,x)$ for $U_t = X_t$ with $1 \le k' \le k$ or $U_t = Y_t, Z_t, \dot{X}_t$ with $1 \le k' \le k-1$. Besides, $\mathcal{D}_s X_t$ is given by

$$\mathcal{D}_s X_t = Y_t\ Z_s \sigma(s, X_s)\ \mathbf{1}_{s\le t}. \tag{2.3}$$

Finally, we recall some standard results related to the integration-by-parts formulae. The Malliavin covariance matrix of a smooth random variable F is defined by

$$\gamma^F = \int_0^T \mathcal{D}_t F\ [\mathcal{D}_t F]^*\ dt. \tag{2.4}$$

PROPOSITION 2.4. *Let $\bar{\gamma}$ be a multi-index, F be a random variable in $\mathbb{D}^{k_1,\infty}$ such that $\det(\gamma^F)$ is almost surely positive with $1/\det(\gamma^F) \in \cap_{p\ge1}\mathbf{L}^p$ and G belongs to $\mathbb{D}^{k_2,\infty}$. Then for any smooth function g with polynomial growth, provided that k_1 and k_2 are large enough (depending on $\bar{\gamma}$), there exists a random variable $H_{\bar{\gamma}}(F,G)$ in any $\mathbf{L}^p$ such that*

$$\mathbb{E}[\partial^{\bar{\gamma}} g(F) G] = \mathbb{E}[g(F) H_{\bar{\gamma}}(F,G)].$$

Moreover, for any arbitrary event A we have

$$\|H_{\bar\gamma}(F,G)\mathbf{1}_A\|_{\mathbf{L}^p} \le C\|[\gamma^F]^{-1}\mathbf{1}_A\|_{\mathbf{L}^{q_3}}^{p_3}\|F\|_{k_1,q_1}^{p_1}\|G\|_{k_2,q_2}$$

for some constants $C, p_1, p_3, q_1, q_2, q_3$ depending on p and $\bar\gamma$.

Proof. See Propositions 3.2.1 and 3.2.2 in Nualart [Nua98, pp. 160–161] when $A = \Omega$. For any other event A, see Proposition 2.4 from Bally and Talay [BT96a]. □

The construction of $H_{\bar\gamma}(F,G)$ is based on the equality (2.1) and involves iterated Skorohod integrals. We do not really need to make it explicit at this stage.

2.2. First approach: Direct Malliavin calculus computations. Here, the guiding idea is to start from Proposition 1.1 and apply results like Proposition 2.4 to get $\dot J(\alpha) = \mathbb{E}(f(X_T)H)$. Nevertheless, there are several ways to do this, depending on whether the diffusion coefficient is elliptic (see also [FLL+99] in that situation) or not.

2.2.1. Elliptic case. Consider first that the assumption (E) is fulfilled.

PROPOSITION 2.5. *Assume* (R_2), (E), *and* (H). *One has* $\dot J(\alpha) = \mathbb{E}\big(H_T^{Mall.Ell.}\big)$, *where*

$$H_T^{Mall.Ell.} = \frac{1}{T} f(X_T)\delta\big([\sigma_\cdot^{-1}\, Y_\cdot\, Z_T\, \dot X_T]^*\big)$$

belongs to $\cap_{p\ge1}\mathbf{L}^p$.

Proof. We can consider that f is smooth, the general case being obtained using an $\mathbf{L}^2$-approximation of f with some smooth and compactly supported functions. As a consequence of (2.3) and Assumption (E), $\mathcal{D}_t X_T$ is invertible for any $t \in [0,T]$: thus, for such t, using the chain rule (Proposition 2.1), one gets that $f'(X_T) = \mathcal{D}_t(f(X_T))\sigma_t^{-1}\, Y_t\, Z_T$. Integrating in time over $[0,T]$ and using Proposition 1.1, one gets that $\dot J(\alpha) = \frac{1}{T}\int_0^T dt\, \mathbb{E}(\mathcal{D}_t(f(X_T))\sigma_t^{-1}\, Y_t\, Z_T\, \dot X_T)$. An application of the relation (2.1) completes the proof of Proposition 2.5 (the $\mathbf{L}^p$-estimates follow from Proposition 2.4). □

When the parameter enters the drift coefficient only, the laws of $(X_t)_{0\le t\le T}$ for two different values of α are equivalent owing to the Girsanov theorem. Exploiting this possible change of measure directly, a simplified expression for $\dot J(\alpha)$ can be found: this is the *likelihood ratio method* or *score method* from Kushner and Yang [YK91].

PROPOSITION 2.6. *Assume* (R_2), (E), *and* (H). *Suppose that the parameter of interest α is not in the diffusion coefficient. Then, one has*

$$\dot J(\alpha) = \mathbb{E}\left(f(X_T)\int_0^T [\sigma_t^{-1}\dot b_t]^* dW_t\right).$$

Proof. Instead of using the Girsanov theorem, we leverage the particular form of $\dot X_T$ given in (1.6) to prove this. Indeed, $f'(X_T)\dot X_T = f'(X_T)Y_T\int_0^T Z_t\dot b_t dt = \int_0^T dt\, \mathcal{D}_t(f(X_T))[\sigma_t^{-1}\dot b_t]$, and the result follows using (2.1). □

2.2.2. General nondegenerate case. There are many situations where the ellipticity Assumption (E) is too stringent and cannot be fulfilled. To illustrate this, let us rewrite the SDE in the following way, splitting its structure into two parts:

$$dX_t = \begin{pmatrix} dS_t \\ dV_t \end{pmatrix} = \begin{pmatrix} b_S(t,X_t,\alpha) \\ b_V(t,X_t,\alpha) \end{pmatrix} dt + \begin{pmatrix} \sigma_S(t,X_t,\alpha) \\ \sigma_V(t,X_t,\alpha) \end{pmatrix} dW_t. \tag{2.5}$$

Here, $(S_t)_{t\geq 0}$ is $(d-r)$-dimensional, $(V_t)_{t\geq 0}$ r-dimensional, and the dimension of W is arbitrary. The cost function of interest may involve only the value of V_T: $J(\alpha) = \mathbb{E}(f(V_T))$. Note that considering $r = d$ reduces to the previous situation. We now give two examples that motivate the statement of Proposition 2.7 below.

(a) In random mechanics (see Krée and Soize [KS86]), the pair position/velocity $dX_t = \begin{pmatrix} dx_t \\ dv_t \end{pmatrix} = \begin{pmatrix} v_t dt \\ \cdots \end{pmatrix}$ cannot satisfy an ellipticity condition, but weaker assumptions such as hypoellipticity are more realistic.

(b) For portfolio optimization in finance (for a recent review, see, e.g., Runggaldier [Run02]), r usually equals 1. $(S_t)_{t\geq 0}$ describes the dynamic of the risky assets, while $(V_t)_{t\geq 0}$ is the wealth process, corresponding to the value of a self-financed portfolio invested in a nonrisky asset with instantaneous return $r(t, S_t)$ and in the assets $(S_t)_{t\geq 0}$ w.r.t. the strategy $(\xi_t = \{\xi_i(t, X_t) : 1 \leq i \leq d-1\})_{t\geq 0}$: $dV_t = \xi(t, X_t) \cdot dS_t + (V_t - \xi(t, X_t) \cdot S_t) r(t, S_t) dt$ (see e.g. Karatzas and Shreve [KS98]). It is clear that the resulting diffusion coefficient for the whole process $X_t = \begin{pmatrix} S_t \\ V_t \end{pmatrix}$ cannot satisfy an ellipticity condition. Nevertheless, requiring that the matrix $\sigma_V \sigma_V^*(t, x)$ satisfy an ellipticity type condition is not very restricting in that framework.

We set γ_T for the Malliavin covariance matrix of V_T: $\gamma_T = \int_0^T \mathcal{D}_t V_T \, [\mathcal{D}_t V_T]^* \, dt$. This allows to reformulate Assumption (E) as the following.

ASSUMPTION (E′). $\det(\gamma_T)$ *is almost surely positive and for any* $p \geq 1$, *one has*

$$\|1/\det(\gamma_T)\|_{\mathbf{L}^p} < +\infty.$$

We now bring together standard results related to Assumption (E′).

PROPOSITION 2.7. *Assumption* (E′) *is fulfilled in the following situations.*

1. Hypoelliptic case *(with* $r = d$*) under* (R_∞). *The Lie algebra generated by the vector fields* $\partial_t + A_0(t, x) := \partial_t + \sum_{i=1}^d (b - \frac{1}{2} \sum_{j=1}^q \sigma_j' \sigma_j)_i(t, x) \partial_{x_i}$, $A_j(t, x) := \sum_{i=1}^d \sigma_{i,j}(t, x) \partial_{x_i}$ *for* $1 \leq j \leq q$ *spans* $\mathbb{R}^{d+1}$ *at the point* $(0, X_0)$*:*

$$\dim \operatorname{span} \operatorname{Lie}(\partial_t + A_0, A_j, 1 \leq j \leq q)(0, X_0) = d + 1.$$

2. Partially elliptic case *(with* $r \geq 1$*) under* (R_2). *For a real number* $\mu_{\min} > 0$, *one has*

$$\forall x \in \mathbb{R}^d, \qquad [\sigma_V \sigma_V^*](T, x, \alpha) \geq \mu_{\min} \, \mathrm{I}_d.$$

Proof. The statement 1 is standard and we refer to Cattiaux and Mesnager [CM02] for a recent account on the subject. The statement 2 is also standard: see, for instance, the arguments in Nualart [Nua98, pp. 158–159]. □

Now, we are in a position to give a sensitivity formula under (E′).

PROPOSITION 2.8. *Assume* (R_2), (E′), *and* (H). *One has* $\dot{J}(\alpha) = \mathbb{E}\big(H_T^{Mall.Gen.}\big)$ *where*

$$H_T^{Mall.Gen.} = f(V_T) \, \delta\big(\dot{V}_T^* \gamma_T^{-1} \mathcal{D}.V_T\big)$$

belongs to $\cap_{p\geq 1} \mathbf{L}^p$.

Proof. Assumption (E′) validates (see Nualart [Nua98, Proposition 3.2.1]) the following computations, adapted from the ones used for Proposition 2.5. The chain

rule property yields $f'(V_T) = \int_0^T \mathcal{D}_t(f(V_T))[\mathcal{D}_t V_T]^* \gamma_T^{-1}\, dt$, and thus $\mathbb{E}(f'(V_T)\dot{V}_T) = \mathbb{E}(\int_0^T \mathcal{D}_t(f(V_T))[\mathcal{D}_t V_T]^* \gamma_T^{-1}\dot{V}_T\, dt)$. Proposition 2.8 now follows from (2.1). □

Proposition 2.8 is also valid under (E) in the case $r = d$, but the formula in Proposition 2.5 is actually a bit simpler to implement.

2.3. A second approach based on the adjoint point of view.

2.3.1. Another representation of the sensitivity of $J(\alpha)$. If we set $u(t,x) = \mathbb{E}\big(f(X_T)|X_t = x\big)$, omitting to indicate the dependence w.r.t. α, we have that $J(\alpha)$ defined in (1.2) equals $u(0, X_0)$. Under smoothness assumptions on b and σ and the nondegeneracy hypothesis on the infinitesimal generator of $(X_t)_{t\geq 0}$, it is well known (see Cattiaux and Mesnager [CM02]) that u is the smooth solution of the partial differential equation (PDE)

$$\begin{cases} \partial_t u(t,x) + \sum_{i=1}^d b_i(t,x)\partial_{x_i} u(t,x) + \frac{1}{2}\sum_{i,j=1}^d [\sigma\sigma^*]_{i,j}(t,x)\partial^2_{x_i,x_j} u(t,x) = 0 \ \text{ for } t < T, \\ u(T,x) = f(x). \end{cases}$$

Our purpose is to give another expression for $\dot{J}(\alpha)$ of Proposition 1.1. The idea is simple: it consists in formally differentiating the PDE above w.r.t. α and in reinterpreting the derivative as an expectation. This is now stated and justified rigorously.

LEMMA 2.9. *Assume* (R_3), (E), *and* (H). *One has*

$$\dot{J}(\alpha) = \int_0^T \mathbb{E}\left(\sum_{i=1}^d \dot{b}_{i,t}\partial_{x_i} u(t, X_t) + \frac{1}{2}\sum_{i,j=1}^d [\dot{\sigma\sigma^*}]_{i,j,t}\partial^2_{x_i,x_j} u(t, X_t)\right) dt.$$

Proof. This is a standard fact that under (R_3) and (E), u is twice differentiable w.r.t. x (see the arguments of Lemma 2.10 below, where the proof is sketched). The technical difficulty in the following computations comes from the possible explosion of derivatives of u for t close to T, when f is nonsmooth. For this reason, we first prove useful uniform estimates: for any multi-index $\bar{\gamma}$ with $|\bar{\gamma}| \leq 2$, any smooth random variable $G \in \mathbb{D}^{2,\infty}$ and any parameters α and α', one has

$$\sup_{t\in[0,T[} |\mathbb{E}[G\, \partial_x^{\bar{\gamma}} u(t, X_t^{\alpha'})]| \leq K(T,x)\frac{\|f\|_\infty}{T^{\frac{|\bar{\gamma}|}{2}}}\|G\|_{|\bar{\gamma}|,p'}. \tag{2.6}$$

Indeed, for $t \geq T/2$, first apply Proposition 2.4: then, use $|u(t,x)| \leq \|f\|_\infty$ combined with some specific estimates for $\|H_{\bar{\gamma}}(X_t^{\alpha'}, G)\|_{\mathbf{L}^p} \leq \frac{K(T,x)}{t^{\frac{|\bar{\gamma}|}{2}}}\|G\|_{|\bar{\gamma}|,p'}$ available under the ellipticity condition (E) (see Theorem 1.20 and Corollary 3.7 in Kusuoka and Stroock [KS84], or section 4.1. in [Gob00] for a brief review). For $t \leq T/2$, note that using the Markov property, one has $\partial_x^{\bar{\gamma}} u(t,x) = \partial_x^{\bar{\gamma}}\mathbb{E}\big(u(\frac{T+t}{2}, X_{\frac{T+t}{2}}^{t,x})\big) = \sum_{1\leq|\gamma'|\leq|\bar{\gamma}|}\mathbb{E}\big(\partial_x^{\gamma'} u(\frac{T+t}{2}, X_{\frac{T+t}{2}}^{t,x}) G_{\frac{T+t}{2}}^{\gamma'}\big)$ with $G_{\frac{T+t}{2}}^{\gamma'} \in \mathbb{D}^{2+|\gamma'|-|\bar{\gamma}|,\infty}$ and $(X_s^{t,y})_{s\geq t}$ standing for the process starting from y at time t. Again applying the integration-by-parts formula with the elliptic estimates gives $|\partial_x^{\bar{\gamma}} u(t,x)| \leq \frac{K(T,x)}{[\frac{T+t}{2}-t]^{\frac{|\bar{\gamma}|}{2}}}\|f\|_\infty$ and (2.6) follows

since $\frac{T+t}{2} - t \geq \frac{T}{4}$. Now, for $\epsilon \in \mathbb{R}$, the difference $J(\alpha+\epsilon) - J(\alpha)$ equals

$$\begin{aligned}
&\mathbb{E}\left(f(X_T^{\alpha+\epsilon}) - f(X_T^{\alpha})\right) = \mathbb{E}\left(u(T, X_T^{\alpha+\epsilon}) - u(0, X_0^{\alpha+\epsilon})\right) \\
&= \int_0^T \mathbb{E}\left(\partial_t u(t, X_t^{\alpha+\epsilon}) + \sum_{i=1}^d b_i(t, X_t^{\alpha+\epsilon}, \alpha+\epsilon) \partial_{x_i} u(t, X_t^{\alpha+\epsilon}) \right. \\
&\qquad \left. + \frac{1}{2} \sum_{i,j=1}^d [\sigma\sigma^*]_{i,j}(t, X_t^{\alpha+\epsilon}, \alpha+\epsilon) \partial^2_{x_i,x_j} u(t, X_t^{\alpha+\epsilon}) \right) dt \\
&= \int_0^T \mathbb{E}\left(\sum_{i=1}^d (b_i(t, X_t^{\alpha+\epsilon}, \alpha+\epsilon) - b_i(t, X_t^{\alpha+\epsilon}, \alpha)) \partial_{x_i} u(t, X_t^{\alpha+\epsilon}) \right. \\
&\qquad \left. + \frac{1}{2} \sum_{i,j=1}^d ([\sigma\sigma^*]_{i,j}(t, X_t^{\alpha+\epsilon}, \alpha+\epsilon) - [\sigma\sigma^*]_{i,j}(t, X_t^{\alpha+\epsilon}, \alpha)) \partial^2_{x_i,x_j} u(t, X_t^{\alpha+\epsilon}) \right) dt,
\end{aligned}$$

where at the last equality we used the PDE solved by u to remove the term $\partial_t u$. Now, divide by ϵ and take its limit to 0: the result follows owing to the uniform estimates (2.6). □

Note that the formulation of Lemma 2.9 is strongly related to a form of the stochastic maximum principle (the Pontryagin principle) for optimal control problems: the processes $([\partial_{x_i} u(t, X_t)]_i)_{0 \leq t < T}$ and $([\partial^2_{x_i,x_j} u(t, X_t)]_{i,j})_{0 \leq t < T}$ are the so-called adjoint processes (see Bensoussan [Ben88] for convex control domains, or more generally Peng [Pen90]) and solve backward SDEs. Usually in these problems, the function f is smooth. Here, since the law of X_t has a smooth density w.r.t. the Lebesgue measure, we can remove the regularity condition on f.

Note also that Lemma 2.9 remains valid under a hypoellipticity hypothesis (condition 1 in Proposition 2.7). However, the derivation of tractable formulae below relies strongly on the ellipticity property.

2.3.2. Transformation using Itô–Malliavin integration-by-parts formulae. The aim of this section is to transform the expression for $\dot{J}(\alpha)$ in terms of explicit quantities. To remove the nonexplicit terms $\partial_{x_i} u$ and $\partial^2_{x_i,x_j} u$, we may use some integration-by-parts formulae, but here, to keep more tractable expressions, we are going to derive Bismut-type formulae, i.e., involving only Itô integrals instead of Skorohod integrals (see Bismut [Bis84]; Elworthy, Le Jan, and Li [EJL99]; and references therein), using a martingale argument (see also Thalmaier [Tha97] or, more recently, Picard [Pic02]). In the cited references, this approach has been used to compute estimates of the gradient of u. Here, we extend it to support higher derivatives. The basic tool is given by the following lemma.

LEMMA 2.10. *Assume* (R_2), (E), *and* (H) *and define* $M_t = u'(t, X_t) Y_t$ *for* $t < T$. *Then* $M = (M_t)_{0 \leq t < T}$ *is an* $\mathbb{R}^1 \otimes \mathbb{R}^d$*-valued martingale.*

Proof. First, we justify that u is continuously differentiable w.r.t x under (R_2) and (E). If f is smooth, this is clear (even without (E)), but (2.10) below also shows that under (E), u' can be expressed without the derivative of f. This easily leads to our assertion (see the proof of Proposition 3.2 in [FLL+99]). Now, the Markov property ensures that $(u(t, X_t^{0,x}))_{0 \leq t < T}$ is a martingale for any $x \in \mathbb{R}^d$. Hence, its derivative w.r.t. x (i.e., $(M_t)_{0 \leq t < T}$) is also a martingale (see Arnaudon and Thalmaier [AT98]). □

We now state a theorem which, if combined with Lemma 2.9, leads to an alternative representation for $\dot{J}(\alpha)$.

THEOREM 2.11. *Assume* (R_3) *and* (E).

Under (H), *one has*

$$\int_0^T \mathbb{E}\Big(\sum_{i=1}^d \dot{b}_{i,t}\partial_{x_i}u(t,X_t)\Big)dt = \mathbb{E}(H_T^{b,Adj.}), \tag{2.7}$$

where $H_T^{b,Adj.} = f(X_T)\displaystyle\int_0^T dt\ \dot{b}_t \cdot \frac{Z_t^*}{T-t}\int_t^T [\sigma_s^{-1}\ Y_s]^*\ dW_s$ *belongs to* $\bigcap_{p\geq 1}\mathbf{L}^p$.

Under (H′), *one has*

$$\int_0^T \mathbb{E}\Big(\sum_{i,j=1}^d [\sigma\dot{\sigma}^*]_{i,j,t}\partial^2_{x_i,x_j}u(t,X_t)\Big)dt = \mathbb{E}(H_T^{\sigma,Adj.}), \tag{2.8}$$

where

$$H_T^{\sigma,Adj.} = \int_0^T dt\ \sum_{i,j=1}^d [\sigma\dot{\sigma}^*]_{i,j,t}[f(X_T)-f(X_t)]\left(\frac{2e^j}{T-t}\cdot\left[Z_t^*\int_{\frac{T+t}{2}}^T [\sigma_s^{-1}Y_s]^* dW_s\right]\right.$$
$$\left.\times\ \frac{2e^i}{T-t}\cdot\left[Z_t^*\int_t^{\frac{T+t}{2}}[\sigma_s^{-1}Y_s]^*dW_s\right] + \frac{2e^i}{T-t}\cdot\left\{\nabla_x\left[Z_t^*\int_t^{\frac{T+t}{2}}[\sigma_s^{-1}Y_s]^*dW_s\right]Z_t e^j\right\}\right)$$

belongs to $\bigcap_{p<p_0}\mathbf{L}^p$.

Proof. Equality (2.7). First, Clark and Ocone's formula [Nua95, p. 42] gives $u(\tau,X_\tau) = u(t,X_t) + \int_t^\tau \mathbb{E}(\mathcal{D}_s[u(\tau,X_\tau)]|\mathcal{F}_s)dW_s$ for $0\leq t\leq\tau<T$. Using (2.3) and the martingale property of Lemma 2.10, we get $\mathbb{E}(\mathcal{D}_s[u(\tau,X_\tau)]|\mathcal{F}_s) = \mathbb{E}(u'(\tau,X_\tau)Y_\tau Z_s\sigma_s|\mathcal{F}_s) = u'(s,X_s)\sigma_s$. Hence, it gives an explicit form to the predictable representation theorem:

$$\forall\ 0\leq t\leq\tau\leq T \qquad u(\tau,X_\tau) = u(t,X_t) + \int_t^\tau u'(s,X_s)\sigma_s dW_s \tag{2.9}$$

(the case $\tau = T$ is obtained by passing to the limit). Note that this representation holds under (R_2). Since $(u'(t,X_t)Y_t)_{0\leq t<T}$ is a martingale, we obtain that

$$\begin{aligned}
u'(t,X_t)Y_t &= \mathbb{E}\left(\frac{1}{T-t}\int_t^T u'(s,X_s)Y_s\ ds|\mathcal{F}_t\right)\\
&= \mathbb{E}\left(\frac{1}{T-t}\left[\int_t^T u'(s,X_s)\sigma_s dW_s\right]\left[\int_t^T[\sigma_s^{-1}Y_s]^*dW_s\right]^*|\mathcal{F}_t\right)\\
&= \mathbb{E}\left(\frac{f(X_T)-u(t,X_t)}{T-t}\left[\int_t^T[\sigma_s^{-1}Y_s]^*dW_s\right]^*|\mathcal{F}_t\right)\\
&= \mathbb{E}\left(\frac{f(X_T)}{T-t}\left[\int_t^T[\sigma_s^{-1}Y_s]^*dW_s\right]^*|\mathcal{F}_t\right),
\end{aligned} \tag{2.10}$$

where for the third equality we used (2.9) with $\tau = T$ and $u(T, X_T) = f(X_T)$. Now the proof of (2.7) is straightforward.

Equality (2.8). Note that a slight modification of the preceding arguments (namely, integrating over $[t,(T+t)/2]$ instead of $[t,T]$ and applying (2.9) with $\tau = (t+T)/2$) leads to $\partial_{x_i} u(t,X_t) = \mathbb{E}\big(u\big(\frac{T+t}{2}, X_{\frac{T+t}{2}}\big)\frac{2e^i}{T-t}\cdot\big[Z_t^*\int_t^{\frac{T+t}{2}}[\sigma_s^{-1}Y_s]^* dW_s\big]|\mathcal{F}_t\big)$. Differentiating w.r.t. x on both sides and using (2.10) yields

$$
\begin{aligned}
(\partial_{x_i} u)'(t,X_t)Y_t &= \mathbb{E}\left(u'\left(\frac{T+t}{2}, X_{\frac{T+t}{2}}\right) Y_{\frac{T+t}{2}} \frac{2e^i}{T-t}\cdot\left[Z_t^*\int_t^{\frac{T+t}{2}}[\sigma_s^{-1}Y_s]^* dW_s\right]|\mathcal{F}_t\right)\\
&\quad+\mathbb{E}\left(u\left(\frac{T+t}{2}, X_{\frac{T+t}{2}}\right)\frac{2e^i}{T-t}\cdot\nabla_x\left\{\left[Z_t^*\int_t^{\frac{T+t}{2}}[\sigma_s^{-1}Y_s]^* dW_s\right]\right\}|\mathcal{F}_t\right)\\
&= \mathbb{E}\left([f(X_T)-f(X_t)]\frac{2}{T-t}\left[\int_{\frac{T+t}{2}}^T[\sigma_s^{-1}Y_s]^* dW_s\right]^*\frac{2e^i}{T-t}\cdot\left[Z_t^*\int_t^{\frac{T+t}{2}}[\sigma_s^{-1}Y_s]^* dW_s\right]\right.\\
&\quad\left.+[f(X_T)-f(X_t)]\frac{2e^i}{T-t}\cdot\nabla_x\left\{\left[Z_t^*\int_t^{\frac{T+t}{2}}[\sigma_s^{-1}Y_s]^* dW_s\right]\right\}|\mathcal{F}_t\right)
\end{aligned}
$$

(note that the $f(X_t)$ terms have no contribution in the expectation). Rearranging this last expression leads to (2.8).

The $\mathbf{L}^p$-estimates can be justified using the generalized Minkowski inequality and standard estimates from the stochastic calculus:

(2.11)

$$
\begin{aligned}
&\|H_T^{b,Adj.}\|_{\mathbf{L}^p} \le \int_0^T \frac{\|f\|_\infty}{T-t}\left\|\dot b(t,X_t)\cdot Z_t^*\int_t^T[\sigma_s^{-1}Y_s]^* dW_s\right\|_{\mathbf{L}^p} dt \le K(T,x)\int_0^T\frac{\|f\|_\infty}{\sqrt{T-t}}dt,\\
&\|H_T^{\sigma,Adj.}\|_{\mathbf{L}^p} \le K(T,x)\int_0^T\frac{\|f(X_T)-f(X_t)\|_{\mathbf{L}^{p'}}}{T-t}\,dt
\end{aligned}
$$

for $p < p' < p_0$. □

Remark 2.1. The $f(X_t)$ terms in $H_T^{\sigma,Adj.}$ seem to be crucial to ensure its $\mathbf{L}^p$ integrability: numerical experiments in section 5 illustrate this fact.

2.4. A third approach using martingales. We emphasize the dependence on α of the expected cost by denoting $u(\alpha,t,x) = \mathbb{E}(f(X_T^\alpha)|X_t^\alpha = x)$: hence, $J(\alpha) = u(\alpha,0,X_0)$. From the estimates proved in Lemma 2.9, this is a differentiable function w.r.t. α and one has $|\dot u(\alpha,t,x)| \le K(T,x)\|f\|_\infty$ and $|u'(\alpha,t,x)| \le \frac{K(T,x)}{\sqrt{T-t}}\|f\|_\infty$. Furthermore, using Theorem 2.11 and the $\mathbf{L}^p$-estimates (2.11) under (H′), one gets

$$
|\dot u(\alpha,t,x)| \le K(T,x)\left[\|f\|_\infty\sqrt{T-t} + \int_t^T\frac{\|f(X_T^{t,x})-f(X_s^{t,x})\|_{\mathbf{L}^{p'}}}{T-s}\,ds\right]
$$

for $p' < p_0$. Consequently, if we put $g(r) = \mathbb{E}\,(\dot u(\alpha,r,X_r))$, we easily obtain $|g(r)| \le K(T,x)[\|f\|_\infty\sqrt{T-r} + \int_r^T \frac{\|f(X_T)-f(X_s)\|_{\mathbf{L}^{p_0}}}{T-s}\,ds]$ and thus, $\lim_{r\to T} g(r) = 0$. For any $0 \le r \le s \le T$, one has $\mathbb{E}(u(\alpha,r,X_r)) = \mathbb{E}(u(\alpha,s,X_s)) = \frac{1}{T-r}\int_r^T \mathbb{E}(u(\alpha,s,X_s))\,ds$

using the Markov property; hence, by differentiation w.r.t. α, one gets

(2.12)
$$\begin{aligned}\mathbb{E}\big(\dot{u}(\alpha,r,X_r)\big) &= \frac{1}{T-r}\int_r^T ds\,\mathbb{E}\big(\dot{u}(\alpha,s,X_s)+u'(\alpha,s,X_s)\dot{X}_s - u'(\alpha,r,X_r)\dot{X}_r\big)\\ &= \frac{1}{T-r}\int_r^T ds\,\mathbb{E}\big(\dot{u}(\alpha,s,X_s)+u'(\alpha,s,X_s)[\dot{X}_s - Y_sZ_r\dot{X}_r]\big),\end{aligned}$$

where we used at the last equality the martingale property of $M_t = u'(\alpha,t,X_t)Y_t$ between $t=s$ and $t=r$ (see Lemma 2.10).

Now, put $h(r) = \frac{1}{T-r}\int_r^T ds\,\mathbb{E}\left(u'(\alpha,s,X_s)[\dot{X}_s - Y_sZ_r\dot{X}_r]\right)$: one has derived the following integral equation:

$$g(t) = \frac{1}{T-t}\int_t^T g(s)\,ds + h(t). \tag{2.13}$$

Before solving it, we express $h(r)$ using only f: for this, we use the predictable representation (2.9), which immediately gives

$$h(r) = \frac{1}{T-r}\mathbb{E}\left((f(X_T)-f(X_r))\int_r^T [\sigma_s^{-1}(\dot{X}_s - Y_sZ_r\dot{X}_r)]^* dW_s\right). \tag{2.14}$$

Note again that the term with $f(X_r)$ has no contribution and is put only to justify that $|h(r)| \le K(T,x)\|f(X_T)-f(X_r)\|_{\mathbf{L}^{p_0}}$ (use the Burkholder–Davis–Gundy inequalities and straightforward upper bounds for $\|\dot{X}_s - Y_sZ_r\dot{X}_r\|_{\mathbf{L}^q} \le K(T,x)\sqrt{s-r}$), from which we deduce that the integral $\int_0^T \frac{h(t)}{T-t}\,dt$ is convergent because of (H′). To solve the integral equation above, note that $[\frac{1}{T-t}\int_t^T g(s)\,ds]' = -\frac{h(t)}{T-t}$, and thus by integration, we have $\frac{1}{T-t}\int_t^T g(s)\,ds = C - \int_t^T \frac{h(r)}{T-r}dr$. The constant C equals 0 since both integrals in the previous equality converge to 0 when t goes to T (use $\lim_{t\to T} g(t) = 0$ and (H′)). Plug this new equality into (2.13), use (2.14), and take $t=0$ (with $\dot{X}_0 = 0$) to get the following representation for $\dot{J}(\alpha)$: this is the main result of this section.

THEOREM 2.12. *Assume* (R_2), (E), *and* (H′). *Then, one has* $\dot{J}(\alpha) = \mathbb{E}(H_T^{Mart.})$ *with*

$$\begin{aligned}H_T^{Mart.} &= \frac{f(X_T)}{T}\int_0^T [\sigma_s^{-1}\dot{X}_s]^* dW_s\\ &\quad + \int_0^T dr\,\frac{[f(X_T)-f(X_r)]}{(T-r)^2}\int_r^T [\sigma_s^{-1}(\dot{X}_s - Y_sZ_r\dot{X}_r)]^* dW_s.\end{aligned} \tag{2.15}$$

Furthermore, the random variable $H_T^{Mart.}$ *belongs to* $\bigcap_{p<p_0}\mathbf{L}^p$.

This method is called the *martingale approach* because it is based on the equality (2.12), which is a consequence of the martingale property of

$$[\dot{u}(\alpha,s,X_s) + u'(\alpha,s,X_s)\dot{X}_s]_{0\le s<T}.$$

Proof. What remains to be proved is the $\mathbf{L}^p$ estimate of $H_T^{Mart.}$: this can be easily obtained by combining Minkowski's inequality, Hölder's inequality, Assumption (H′), and standard stochastic calculus inequalities as before. □

Remark 2.2. When the parameter is not involved in the diffusion coefficient, it is easy to see that the improved estimate $\|\dot{X}_s - Y_s Z_r \dot{X}_r\|_{L_q} \leq K(T,x)(s-r)$ is available: thus, this allows us to remove $f(X_r)$ terms in the expression of $H_T^{Mart.}$ without changing the finiteness of the $\mathbf{L}^p$-norm of the new $H_T^{Mart.}$. In other words, only Assumption (H) is needed.

Besides, when α is only in the drift coefficient and these $f(X_r)$ terms are suppressed, this representation coincides with that of Theorem 2.11. Indeed, let us write $P_r = \int_r^T [\sigma_s^{-1}(\dot{X}_s - Y_s Z_r \dot{X}_r)]^* dW_s = \int_r^T [\sigma_s^{-1}\dot{X}_s]^* dW_s - [Z_r \dot{X}_r]^* \int_r^T [\sigma_s^{-1} Y_s]^* dW_s := P_{1,r} - P_{2,r}$, where $P_{1,r} = \int_0^T [\sigma_s^{-1}\dot{X}_s]^* dW_s - [Z_r\dot{X}_r]^* \int_0^T [\sigma_s^{-1}Y_s]^* dW_s$ and $P_{2,r} = \int_0^r [\sigma_s^{-1}\dot{X}_s]^* dW_s - [Z_r\dot{X}_r]^* \int_0^r [\sigma_s^{-1}Y_s]^* dW_s$. From the fact that $Z_r\dot{X}_r = \int_0^r Z_s \dot{b}_s ds$ (see (1.6)), one gets $dP_{2,r} = [Z_r \dot{b}_r]^* (\int_0^r [\sigma_t^{-1} Y_t]^* dW_t) dr$, hence $P_{2,r}$ is of bounded variation. $P_{1,r}$ is also of bounded variation, since $Z_r\dot{X}_r$ is. Thus, one obtains $dP_r = -\dot{b}_r \cdot Z_r^* (\int_r^T [\sigma_t^{-1} Y_t]^* dW_t) dr$: furthermore, since $P_T = 0$, one has $\|P_r\|_{\mathbf{L}^p} \leq K(T,x)(T-r)^{3/2}$. Using an integration-by-parts formula in (2.15) finally completes our assertion: $H_T^{Mart.} = f(X_T)\big(\frac{1}{T}P_0 + \int_0^T \frac{P_r}{(T-r)^2} dr\big) = f(X_T)\big(-\int_0^T \frac{dP_r}{(T-r)}\big) = H_T^{b,Adj.}$.

Consequently, this martingale approach does not provide any new elements when the parameter is not in the diffusion coefficient. On the contrary, if σ depends on α, the representation with the adjoint point of view is different from the martingale one (see numerical experiments). However, we must admit that this martingale approach remains somewhat mysterious to us.

3. Monte Carlo simulation and analysis of the discretization error. In this section, we discuss the numerical implementation of the formulae derived in this paper to compute the sensitivity of $J(\alpha)$ w.r.t. α. These formulae are written as expectations of some functionals of the process $(X_t)_{0\leq t\leq T}$ and related ones: a standard way to proceed consists in drawing independent simulations, approximating the functional using Euler schemes, and averaging independent samples of the resulting functional to get an estimation of the expectation (see section 5).

Here, we focus on the impact of the time step $h = T/N$ (N is the number of discretization times in the regular mesh of the interval $[0,T]$) in the simulation of the functional: it is well known that for the evaluation of $\mathbb{E}(f(X_T))$, the discretization error using an Euler scheme is of order h (see Bally and Talay [BT96a] for measurable functions f, or Kohatsu-Higa and Pettersson [KHP02] if f is a distribution and for more general discretization schemes). We recall that the error on the processes (called the *strong* error) is much easier to analyze than the one on the expectations (the *weak* error): the first one is essentially of order $\sqrt{h}$ (see [KP95]) but this is not relevant for the current issues.

Besides, the quantity of interest here has a more complex structure that is essentially $\mathbb{E}(f(X_T)H)$, where H is one of the random variables resulting from our computations. In general, H involves Itô or Skorohod integrals: our first purpose is to give some approximation procedure to simulate these weights using only the increments of the Brownian motion computed along the regular mesh with time step h.

Our second purpose is to analyze the error induced by this discretization procedure: generally speaking, the weak error is still at most linear w.r.t. h, as for $\mathbb{E}(f(X_T))$. The proofs are quite intricate and we postpone them to section 4. For the sake of clarity, we assume (R_∞), that is, b and σ of class C^∞, but approximation results only depend on a finite number of coefficients' derivatives.

Approximation procedure. We consider a regular mesh of the interval $[0,T]$, with N discretization times $t_i = ih$, where $h = T/N$ is the time step. Denote $\phi(t) = \sup\{t_i : t_i \le t\}$. The processes we need to simulate are essentially $(X_t)_{0\le t\le T}$, $(Y_t)_{0\le t\le T}$, $(Z_t)_{0\le t\le T}$, $(\dot X_t)_{0\le t\le T}$, and we approximate them using a standard Euler scheme as follows:

(3.1)
$$X_t^N = x + \int_0^t b(\phi(s), X_{\phi(s)}^N)\, ds + \sum_{j=1}^q \int_0^t \sigma_j(\phi(s), X_{\phi(s)}^N)\, dW_s^j,$$

(3.2)
$$Y_t^N = \mathrm{I}_d + \int_0^t b'(\phi(s), X_{\phi(s)}^N)\, Y_{\phi(s)}^N\, ds + \sum_{j=1}^q \int_0^t \sigma_j'(\phi(s), X_{\phi(s)}^N)\, Y_{\phi(s)}^N\, dW_s^j,$$

(3.3)
$$Z_t^N = \mathrm{I}_d - \int_0^t Z_{\phi(s)}^N (b' - \sum_{j=1}^q (\sigma_j')^2)(\phi(s), X_{\phi(s)}^N)\, ds - \sum_{j=1}^q \int_0^t Z_{\phi(s)}^N \sigma_j'(\phi(s), X_{\phi(s)}^N)\, dW_s^j,$$

(3.4)
$$\begin{aligned}\dot X_t^N = {} & \int_0^t \left(\dot b(\phi(s), X_{\phi(s)}^N) + b'(\phi(s), X_{\phi(s)}^N)\, \dot X_{\phi(s)}^N\right) ds \\ & + \sum_{j=1}^q \int_0^t \left(\dot\sigma_j(\phi(s), X_{\phi(s)}^N) + \sigma_j'(\phi(s), X_{\phi(s)}^N)\, \dot X_{\phi(s)}^N\right) dW_s^j.\end{aligned}$$

Note that only the increments $(W_{t_{i+1}}^j - W_{t_i}^j; 1 \le j \le q)_{0\le i\le N-1}$ of the Brownian motion are needed to get values of X^N, Z^N, Y^N, $\dot X^N$ at times $(t_i)_{0\le i\le N}$.

3.1. Pathwise approach.

THEOREM 3.1. *Assume* (R_∞). *Then, one has*

$$\left|\dot J(\alpha) - \mathbb{E}\left(f'(X_T^N)\dot X_T^N\right)\right| \le C(T,x,f)h,$$

under either one of the two following assumptions on f and X:

(A1) *f is of class C_b^4: one may put $C(T,x,f) = K(T,x)\sum_{1\le|\alpha|\le 4}\|\partial^\alpha f\|_\infty$ in that case.*

(A2) *f is continuously differentiable with a bounded gradient and the nondegeneracy condition* (E′) *holds: in that case, $C(T,x,f)$ may be set to*

$$K(T,x)\|f'\|_\infty \|1/\det(\gamma_T)\|_{\mathbf{L}^p}^q$$

for some positive numbers p and q.

Note that in the case (A1), only three additional derivatives of the function f' are required to get the order 1 w.r.t. h: this is a slight improvement compared to results in Talay and Tubaro [TL90], where four derivatives are needed.

3.2. Malliavin calculus approach.

3.2.1. Elliptic case. One needs to define the approximation for the random variable $H_T^{Mall.Ell.} := \delta\big([\sigma^{-1}(\cdot, X.)\ Y.\ Z_T\ \dot X_T]^*\big)$ involved in Proposition 2.5. Basic

algebra using the equality (2.2) gives

$$H_T^{Mall.Ell.} = \sum_{i=1}^{d} \delta\big([\sigma^{-1}(\cdot, X_\cdot)\, Y_\cdot]_i^* \, [Z_T \; \dot{X}_T]_i\big)$$
$$= \sum_{i=1}^{d} [Z_T \; \dot{X}_T]_i \int_0^T [\sigma^{-1}(s, X_s)\, Y_s]_i^* \, dW_s - \sum_{i=1}^{d} \int_0^T \mathcal{D}_s([Z_T \; \dot{X}_T]_i)[\sigma^{-1}(s, X_s)\, Y_s]_i \, ds.$$

The new quantities involved are $\mathcal{D}_s Z_{j,k,T}$ and $\mathcal{D}_s \dot{X}_{k,T}$. We now indicate how to simulate them. The $\mathbb{R}^{2d}$-valued process $\binom{X_t}{\dot{X}_t}_{t\geq 0}$ forms a new stochastic differential equation (see (1.5)): we denote the flow of this extended system by $\hat{Y}_t$ and its inverse by $\hat{Z}_t$. As we did for Y_t and Z_t, we can define their Euler scheme (as in (3.2) and (3.3)), which we denote $\hat{Y}_t^N$ and $\hat{Z}_t^N$. The Malliavin derivative of this system follows from (2.3). Hence, one has

$$\mathcal{D}_s \dot{X}_T = \Pi_d^R \left(\hat{Y_T} \hat{Z}_s \begin{pmatrix} \vdots & \sigma_j(s, X_s) & \vdots \\ \vdots & \dot{\sigma}_j(s, X_s) + \sigma_j'(s, X_s)\, \dot{X}_s & \vdots \end{pmatrix} \right), \tag{3.5}$$

and we naturally approximate it by

$$[\mathcal{D}_s \dot{X}_T]^N = \Pi_d^R \left(\hat{Y}_T^N \hat{Z}_s^N \begin{pmatrix} \vdots & \sigma_j(s, X_s^N) & \vdots \\ \vdots & \dot{\sigma}_j(s, X_s^N) + \sigma_j'(s, X_s^N)\, \dot{X}_s^N & \vdots \end{pmatrix} \right). \tag{3.6}$$

The same approach can be developed for the cth column of the transpose of Z_T, since $\binom{X_t}{(Z_t{}^*)_c}_{t\geq 0}$ forms a new SDE (see (1.4)): the associated flow and its inverse, respectively denoted $\hat{Y}_t^c$ and $\hat{Z}_t^c$, enable us to derive a simple expression for $\mathcal{D}_s[(Z_t{}^*)_c]$ analogously to (3.5) and (3.6). As a consequence, one gets

$$\mathcal{D}_s([Z_T \; \dot{X}_T]_i) = \mathbf{1}_{s\leq T} \sum_j A_{\beta(j,i),T} B_{\beta(j,i),s}, \tag{3.7}$$

where $A_{\beta(j,i),T}$ and $B_{\beta(j,i),s}$ are given by some appropriate coordinates of the processes $\hat{Y}_T$, $(\hat{Y}_T^c)_{1\leq c\leq d}$ on one hand; and $\hat{Z}_s$, $(\hat{Z}_s^c)_{1\leq c\leq d}$, $\sigma_j(s, X_s)$, $\dot{\sigma}_j(s, X_s)$, $\sigma_j'(s, X_s)$, $\dot{X}_s$, Z_s on the other hand; in order to keep things clear, we do not develop their expression further (we refer to a technical report [GM03] for full details). Finally, we approximate $H_T^{Mall.Ell.}$ by

$$H_T^{Mall.Ell.,N} = \sum_{i=1}^{d} [Z_T^N \; \dot{X}_T^N]_i \int_0^T [\sigma^{-1}(\phi(s), X_{\phi(s)}^N)\, Y_{\phi(s)}^N]_i^* \, dW_s$$
$$- \sum_{i=1}^{d} \int_0^T \left(\sum_j A_{\beta(j,i),T}^N B_{\beta(j,i),\phi(s)}^N \right) [\sigma^{-1}(\phi(s), X_{\phi(s)}^N)\, Y_{\phi(s)}^N]_i \, ds,$$

which can be simulated using only the Brownian increments as before. We now state that the approximation above converges at order 1 w.r.t. the time step.

THEOREM 3.2. *Assume* (R_∞), (E), *and* (H). *For some* $q \geq 0$, *one has*

$$\left| \dot{J}(\alpha) - \mathbb{E}\left(f(X_T^N) H_T^{Mall.Ell.,N} \right) \right| \leq K(T, x) \frac{\|f\|_\infty}{T^q} h.$$

Remark 3.1. Instead of basing the computations of Malliavin derivatives for different adapted processes U, $(\mathcal{D}_{t_i}U_{t_j})_{0\le i\le j\le N}$, on the equality (2.3), an alternative approach would be to derive equations solved by $(\mathcal{D}_{t_i}U_t)_{t_i\le t\le T}$ and then approximate them with a discretization scheme (for each t_i). However, this approach would require essentially $O(N^2)$ operations, instead of $O(N)$ in our case.

3.2.2. General nondegenerate case. Denote by $0_{d_1,d_2}$ the $d_1\times d_2$ matrix with 0 for each element. Simple algebra yields that $\dot{V}_T^*\gamma_T^{-1}\mathcal{D}_sV_T$ is equal to

$$\dot{V}_T^*\gamma_T^{-1}\Pi_r^R(Y_TZ_s\sigma(s,X_s)) = (0_{1,d-r}\dot{V}_T^*)\begin{pmatrix} 0_{d-r,d-r} & 0_{d-r,r} \\ 0_{r,d-r} & \gamma_T^{-1}\end{pmatrix} Y_TZ_s\sigma(s,X_s)$$
$$= \sum_{i=1}^d F_i[(Z_s\sigma(s,X_s))^*]_i,$$

where

$$F_i = \left(Y_T^*\begin{pmatrix} 0_{d-r,d-r} & 0_{d-r,r} \\ 0_{r,d-r} & \gamma_T^{-1}\end{pmatrix}\begin{pmatrix} 0_{d-r,1} \\ \dot{V}_T\end{pmatrix}\right)_i = \sum_j U_{\kappa(i,j),T}(\gamma_T^{-1})_{\beta(i,j),\gamma(i,j)},$$

with the random variables $(U_{\kappa(i,j),T})_{i,j}$ being expressed as a product of coordinates of Y_T and $\dot{V}_T$. As before, we do not develop their expression to keep the formulae easy to manipulate, and we refer to [GM03] for more details.

Hence, the random variable of interest in Proposition 2.8, i.e., $H_T^{Mall.Gen.}$, is

$$\delta\left(\dot{V}_T^*\gamma_T^{-1}\mathcal{D}_{\cdot}V_T\right) = \sum_{i=1}^d F_i\int_0^T[(Z_s\sigma(s,X_s))^*]_i^*dW_s - \sum_{i=1}^d\int_0^T\mathcal{D}_sF_i[(Z_s\sigma(s,X_s))^*]_ids.$$

By the chain rule, the Malliavin derivative of F_i is related to that of $U_{\kappa(i,j),T}$ (i.e., coordinates of Y_T and $\dot{V}_T$) and that of $(\gamma_T^{-1})_{\beta(i,j),\gamma(i,j)}$: the latter can be expressed in terms of γ_T^{-1} and $\mathcal{D}_s\gamma_T$ (see Lemma 2.1.6 in Nualart [Nua95, p. 89]) and we obtain

$$H_T^{Mall.Gen.} = \sum_{i,j}U_{\kappa(i,j),T}(\gamma_T^{-1})_{\beta(i,j),\gamma(i,j)}\int_0^T[(Z_s\sigma(s,X_s))^*]_i^*dW_s \tag{3.8}$$

$$-\sum_{i,j}(\gamma_T^{-1})_{\beta(i,j),\gamma(i,j)}\int_0^T\mathcal{D}_sU_{\kappa(i,j),T}[(Z_s\sigma(s,X_s))^*]_ids \tag{3.9}$$

$$+\sum_{i,j,k,l}U_{\kappa(i,j),T}(\gamma_T^{-1})_{\beta(i,j),k}(\gamma_T^{-1})_{l,\gamma(i,j)}\int_0^T\mathcal{D}_s(\gamma_{k,l,T})[(Z_s\sigma(s,X_s))^*]_ids. \tag{3.10}$$

Analogously to the elliptic case, the integrals above may be discretized. Furthermore, the random variables $U_{\kappa(i,j),T}$ may be approximated by $U^N_{\kappa(i,j),T}$, defined by the same product of coordinates of Y_T^N and $\dot{V}_T^N$ as the one defining $U_{\kappa(i,j),T}$. Its weak derivative can be computed as in (3.7): indeed, with the same arguments, one may prove that

$$\mathcal{D}_sU_{\kappa(i,j),T} = \mathbf{1}_{s\le T}\sum_k\hat{U}_{\kappa(i,j,k),T}\check{U}_{\beta(i,j,k),s}, \tag{3.11}$$

where $(\hat{U}_{\kappa(i,j,k),T})_{i,j,k}$ (resp., $(\check{U}_{\beta(i,j,k),s})_{i,j,k}$) are appropriate real values (resp., vectors) at time T (resp., at time s) of some extended systems of SDEs. Then, the

natural approximation is

$$[\mathcal{D}_s U_{\kappa(i,j),T}]^N = \mathbf{1}_{s\le T} \sum_k \hat{U}^N_{\kappa(i,j,k),T} \check{U}^N_{\beta(i,j,k),s}. \tag{3.12}$$

Actually, what differs from the elliptic case are the Malliavin covariance matrix γ_T and its weak derivative. Even if $\gamma_T = \int_0^T \Pi^R_r(Y_T Z_s \sigma(s, X_s))[\Pi^R_r(Y_T Z_s \sigma(s, X_s))]^* ds$ is almost surely invertible with an inverse in any $\mathbf{L}^p$, a naive approximation may not satisfy these invertibility properties: for this reason, we add a small perturbation in its discretization as follows:

$$\gamma^N_T = \int_0^T \Pi^R_r(Y^N_T Z^N_{\phi(s)} \sigma(\phi(s), X^N_{\phi(s)}))[\Pi^R_r(Y^N_T Z^N_{\phi(s)} \sigma(\phi(s), X^N_{\phi(s)}))]^* ds + \frac{T}{N} \mathrm{I}_d. \tag{3.13}$$

This allows us to state the following result.

LEMMA 3.3. *Assume* (R_∞) *and* (E'). *Then, for any* $p \ge 1$, *one has for some positive numbers* p_1 *and* q_1: $\|1/\det(\gamma^N_T)\|_{\mathbf{L}^p} \le K(T,x) \|1/\det(\gamma_T)\|^{q_1}_{\mathbf{L}^{p_1}}$ *with a constant* $K(T,x)$ *independent of* N.

Proof. It is easy to check that $\|\gamma^N_T - \gamma_T\|_{\mathbf{L}^p} \le K(T,x)\sqrt{h}$ (use Lemma 4.3 below). Moreover, the eigenvalues of γ^N_T are all greater than h; hence $\det(\gamma^N_T) \ge h^r$, and one deduces

$$\begin{aligned}
\mathbb{E}(\det(\gamma^N_T)^{-p}) &= \mathbb{E}\left(\det(\gamma^N_T)^{-p} \mathbf{1}_{\det(\gamma^N_T) \le \frac{1}{2}\det(\gamma_T)}\right) + \mathbb{E}\left(\det(\gamma^N_T)^{-p} \mathbf{1}_{\det(\gamma^N_T) > \frac{1}{2}\det(\gamma_T)}\right) \\
&\le h^{-rp} \mathbb{P}\left(\frac{\det(\gamma_T) - \det(\gamma^N_T)}{\det(\gamma_T)} \ge \frac{1}{2}\right) + 2^p \mathbb{E}(\det(\gamma_T)^{-p}) \\
&\le h^{-rp} 2^q \| \, |\det(\gamma_T) - \det(\gamma^N_T)|^q \|_{\mathbf{L}^{p_1}} \|\det(\gamma_T)^{-q}\|_{\mathbf{L}^{p_2}} + 2^p \mathbb{E}(\det(\gamma_T)^{-p})
\end{aligned}$$

where p_1 and p_2 are conjugate numbers. Take $q = 2rp$ to get the result. ∎

To deal with the weak derivative of γ_T, one needs to rewrite

$$\gamma_{k,l,T} = \sum_{i'} A_{\epsilon(k,l,i'),T} \int_0^T B_{\eta(k,l,i'),u} du,$$

where $A_{\epsilon(k,l,i'),T}$ (resp., $B_{\eta(k,l,i'),u}$) are products of coordinates of Y_T (resp., Z_u and $\sigma(u, X_u)$). As for (3.7), the Malliavin derivative of $A_{\epsilon(k,l,i'),T}$ and $B_{\eta(k,l,i'),u}$ can be expressed as

$$\begin{aligned}
\mathcal{D}_s A_{\epsilon(k,l,i'),T} &= \mathbf{1}_{s\le T} \sum_{j'} C_{\epsilon(k,l,i',j'),T} D_{\epsilon(k,l,i',j'),s}, \\
\mathcal{D}_s B_{\eta(k,l,i'),u} &= \mathbf{1}_{s\le u} \sum_{j'} E_{\eta(k,l,i',j'),u} F_{\eta(k,l,i',j'),s}.
\end{aligned}$$

Hence, for $s \le T$, one has

$$\begin{aligned}
\mathcal{D}_s \gamma_{k,l,T} = &\sum_{i',j'} C_{\epsilon(k,l,i',j'),T} \left(\int_0^T B_{\eta(k,l,i'),u} du\right) D_{\epsilon(k,l,i',j'),s} \\
&+ \sum_{i',j'} A_{\epsilon(k,l,i'),T} F_{\eta(k,l,i',j'),s} \int_s^T E_{\eta(k,l,i',j'),u} du,
\end{aligned} \tag{3.14}$$

which can be approximated by

$$[\mathcal{D}_s\gamma_{k,l,T}]^N = \sum_{i',j'} C^N_{\epsilon(k,l,i',j'),T}\left(\int_0^T B^N_{\eta(k,l,i'),\phi(u)}du\right) D^N_{\epsilon(k,l,i',j'),s}$$

$$+\sum_{i',j'} A^N_{\epsilon(k,l,i'),T}\; F^N_{\eta(k,l,i',j'),s}\int_s^T E^N_{\eta(k,l,i',j'),\phi(u)}du. \tag{3.15}$$

We now turn to the global approximation of the weight $H_T^{Mall.Gen.}$:

(3.16)

$$H_T^{Mall.Gen.,N} = \sum_{i,j} U^N_{\kappa(i,j),T}[(\gamma_T^N)^{-1}]_{\beta(i,j),\gamma(i,j)}\int_0^T [(Z^N_{\phi(s)}\sigma(\phi(s),X^N_{\phi(s)}))^*]^*_i dW_s$$

$$-\sum_{i,j}[(\gamma_T^N)^{-1}]_{\beta(i,j),\gamma(i,j)}\int_0^T [\mathcal{D}_{\phi(s)}U_{\kappa(i,j),T}]^N[(Z^N_{\phi(s)}\sigma(\phi(s),X^N_{\phi(s)}))^*]_i ds \tag{3.17}$$

$$+\sum_{i,j,k,l} U^N_{\kappa(i,j),T}[(\gamma_T^N)^{-1}]_{\beta(i,j),k}[(\gamma_T^N)^{-1}]_{l,\gamma(i,j)}$$

$$\int_0^T [\mathcal{D}_{\phi(s)}(\gamma_{k,l,T})]^N[(Z^N_{\phi(s)}\sigma(\phi(s),X^N_{\phi(s)}))^*]_i ds. \tag{3.18}$$

We are now in a position to state the following approximation result.

THEOREM 3.4. *Assume* (R_∞), (E′), *and* (H). *For some positive numbers* p *and* q, *one has:*

$$\left|\dot{J}(\alpha) - \mathbb{E}(f(V_T^N)H_T^{Mall.Gen.,N})\right| \le K(T,x)\|f\|_\infty \|1/\det(\gamma_T)\|^q_{\mathbf{L}^p} h.$$

In the hypoelliptic case (case 1) in Proposition 2.7), note that the weak approximation result above holds true under a nondegeneracy condition stated only at the initial point $(0, X_0)$, which is a significant improvement compared to [BT96a] (or more recently in [TZ04]), where the condition was stated in the whole space.

3.3. Adjoint approach. To approximate $H_T^{b,Adj.}$ and $H_T^{\sigma,Adj.}$ from Theorem 2.11, we propose the following natural estimates:

(3.19)

$$H_T^{b,Adj.,N} = f(X_T^N)h\sum_{k=0}^{N-1}\dot{b}(t_k,X^N_{t_k})\cdot\frac{Z^{N*}_{t_k}}{T-t_k}\int_{t_k}^T [\sigma^{-1}(\phi(s),X^N_{\phi(s)})\; Y^N_{\phi(s)}]^*\; dW_s,$$

$$H_T^{\sigma,Adj.,N} = h\sum_{k=0}^{N-1}\sum_{i,j=1}^d [\sigma\dot{\sigma}^*]_{i,j}(t_k,X^N_{t_k})[f(X_T^N)-f(X^N_{t_k})]$$

$$\times\left(\frac{2e^j}{T-t_k}\cdot\left[Z^{N*}_{t_k}\int_{\phi(\frac{T+t_k}{2})}^T [\sigma^{-1}(\phi(s),X^N_{\phi(s)})Y^N_{\phi(s)}]^* dW_s\right]\right.$$

$$\left.\times\frac{2e^i}{T-t_k}\cdot\left[Z^{N*}_{t_k}\int_{t_k}^{\phi(\frac{T+t_k}{2})} [\sigma^{-1}(\phi(s),X^N_{\phi(s)})Y^N_{\phi(s)}]^* dW_s\right]\right.$$

$$\left.+\frac{2e^i}{T-t_k}\cdot\left\{\nabla_x\left[Z^{N*}_{t_k}\int_{t_k}^{\phi(\frac{T+t_k}{2})} [\sigma^{-1}(\phi(s),X^N_{\phi(s)})Y^N_{\phi(s)}]^* dW_s\right]Z^N_{t_k}e^j\right\}\right). \tag{3.20}$$

Derivatives $\nabla_x Y^N_{\phi(s)}$ and $\nabla_x Z^N_{t_k}$ are obtained by a direct differentiation in (3.2) and (3.3): we do not make the equations explicit; they coincide with those of the Euler procedure applied to $\nabla_x Y_t$ and $\nabla_x Z_t$ defined in (1.3) and (1.4).

These approximations also induce a discretization error in the computation of $\dot{J}(\alpha)$ of order 1 w.r.t. h.

THEOREM 3.5. *Assume* (R_∞), (E), *and* (H). *For some* $p \geq 0$, *one has*

$$\left| \dot{J}(\alpha) - \mathbb{E}\left(H_T^{b,Adj.,N} + H_T^{\sigma,Adj.,N} \right) \right| \leq K(T,x) \frac{\|f\|_\infty}{T^p} h.$$

The proof is postponed to section 4.4.

3.4. Martingale approach. The natural approximation of $H_T^{Mart.}$ defined in Theorem 2.12 may be given by

$$H_T^{Mart.,N} = \frac{f(X_T^N)}{T} \int_0^T [\sigma^{-1}(\phi(s), X^N_{\phi(s)}) \dot{X}^N_{\phi(s)}]^* dW_s + \int_0^T dr \, \frac{[f(X_T^N) - f(X^N_{\phi(r)})]}{(T-\phi(r))^2}$$

$$\times \int_{\phi(r)}^T [\sigma^{-1}(\phi(s), X^N_{\phi(s)})(\dot{X}^N_{\phi(s)} - Y^N_{\phi(s)} Z^N_{\phi(r)} \dot{X}^N_{\phi(r)})]^* dW_s.$$

Unfortunately, we have not been able to analyze the approximation error $\dot{J}(\alpha) - \mathbb{E}(H_T^{Mart.,N})$ under the fairly general assumption (H′). Indeed, an immediate issue to handle would be to quantify the quality of the approximation of $\int_0^T dr \, \mathbb{E}\big(\frac{[f(X_T) - f(X_r)]}{(T-r)^2} \int_r^T [\sigma_s^{-1}(\dot{X}_s - Y_s Z_r \dot{X}_r)]^* dW_s \big)$ by its Riemann sum, which seems to be far from obvious under (H′).

4. Proof of the results on the discretization error analysis. This section is devoted to the proof of section 3's theorems analyzing the discretization error.

The trick to prove these estimates for $\mathbb{E}(f(X_T))$ usually relies on the Markov property: one decomposes the error using the PDE solved by the function $(t,x) \mapsto \mathbb{E}(f(X^x_{T-t}))$ (see Bally and Talay [BT96a]), but this makes no sense in our situation. Another way to proceed consists in cleverly using the duality relationship (2.1) with some stochastic expansion to get the right order (see Kohatsu-Higa [KH01] or [KHP02]). During the revision of this work, Kohatsu-Higa brought to our attention another paper [KHP00] where the approximation of some smooth functionals of SDEs is successfully analyzed in this way. Here, we also adopt this approach. However, the functionals of interest are much more complex. Moreover, extra technicalities compared to [KHP02] are required, because of the necessity for our Malliavin calculus computations to introduce a localization factor $\psi_T^{N,\epsilon}$.

To clarify the arguments, we first state a quite general result, whose statement enables us to reduce the proof of our theorems to check that a stochastic expansion holds true.

4.1. A more general result. By convention, we set $dW_s^0 = ds$. First, we need to define some particular forms of stochastic expansions.

DEFINITION 4.1. *The real random variable U_T (which in general depends on N) satisfies property $(\mathcal{P})$ if it can be written as*

$$U_T = \sum_{i,j=0}^{q} c_{i,j}^{U,0}(T) \int_0^T c_{i,j}^{U,1}(t) \left(\int_{\phi(t)}^t c_{i,j}^{U,2}(s) dW_s^i \right) dW_t^j + \sum_{i,j,k=0}^{q} c_{i,j,k}^{U,0}(T) \int_0^T c_{i,j,k}^{U,1}(t) \left[\int_0^t c_{i,j,k}^{U,2}(s) \left(\int_{\phi(s)}^s c_{i,j,k}^{U,3}(u) dW_u^i \right) dW_s^j \right] dW_t^k$$

for some adapted processes $\{(c_{i,j}^{U,i_1}(t), c_{i,j,k}^{U,i_2}(t))_{t\geq 0}\colon\ 0 \leq i,j,k \leq q, 0 \leq i_1 \leq 2, 0 \leq i_2 \leq 3\}$ (possibly depending on N) and if, for each $t \in [0,T]$, they belong to $\mathbb{D}^\infty$ with Sobolev norms satisfying $\sup_{N,t\in[0,T]} \left(\|c_{i,j}^{U,i_1}(t)\|_{k',p} + \|c_{i,j,k}^{U,i_2}(t)\|_{k',p} \right) < \infty$ for any $k', p \geq 1$.

THEOREM 4.2. *Assume* (R_∞) *and that $H_T^N - H_T$ satisfies property $(\mathcal{P})$. Then,*

1. *if f is of class C_b^3, one has*

$$\left| \mathbb{E}(f(V_T)H_T) - \mathbb{E}(f(V_T^N)H_T^N) \right| \leq K(T,x) \Big(\sum_{0\leq|\alpha|\leq 3} \|\partial^\alpha f\|_\infty \Big) h;$$

2. *under* (E') *and* (H), *one has*

$$\left| \mathbb{E}(f(V_T)H_T) - \mathbb{E}(f(V_T^N)H_T^N) \right| \leq K(T,x)\ \|f\|_\infty\ \|1/\det(\gamma_T)\|_{\mathbf{L}^p}^q\ h.$$

In the statement above, $(V_t)_{0\leq t\leq T}$ corresponds to some coordinates of a SDE $(X_t)_{0\leq t\leq T}$ as it is defined in (2.5), but we can also simply consider $V = X$.

Theorem 4.2 is proved at the end of this section, and for a while, we focus on its applications to derive the announced results about the discretization errors. Remember that the approximation of the weights H is essentially based on an Euler scheme applied to a system of SDEs. For this reason, the verification of property $(\mathcal{P})$ is tightly connected to the decomposition of the error, between a Brownian SDE and its Euler approximation, in terms of a stochastic expansion. This is the purpose of the following standard lemma (for more general driven semimartingales, see Jacod and Protter [JP98]).

LEMMA 4.3. *Consider a general d'-dimensional SDE $(\bar{X}_t)_{t\geq 0}$ defined by C^∞ coefficients with bounded derivatives, and $(\bar{X}_t^N)_{t\geq 0}$ its Euler approximation:*

$$\bar{X}_t = x + \int_0^t \bar{b}(s, \bar{X}_s)\ ds + \sum_{j=1}^q \int_0^t \bar{\sigma}_j(s, \bar{X}_s)\ dW_s^j,$$

$$\bar{X}_t^N = x + \int_0^t \bar{b}(\phi(s), \bar{X}_{\phi(s)}^N)\ ds + \sum_{j=1}^q \int_0^t \bar{\sigma}_j(\phi(s), \bar{X}_{\phi(s)}^N)\ dW_s^j.$$

Then, for each t, each component of $\bar{X}_t - \bar{X}_t^N$ satisfies $(\mathcal{P})$. Namely, for $1 \leq k \leq d'$, one has

$$\bar{X}_{k,t} - \bar{X}_{k,t}^N = \sum_{i,j=0}^q c_{i,j,k}^{\bar{X},0}(t) \int_0^t c_{i,j,k}^{\bar{X},1}(s) \Big(\int_{\phi(s)}^s c_{i,j,k}^{\bar{X},2}(u) dW_u^i \Big) dW_s^j$$

for some adapted processes $\{(c_{i,j,k}^{\bar{X},i_1}(t))_{t\geq 0} : 0 \leq i,j \leq q, 1 \leq k \leq d', 0 \leq i_1 \leq 2\}$ satisfying $\sup_{N,t\in[0,T]} \|c_{i,j,k}^{\bar{X},i_1}(t)\|_{k',p} < \infty$ for any $k', p \geq 1$.

Proof. One has $\bar{X}_t - \bar{X}_t^N = \int_0^t \bar{b}'(s)(\bar{X}_s - \bar{X}_s^N)\,ds + \sum_{j=1}^q \int_0^t \bar{\sigma}_j'(s)(\bar{X}_s - \bar{X}_s^N)\,dW_s^j + \int_0^t [\bar{b}(s,\bar{X}_s^N) - \bar{b}(\phi(s),\bar{X}_{\phi(s)}^N)]\,ds + \sum_{j=1}^q \int_0^t [\bar{\sigma}_j(s,\bar{X}_s^N) - \bar{\sigma}_j(\phi(s),\bar{X}_{\phi(s)}^N)]\,dW_s^j$ with $a'(s) = \int_0^1 \nabla_x a(s, \bar{X}_s^N + \lambda(\bar{X}_s - \bar{X}_s^N))d\lambda$ for $a = \bar{b}$ or $a = \bar{\sigma}_j$. Now, consider the unique solution of the linear equation $\mathcal{E}_t = \mathrm{I}_d + \int_0^t \bar{b}'(s)\mathcal{E}_s\,ds + \sum_{j=1}^q \int_0^t \bar{\sigma}_j'(s)\mathcal{E}_s\,dW_s^j$. From Theorem 56 p. 271 in Protter [Pro90], one deduces that

$$\begin{aligned}\bar{X}_t - \bar{X}_t^N = \mathcal{E}_t \int_0^t {\mathcal{E}_s}^{-1} \Bigg\{ & [\bar{b}(s,\bar{X}_s^N) - \bar{b}(\phi(s),\bar{X}_{\phi(s)}^N)] \\ & - \sum_{j=1}^q \bar{\sigma}_j'(s)[\bar{\sigma}_j(s,\bar{X}_s^N) - \bar{\sigma}_j(\phi(s),\bar{X}_{\phi(s)}^N)] \Bigg\} ds \\ + \sum_{j=1}^q \mathcal{E}_t \int_0^t {\mathcal{E}_s}^{-1} & [\bar{\sigma}_j(s,\bar{X}_s^N) - \bar{\sigma}_j(\phi(s),\bar{X}_{\phi(s)}^N)]\,dW_s^j;\end{aligned}$$

then, by applying Itô's formula between $\phi(s)$ and s, we can easily complete the proof of Lemma 4.3. □

4.2. Proof of Theorem 3.4 (general nondegenerate case). Owing to Theorem 4.2, we only have to prove that $H_T^{Mall.Gen.} - H_T^{Mall.Gen.,N}$ satisfies property $(\mathcal{P})$. Thus, it is enough to separately look at each factor in $H_T^{Mall.Gen.}$ and $H_T^{Mall.Gen.,N}$, by proving that their difference is of the form $c_{i,j}^{U,0}(T)\int_0^T c_{i,j}^{U,1}(t)\big(\int_{\phi(t)}^t c_{i,j}^{U,2}(s)dW_s^i\big)dW_t^j$ or $c_{i,j,k}^{U,0}(T)\int_0^T c_{i,j,k}^{U,1}(t)\big[\int_0^t c_{i,j,k}^{U,2}(s)\big(\int_{\phi(s)}^s c_{i,j,k}^{U,3}(u)dW_u^i\big)dW_s^j\big]dW_t^k$, while the other factors just belong to $\mathbb{D}^\infty$ with uniformly bounded Sobolev norms.

(a) The fact that the difference $U_{\kappa(i,j),T} - U_{\kappa(i,j),T}^N$ (involved in (3.8), (3.10), (3.16), and (3.18)) satisfies $(\mathcal{P})$ can be derived from an application of Lemma 4.3 by noticing that $U_{\kappa(i,j),T}$ is the product of coordinates of Y_T and $\dot{V}_T$.

(b) Using the expressions of γ_T and γ_T^N, one gets $\gamma_{k,l,T} - \gamma_{k,l,T}^N = -\delta_{k,l}h + \mathcal{E}_{3,1,k,l} + \mathcal{E}_{3,2,k,l}$ with

$$\begin{aligned}\mathcal{E}_{3,1,k,l} &= \int_0^T \big[\Pi_r^R(Y_T Z_s \sigma(s,X_s))[\Pi_r^R(Y_T Z_s \sigma(s,X_s))]^*\big]_{k,l}\,ds \\ &\quad - \int_0^T \big[\Pi_r^R(Y_T^N Z_s^N \sigma(s,X_s^N))[\Pi_r^R(Y_T^N Z_s^N \sigma(s,X_s^N))]^*\big]_{k,l}\,ds, \\ \mathcal{E}_{3,2,k,l} &= \int_0^T \big[\Pi_r^R(Y_T^N Z_s^N \sigma(s,X_s^N))[\Pi_r^R(Y_T^N Z_s^N \sigma(s,X_s^N))]^*\big]_{k,l}\,ds \\ &\quad - \int_0^T \Big[\Pi_r^R(Y_T^N Z_{\phi(s)}^N \sigma(\phi(s),X_{\phi(s)}^N))[\Pi_r^R(Y_T^N Z_{\phi(s)}^N \sigma(\phi(s),X_{\phi(s)}^N))]^*\Big]_{k,l}\,ds.\end{aligned}$$

Using Lemma 4.3 and the relation $a(s,X_s) - a(s,X_s^N) = a'(s)(X_s - X_s^N)$ with $a'(s) = \int_0^1 \nabla_x a(s, X_s^N + \lambda(X_s - X_s^N))d\lambda$ available for smooth functions a, it is straightforward to see that $\mathcal{E}_{3,1,k,l}$ can be written as a sum of terms satisfying $(\mathcal{P})$. The same conclusion holds for $\mathcal{E}_{3,2,k,l}$ if we apply Itô's formula between $\phi(s)$ and s.

Finally, as $1/\det(\gamma_T)$ and $1/\det(\gamma_T^N)$ belong to any $\mathbf{L}^p$ $(p \geq 1)$ according to Lemma 3.3, it follows that the difference $[\gamma_T^{-1}]_{k',l'} - [(\gamma_T^N)^{-1}]_{k',l'}$ (involved in (3.8), (3.9), (3.10), (3.16), (3.17), and (3.18)) satisfies $(\mathcal{P})$.

(c) Concerning (3.8) and (3.16), the difference $\int_0^T [(Z_s\sigma(s,X_s))^*]_i^* dW_s - \int_0^T [(Z^N_{\phi(s)}\sigma(\phi(s),X^N_{\phi(s)}))^*]_i^* dW_s$ is equal to a sum of two terms:

$$\int_0^T [(Z_s\sigma(s,X_s))^*]_i^* dW_s - \int_0^T [(Z^N_s\sigma(s,X^N_s))^*]_i^* dW_s,$$

$$\int_0^T [(Z^N_s\sigma(s,X^N_s))^*]_i^* dW_s - \int_0^T [(Z^N_{\phi(s)}\sigma(\phi(s),X^N_{\phi(s)}))^*]_i^* dW_s.$$

It is straightforward to check that both contributions satisfy $(\mathcal{P})$, the first one because of Lemma 4.3 and the second one as an application of Itô's formula.

(d) The approximation error between terms (3.9) and (3.17) also comes from the difference $\int_0^T \mathcal{D}_s U_{\kappa(i,j),T}[(Z_s\sigma(s,X_s))^*]_i ds - \int_0^T [\mathcal{D}_{\phi(s)}U_{\kappa(i,j),T}]^N[(Z^N_{\phi(s)}\sigma(\phi(s),X^N_{\phi(s)}))^*]_i ds := \mathcal{E}_{4,1,i,j} + \mathcal{E}_{4,2,i,j}$, where

$$\begin{aligned}\mathcal{E}_{4,1,i,j} &= \int_0^T \mathcal{D}_s U_{\kappa(i,j),T}[(Z_s\sigma(s,X_s))^*]_i ds \\ &\qquad - \int_0^T [\mathcal{D}_s U_{\kappa(i,j),T}]^N[(Z^N_s\sigma(s,X^N_s))^*]_i ds, \\ \mathcal{E}_{4,2,i,j} &= \int_0^T [\mathcal{D}_s U_{\kappa(i,j),T}]^N[(Z^N_s\sigma(s,X^N_s))^*]_i ds \\ &\qquad - \int_0^T [\mathcal{D}_{\phi(s)}U_{\kappa(i,j),T}]^N[(Z^N_{\phi(s)}\sigma(\phi(s),X^N_{\phi(s)}))^*]_i ds.\end{aligned}$$

The error induced by the approximation between $Z_s\sigma(s,X_s)$, $Z^N_s\sigma(s,X^N_s)$, and $Z^N_{\phi(s)}\sigma(\phi(s),X^N_{\phi(s)})$ can be handled as before using Lemma 4.3 and Itô's formula. To deal with $\mathcal{D}_s U_{\kappa(i,j),T}$, $[\mathcal{D}_s U_{\kappa(i,j),T}]^N$ and $[\mathcal{D}_{\phi(s)}U_{\kappa(i,j),T}]^N$, we may recall their particular forms given by equations (3.11) and (3.12). Again, Lemma 4.3 applies to the extended systems which help in defining $\mathcal{D}_s U_{\kappa(i,j),T}$. This provides a contribution error equal to a sum of terms satisfying $(\mathcal{P})$.

(e) The difference $\int_0^T \mathcal{D}_s(\gamma_{k,l,T})[(Z_s\sigma(s,X_s))^*]_i ds - \int_0^T [\mathcal{D}_{\phi(s)}(\gamma_{k,l,T})]^N [(Z^N_{\phi(s)}\sigma(\phi(s),X^N_{\phi(s)}))^*]_i ds$ coming from (3.10) and (3.18) can be analyzed with the same arguments as before, if we take into account the specific form of the derivative $\mathcal{D}_s(\gamma_{k,l,T})$ and its approximation given by (3.14) and (3.15).

The proof of Theorem 3.4 is complete.

4.3. Proof of Theorems 3.1 (pathwise approach) and 3.2 (elliptic case).

Proof of Theorem 3.1. By an application of Theorem 4.2, it is enough to check that $\dot{X}_T - \dot{X}^N_T$ satisfies $(\mathcal{P})$, which is actually a direct consequence of Lemma 4.3.

Proof of Theorem 3.2. As for Theorem 3.4, we can check that $H_T^{Mall.Ell.} - H_T^{Mall.Ell.,N}$ satisfies $(\mathcal{P})$. Thus, Theorem 4.2 with $V_T = X_T$ and $V^N_T = X^N_T$ yields $\left|\dot{J}(\alpha) - \mathbb{E}(f(X^N_T)H_T^{Mall.Ell.,N})\right| \le \frac{K(T,x)}{T}\ \|f\|_\infty\ \|1/\det(\gamma_T)\|^q_{\mathbf{L}^p}\ h$, for some positive numbers p and q. Invoking the following well-known upper bound (see Theorem 3.5 in [KS84]) $\|1/\det(\gamma_T)\|_{\mathbf{L}^p} \le K(T,x)/T^d$ completes the estimate given in Theorem 3.2.

4.4. Theorem 3.5 (adjoint approach). The first approximation which is easy to justify is the time discretization of the integral involved in Lemma 2.9. For this, note that the function $t \mapsto \mathbb{E}(\sum_{i=1}^d \dot{b}_i(t,X_t)\partial_{x_i}u(t,X_t) + \frac{1}{2}\sum_{i,j=1}^d [\sigma\dot{\sigma}^*]_{i,j}(t,X_t)\partial^2_{x_i,x_j}u(t,X_t))$

is of class $C_b^1([0,T],\mathbb{R})$: indeed, it is a smooth function in particular because u is, and its derivatives are uniformly bounded thanks to estimates of type (2.6). Hence, it remains to prove the following upper bounds, uniformly in i,j:

(4.1)
$$\left|\mathbb{E}\left(f(X_T)\dot{b}(t_k,X_{t_k})\cdot {Z_{t_k}}^* \int_{t_k}^T [\sigma^{-1}(s,X_s)\, Y_s]^*\, dW_s - f(X_T^N)\dot{b}(t_k,X_{t_k}^N)\cdot Z_{t_k}^{N*}\right.\right.$$
$$\left.\left.\times \int_{t_k}^T [\sigma^{-1}(\phi(s),X_{\phi(s)}^N)\, Y_{\phi(s)}^N]^*\, dW_s\right)\right| \le K(T,x)\frac{\|f\|_\infty}{T^q}(T-t_k)h,$$

(4.2)
$$\left|\mathbb{E}\left([\sigma\dot{\sigma}^*]_{i,j}(t_k,X_{t_k})f(X_T)e^j\cdot\left[{Z_{t_k}}^*\int_{\frac{T+t_k}{2}}^T[\sigma^{-1}(s,X_s)Y_s]^*dW_s\right]\right.\right.$$
$$\times\, e^i\cdot\left[{Z_{t_k}}^*\int_{t_k}^{\frac{T+t_k}{2}}[\sigma^{-1}(s,X_s)Y_s]^*dW_s\right] - [\sigma\dot{\sigma}^*]_{i,j}(t_k,X_{t_k}^N)f(X_T^N)$$
$$\times\, e^j\cdot\left[Z_{t_k}^{N*}\int_{\phi(\frac{T+t_k}{2})}^T[\sigma^{-1}(\phi(s),X_{\phi(s)}^N)Y_{\phi(s)}^N]^*dW_s\right]$$
$$\left.\left.\times\, e^i\cdot\left[Z_{t_k}^{N*}\int_{t_k}^{\phi(\frac{T+t_k}{2})}[\sigma^{-1}(\phi(s),X_{\phi(s)}^N)Y_{\phi(s)}^N]^*dW_s\right]\right)\right| \le K(T,x)\frac{\|f\|_\infty}{T^q}(T-t_k)^2h,$$

(4.3)
$$\left|\mathbb{E}\left([\sigma\dot{\sigma}^*]_{i,j}(t_k,X_{t_k})f(X_T)e^i\cdot\left\{\nabla_x\left[{Z_{t_k}}^*\int_{t_k}^{\frac{T+t_k}{2}}[\sigma^{-1}(s,X_s)Y_s]^*dW_s\right]Z_{t_k}e^j\right\}\right.\right.$$
$$- [\sigma\dot{\sigma}^*]_{i,j}(t_k,X_{t_k}^N)f(X_T^N)e^i$$
$$\left.\left.\cdot\left\{\nabla_x\left[Z_{t_k}^{N*}\int_{t_k}^{\phi(\frac{T+t_k}{2})}[\sigma^{-1}(\phi(s),X_{\phi(s)}^N)Y_{\phi(s)}^N]^*dW_s\right]Z_{t_k}^Ne^j\right\}\right)\right|$$
$$\le K(T,x)\frac{\|f\|_\infty}{T^q}(T-t_k)h.$$

Note that terms with $f(X_{t_k})$ and $f(X_{t_k}^N)$ have been removed since they do not contribute in the expectation. The three errors above can be analyzed by applying Theorem 4.2, except that the upper bounds have to include factors $(T-t_k)$ or $(T-t_k)^2$: this is a simple improvement that we won't detail here. □

4.5. Proof of Theorem 4.2.

4.5.1. When f is of class C_b^3. Set $V_T^{\lambda,N} = V_T^N + \lambda(V_T - V_T^N)$ for $\lambda\in[0,1]$. Then, the error to analyze is

(4.4)
$$\mathbb{E}(f(V_T)H_T) - \mathbb{E}(f(V_T^N)H_T^N) = \mathbb{E}([f(V_T)-f(V_T^N)]H_T) + \mathbb{E}(f(V_T^N)[H_T - H_T^N]).$$

Note that the difference $V_T - V_T^N$ can be expressed componentwise using Lemma 4.3; using a Taylor expansion, it follows that the first contribution in the right-hand side

above can be split in a sum of terms

$$\mathcal{E}_{i,j,k} = \int_0^1 d\lambda\, \mathbb{E}\left(\partial_{x_k} f(V_T^{\lambda,N}) H_T c_{i,j,k}^0(T) \int_0^T c_{i,j,k}^1(t) \left[\left(\int_{\phi(t)}^t c_{i,j,k}^2(s) dW_s^i\right) dW_t^j\right]\right),$$

for $0 \le i,j \le q$. If i and j are different from 0, we twice apply the duality relation (2.1) combined with Fubini's theorem to obtain that $\mathcal{E}_{i,j,k}$ equals

$$\begin{aligned}
&\int_0^1 d\lambda \int_0^T dt\, \mathbb{E}\left(\mathcal{D}_t^j\left[\partial_{x_k} f(V_T^{\lambda,N}) H_T c_{i,j,k}^0(T)\right] c_{i,j,k}^1(t) \left(\int_{\phi(t)}^t c_{i,j,k}^2(s) dW_s^i\right)\right)\\
&= \int_0^1 d\lambda \int_0^T dt \int_{\phi(t)}^t ds\, \mathbb{E}\big(\mathcal{D}_s^i\big\{\mathcal{D}_t^j\left[\partial_{x_k} f(V_T^{\lambda,N}) H_T c_{i,j,k}^0(T)\right] c_{i,j,k}^1(t)\big\} c_{i,j,k}^2(s)\big)\\
&= \sum_l \int_0^1 d\lambda \int_0^T dt \int_{\phi(t)}^t ds\, \mathbb{E}\big(\partial_x^{\gamma_1(l)} f(V_T^{\lambda,N}) G_{s,t,T}^{1,l,\lambda,N}\big), \qquad (4.5)
\end{aligned}$$

where the length of the differentiation index $\gamma_1(l)$ is less than 3. If i and/or j equals 0, an analogous formula holds with $|\gamma_1(l)| \le 2$. The random variable $G_{s,t,T}^{1,l,\lambda,N}$ is integrable with an $\mathbf{L}^1$-norm uniformly bounded w.r.t. λ, N, s, t. We have proved that this first contribution in the right-hand side of (4.4) meets the estimate of Theorem 4.2.

For the second contribution, taking into account that $H_T - H_T^N$ satisfies $(\mathcal{P})$ and using the same techniques as before, we obtain that $\mathbb{E}(f(V_T^N)[H_T - H_T^N])$ can be decomposed as a sum of terms of type

(4.6)
$$\int_0^T dt \int_{\phi(t)}^t ds \mathbb{E}\big(\partial_x^{\gamma_2} f(V_T^N) G_{s,t,T}^{\gamma_2,N}\big) \text{ and } \int_0^T dt \int_0^t ds \int_{\phi(s)}^s du \mathbb{E}\big(\partial_x^{\gamma_3} f(V_T^N) G_{u,s,t,T}^{\gamma_3,N}\big)$$

with $|\gamma_2| \le 2$ and $|\gamma_3| \le 3$. Uniform $\mathbf{L}^p$ estimates are available for $G_{s,t,T}^{\gamma_2,N}$ and $G_{u,s,t,T}^{\gamma_3,N}$ and the proof is complete for the case of functions f of class C_b^3.

4.5.2. When f is only measurable. Formally, techniques are identical, but to remove the *derivatives* of f in (4.5) and (4.6), we may integrate by parts. This step is not directly possible since the Malliavin covariance matrix of V_T^N or $V_T^{\lambda,N}$ may have bad properties, even under Assumption (E′). We do not encounter this problem in the one-dimensional elliptic case developed in [KHP02], since the convex combination of positive diffusion coefficients is still positive: this argument fails in higher dimensions with elliptic matrices, and a fortiori in a hypoelliptic framework. To circumvent this difficulty, we introduce a series of localization and approximation arguments, which unfortunately make the reading more tedious.

We put $V_T^\epsilon = V_T + \epsilon \tilde{W}_T$ and $V_T^{N,\epsilon} = V_T^N + \epsilon \tilde{W}_T$, where $(\tilde{W}_T)_{t\ge 0}$ is an extra independent r-dimensional Brownian motion and we define $V_T^{\lambda,N,\epsilon} = V_T^{N,\epsilon} + \lambda(V_T^\epsilon - V_T^{N,\epsilon})$ for $\lambda \in [0,1]$. In the following computations, the Malliavin calculus will be made w.r.t. the $(q+r)$-dimensional Brownian motion $\binom{W_t}{\tilde{W}_t}_{0\le t\le T}$.

Denote by $\bar{\mu}$ the measure defined by $\int_{\mathbb{R}^r} g(x)\bar{\mu}(dx) = \mathbb{E}(g(V_T^{0,N,0})) + \mathbb{E}(g(V_T^{1,N,0})) + \int_0^1 \mathbb{E}(g(V_T^{\lambda,N,0}))d\lambda$ and consider $(f_m)_{m\ge 1}$ a sequence of smooth functions with compact support, which converges to f in $\mathbf{L}^2(\bar{\mu})$. Thus, one easily gets that

$$\lim_{m\uparrow\infty} \lim_{\epsilon\downarrow 0} \|f_m(V_T^{\lambda,N,\epsilon})\|_{\mathbf{L}^2} = \lim_{m\uparrow\infty} \|f_m(V_T^{\lambda,N,0})\|_{\mathbf{L}^2} = \|f(V_T^{\lambda,N,0})\|_{\mathbf{L}^2} \le \|f\|_\infty \qquad (4.7)$$

for $\lambda = 0$ or 1, and

$$\lim_{m\uparrow\infty}\lim_{\epsilon\downarrow 0}\left(\int_0^1 \|f_m(V_T^{\lambda,N,\epsilon})\|_{\mathbf{L}^2}d\lambda\right) = \lim_{m\uparrow\infty}\left(\int_0^1 \|f_m(V_T^{\lambda,N})\|_{\mathbf{L}^2}d\lambda\right)$$

$$\leq \lim_{m\uparrow\infty}\sqrt{\int_0^1 \mathbb{E}(f_m^2(V_T^{\lambda,N}))d\lambda} = \sqrt{\int_0^1 \mathbb{E}(f^2(V_T^{\lambda,N,0}))d\lambda} \leq \|f\|_\infty. \tag{4.8}$$

Then, the error to analyze is equal to $\mathbb{E}(f(V_T)H_T) - \mathbb{E}(f(V_T^N)H_T^N) = \lim_{m\uparrow\infty,\epsilon\downarrow 0}[\mathcal{E}_1(m,\epsilon) + \mathcal{E}_2(m,\epsilon)]$ with

$$\mathcal{E}_1(m,\epsilon) = \mathbb{E}\left(f_m(V_T^\epsilon)H_T - f_m(V_T^{N,\epsilon})H_T\right),$$
$$\mathcal{E}_2(m,\epsilon) = \mathbb{E}\left(f_m(V_T^{N,\epsilon})H_T - f_m(V_T^{N,\epsilon})H_T^N\right).$$

In view of (4.7) and (4.8), it is enough to prove the following estimates, with some constants $K(T,x)$, p and q uniform in m, and $\epsilon \leq 1$:

$$|\mathcal{E}_1(m,\epsilon)| \leq K(T,x)\bigg(\|f_m(V_T^{0,N,\epsilon})\|_{\mathbf{L}^2} + \|f_m(V_T^{1,N,\epsilon})\|_{\mathbf{L}^2}$$
$$+ \int_0^1 \|f_m(V_T^{\lambda,N,\epsilon})\|_{\mathbf{L}^2}d\lambda\bigg)\|1/\det(\gamma_T)\|_{\mathbf{L}^p}^q h, \tag{4.9}$$

$$|\mathcal{E}_2(m,\epsilon)| \leq K(T,x)\left(\|f_m(V_T^{0,N,\epsilon})\|_{\mathbf{L}^2} + \|f_m(V_T^{1,N,\epsilon})\|_{\mathbf{L}^2}\right)\|1/\det(\gamma_T)\|_{\mathbf{L}^p}^q h. \tag{4.10}$$

We introduce a localization factor $\psi_T^{N,\epsilon} \in [0,1]$, satisfying the following properties:

(a) $\psi_T^{N,\epsilon} \in \mathbb{D}^\infty$ and $\sup_{N,\epsilon}\|\psi_T^{N,\epsilon}\|_{k,p} \leq K(T,x)\|1/\det(\gamma_T)\|_{\mathbf{L}^{q_1}}^{q_2}$ for any integers k,p;

(b) $\mathbb{P}(\psi_T^{N,\epsilon} \neq 1) \leq K(T,x)\|1/\det(\gamma_T)\|_{\mathbf{L}^p}^q h^k$ for any $k \geq 1$, uniformly in ϵ;

(c) $\{\psi_T^{N,\epsilon} \neq 0\} \subset \{\forall\lambda \in [0,1] : \det(\gamma^{V_T^{\lambda,N,\epsilon}}) \geq \frac{1}{2}\det(\gamma^{V_T})\}$.

Its construction is given at the end of this section.

Error $\boldsymbol{\mathcal{E}_1(m,\epsilon)}$. Clearly, one has $\mathcal{E}_1(m,\epsilon) = \mathcal{E}_{1,1}(m,\epsilon) + \mathcal{E}_{1,2}(m,\epsilon)$ with

$$\mathcal{E}_{1,1}(m,\epsilon) = \mathbb{E}\left([f_m(V_T^\epsilon) - f_m(V_T^{N,\epsilon})](1-\psi_T^{N,\epsilon})H_T\right), \tag{4.11}$$

$$\mathcal{E}_{1,2}(m,\epsilon) = \mathbb{E}\left([f_m(V_T^\epsilon) - f_m(V_T^{N,\epsilon})]\psi_T^{N,\epsilon}H_T\right). \tag{4.12}$$

The first term can easily be bounded by $K(T,x)\big(\|f_m(V_T^{0,N,\epsilon})\|_{\mathbf{L}^2} + \|f_m(V_T^{1,N,\epsilon})\|_{\mathbf{L}^2}\big)$ $\|1/\det(\gamma_T)\|_{\mathbf{L}^p}^q h^k$ for any $k \geq 1$, using property (b) of $\psi_T^{N,\epsilon}$.

Now, to deal with the term $\mathcal{E}_{1,2}(m,\epsilon)$, we proceed as for the first term of the right-hand side of (4.4), that is, by decomposing $V_T^\epsilon - V_T^{N,\epsilon} = V_T - V_T^N$ using Lemma 4.3 and applying the duality relationship. Consequently, $\mathcal{E}_{1,2}(m,\epsilon)$ can be written as a sum of terms

$$\mathcal{E}_{1,2,\gamma_1}(m,\epsilon) = \int_0^1 d\lambda \int_0^T dt \int_{\phi(t)}^t ds\, \mathbb{E}\big(\partial_x^{\gamma_1} f_m(V_T^{\lambda,N,\epsilon})G_{s,t,T}^{\gamma_1,\lambda,N}\big) \tag{4.13}$$

with $|\gamma_1| \leq 3$. The random variable $G_{s,t,T}^{\gamma_1,\lambda,N}$ does not depend on ϵ and belongs to $\mathbb{D}^\infty$ with Sobolev norms uniformly bounded w.r.t. λ, N, s, t. Owing to the factor $\psi_T^{N,\epsilon}$,

note that $G^{\gamma_1,\lambda,N}_{s,t,T} = 0$ when $\psi^{N,\epsilon}_T = 0$, because of the local property of the derivative operator (see Proposition 1.3.7 in [Nua95, p. 44]). Since $\det(\gamma^{V^{\lambda,N,\epsilon}_T}) \geq \epsilon^{2r}$, one can apply Proposition 2.4, which yields

$$\mathbb{E}\big(\partial^{\gamma_1}_x f_m(V^{\lambda,N,\epsilon}_T) G^{\gamma_1,\lambda,N}_{s,t,T}\big) = \mathbb{E}\big(f_m(V^{\lambda,N,\epsilon}_T) H_{\gamma_1}(V^{\lambda,N,\epsilon}_T, G^{\gamma_1,\lambda,N}_{s,t,T})\big)$$

for some iterated Skorohod integral $H_{\gamma_1}(V^{\lambda,N,\epsilon}_T, G^{\gamma_1,\lambda,N}_{s,t,T})$. Due to the local property of the Skorohod integral (see Proposition 1.3.6 in [Nua95, p. 43]), one has $H_{\gamma_1}(V^{\lambda,N,\epsilon}_T, G^{\gamma_1,\lambda,N}_{s,t,T}) = H_{\gamma_1}(V^{\lambda,N,\epsilon}_T, G^{\gamma_1,\lambda,N}_{s,t,T})\mathbf{1}_{\psi^{N,\epsilon}_T \neq 0}$, and applying the estimate from Proposition 2.4, one gets

$$\|H_{\gamma_1}(V^{\lambda,N,\epsilon}_T, G^{\gamma_1,\lambda,N}_{s,t,T})\|_{\mathbf{L}^2} \leq C \|[\gamma^{V^{\lambda,N,\epsilon}_T}]^{-1} \mathbf{1}_{\psi^{N,\epsilon}_T \neq 0}\|^{p_3}_{\mathbf{L}^{q_3}} \|V^{\lambda,N,\epsilon}_T\|^{p_1}_{k_1,q_1} \|G^{\gamma_1,\lambda,N}_{s,t,T}\|_{k_2,q_2}$$

for some integers $p_1, p_3, q_1, q_2, q_3, k_1, k_2$. It is easy to upper bound $\|V^{\lambda,N,\epsilon}_T\|_{k_1,q_1}$ and $\|G^{\gamma_1,\lambda,N}_{s,t,T}\|_{k_2,q_2}$, uniformly in λ, N, s, t, and $\epsilon \leq 1$. It is straightforward to derive the estimation of $\|[\gamma^{V^{\lambda,N,\epsilon}_T}]^{-1} \mathbf{1}_{\psi^{N,\epsilon}_T \neq 0}\|_{\mathbf{L}^{q_3}}$ since on $\{\psi^{N,\epsilon}_T \neq 0\}$, $\det(\gamma^{V^{\lambda,N,\epsilon}_T}) \geq \frac{1}{2}\det(\gamma^{V_T})$, which has an inverse in any $\mathbf{L}^p$. One has proved that

$$|\mathcal{E}_{1,2,\gamma_1}(m,\epsilon)| \leq K(T,x) \left(\int_0^1 \|f_m(V^{\lambda,N,\epsilon}_T)\|_{\mathbf{L}^2} d\lambda \right) \|1/\det(\gamma_T)\|^q_{\mathbf{L}^p} h;$$

this completes the estimation (4.9).

Error $\boldsymbol{\mathcal{E}_2(m,\epsilon)}$. As before, this error can be split into two parts $\mathcal{E}_2(m,\epsilon) = \mathbb{E}\big(f_m(V^{N,\epsilon}_T)(1-\psi^{N,\epsilon}_T)(H_T - H^N_T)\big) + \mathbb{E}\big(f_m(V^{N,\epsilon}_T)\psi^{N,\epsilon}_T(H_T - H^N_T)\big)$. The first contribution can be neglected using property (b) about $\psi^{N,\epsilon}_T$. The other contribution is analyzed as the second term in the right-hand side of (4.4): it gives a sum of terms of type (4.6) with $V^{N,\epsilon}_T$ instead of V^N_T and some random variables $G^{\gamma_2,N}_{s,t,T}$ and $G^{\gamma_3,N}_{u,s,t,T}$ vanishing when $\psi^{N,\epsilon}_T = 0$. Then, the rest of the proof is identical to that for (4.13); we omit details. The inequality (4.10) follows and Theorem 4.2 is proved.

Construction of $\boldsymbol{\psi^{N,\epsilon}_T}$. Set $d(\mu) = \det(\gamma^{V^\epsilon_T + \mu(V^N_T - V_T)})$ for $\mu \in [0,1]$. Since $\det(\gamma^{V^{\lambda,N,\epsilon}_T}) = d(1-\lambda)$, one has

$$\det(\gamma^{V^{\lambda,N,\epsilon}_T}) = \det(\gamma^{V^\epsilon_T}) - \int_{1-\lambda}^1 d'(\mu) d\mu. \tag{4.14}$$

Assume that for some $C > 0$, one has for any $\mu \in [0,1]$, $|d'(\mu)|^2 \leq R^N_T$ with

$$R^N_T := C \left(\int_0^T \|\mathcal{D}_t(V^N_T - V_T)\|^2 dt \right) \left(\int_0^T \big[\|\mathcal{D}_t V^\epsilon_T\|^2 + \|\mathcal{D}_t(V^N_T - V_T)\|^2\big] dt \right)^3. \tag{4.15}$$

Then, if we put $\psi^{N,\epsilon}_T = \psi(\frac{R^N_T}{\det^2(\gamma^{V^\epsilon_T})})$ with $\psi \in C^\infty_b(\mathbb{R},\mathbb{R})$ such that $\mathbf{1}_{[0,\frac{1}{8}]} \leq \psi \leq \mathbf{1}_{[0,\frac{1}{4}]}$, it is now clear that statement (a) is fulfilled using $\gamma^{V^\epsilon_T} = \gamma^{V_T} + \epsilon^2 \mathrm{I}_d$. Besides, $\psi^{N,\epsilon}_T \neq 1 \Rightarrow R^N_T > \frac{1}{8}\det^2(\gamma^{V^\epsilon_T})$, and thus estimates (b) follow using techniques of Lemma 3.3. Finally, $\psi^{N,\epsilon}_T \neq 0 \Rightarrow R^N_T < \frac{1}{4}\det^2(\gamma^{V^\epsilon_T}) \Rightarrow \det(\gamma^{V^{\lambda,N,\epsilon}_T}) \geq \frac{1}{2}\det(\gamma^{V^\epsilon_T}) \geq \frac{1}{2}\det(\gamma^{V_T})$ using (4.14) and (c) holds true.

It remains to prove (4.15). For this, using for any invertible matrix A the relations $\partial_A \det(A) = \det(A)[A^*]^{-1}$ (see Theorem A.98 of [RT99]) and $A^{-1} = \frac{1}{\det(A)}[\mathrm{Cof}(A)]^*$ ($\mathrm{Cof}(A)$ being the matrix of cofactors of A), one gets

$$\begin{aligned}[\det(A(\mu))]' &= \sum_{i,j} \partial_{a_{i,j}}[\det(A)] a'_{i,j}(\mu) = \sum_{i,j} \det(A)[(A^*)^{-1}]_{i,j} a'_{i,j}(\mu) \\ &= \mathrm{Tr}\big(\mathrm{Cof}(A(\mu)) A'^*(\mu)\big).\end{aligned}$$

Put $A(\mu) = \gamma^{V_T^\epsilon + \mu(V_T^N - V_T)} = \gamma^{V_T^\epsilon} + \mu\big(\int_0^T \mathcal{D}_t V_T^\epsilon [\mathcal{D}_t(V_T^N - V_T)]^* dt + \int_0^T \mathcal{D}_t(V_T^N - V_T)[\mathcal{D}_t V_T^\epsilon]^* dt\big) + \mu^2 \gamma^{V_T^N - V_T}$. We now easily prove that

$$[(\mathrm{Cof}(A))_{i,j}(\mu)]^2 \le C_1 \left(\int_0^T \big[\|\mathcal{D}_t V_T^\epsilon\|^2 + \|\mathcal{D}_t(V_T^N - V_T)\|^2 \big] dt \right)^2,$$

$$[(A'_{i,j})(\mu)]^2 \le C_2 \left(\int_0^T \|\mathcal{D}_t(V_T^N - V_T)\|^2 dt \right) \left(\int_0^T \big[\|\mathcal{D}_t(V_T^N - V_T)\|^2 + \|\mathcal{D}_t V_T^\epsilon\|^2 \big] dt \right).$$

Thus, (4.15) immediately follows using $d'(\mu) = [\det(A(\mu))]'$.

5. Numerical experiments.

5.1. Analysis of computational complexity. In this paragraph we indicate the first-order approximation of the number of elementary operations (multiplications) needed for computing the different estimators w.r.t. the quantities m (number of parameters), d (dimension of the space), q (dimension of the Brownian motion), and N (number of discretization times).

In previous sections we derived estimators of the gradient of the performance measure $J(\alpha)$ w.r.t. α when J is defined by a terminal cost (see (1.2)). However, these results may be extended to functionals with instantaneous costs such as $J(\alpha) = \mathbb{E}\left(\int_0^T g(t, X_t) dt + f(X_T)\right)$ for which an estimator of the gradient may be $\frac{T}{N}\sum_{i=1}^N H_{t_i}^N(g) + H_T^N(f)$, where $H_{t_i}^N(g)$ (resp., $H_T^N(f)$) is an approximated estimator of the gradient of $\mathbb{E}(g(t_i, X_{t_i}))$ (resp., $\mathbb{E}(f(X_T))$). This case is illustrated in the first numerical experiment considered below.

The computational complexity of the different estimators depends on whether the payoff has instantaneous costs (in which case an estimator $H_{t_i}^N(g)$ for all $i \in \{1, \dots, N\}$ is needed) or if it has only a terminal cost (for which only $H_T^N(f)$ is needed). In the pathwise and Malliavin calculus methods, the cost of computing $H_{t_i}^N(g)$ for all $i \in \{1, \dots, N\}$ is the same as just computing $H_T^N(f)$, whereas in the adjoint and martingale methods, there is an additional computational burden.

- Complexity of the pathwise method: $d^2 qmN$ operations for computing the pathwise estimator $H_{t_i}^{Path.,N}(g)$ (see Proposition 1.1) for all $i \in \{1, \dots, N\}$ (required for computing $\dot{X}_{t_i}^N$, for all $i \in \{1, \dots, N\}$ and all m parameters).
- Complexity of the Malliavin calculus method, in the elliptic case ($q = d$): $3d^4(d + m)N$ operations, for computing the Malliavin calculus estimator $H_{t_i}^{Mall.Ell.,N}(g)$ (see Proposition 2.5) for all $i \in \{1, \dots, N\}$. Indeed, the complexity of computing the Malliavin derivative of each column c (among d) of Z_{c,t_i}^N is $3d^4 N$ and computing the Malliavin derivative of $\dot{X}_{t_i}^N$ for all m parameters requires $3d^3 mN$ operations.

- Complexity of the adjoint method: $d^4N^2 + d^2mN^2/2$ operations are needed to compute the adjoint estimator $H_{t_i}^{Adj.,N}(g) = H_{t_i}^{b,Adj.,N}(g) + \frac{1}{2}H_{t_i}^{\sigma,Adj.,N}(g)$ (see Lemma 2.9 and Theorem 2.11) for all $i \in \{1,\dots,N\}$. Our implementation memorizes $Z_{t_i}^N$ (and other data) along the trajectory and computes $H_{t_i}^{b,Adj.,N}(g)$ and $H_{t_i}^{\sigma,Adj.,N}(g)$ for all $i \in \{1,\dots,N\}$ afterwards. Such an implementation allows to treat problems with instantaneous costs. If we consider a problem with a terminal cost only, the complexity is reduced to $4d^4N + 3d^2mN$.
- Complexity of the martingale method: $d^2N^2/2 + dmN^2/2 + d^3mN$ for computing the martingale estimator $H_{t_i}^{Mart.,N}(g)$ (see Theorem 2.12) for all $i \in \{1,\dots,N\}$. For problems with terminal cost only the complexity of computing $H_T^{Mart.,N}(f)$ is d^3mN.

These results are summarized in Table 5.1. Note that they are strongly related to the way we have implemented the methods and they are not guaranteed to be optimal.

TABLE 5.1
Complexity (in terms of number of elementary operations) of the different estimators for payoff with instantaneous costs or with terminal cost only.

	Pathwise	Malliavin	Adjoint	Martingale
Instantaneous costs	d^3mN	$3d^4(d+m)N$	$d^4N^2 + d^2m\frac{N^2}{2}$	$d(d+m)\frac{N^2}{2} + d^3mN$
Terminal cost	Same	Same	$4d^4N + 3d^2mN$	$d^2N + d^3mN$

5.2. Stochastic linear quadratic optimal control. We consider a simple one-dimensional stochastic linear quadratic (SLQ) control problems (see [CY01] and [YZ99] for an extensive study on SLQ problems) for which the control $u(\cdot)$ appears in particular in the diffusion term: $dX_t = u(t)dt + \delta u(t)dW_t$. The cost functional to be minimized is $J(u(\cdot)) = \mathbb{E}\big[\int_0^1 X_t^2 dt\big]$. This problem admits an optimal control (see references above) given by the state feedback $u^*(t) = -\frac{X_t}{\delta^2}$.

We consider a class of feedback controllers $u(t,x,\alpha)$ linearly parameterized by a three-dimensional vector α with basis functions 1, x, and t (i.e., $u(t,x,\alpha) = \alpha_1 + \alpha_2 x + \alpha_3 t$) and we write $J(\alpha)$ for $J(u(\cdot,X_\cdot,\alpha))$. In that case, the optimal control u^* belongs to the class of parameterized feedback controllers and corresponds to the parameter $\alpha^* = (0, -1/\delta^2, 0)$.

As explained before, since the payoff involves instantaneous costs, we evaluate $\nabla_\alpha J(\alpha)$ using a quantity of type $\frac{T}{N}\sum_{i=1}^N H_{t_i}^N(x^2)$. We check that the different estimators (pathwise, Malliavin calculus, adjoint, martingale) return a zero gradient for the value α^* of the parameter and we compare their variance and time for computation. Table 5.2 shows the empirical variance of the different estimators obtained for 1000 trajectories, with $h = 0.05$, $\delta = 1$. These simulations have been performed on a Pentium III, 700Mhz processor.

TABLE 5.2
Variance of the estimators $H^{Path.}$, $H^{Mall.Ell.}$, $H^{Adj.}$, and $H^{Mart.}$ of $\nabla_\alpha J(\alpha)$ at the optimal setting of the parameter: $\alpha_1 = 0$, $\alpha_2 = -1$, $\alpha_3 = 0$.

Var(H)	Pathwise	Malliavin	Adjoint	Martingale
α_1	0.1346	0.3754	0.6669	0.1653
α_2	0.0525	0.1188	0.1707	0.0480
α_3	0.0136	0.0446	0.0612	0.0148
CPU time	0.44s	1.95s	2.89s	0.89s

Note that the estimator used in the adjoint approach includes the term $f(X_T) - f(X_t)$ in the computation of $H_T^{\sigma,Adj.}$. Table 5.3 shows similar results for a suboptimal setting of the parameter (here α_1, α_2, and α_3 are chosen randomly within the range $[-0.1, 0.1]$). The columns Adjoint2 and Martingale2 describe simulations of the adjoint and martingale methods when the term $f(X_t)$ is omitted from the computation of the estimators $H_T^{\sigma,Adj.}$ and $H_T^{Mart.}$. We note that the variance of these estimators is significantly larger than when the term $f(X_t)$ is included, which corroborates Remark 2.1.

TABLE 5.3
Variance of the different estimators of $\nabla_\alpha J(\alpha)$ *for* $\alpha_1 = -0.0789$, $\alpha_2 = 0.0156$, $\alpha_3 = 0.0648$.

Var(H)	Pathwise	Malliavin	Adjoint	Adjoint2	Mart.	Mart.2
α_1	0.2005	4.0347	1.0287	9.5535	1.5085	4.6029
α_2	0.0252	0.6597	0.1433	1.6781	0.2360	0.7894
α_3	0.0174	0.3869	0.1051	2.2337	0.1407	1.0185
CPU time	0.44s	1.97s	2.94s	2.94s	0.90s	0.90s

For this problem with smooth cost functions, the pathwise approach provides the best performance in terms of the estimator's variance. This nice behavior for smooth costs compared to other methods has been previously observed in [FLL+99].

5.2.1. Stochastic approximation algorithm. The computation of an estimator H of $\nabla_\alpha J(\alpha)$ may be used in a stochastic approximation algorithm (see, e.g., [KY97] or [BMP90]) to search a locally optimal parameterization of the controller. The algorithm begins with an initial setting of the parameter α^0. Then, if α^k denotes the value of the parameter at iteration k, the algorithm proceeds by computing an estimator $\widehat{\nabla_\alpha J(\alpha^k)}$ of $\nabla_\alpha J(\alpha^k)$ and then by performing a stochastic gradient ascent

$$\alpha^{k+1} = \alpha^k + \eta_k \widehat{\nabla_\alpha J(\alpha^k)}, \tag{5.1}$$

where the learning steps η_k satisfy a decreasing condition (for example, $\sum_k \eta_k = \infty$ and $\sum_k \eta_k^2 < \infty$; see [Pol87]). Assuming smoothness conditions on $J(\alpha)$ and a bounded variance for $\widehat{\nabla_\alpha J(\alpha^k)}$, one proves that if α^k converges, then the limit is a point of local minimum for $J(\alpha)$ (see references above for several sets of hypotheses for which the convergence is guaranteed).

Figure 5.1 illustrates this algorithm on the SLQ problem described previously, where the initial parameter is chosen randomly (same value as in Table 5.3). At iteration k, one trajectory is simulated using the controller parameterized by α^k, and an estimation $\widehat{\nabla_\alpha J(\alpha^k)}$ of $\nabla J(\alpha_k)$ (using the pathwise method) is obtained. The parameter is updated according to (5.1) with a learning step $\eta_k = \frac{K}{K+k}$. We take $K = 200$ to avoid a too rapid decreasing of $(\eta_k)_k$ at the beginning: this trick usually speeds up the numerical optimization as mentioned in [BT96b]. We note that the parameter converges to $\alpha^* = (0, -1, 0)$.

The speed of convergence for such algorithms is closely related to the gradient estimator's variance, which motivates our variance analysis for the different estimators.

5.2.2. Discretization error. Here, we report the impact of the number of discretization times in the regular mesh of the interval $[0, T]$, in the computation of the gradient $\nabla_\alpha J(\alpha)$ for the SLQ problem. Figure 5.2 reports the sensitivity of

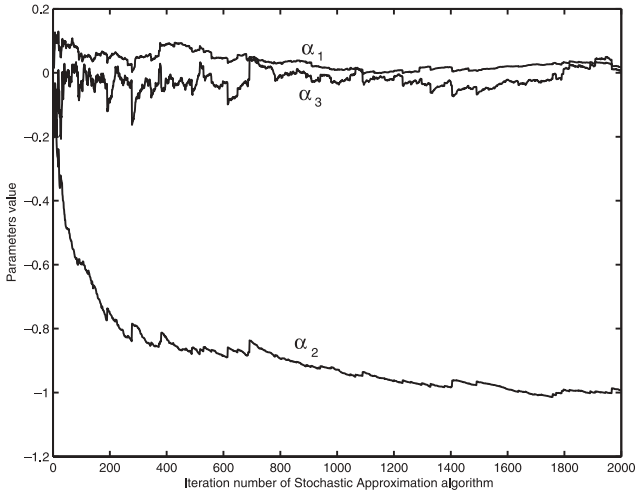


FIG. 5.1. *Stochastic approximation of the control parameters. The gradient* $\nabla_\alpha J(\alpha_k)$ *is estimated using the pathwise method.*

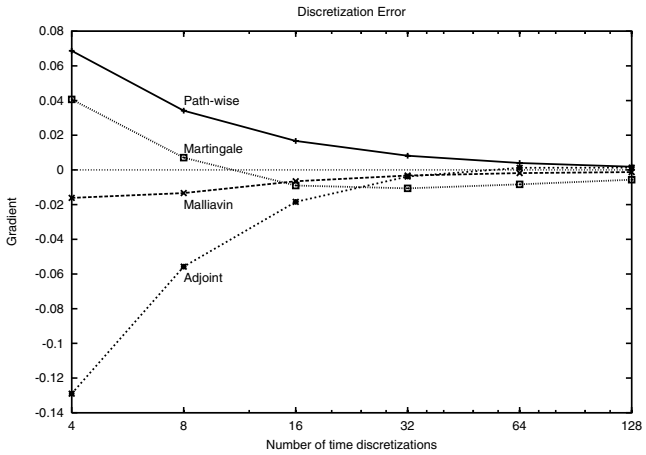


FIG. 5.2. *Discretization error as a function of the number of discretization times.*

$J(\alpha)$ (for $\alpha = \alpha^*$) w.r.t. the parameter α_1, computed with different estimators, with $N = 8, 16, 32, 64$, and 128 discretization times. Recall that, for this setting of the parameter, the true gradient is zero. To get relevant results, we have run 10^7 sample paths, which ensures that the confidence interval's width is less than 10^{-3} for all methods. We can empirically check that the convergence holds at rate $1/N$ (as previously proved), except for the martingale method, for which the rate of convergence is not clear because of the sign change (more discretization times would be needed to clarify the speed of approximation). Note that the discretization error for the Malliavin calculus estimator is smaller than for the other ones, although we have not found any explanation for this.

5.3. Sensitivity analysis in a financial market. We consider two risky assets with price process evolving according to the following SDE under the so-called risk-

neutral probability:

$$\frac{dS_t^1}{S_t^1} = r\,dt + \sigma(S_t^1, \lambda_1)\,dW_t^1,$$
$$\frac{dS_t^2}{S_t^2} = r\,dt + \sigma(S_t^2, \lambda_2)\left(\rho\,dW_t^1 + \sqrt{1-\rho^2}\,dW_t^2\right),$$

with constant interest rate r and volatility function $\sigma(x,\lambda) = 0.25(1 + \frac{1}{1+e^{-\lambda x}})$. The parameters of this dynamics are λ_1, λ_2, and the correlation coefficient ρ. Suppose that the true model is given by a set of parameters and that we are interested in the impact of the inaccuracy on these parameters (due to a previous statistical procedure) on option prices. For instance, we may consider digital options with payoff $\chi(S_T^1 - S_T^2)$ (where $\chi(x) = \mathbf{1}_{x\geq 0}$) whose prices are given by $J(\lambda_1,\lambda_2,\rho) = \mathbb{E}[\chi(S_T^1 - S_T^2)]$ up to the discount factor.

TABLE 5.4
Variance of the estimators $H_T^{Mall.Ell.}$, $H_T^{Adj.}$, $H_T^{Mart.}$, $H_T^{\varepsilon,Path.}$.

Var(H) or Var(H^ε)	Malliavin	Adjoint	Martingale	Pathwise $\varepsilon = 10^{-2}$	Pathwise $\varepsilon = 10^{-3}$	Pathwise $\varepsilon = 10^{-4}$
λ_1	0.0011	0.0022	0.0012	0.0053	0.0378	3.8951
λ_2	0.0048	0.0030	0.0018	0.0042	0.0296	4.9427
ρ	1.5788	2.0829	1.4323	1.6523	14.923	100.86
CPU time	20.8s	18.6s	7.31s	2.97s	2.97s	2.97s

We estimate the sensitivity of J w.r.t. the parameters λ_1, λ_2, and ρ. Table 5.4 reports the empirical variance of the estimators ($H_T^{Mall.Ell.}$, $H_T^{Adj.}$, and $H_T^{Mart.}$) of the sensitivity of J w.r.t. the parameters for the Malliavin calculus, adjoint, and martingale methods. Since the payoff function is not differentiable, we cannot directly apply the pathwise method; instead, we use χ^ε, a regularization of χ defined by $\chi^\varepsilon(x) = 1$ if $x > \varepsilon$, 0 if $x < -\varepsilon$, and $(x+\varepsilon)/(2\varepsilon)$ otherwise. Note that this induces a bias on the true value of the gradient, bias which vanishes when ε goes to 0. The pathwise estimator that we obtain with this regularization is denoted by $H^{\varepsilon,Path.}$ and Table 5.4 also reports its variance for different values of ε.

For this experiment, we ran 1000 trajectories with initial values $S_0^1 = S_0^2 = 1$, $r = 0.04$, $T = 1$, $h = 0.01$ and parameters setting $\lambda_1 = 2$, $\lambda_2 = 2$, and $\rho = 0.6$.

We note that the variance obtained by the pathwise methods is significantly larger than those obtained by the other methods (especially when ε is small), which motivates the use of the Malliavin calculus, adjoint, or martingale estimators for non-smooth cost functions. To further reduce the variance in the case of piecewise smooth cost functions, we could combine two methods as suggested in [FLL$^+$99]: the pathwise method where the cost function is smooth and one of the other methods where it is not.

5.4. Neurocontrol for a stochastic target problem. We consider a two-dimensional stochastic target (for example, that models the displacement of a fly) moving according to a diffusion. We control a squared fly-swatter with a two-dimensional bounded force $(b(u_1), b(u_2))$ (where $u = (u_1, u_2)$ is the control), and our goal is to hit the fly at time T. Let $X = (X_1, X_2)$ be the relative coordinates of

the fly w.r.t. the fly-swatter, and $V = (V_1, V_2)$ be the velocity of the fly-swatter. A simple model of the dynamics is

$$\begin{aligned} dX_{1,t} &= V_{1,t}dt + \sigma_{fly}dW_t^1, \\ dX_{2,t} &= V_{2,t}dt + \sigma_{fly}dW_t^2, \\ dV_{1,t} &= b(u_{1,t})dt + \sigma_{swat}(1 + ||u_t||)dW_t^3, \\ dV_{2,t} &= b(u_{2,t})dt + \sigma_{swat}(1 + ||u_t||)dW_t^4, \end{aligned}$$

where $b(x) = [1 - e^{-x}]/[1 + e^{-x}]$. $[(W_t^i)_{t\geq 0}]_i$ are independent standard Brownian motions; the coefficients σ_{fly} and σ_{swat} are constant. The factor $(1 + ||u||)$ (where $||u|| = \sqrt{u_1^2 + u_2^2}$) adds uncertainty on highly forced movements. The goal is to reach the fly with the fly-swatter at time T: hence, $J(u(\cdot); X_0, V_0) = \mathbb{E}[\mathbf{1}_{(X_{1,T}, X_{2,T})\in A}]$, where $A = [-a, a] \times [-a, a]$ is the squared fly-swatter.

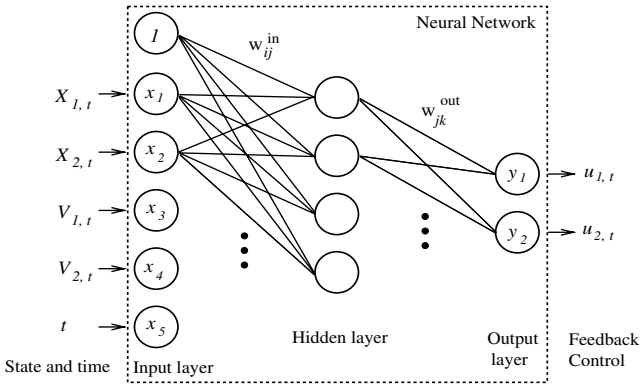


FIG. 5.3. *The architecture of the network.*

We implement the feedback controller using a one-hidden-layer neural network (see [Hay94] or [RM86] for general references on neural networks) whose architecture is given in Figure 5.3. The input layer $(x_i)_{1\leq 5}$ is connected to the state and time variables: $x_1 = X_{1,t}$, $x_2 = X_{2,t}$, $x_3 = V_{1,t}$, $x_4 = V_{2,t}$, and $x_5 = t$. There is one hidden layer with n neurons, and the output layer $(y_k)_{1\leq k\leq 2}$ returns the feedback control $y_1 = u_1(t)$, $y_2 = u_2(t)$. The network is defined by two matrices of weights (the parameters): the *input weights* $\{w_{ij}^{in}\}$ and the *output weights* $\{w_{jk}^{out}\}$. The network's output is given by $y_k = \sum_{j=1}^n w_{jk}^{out}\varphi(\sum_{i=0}^5 w_{ij}^{in}x_i)$ (for $1 \leq k \leq 2$), where the w_{0j}^{in}'s are the *bias weights* (and we set $x_0 = 1$) and $\varphi(s) = 1/(1 + e^{-s})$ is the sigmoid function.

In this experiment, we use a hidden layer with four neurons (thus there are $6 \times 4 + 4 \times 2 = 32$ control parameters); we have run 1000 trajectories with initial values of the weights chosen randomly within the range $[-0.1, 0.1]$. Here, $T = 1$, $h = 0.05$, $\sigma_{fly} = \sigma_{swat} = 0.1$, and $a = 0.1$. Each trajectory starts from a initial state chosen randomly within the range $\Omega = [-0.5, 0.5]^4$. Thus, we actually estimate $\nabla_w \mathbb{E}[J(\cdot; X_0, V_0) \mid (X_0, V_0) \sim \frac{1}{|\Omega|}\mathbf{1}_\Omega(d\omega)]$, for each weight w.

Table 5.5 reports the empirical variance of the estimators ($H^{Mall.Ell.}$, $H^{Adj.}$, and $H^{Mart.}$) of the gradient of J w.r.t. the parameters (the set of input and output weights). Here again, the function to be maximized is not differentiable and to apply the pathwise method, we use a regularization of the indicator function of A (i.e., $J^\varepsilon(\alpha) = \mathbb{E}[(\chi^\varepsilon(X_{1,T}+a)-\chi^\varepsilon(X_{1,T}-a))(\chi^\varepsilon(X_{2,T}+a)-\chi^\varepsilon(X_{2,T}-a))]$). The associated

TABLE 5.5
Variance of the estimators of the gradient of $\mathbb{E}[J(\cdot; X_0, V_0)|(X_0, V_0) \sim \frac{1}{|\Omega|}\mathbf{1}_\Omega(d\omega)]$ *w.r.t. the weights. The values provided are the averaged variances over all* 32 *parameters.*

Var(H) or Var($H_\varepsilon^{Path.}$)	Malliavin	Adjoint	Martingale	Pathwise $\varepsilon = 10^{-3}$	Pathwise $\varepsilon = 10^{-4}$
Average over all parameters	0.1917	0.2550	0.1701	3.364	187.48
CPU time	70.44s	22.04s	5.73s	2.88s	2.88s

pathwise estimator is denoted $H_T^{\varepsilon,Path.}$: its variance for some values of ε is also given in Table 5.5. Although its computational time is the lowest one, the pathwise approach is not appropriate because of its large variance. On the other hand, the martingale method is the most attractive.

Stochastic approximation of an optimal controller. We run a stochastic approximation algorithm (5.1) with a learning rate $\eta_k = \frac{K}{K+k}$ (with $K = 1000$) using a neural network with four hidden neurons. At each iteration, the SA algorithm uses an estimator of the gradient of J w.r.t. the weights, which averages 50 samples of the martingale estimator.

On Figure 5.4, we plot the parameter and performance evolutions w.r.t. the iteration number: we obtain a series of weights that provide a locally optimal performance, although there is no guarantee of global optimality of the controller.

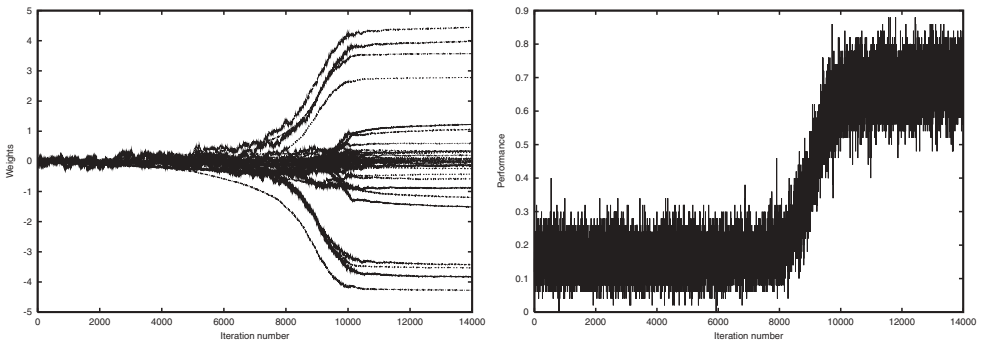


FIG. 5.4. *Stochastic approximation of the parameters (the weights of the neural network) and performance of the parameterized controller. The gradient is estimated using the martingale method.*

This stochastic gradient algorithm in the space of parameterized policies is often called *policy search* (about which an abundant literature exists in the discrete-time case; see, e.g., [BB01]), as opposed to *value search* for which some approximate dynamic programming algorithm is performed on a parameterized value function (see, e.g., [BT96b]). One may also combine these approaches and learn an approximate value function to perform a policy search (the so-called *Actor-Critic* algorithms, see e.g. [KB99]).

6. Conclusion. In this work, we have derived three new types of formulae to compute $\nabla_\alpha \mathbb{E}(f(X_T^\alpha))$ or $\nabla_\alpha \mathbb{E}(\int_0^T g(t, X_t^\alpha)dt + f(X_T^\alpha))$ using Monte Carlo methods. Our computations rely on Itô–Malliavin calculus and martingale techniques: the representations derived are simple to implement using Euler-type schemes and the associate weak error is in most of the cases linear w.r.t. the time step. We have assumed that f is bounded, but all results remain valid if f satisfies some polynomial growth.

The numerical experiments enable us to draw the following conclusions on how to select the appropriate method to use.

1. *Pathwise approach.* This can be used only if the instantaneous and terminal costs are differentiable. Otherwise, regularization procedures lead to high variances. It provides the smallest computational time. Note also that no condition on the nondegeneracy of the diffusion coefficient is needed. For the implementation, only the first derivatives of the coefficient are required.
2. *Malliavin calculus approach.* This handles the case of nonsmooth costs, but the computational time is rather large. A nondegeneracy assumption has to be satisfied but it may be not stringent (hypoellipticity, e.g.). Note that the simulation procedures require the computations of the second derivatives $\partial^2_{x,x}$ and $\partial^2_{x,\alpha}$ of the coefficients.
3. *Adjoint approach.* It can be applied in the elliptic case and is particularly efficient in terms of computational time for a large number of parameters. However, it is quite slow, especially when there are instantaneous costs (because of double time integrals and a possible large number of discretization times). The second derivatives required for the simulations concern only $\partial^2_{x,x}$.
4. *Martingale approach.* The diffusion coefficient has to be elliptic. As for the adjoint approach, it handles situations with nonsmooth costs. It appears to be very fast (almost as fast as the pathwise approach), but it is slower for instantaneous cost problems (same reason as for the adjoint approach). Note also that only the first derivatives of the coefficient are needed.

In future research, we will consider the analysis of the weak error for the martingale method and numerical optimizations in the general nondegenerate case (such as portfolio optimization problems in finance).

Acknowledgment. We thank the two referees and the associate editor for their remarks and suggestions, which helped to improve the clarity of the paper.

REFERENCES

[AT98] M. Arnaudon and A. Thalmaier, *Stability of stochastic differential equations in manifolds*, in Séminaire de Probabilités, XXXII, Springer-Verlag, Berlin, 1998, pp. 188–214.

[BB01] J. Baxter and P.L. Bartlett, *Infinite-horizon gradient-based policy search*, J. Artificial Intelligence Res., 15 (2001), pp. 319–350.

[Ben88] A. Bensoussan, *Perturbation Methods in Optimal Control*, Wiley/Gauthier-Villars Series in Modern Applied Mathematics, John Wiley & Sons, Chichester, UK, 1988.

[BG96] M. Broadie and P. Glasserman, *Estimating security price derivatives using simulation*, Management Sci., 42 (1996), pp. 269–285.

[Bis84] J. M. Bismut, *Large Deviations and the Malliavin Calculus*, Birkhäuser Boston, Boston, 1984.

[BMP90] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, New York, 1990.

[BT96a] V. Bally and D. Talay, *The law of the Euler scheme for stochastic differential equations:* I. *Convergence rate of the distribution function*, Probab. Theory Related Fields, 104 (1996), pp. 43–60.

[BT96b] D. P. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Nashua, NH, 1996.

[CK99] J. Cvitanić and I. Karatzas, *On dynamic measures of risk*, Finance Stoch., 3 (1999), pp. 451–482.

[CM02] P. Cattiaux and L. Mesnager, *Hypoelliptic non-homogeneous diffusions*, Probab. Theory Related Fields, 123 (2002), pp. 453–483.

[CY01] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems*, Appl. Math. Optim., 43 (2001), pp. 21–45.

[EJL99] K. D. Elworthy, Y. Le Jan, and X.-M. Li, *On the Geometry of Diffusion Operators and Stochastic Flows*, Springer-Verlag, Berlin, 1999.

[FLL+99] E. Fournié, J. M. Lasry, J. Lebuchoux, P. L. Lions, and N. Touzi, *Applications of Malliavin calculus to Monte Carlo methods in finance*, Finance and Stochastics, 3 (1999), pp. 391–412.

[Fri64] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[Gly86] P. W. Glynn, *Stochastic approximation for Monte Carlo optimization*, in Proceedings of the 1986 Winter Simulation Conference, J. Wilson, J. Henriksen, and S. Roberts, eds., 1986, pp. 356–365.

[Gly87] P. W. Glynn, *Likelihood ratio gradient estimation: An overview*, in Proceedings of the 1987 Winter Simulation Conference, A. Thesen, H. Grant, and W. D. Kelton, eds., 1987, pp. 366–375.

[GM03] E. Gobet and R. Munos, *Sensitivity Analysis using Itô-Malliavin Calculus and Martingales. Numerical Implementation*, Technical report 520, CMAP, Ecole Polytechnique, Palaiseau, France, 2003.

[Gob00] E. Gobet, *Euler schemes for the weak approximation of killed diffusion*, Stochastic Process. Appl., 87 (2000), pp. 167–197.

[Gob01a] E. Gobet, *Euler schemes and half-space approximation for the simulation of diffusions in a domain*, ESAIM Probab. Statist., 5 (2001), pp. 261–297.

[Gob01b] E. Gobet, *Local asymptotic mixed normality property for elliptic diffusion: A Malliavin calculus approach*, Bernoulli, 7 (2001), pp. 899–912.

[Gob02] E. Gobet, *LAN property for ergodic diffusion with discrete observations*, Ann. Inst. H. Poincaré Probab. Statist., 38 (2002), pp. 711–737.

[GT77] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 1st ed., Springer-Verlag, New York, 1977.

[GY92] P. Glasserman and D. D. Yao, *Some guidelines and guarantees for common random numbers*, Management Sci., 38 (1992), pp. 884–908.

[Hay94] S. Haykin, *Neural Networks: A Comprehensive Foundation*, MacMillan, New York, 1994.

[JP98] J. Jacod and P. Protter, *Asymptotic error distributions for the Euler method for stochastic differential equations*, Ann. Probab., 26 (1998), pp. 267–307.

[KB99] V. R. Konda and V. S. Borkar, *Actor-critic–type learning algorithms for Markov decision processes*, SIAM J. Control Optim., 38 (1999), pp. 94–123.

[KD01] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed., Appl. Math. 24, Springer-Verlag, New York, 2001.

[KH01] A. Kohatsu-Higa, *Weak approximations. A Malliavin calculus approach*, Math. Comp., 70 (2001), pp. 135–172.

[KHP00] A. Kohatsu-Higa and R. Pettersson, *On the simulation of some functionals of diffusion processes*, Sūrikaisekikenkyūsho Kōkyūroku, 1127 (2000), pp. 153–170.

[KHP02] A. Kohatsu-Higa and R. Pettersson, *Variance reduction methods for simulation of densities on Wiener space*, SIAM J. Numer. Anal., 40 (2002), pp. 431–450.

[KP95] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, New York, 1995.

[KS84] S. Kusuoka and D. Stroock, *Applications of the Malliavin calculus* I, in Stochastic Analysis, K. Itô, ed., Kinokuniya, Tokyo, 1984, pp. 271–306.

[KS86] P. Krée and C. Soize, *Mathematics of Random Phenomena: Random Vibrations of Mechanical Structures*, Math. Appl. 32, D. Reidel, Dordrecht, 1986.

[KS98] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.

[Kun84] H. Kunita, *Stochastic differential equations and stochastic flows of diffeomorphisms*, in Ecole d'Eté de Probabilités de St-Flour XII, 1982, Lecture Notes in Math. 1097, Springer-Verlag, New York, 1984, pp. 144–305.

[KY97] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, Berlin, New York, 1997.

[LP94] P. L'Ecuyer and G. Perron, *On the convergence rates of IPA and FDC derivative estimators*, Oper. Res., 42 (1994), pp. 643–656.

[Nua95] D. Nualart, *Malliavin Calculus and Related Topics*, Springer-Verlag, New York, 1995.

[Nua98] D. Nualart, *Analysis on Wiener space and anticipating stochastic calculus*, in Lectures on Probability Theory and Statistics (Saint-Flour, 1995), Springer-Verlag, Berlin, 1998, pp. 123–227.

[Pen90] S. G. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.

[Pic02] J. Picard, *Gradient estimates for some diffusion semigroups*, Probab. Theory Related Fields, 122 (2002), pp. 593–612.

[Pol87] B. T. Polyak, *Introduction to Optimization*, Optimization Software Inc., New York, 1987.

[Pro90] P. Protter, *Stochastic Integration and Differential Equations*, Springer-Verlag, New York, 1990.

[RM86] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, Vols. I and II, MIT Press, Cambridge, MA, 1986.

[RT99] C. R. Rao and H. Toutenburg, *Linear Models*, Springer Series in Statistics, 2nd ed., Springer-Verlag, New York, 1999.

[Run02] W. J. Runggaldier, *On stochastic control in finance*, in Mathematical Systems Theory in Biology, Communication, Computation and Finance (MTNS-2002), Springer-Verlag, New York, 2002.

[RW86] M. I. Reiman and A. Weiss, *Sensitivity analysis via likelihood ratios*, in Proceedings of the 1986 Winter Simulation Conference, J. Wilson, J. Henriksen, and S. Roberts, eds., 1986, pp. 285–289.

[Tha97] A. Thalmaier, *On the differentiation of heat semigroups and Poisson integrals*, Stochastics Stochastics Rep., 61 (1997), pp. 297–321.

[TL90] D. Talay and L.Tubaro, *Expansion of the global error for numerical schemes solving stochastic differential equations*, Stochastic Anal. Appl., 8 (1990), pp. 94–120.

[TZ04] D. Talay and Z. Zheng, *Approximation of quantiles of components of diffusion processes*, Stochastic Process. Appl., 109 (2004), pp. 23–46.

[YK91] J. Yang and H. J. Kushner, *A Monte Carlo method for sensitivity analysis and parametric optimization of nonlinear stochastic systems*, SIAM J. Control Optim., 29 (1991), pp. 1216–1249.

[YZ99] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Appl. Math. 43, Springer-Verlag, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1714–1736

A REAL-TIME ITERATION SCHEME FOR NONLINEAR OPTIMIZATION IN OPTIMAL FEEDBACK CONTROL*

MORITZ DIEHL†, HANS GEORG BOCK†, AND JOHANNES P. SCHLÖDER†

Abstract. An efficient Newton-type scheme for the approximate on-line solution of optimization problems as they occur in optimal feedback control is presented. The scheme allows a fast reaction to disturbances by delivering approximations of the exact optimal feedback control which are iteratively refined *during the runtime* of the controlled process. The contractivity of this *real-time iteration* scheme is proven, and a bound on the loss of optimality—compared with the theoretical optimal solution—is given. The robustness and excellent real-time performance of the method is demonstrated in a numerical experiment, the control of an unstable system, namely, an airborne kite that shall fly loops.

Key words. direct multiple shooting, Newton-type optimization, optimal feedback control, nonlinear model predictive control, ordinary differential equations, real-time optimization

AMS subject classifications. 34B15, 34H05, 49N35, 49N90, 90C06, 90C30, 90C55, 90C59, 90C90, 93C55

DOI. 10.1137/S0363012902400713

1. Introduction.

Feedback control based on the real-time optimization of nonlinear dynamic process models, also referred to as *nonlinear model predictive control (NMPC)*, has attracted increasing attention over the past decade, particularly in chemical engineering [4, 27, 1, 28]. Based on the current system state, feedback is provided by an online optimization of the predicted system behavior, using the mathematical model. The first part of the optimized control trajectory is implemented at the real system, and a sampling time later the optimization procedure is repeated. Among the advantages of this approach are the flexibility provided in formulating the objective and in modeling the process using ordinary or partial differential equations (ODEs or PDEs), the capability of directly handling equality and inequality constraints, and the possibility of treating large disturbances quickly.

One important precondition, however, is the availability of reliable and efficient numerical optimal control algorithms. One particularly successful algorithm that is designed to achieve this aim, the recently developed *real-time iteration* scheme, will be the focus of this paper. In the literature, several suggestions have been made on how to adapt off-line optimal control algorithms for use in on-line optimization. For an overview and comparison of important approaches, see, e.g., Binder et al. [6]. We particularly mention here the "Newton-type control algorithm" proposed by Li and Biegler [32] and de Oliveira and Biegler [15] and the "feasibility-perturbed SQP" approach to NMPC by Tenny, Wright, and Rawlings [38]. Both approaches keep even intermediate optimization iterates feasible. This is in contrast to the *simultaneous* dynamic optimization methods, as the collocation method proposed in Biegler [5] or the direct multiple shooting method in Bock et al. [7] and in Santos [37], which allow

*Received by the editors July 30, 2002; accepted for publication (in revised form) April 12, 2004; published electronically March 11, 2005. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) within the priority program 469 "Online-Optimization of Large Scale Systems" and within the research project "Optimization Based Control of Chemical Processes."
http://www.siam.org/journals/sicon/43-5/40071.html

†Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany (m.diehl@iwr.uni-heidelberg.de, scicom@iwr.uni-heidelberg.de, johannes.schloeder@iwr.uni-heidelberg.de).

infeasible state trajectories and are more suitable for trajectory following problems and problems with final state constraints. The real-time iteration scheme belongs to this latter class.

Most approaches in the literature try to solve quickly but exactly an optimal control problem. However, if the time scale for feedback is too short for exact computation, some approximations must be made: for this aim an "instantaneous control" technique has been proposed in the context of PDE models that approximates the optimal feedback control problem by regarding one future time step only (Choi et al. [14, 13]). By construction, this "greedy" approach to optimal control is based on immediate gains only and neglects future costs; thus it may result in poor performance when future costs matter. A somewhat opposed approach to derive a feedback approximation (formulated for ODE models) is based on a system linearization along a fixed optimal trajectory over the whole time horizon and can, e.g., be found in Krämer-Eis and Bock [29] or Kugelmann and Pesch [30]. The approach works well when the nonlinear system is not too largely disturbed and stays close to the nominal trajectory.

The real-time iteration scheme presented in this paper is a different approximation technique for optimal feedback control. It regards the complete time horizon and performs successive linearizations along (approximately) optimal trajectories to provide feedback approximations. Using these linearizations, it iterates toward the rigorous optimal solutions *during the runtime of the process*. In this way a truly nonlinear optimal feedback control is provided whose accuracy is limited, however, by the time needed to converge to the current optimal solutions. In contrast to a somewhat similar idea mentioned in [32], the real-time iteration scheme is based on the direct multiple shooting method [12], a simultaneous optimization technique, which offers excellent convergence properties, particularly for tracking problems and problems with state constraints.

The scheme was introduced in its present form in Diehl et al. [19] going back to ideas presented in Bock et al. [10]. In its actual implementation it is able to treat *differential algebraic equation (DAE)* models (Leineweber [31]), as they often arise in practical applications. It has already been successfully tested for the feedback control of large-scale DAE models with inequality constraints, particularly a binary distillation column [11, 33, 19]. Moreover, it has been applied for the NMPC of a real pilot plant distillation column situated at the *Institut für Systemdynamik und Regelungstechnik (ISR)* of the University of Stuttgart [25, 16, 21].

However, to concentrate on the essential features of the method and on a new proof of contractivity of the scheme, we will restrict the presentation in this paper to ODE models and optimization problems of a simplified type. Moreover, we will start the paper by regarding (nonlinear) *discrete*-time systems first; the multiple shooting technique, which allows us to formulate a discrete-time system from an ODE system, is only introduced later, and very briefly, when the numerical example is presented. Part of the material is also covered in [18]; for technical details about the real-time iteration scheme we refer to [16, 20].

1.1. Real-time optimal feedback control. Throughout this paper, let us consider the simplified nonlinear controlled discrete-time system

$$x_{k+1} = f_k(x_k, u_k), \quad k = 0, \dots, N-1, \tag{1.1}$$

with system states $x_k \in \mathbb{R}^{n_x}$ and controls $u_k \in \mathbb{R}^{n_u}$. The aim of optimal feedback control is to find controls u_k that depend on the current system state x_k and that are

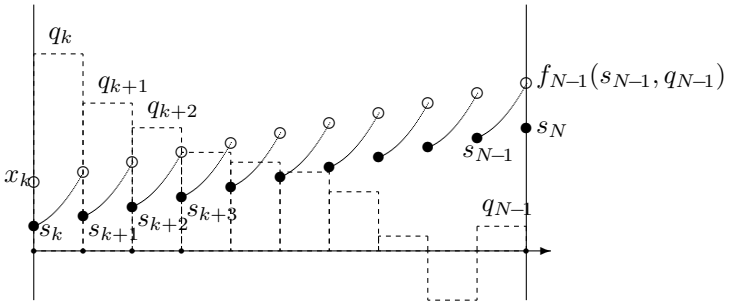


FIG. 1.1. *Problem $P_k(x_k)$: Initial value x_k and problem variables $s_k, \dots, s_N$ and $q_k, \dots, q_{N-1}$.*

optimal with respect to a specified objective. As time advances, we proceed by solving a sequence of nonlinear programming problems $P_k(x_k)$ on shrinking horizons, each with the current system state x_k as initial value (for a visualization, see Figure 1.1). Let us define $P_k(x_k)$ to be the problem

$$\min_{\substack{s_k,\dots,s_N,\\ q_k,\dots,q_{N-1}}} \quad \sum_{i=k}^{N-1} L_i(s_i, q_i) + E(s_N) \tag{1.2a}$$

subject to

$$x_k - s_k = 0, \tag{1.2b}$$

$$f_i(s_i, q_i) - s_{i+1} = 0, \quad i = k, \dots, N-1. \tag{1.2c}$$

The control part $(q_k^*, \dots, q_{N-1}^*)$ of the solution of problem $P_k(x_k)$ allows us to define the optimal feedback control

$$u_k := q_k^*.$$

Note that, due to the dynamic programming property, the optimal control trajectory $(q_0^*, \dots, q_{N-1}^*)$ of the first problem $P_0(x_0)$ would already give all later closed-loop controls $u_0, u_1, \dots u_{N-1}$, if the system behaves as predicted by the model. The practical reason to introduce the closed-loop optimal feedback control is, of course, that it allows us to optimally respond to disturbances.

We will now assume that we know each initial value x_k only at the time when the corresponding control u_k is already needed for application to the real process, and that the solution time for each problem $P_k(x_k)$ is not negligible compared with the runtime of the process. This is a typical situation in realistic applications: ideally, we would like to have the solution of each problem $P_k(x_k)$ instantaneously, but due to finite computing power this usually cannot be accomplished in practice. In this paper we propose and investigate an efficient Newton-type scheme that allows us to approximately solve the optimization problems $P_k(x_k)$ *during the runtime of the real process*.

Remark. In practical applications, inequality path constraints of the form $h(s_i, q_i) \geq 0$, like bounds on controls or states, are of major interest and are usually present in the formulation of the optimization problems $P_k(x_k)$. For the purpose of this paper we leave such constraints unconsidered, since general convergence results for Newton-type methods with changing active sets are difficult to establish. However, we note that in the practical implementation of the real-time iteration scheme they are included and pose no difficulty for the performance of the algorithm.

1.2. Overview. The paper is organized as follows:

- In section 2 we give a review of Newton-type optimization methods for the solution of optimal control problems of type (1.2) and discuss the problem structure.
- The real-time iteration scheme is presented in section 3 building on the previously introduced Newton-type methods. It performs only one Newton-type iteration per optimization problem $P_k(x_k)$, applies the obtained feedback control to the real system, and then proceeds already to the following problem, $P_{k+1}(x_{k+1})$, until the end of the horizon is reached. This allows a particularly fast reaction to disturbances.
- A new contractivity result for the scheme is presented and proven in section 4. The theorem guarantees that the real-time iteration scheme is contracting under mild conditions and delivers approximations to the optimal feedback control with diminishing error.
- Based on the contractivity result, a bound on the loss of optimality of the scheme compared to exact optimal feedback control is established in section 5.
- In section 6 we finally present a numerical example, the real-time control of a kite that shall start to fly loops. A new kite model is developed and a periodic reference orbit is defined. The optimal control problem is to steer the kite into the periodic orbit, starting at an a priori unknown initial value. This example, though of small state dimension, is particularly challenging, because the system is highly nonlinear and unstable.

2. Newton-type optimization methods. In order to solve an optimization problem $P_k(x_k)$, let us first introduce the Lagrange multipliers $\lambda_k, \dots, \lambda_N$ and define the Lagrangian function $\mathcal{L}^k(\lambda_k, s_k, q_k, \dots)$ of problem $P_k(x_k)$ to be

$$\mathcal{L}^k(\cdot) = \sum_{i=k}^{N-1} L_i(s_i, q_i) + E(s_N) + {\lambda_k}^T(x_k - s_k) + \sum_{i=k}^{N-1} {\lambda_{i+1}}^T(f_i(s_i, q_i) - s_{i+1}).$$

Summarizing all variables in a vector $y := (\lambda_k, s_k, q_k, \lambda_{k+1}, s_{k+1}, q_{k+1}, \dots, \lambda_N, s_N) \in \mathbb{R}^{n_k}$,[1] we can formulate necessary optimality conditions of first order (also called *Karush–Kuhn–Tucker* conditions):

$$\nabla_y \mathcal{L}^k(y) = 0. \tag{2.1}$$

To solve this system, the exact full-step Newton–Raphson method would start at an initial guess y_0 and compute a sequence of iterates $y_1, y_2, \dots$ according to

$$y_{i+1} = y_i + \Delta y_i, \tag{2.2}$$

where each Δy_i is the solution of the linearized system

$$\nabla_y \mathcal{L}^k(y_i) + \nabla_y^2 \mathcal{L}^k(y_i)\Delta y_i = 0. \tag{2.3}$$

The Newton-type methods considered in this paper differ from the exact Newton–Raphson method in the way that a part of the exact second derivative $\nabla_y^2 \mathcal{L}^k$, namely, the Hessian $\nabla^2_{(q,s)} \mathcal{L}^k$, is replaced by a (symmetric) approximation. We denote the resulting approximation of $\nabla_y^2 \mathcal{L}^k(y)$ by $J^k(y)$. For our Newton-type method, (2.3) is replaced by the approximation

$$\nabla_y \mathcal{L}^k(y_i) + J^k(y_i)\Delta y_i = 0. \tag{2.4}$$

[1] For simplicity, we omit the index k for the variable y and implicitly assume that $y \in \mathbb{R}^{n_k}$ when not specified otherwise. Note that $n_k = (2n_x + n_u)(N - k) + 2n_x$.

The matrix $\nabla_y^2 \mathcal{L}^k$, respectively, its approximation J^k, is often referred to as the *Karush–Kuhn–Tucker (KKT)* matrix.

2.1. Structure of the Karush–Kuhn–Tucker matrix. The Lagrangian function $\mathcal{L}^k$ of the optimal control problem is *partially separable* [12], and its second derivative has a block diagonal structure:

$$\nabla_y^2 \mathcal{L}^k(y) = \begin{pmatrix}
 & -\mathbb{I} & & & & & & & & & \\
-\mathbb{I} & Q_k & M_k & A_k^T & & & & & & & \\
 & M_k^T & R_k & B_k^T & & & & & & & \\
 & A_k & B_k & & -\mathbb{I} & & & & & & \\
 & & & -\mathbb{I} & Q_{k+1} & M_{k+1} & A_{k+1}^T & & & & \\
 & & & & M_{k+1}^T & R_{k+1} & B_{k+1}^T & & & & \\
 & & & & A_{k+1} & B_{k+1} & & & & & \\
 & & & & & & & \ddots & & & \\
 & & & & & & \ddots & Q_{N-1} & M_{N-1} & A_{N-1}^T & \\
 & & & & & & & M_{N-1}^T & R_{N-1} & B_{N-1}^T & \\
 & & & & & & & A_{N-1} & B_{N-1} & & -\mathbb{I} \\
 & & & & & & & & & -\mathbb{I} & Q_N
\end{pmatrix},$$

where we have set

$$A_i := \frac{\partial f_i}{\partial s_i}, \quad B_i := \frac{\partial f_i}{\partial q_i}, \quad \begin{pmatrix} Q_i & M_i \\ M_i^T & R_i \end{pmatrix} := \nabla^2_{(s_i,q_i)} \mathcal{L}^k, \quad \text{and} \quad Q_N := \nabla^2_{s_N} \mathcal{L}^k.$$

In the approximation $J^k(y)$ of this second derivative, we replace Q_i, M_i, and R_i by approximations $Q_i^H(s_i, q_i, \lambda_{k+1}), M_i^H(s_i, q_i, \lambda_{k+1})$, and $R_i^H(s_i, q_i, \lambda_{k+1})$.

Remark 1. Note that the KKT matrix of each problem $P_k(x_k)$ is completely independent of the value of x_k.

Remark 2. If we split the variables $y = (\lambda_k, s_k, q_k, \dots) = (\lambda_k, s_k, q_k, \tilde{y})$ into a first and a second part, the second part $\tilde{y} \in \mathbb{R}^{n_{k+1}}$ corresponds directly to the variable space of the next, shrunken problem $P_{k+1}(x_{k+1})$, and we can see that the KKT matrix contains the KKT matrix of the next problem as a submatrix, as

$$\nabla_y^2 \mathcal{L}^k(y) = \left(\begin{array}{ccc|c}
 & -\mathbb{I} & & \\
-\mathbb{I} & Q_k & M_k & A_k^T \\
 & M_k^T & R_k & B_k^T \\
\hline
 & A_k & B_k & \\
 & & & \nabla_{\tilde{y}}^2 \mathcal{L}^{k+1}(\tilde{y}) \\
 & & &
\end{array}\right).$$

Remark 3. The favorable structure of the matrix $\nabla_y^2 \mathcal{L}^k(y)$, respectively, its approximation $J^k(y)$, allows an efficient solution of the linear system $J^k(y)x = b$ by a Riccati recursion proposed independently by Pantoja [34] and Dunn and Bertsekas [26]; cf. also [36, 16].

2.2. The constrained Gauss–Newton method. An important special case of the Newton-type methods considered in this paper is the constrained Gauss–Newton method, which is applicable for problems with a least squares form of the objective function

$$\sum_{i=k}^{N-1} \frac{1}{2} \|l_i(s_i, q_i)\|_2^2 + \frac{1}{2} \|e(s_N)\|_2^2. \tag{2.5}$$

For this case, the Hessian block approximations Q_i^H, M_i^H, and R_i^H are defined to be

$$\begin{pmatrix} Q_i^H & M_i^H \\ (M_i^H)^T & R_i^H \end{pmatrix} := \left(\frac{\partial l_i(s_i,q_i)}{\partial(s_i,q_i)}\right)^T \frac{\partial l_i(s_i,q_i)}{\partial(s_i,q_i)}, \quad Q_N^H := \left(\frac{\partial e(s_N)}{\partial s_N}\right)^T \frac{\partial e(s_N)}{\partial s_N}. \tag{2.6}$$

Note that these Hessian block approximations do not depend on the values of the Lagrange multipliers.

2.3. Local convergence. It is well known that the Newton-type scheme (2.4) for the solution of (2.1) converges in a neighborhood $D_k \subset \mathbb{R}^{n_k}$ of a solution y_*^k that satisfies the second order sufficient conditions for optimality of problem $P_k(x_k)$ if $J^k(y)$ approximates $\nabla_y^2 \mathcal{L}^k(y)$ sufficiently well on D_k.

3. Real-time iterations. Let us now go back to the real-time scenario described in section 1.1, where we want to solve the sequence of optimization problems $P_k(x_k)$, but where we do not have the time to iterate each problem to convergence. Let us more specifically assume that each Newton-type iteration needs exactly as much computation time as corresponds to the time that the real process needs for the transition from one system state to the next. Thus, we can only perform *one single Newton-type iteration* for each problem $P_k(x_k)$, and then we have to proceed already to the next problem $P_{k+1}(x_{k+1})$. The *real-time iteration* scheme that we will investigate here is based on a carefully designed transition between subsequent problems. After an initial disturbance, it subsequently delivers approximations u_k for the optimal feedback control that become better and better if no further disturbance occurs, as will be shown in section 4.

It turns out that the computations of the real-time iteration belonging to problem $P_k(x_k)$ can largely be prepared *without knowledge of the value of* x_k so that we can assume that the approximation u_k of the optimal feedback control is instantly available at the time that x_k is known. However, after this feedback has been delivered, we need to *prepare* the next real-time iteration (belonging to problem $P_{k+1}(x_{k+1})$) which needs the full computing time.

In the framework for optimal feedback control on shrinking horizons (1.2), we reduce the number of remaining intervals from one problem $P_k(x_k)$ to the next $P_{k+1}(x_{k+1})$, in order to keep pace with the process development. Therefore, we have to perform real-time iterates in primal-dual variable spaces $\mathbb{R}^{n_0} \supset \cdots \supset \mathbb{R}^{n_k} \supset \mathbb{R}^{n_{k+1}} \supset \cdots \supset \mathbb{R}^{n_{N-1}}$ of different sizes. Let us denote by Π^{k+1} the projection from $\mathbb{R}^{n_k}$ onto $\mathbb{R}^{n_{k+1}}$; i.e., if $y = (\lambda_k, s_k, q_k, \tilde{y}) \in \mathbb{R}^{n_k}$, then $\Pi^{k+1} y = \tilde{y} \in \mathbb{R}^{n_{k+1}}$.

3.1. The real-time iteration algorithm. Let us assume that we have an initial guess $y^0 \in \mathbb{R}^{n_0}$ for the primal-dual variables of problem $P_0(\cdot)$. We set the iteration index k to zero and perform the following steps:

1. Preparation. Based on the initial guess $y^k \in \mathbb{R}^{n_k}$, compute the vector $\nabla_y \mathcal{L}^k(y^k)$ and the matrix $J^k(y^k)$: Note that $J^k(y^k)$ is completely independent of the value of x_k, and that of the vector $\nabla_y \mathcal{L}^k(y^k)$ only the first component $(\nabla_{\lambda_k} \mathcal{L}^k = x_k - s_k)$ depends on x_k. This component will only be needed in the second step. Therefore, prepare the linear algebra computation of $J^k(y^k)^{-1} \nabla_y \mathcal{L}^k(y^k)$ as much as possible without knowledge of the value of x_k (a detailed description how this can be achieved is given in [19] or [16]).
2. Feedback response. At the time when x_k is exactly known, finish the computation of the step vector $\Delta y^k = -J^k(y)^{-1} \nabla_y \mathcal{L}^k(y^k)$ and give the control $u_k := q_k + \Delta q_k$ immediately to the real system.

3. Transition. If $k = N-1$, stop. Otherwise, compute the next initial guess y^{k+1} by adding the step vector to y^k and "shrinking" the resulting variable vector onto $\mathbb{R}^{n_{k+1}}$; i.e., $y^{k+1} := \Pi^{k+1}(y^k + \Delta y^k)$. Set $k = k+1$ and go to 1.

Note that after one iteration belonging to system state x_k we expect the next system state to be $x_{k+1} = f_k(x_k, u_k)$, but this may not be true due to disturbances. The scheme allows an immediate feedback to such disturbances, due to the separation of steps 1 and 2. This separation is only possible because we do *not* require the guess of initial value, s_k, to be equal to the real initial value, x_k. This formulation may be regarded as an *initial value embedding* of each problem into the manifold of perturbed problems. Though this formulation comes quite naturally in the framework of an *infeasible path* (also *simultaneous*) solution strategy, as presented, where optimality and constraints are treated simultaneously, it deserves strong emphasis as it is a feature that is crucial for the success of the method in practice.

We will in the following investigate the contraction properties of the real-time iteration scheme. Though a principal advantage of the scheme lies in this immediate response to disturbances, we will investigate contractivity only under the assumption that after an initial disturbance the system behaves according to the model. This is analogous to the notion of "nominal stability" for an infinite horizon steady state tracking problem.

4. Contractivity of the real-time iterations. In this subsection we investigate the contraction properties of the real-time iteration scheme. The system starts at an initial state x_0, and the real-time algorithm is initialized with an initial guess $y^0 \in D_0 \subset \mathbb{R}^{n_0}$. Let us define the projections D_k of the neighborhood D_0 onto the primal-dual subspaces $\mathbb{R}^{n_k}$; i.e., $D_{k+1} := \Pi^{k+1} D_k$.

We will in the following make use of vector and corresponding matrix norms $\|\cdot\|_k$ defined on the subspaces $\mathbb{R}^{n_k}$. These norms are assumed to be compatible in the sense that $\|\Pi^{k+1} y\|_{k+1} \le \|y\|_k$ and that $\|{\Pi^{k+1}}^T \tilde{y}\|_k = \|\tilde{y}\|_{k+1}$.

THEOREM 4.1 (local contractivity of the real-time iterations). *Let us assume that the Lagrangian functions $\mathcal{L}^k : D_k \to \mathbb{R}$ for all $k = 0, \dots, N$ are twice continuously differentiable and that their second derivative approximations $J^k : D_k \to \mathbb{R}^{n_k \times n_k}$ are continuous and have a bounded inverse $(J^k)^{-1} : D_k \to \mathbb{R}^{n_k \times n_k}$.*

Furthermore, let us assume that there exist a $\kappa < 1$ and an $\omega < \infty$ such that for each $k = 0, \dots, N$ and all $y', y \in D_k$, $\Delta y = y' - y$, and all $t \in [0,1]$ it holds that

$$\left\| \left(J^k(y')\right)^{-1} \left(J^k(y + t\Delta y) - \nabla_y^2 \mathcal{L}^k(y + t\Delta y)\right) \Delta y \right\|_k \le \kappa \|\Delta y\|_k \tag{4.1a}$$

and that

$$\left\| \left(J^k(y')\right)^{-1} \left(J^k(y + t\Delta y) - J^k(y)\right) \Delta y \right\|_k \le \omega t \|\Delta y\|_k^2, \tag{4.1b}$$

and such that for each $k = 0, \dots, N-1$ it additionally holds that

$$\left\| \left(J^{k+1}(\Pi^{k+1} y')\right)^{-1} \Pi^{k+1} \left(J^k(y + t\Delta y) - J^k(y)\right) \Delta y \right\|_{k+1} \le \omega t \|\Delta y\|_k^2. \tag{4.1c}$$

We suppose that the first step $\Delta y^0 := J^0(y^0)^{-1} \nabla_y \mathcal{L}^0(y^0)$ starting at the initial guess y^0 is sufficiently small so that

$$\delta_0 := \kappa + \frac{\omega}{2} \|\Delta y^0\|_0 < 1 \tag{4.1d}$$

and that the ball

$$B_0 := \left\{ y \in \mathbb{R}^{n_0} \,|\, \|y - y^0\|_0 \le \frac{\|\Delta y^0\|_0}{1-\delta_0} \right\} \tag{4.1e}$$

is completely contained in D_0. Under these conditions the real-time iterates $y^0, \dots, y^N$ defined by

$$\Delta y^k := -J^k(y^k)^{-1} \nabla_y \mathcal{L}^k(y^k), \quad y^{k+1} := \Pi^{k+1}(y^k + \Delta y^k)$$

(where $\mathcal{L}^k$ is the Lagrangian function corresponding to problem $P_k(x_k)$ with the system state obtained according to the closed-loop dynamics $x_{k+1} = f_k(x_k, u_k)$, $u_k := q_k^k + \Delta q_k^k$) are well defined and stay in the projections of the ball B_0, i.e.,

$$y^k \in \Pi^k \cdots \Pi^1 B_0 \subset D_k, \tag{4.2}$$

and satisfy the contraction condition

$$\|\Delta y^{k+1}\|_{k+1} \le \left(\kappa + \frac{\omega}{2} \|\Delta y^k\|_k \right) \|\Delta y^k\|_k =: \delta_k \|\Delta y^k\|_k \le \delta_0 \|\Delta y^k\|_k. \tag{4.3}$$

Furthermore, the iterates y^k approach the exact stationary points y_*^k of the corresponding problems $P_k(x_k)$:

$$\|y^k - y_*^k\|_k \le \frac{\|\Delta y^k\|_k}{1-\delta_k} \le \frac{(\delta_0)^k \|\Delta y^0\|_0}{1-\delta_0}. \tag{4.4}$$

Proof. We divide the proof into three parts, corresponding to the properties (4.3), (4.2), and (4.4).

Contraction property. We will first show that the contraction property (4.3) holds. By adding zero to the defining equation of Δy^{k+1} we get

$$\begin{aligned} -\Delta y^{k+1} &= J^{k+1}(y^{k+1})^{-1} \nabla_y \mathcal{L}^{k+1}(y^{k+1}) \\ &= J^{k+1}(y^{k+1})^{-1} \left(\nabla_y \mathcal{L}^{k+1}(y^{k+1}) - \Pi^{k+1} \left(\nabla_y \mathcal{L}^k(y^k) + J^k(y^k) \Delta y^k \right) \right). \end{aligned} \tag{4.5}$$

Using the notation $y^k = (\lambda_k^k, s_k^k, q_k^k, \lambda_{k+1}^k, s_{k+1}^k, q_{k+1}^k, \dots)$ we observe that

$$\begin{aligned} \nabla_y \mathcal{L}^{k+1}(y^{k+1}) &= \nabla_y \mathcal{L}^{k+1}(\Pi^{k+1}(y^k + \Delta y^k)) = \begin{pmatrix} x_{k+1} - (s_{k+1}^k + \Delta s_{k+1}^k) \\ \vdots \end{pmatrix} \\ &= \begin{pmatrix} f_k(s_k^k + \Delta s_k^k, q_k^k + \Delta q_k^k) - (s_{k+1}^k + \Delta s_{k+1}^k) \\ \vdots \end{pmatrix} \\ &= \Pi^{k+1} \nabla_y \mathcal{L}^k(y^k + \Delta y^k), \end{aligned}$$

because $x_{k+1} = f_k(x_k, u_k) = f_k(s_k^k + \Delta s_k^k, q_k^k + \Delta q_k^k)$ if the system was undisturbed.[2] Therefore, we can continue to transform Δy^{k+1} and write

$$\begin{aligned} -\Delta y^{k+1} &= J^{k+1}(y^{k+1})^{-1} \Pi^{k+1} \left(\nabla_y \mathcal{L}^k(y^k + \Delta y^k) - \nabla_y \mathcal{L}^k(y^k) - J^k(y^k) \Delta y^k \right) \\ &= J^{k+1}(y^{k+1})^{-1} \Pi^{k+1} \int_0^1 (\nabla_y^2 \mathcal{L}^k(y^k + t\Delta y^k) - J^k(y^k)) \Delta y^k \, dt \\ &= J^{k+1}(y^{k+1})^{-1} \Pi^{k+1} \int_0^1 (\nabla_y^2 \mathcal{L}^k(y^k + t\Delta y^k) - J^k(y^k + t\Delta y^k)) \Delta y^k \, dt \\ &\quad + J^{k+1}(y^{k+1})^{-1} \Pi^{k+1} \int_0^1 (J^k(y^k + t\Delta y^k) - J^k(y^k)) \Delta y^k \, dt. \end{aligned} \tag{4.6}$$

[2] Note that $s_k^k + \Delta s_k^k = x_k$ due to the linearity of the constraint $x_k - s_k = 0$ and that $u_k = q_k^k + \Delta q_k^k$ by definition.

Noting that, with $\tilde{y} := \Pi^{k+1} y$,

$$\begin{aligned}\Pi^{k+1}\left(\nabla_y^2 \mathcal{L}^k(y) - J^k(y)\right) &= \Pi^{k+1}\left(\begin{array}{ccc|c} & 0 & & \\ 0 & \Delta Q_k & \Delta M_k & 0 \\ & \Delta M_k^T & \Delta R_k & 0 \\ \hline & 0 & 0 & \\ & & & \nabla_{\tilde{y}}^2 \mathcal{L}^{k+1}(\tilde{y}) - J^{k+1}(\tilde{y}) \end{array}\right) \\ &= \left(\begin{array}{ccc|c} 0 & 0 & 0 & \\ \vdots & \vdots & \vdots & \nabla_{\tilde{y}}^2 \mathcal{L}^{k+1}(\tilde{y}) - J^{k+1}(\tilde{y}) \\ 0 & 0 & 0 & \end{array}\right)\end{aligned}$$

and abbreviating $\tilde{y}^k := \Pi^{k+1} y^k$, $\Delta\tilde{y}^k := \Pi^{k+1}\Delta y^k$, we can exploit assumption (4.1a) to obtain

$$\begin{aligned}&\|J^{k+1}(y^{k+1})^{-1}\,\Pi^{k+1}(\nabla_y^2 \mathcal{L}^k(y^k + t\Delta y^k) - J^k(y^k + t\Delta y^k))\Delta y^k\|_{k+1} \\ &\quad = \|J^{k+1}(y^{k+1})^{-1}(\nabla_y^2 \mathcal{L}^{k+1}(\tilde{y}^k + t\Delta\tilde{y}^k) - J^{k+1}(\tilde{y}^k + t\Delta\tilde{y}^k))\Delta\tilde{y}^k\|_{k+1} \\ &\quad \le \kappa\|\Delta\tilde{y}^k\|_{k+1} = \kappa\|\Pi^{k+1}\Delta y^k\|_{k+1} \le \kappa\|\Delta y^k\|_k.\end{aligned}$$

Making also use of assumption (4.1c), we can, building on (4.6), prove the left inequality of the contraction property (4.3):

$$\|\Delta y^{k+1}\|_{k+1} \le \kappa\|\Delta y^k\|_k + \int_0^1 \omega t\|\Delta y^k\|_k^2\, dt. = \kappa\|\Delta y^k\|_k + \frac{1}{2}\omega\|\Delta y^k\|_k^2 =: \delta_k\|\Delta y^k\|_k.$$

With the help of condition (4.1d) ($\delta_0 < 1$) it is straightforward to deduce inductively that

$$\delta_{k+1} = \kappa + \frac{\omega}{2}\|\Delta y^{k+1}\|_{k+1} \le \kappa + \frac{\omega}{2}\delta_k\|\Delta y^k\|_k \le \delta_k \le \delta_0,$$

which proves the remaining part of (4.3).

Well definedness. To show that the iterates remain inside the domains of definition as stated in (4.2) we first observe that

$$\|\Delta y^k\|_k \le \delta_{k-1}\delta_{k-2}\dots\delta_0\|\Delta y^0\|_0 \le (\delta_0)^k\|\Delta y^0\|_0.$$

Using the representation

$$\begin{aligned}y^k &= \Pi^k(y^{k-1} + \Delta y^{k-1}) = \Pi^k(\Pi^{k-1}(y^{k-2} + \Delta y^{k-2}) + \Delta y^{k-1}) \\ &= \Pi^k(\Pi^{k-1}(\dots\Pi^1(y^0 + \Delta y^0)\dots) + \Delta y^{k-1}) \\ &= \Pi^k\cdots\Pi^1 y^0 + \Pi^k\cdots\Pi^1\Delta y^0 + \cdots + \Pi^k\Delta y^{k-1},\end{aligned}$$

we can find $y' := y^0 + (\Pi^k\cdots\Pi^1)^T(y^k - \Pi^k\cdots\Pi^1 y^0)$ such that

$$\begin{aligned}\|y' - y^0\|_0 &= \|(\Pi^k\cdots\Pi^1)^T(y^k - \Pi^k\cdots\Pi^1 y^0)\|_0 = \|y^k - \Pi^k\cdots\Pi^1 y^0\|_k \\ &\le \textstyle\sum_{i=0}^{k-1}\|\Delta y^i\|_i \le \|\Delta y^0\|_0\sum_{i=0}^{k-1}(\delta_0)^i \le \frac{\|\Delta y^0\|_0}{1-\delta_0},\end{aligned}$$

i.e., $y' \in B_0$ and $y^k = \Pi^k\dots\Pi^1 y'$, i.e., $y^k \in \Pi^k\dots\Pi^1 B_0$, as desired.

Distance to optimal solutions. It remains to be shown that the iterates y^k approach the exact solutions of the corresponding problems $P_k(x_k)$ as stated in (4.4). For this aim we devise a hypothetical standard Newton-type algorithm as introduced in (2.4) that allows us to compute the exact solution y_*^k of $P_k(x_k)$. As a by-product, we obtain a bound on the distance of y_*^k from y^k.

The hypothetical algorithm would proceed by starting at $y_0^k := y^k$ and iterating with iterates $y_1^k, y_2^k, \dots$ according to

$$y_{i+1}^k := y_i^k + \Delta y_i^k, \quad \Delta y_i^k := -J^k(y_i^k)^{-1} \nabla_y \mathcal{L}^k(y_i^k).$$

Note that the first step Δy_0^k is identical to Δy^k. It is for this hypothetical algorithm only that we need assumption (4.1b). Due to this condition and (4.1a), we have the contraction property

$$\|\Delta y_{i+1}^k\|_k \le \left(\kappa + \frac{\omega}{2}\|\Delta y_i^k\|_k\right) \|\Delta y_i^k\|_k$$

as can be shown by a well-known technique for Newton-type methods (see, e.g., [9]):

(4.7)

$$\begin{aligned}
\|\Delta y_{i+1}^k\|_k &= \|J^k(y_{i+1}^k)^{-1} \cdot \nabla_y \mathcal{L}^k(y_{i+1}^k)\|_k \\
&= \|J^k(y_{i+1}^k)^{-1} \cdot (\nabla_y \mathcal{L}^k(y_{i+1}^k) - \nabla_y \mathcal{L}^k(y_i^k) - J^k(y_i^k) \cdot \Delta y_i^k)\|_k \\
&= \|J^k(y_{i+1}^k)^{-1} \cdot \textstyle\int_0^1 (\nabla_y^2 \mathcal{L}^k(y_i^k + t\Delta y_i^k) - J^k(y_i^k)) \cdot \Delta y_i^k \, dt\|_k \\
&= \|J^k(y_{i+1}^k)^{-1} \cdot \textstyle\int_0^1 (\nabla_y^2 \mathcal{L}^k(y_i^k + t\Delta y_i^k) - J^k(y_i^k + t\Delta y_i^k))\Delta y_i^k \, dt \\
&\quad + J^k(y_{i+1}^k)^{-1} \cdot \textstyle\int_0^1 (J^k(y_i^k + t\Delta y_i^k) - J^k(y_i^k))\Delta y_i^k \, dt\|_k \\
&\le \textstyle\int_0^1 \|J^k(y_{i+1}^k)^{-1} (\nabla_y^2 \mathcal{L}^k(y_i^k + t\Delta y_i^k) - J^k(y_i^k + t\Delta y_i^k))\Delta y_i^k\|_k \, dt \\
&\quad + \textstyle\int_0^1 \|J^k(y_{i+1}^k)^{-1} (J^k(y_i^k + t\Delta y_i^k) - J^k(y_i^k))\Delta y_i^k\|_k \, dt \\
&\le \kappa\|\Delta y_i^k\|_k + \textstyle\int_0^1 \omega t \|\Delta y_i^k\|_k^2 \, dt \\
&= \left(\kappa + \tfrac{\omega}{2}\|\Delta y_i^k\|_k\right) \|\Delta y_i^k\|_k.
\end{aligned}$$

Together with the property

$$\kappa + \frac{\omega}{2}\|\Delta y_0^k\|_k = \kappa + \frac{\omega}{2}\|\Delta y^k\|_k = \delta_k < 1,$$

this leads again to the conclusion that $\|\Delta y_i^k\|_k \le (\delta_k)^i \|\Delta y^k\|_k$, so that $y_0^k, y_1^k, y_2^k, \dots$ is a Cauchy sequence and remains in the ball

$$B_k := \left\{ y \in \mathbb{R}^{n_k} \,|\, \|y - y^k\|_k \le \frac{\|\Delta y^k\|_k}{1 - \delta_k} \right\}$$

and thus converges toward a point $y_*^k \in B_k$, which satisfies $\nabla_y \mathcal{L}^k(y_*^k) = 0$ due to the boundedness of J^k on the (compact) ball B_k, as $\nabla_y \mathcal{L}^k(y_i^k) = -J^k(y_i^k)\Delta y_i^k \to 0$ for $i \to \infty$. ∎

5. Comparison with optimal feedback control. To assess the performance of the proposed real-time iteration scheme, we will compare the resulting system trajectory with the one which would have been obtained by exact optimal feedback control. For this aim, we denote by $u_0, \dots, u_{N-1}$ and $x_1, \dots, x_N$ the control and system state trajectories obtained by an application of the real-time iteration scheme to the system starting at the state x_0, when the iteration scheme was initialized with an initial guess $y^0 = (\lambda_0^0, s_0^0, q_0^0, \dots, \lambda_N^0, s_N^0)$, as in Theorem 4.1. On the other hand,

we denote by $q_0^*, \dots, q_{N-1}^*$ and $s_1^*, \dots, s_N^*$ the corresponding trajectories which would have been obtained by an application of exact optimal feedback control, starting at the same initial state x_0. Note that this trajectory is contained in the exact primal-dual solution vector $y_*^0 = (\lambda_0^*, s_0^*, q_0^*, \lambda_1^*, s_1^*, q_1^*, \dots, \lambda_N^*, s_N^*)$ of problem $P_0(x_0)$, as already pointed out in section 1.1.

THEOREM 5.1 (loss of optimality). *Let us in addition to the assumptions of Theorem* 4.1 *suppose that the Hessian of the Lagrangian $\mathcal{L}^0(\cdot)$ of problem $P_0(\cdot)$ is bounded on B_0, i.e.,*

$$\|\nabla_y^2 \mathcal{L}^0(y)\|_0 \le C \quad \forall y \in B_0. \tag{5.1}$$

Then the objective values, on the one hand evaluated at the closed-loop trajectory resulting from the real-time iteration scheme and on the other hand at the trajectory resulting from optimal feedback control

$$F_{\text{real}} := \sum_{i=k}^{N-1} L_i(x_i, u_i) + E(x_N) \quad \textit{and} \quad F_{\text{opt}} := \sum_{i=k}^{N-1} L_i(s_i^*, q_i^*) + E(s_N^*),$$

can be compared by

$$F_{\text{real}} \le F_{\text{opt}} + 2C \left(\frac{\delta_0}{1-\delta_0} \right)^2 \|\Delta y^0\|_0^2. \tag{5.2}$$

In particular, if $\kappa = 0$ (as for the exact Newton method), the loss of optimality is of fourth order in the size of the first step Δy^0:

$$F_{\text{real}} \le F_{\text{opt}} + \frac{C}{2} \left(\frac{\omega}{1 - \frac{\omega}{2}\|\Delta y^0\|_0} \right)^2 \|\Delta y^0\|_0^4. \tag{5.3}$$

Proof. First note that both the real-time iteration trajectory $(x_0, u_0, x_1, \dots, x_N)$ and the optimal feedback control trajectory $(s_0^*, q_0^*, s_1^*, \dots) = (x_0, q_0^*, s_1^*, \dots)$ are feasible "points" for the optimization problem $P_0(x_0)$. Let us augment the real-time iteration trajectory to a primal-dual point $y_{\text{real}} := (\lambda_0, x_0, u_0, \dots, \lambda_N, x_N)$, which is obtained by

$$y_{\text{real}} := y^0 + \Delta y^0 + {\Pi^1}^T \Delta y^1 + (\Pi^2 \Pi^1)^T \Delta y^2 + \cdots + (\Pi^N \cdots \Pi^1)^T \Delta y^N.$$

From the contractivity condition (4.3), it can easily be verified that $\|y_{\text{real}} - y^0\|_0 \le \frac{\|\Delta y^0\|_0}{1-\delta_0}$, i.e., that $y_{\text{real}} \in B_0$. We similarly see that

$$\|y_{\text{real}} - (y^0 + \Delta y^0)\|_0 \le \frac{\delta_0 \|\Delta y^0\|_0}{1-\delta_0} \quad \text{and that} \quad \|y_*^0 - (y^0 + \Delta y^0)\|_0 \le \frac{\delta_0 \|\Delta y^0\|_0}{1-\delta_0},$$

where the latter bound is due to contraction property (4.7) for the hypothetical Newton-type iterations toward the solution of $P_0(x_0)$, and the fact that the first step Δy_0^0 of these iterations coincides with the step vector Δy^0 of the real-time iterations. We can conclude that

$$\|y_{\text{real}} - y_*^0\|_0 \le \|y_{\text{real}} - (y^0 + \Delta y^0)\|_0 + \|y_*^0 - (y^0 + \Delta y^0)\|_0 \le 2 \frac{\delta_0 \|\Delta y^0\|_0}{1-\delta_0}. \tag{5.4}$$

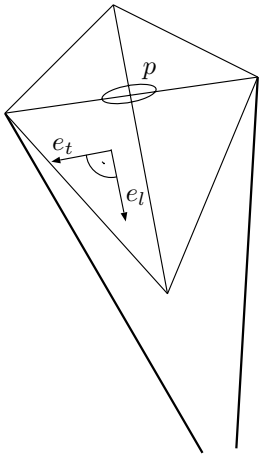


FIG. 6.1. *A picture of the kite from the pilot's point of view.*

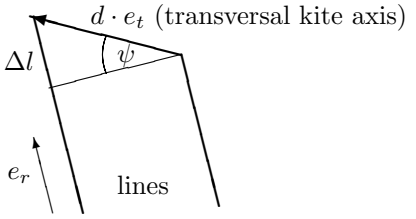


FIG. 6.2. *The kite seen from the top and visualization of the roll angle ψ.*

Because of feasibility of the primal-dual points y_{real} and y_*^0 the values of the Lagrangian function coincide with those of the objective, so that we can deduce

$$\begin{aligned} F_{\text{real}} - F_{\text{opt}} &= \mathcal{L}^0(y_{\text{real}}) - \mathcal{L}^0(y_*^0) = \int_0^1 \nabla_y \mathcal{L}^0(y_*^0 + t_1(y_{\text{real}} - y_*^0))^T \, (y_{\text{real}} - y_*^0) \, dt_1 \\ &= \int_0^1 \Big(\int_0^{t_1} \nabla_y^2 \mathcal{L}^0(y_*^0 + t_2(y_{\text{real}} - y_*^0))(y_{\text{real}} - y_*^0) \, dt_2 \Big)^T (y_{\text{real}} - y_*^0) \, dt_1 \\ &= (y_{\text{real}} - y_*^0)^T \Big(\int_0^1 \int_0^{t_1} \nabla_y^2 \mathcal{L}^0(y_*^0 + t_2(y_{\text{real}} - y_*^0)) \, dt_2 \, dt_1 \Big)^T (y_{\text{real}} - y_*^0), \end{aligned}$$

where we have used the fact that $\nabla_y \mathcal{L}^0(y_*^0) = 0$. We conclude with (5.1) and (5.4) that

$$F_{\text{real}} - F_{\text{opt}} \le \frac{1}{2} C \| y_{\text{real}} - y_*^0 \|_0^2 \le \frac{1}{2} C \left(2 \frac{\delta_0 \| \Delta y^0 \|_0}{1 - \delta_0} \right)^2 . \qquad \square$$

6. Numerical example: Control of a looping kite. In order to demonstrate the versatility of the proposed real-time iteration scheme we present here the control of an airborne kite as a periodic control example. The kite is held by two lines which allow control of the roll angle of the kite; see Figures 6.1 and 6.2. By pulling one line the kite will turn in the direction of the line being pulled. This allows an experienced kite pilot to fly loops or similar figures. The aim of our automatic control is to make the kite fly a figure that may be called a "lying eight," with a cycle time of 8 seconds (see Figure 6.3). The corresponding orbit is not open-loop stable, so that feedback has to be applied during the flight; we will show simulation results where our proposed real-time iteration scheme is used to control the kite, starting at a largely disturbed initial state x_0, over three periods, with a sampling time of one second.

6.1. The dual line kite model. The movement of the kite in the sky can be modeled by Newton's laws of motion and a suitable model for the aerodynamic force. The most difficulty lies in the determination of suitable coordinate systems; we will first describe the kite's motion in polar coordinates, and second we will determine the direction of the aerodynamic forces.

6.1.1. Newton's laws of motion in polar coordinates. The position $p \in \mathbb{R}^3$ of the kite can be modeled in three-dimensional Euclidean space, choosing the position of the kite pilot as the origin, and the third component p_3 to be the height of the kite above the ground. With m denoting the mass of the kite and $F \in \mathbb{R}^3$ the total force acting on the kite, Newton's law of motion reads

$$\ddot{p} = \frac{d^2 p}{dt^2} = \frac{F}{m}.$$

Let us introduce polar coordinates θ, ϕ, r:

$$p = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = \begin{pmatrix} r\sin(\theta)\cos(\phi) \\ r\sin(\theta)\sin(\phi) \\ r\cos(\theta) \end{pmatrix}.$$

Note that the distance r between pilot and kite is usually constant during flight, and θ is the angle that the lines form with the vertical. Let us introduce a local right-handed coordinate system with the three basis vectors

$$e_\theta = \begin{pmatrix} \cos(\theta)\cos(\phi) \\ \cos(\theta)\sin(\phi) \\ -\sin(\theta) \end{pmatrix}, \quad e_\phi = \begin{pmatrix} -\sin(\phi) \\ \cos(\phi) \\ 0 \end{pmatrix}, \quad \text{and} \quad e_r = \begin{pmatrix} \sin(\theta)\cos(\phi) \\ \sin(\theta)\sin(\phi) \\ \cos(\theta) \end{pmatrix}.$$

Defining $F_\theta := F \cdot e_\theta, F_\phi := F \cdot e_\phi$, and $F_r := F \cdot e_r$, we can write Newton's laws of motion in the form

$$\begin{aligned} r\ddot{\theta} - r\sin(\theta)\cos(\theta)\dot{\phi}^2 + 2\dot{r}\dot{\theta} &= \frac{F_\theta}{m}, \\ r\sin(\theta)\ddot{\phi} + 2r\cos(\theta)\dot{\phi}\dot{\theta} + 2\sin(\theta)\dot{r}\dot{\phi} &= \frac{F_\phi}{m}, \\ \ddot{r} - r\dot{\theta}^2 - r\sin^2(\theta)\dot{\phi}^2 &= \frac{F_r}{m}. \end{aligned} \tag{6.1}$$

If the length of the lines, denoted by r, is kept constant, all terms involving time derivatives of r will drop out. Furthermore, the last equation (6.1) will become redundant, as any acting force F_r' in the radial direction will automatically be augmented by a constraint force contribution $F_c := F_r + mr\dot{\theta}^2 + mr\sin^2(\theta)\dot{\phi}^2$ so that (6.1) is satisfied with $F_r := F_r' - F_c$. In this case we can regard only the components F_θ and F_ϕ which are not changed by the constraint force. The equations of motion[3] simplify to

$$\ddot{\theta} = \frac{F_\theta}{rm} + \sin(\theta)\cos(\theta)\dot{\phi}^2, \tag{6.2}$$

$$\ddot{\phi} = \frac{F_\phi}{rm\sin(\theta)} - 2\cot(\theta)\dot{\phi}\dot{\theta}. \tag{6.3}$$

In our model, the force F acting on the kite consists of three contributions, constraint force $-F_c e_r$, gravitational force F^{gra}, and aerodynamic force F^{aer}. In Cartesian coordinates, $F^{\text{gra}} = (0, 0, -mg)^T$ with $g = 9.81$ m s^{-2} being the earth's gravitational acceleration. In local coordinates we therefore have

$$F_\theta = F_\theta^{\text{gra}} + F_\theta^{\text{aer}} = \sin(\theta)mg + F_\theta^{\text{aer}} \quad \text{and} \quad F_\phi = F_\phi^{\text{aer}}.$$

It remains to derive an expression for the aerodynamic force F^{aer}.

[3]Note that the validity of these equations requires that $F_c \geq 0$, as a line can only pull and not push.

6.1.2. Kite orientation and the aerodynamic force. To model the aerodynamic force that is acting on the kite, we first assume that the kite's trailing edge is always pulled by the tail into the direction of the effective wind, as seen from the kite's body-fixed frame. This assumption can be regarded as the limiting case of very large tail force. It crucially simplifies the model by allowing us to disregard angular momentum and the moments acting on the kite. Under this assumption we are able to determine the kite orientation as an explicit function of position and velocity only, as shown in the following.

By the large tail force assumption, the kite's longitudinal axis is always in line with the effective wind vector $w_e := w - \dot{p}$, where $w = (v_w, 0, 0)^T$ is the wind as seen from the earth system, and $\dot{p}$ is the kite velocity. If we introduce a unit vector e_l pointing from the front toward the trailing edge of the kite (cf. Figure 6.1), we therefore assume that

$$e_l = \frac{w_e}{\|w_e\|}.$$

The transversal axis of the kite can be described by a perpendicular unit vector e_t that is pointing from the left to the right wing tip. Clearly, it is orthogonal to the longitudinal axis, i.e.,

$$e_t \cdot e_l = \frac{e_t \cdot w_e}{\|w_e\|} = 0. \tag{6.4}$$

The orientation of the transversal axis e_t against the lines' axis (which is given by the vector e_r) can be influenced by the length difference Δl of the two lines. If the distance between the two lines' fixing points on the kite is d, then the vector from the left to the right fixing point is de_t, and the projection of this vector onto the lines' axis should equal $\Delta l = d\, e_t \cdot e_r$, being positive if the left hand's lines wingtip is farther away from the pilot; cf. Figure 6.2. Let us define the *roll angle* ψ to be

$$\psi = \arcsin\left(\frac{\Delta l}{d}\right).$$

We will assume that we control this angle ψ directly. It determines the orientation of e_t which has to satisfy

$$e_t \cdot e_r = \frac{\Delta l}{d} = \sin(\psi). \tag{6.5}$$

A third requirement that e_t should satisfy is that

$$(e_l \times e_t) \cdot e_r = \frac{w_e \times e_t}{\|w_e\|} \cdot e_r > 0, \tag{6.6}$$

which takes account of the fact that the kite is always in the same orientation with respect to the lines.

How does one find a unit vector e_t that satisfies these requirements (6.4)–(6.6)? Using the projection w_e^p of the effective wind vector w_e onto the tangent plane spanned by e_θ and e_ϕ,

$$w_e^p := e_\theta(e_\theta \cdot w_e) + e_\phi(e_\phi \cdot w_e) = w_e - e_r(e_r \cdot w_e),$$

we can define the orthogonal unit vectors

$$e_w := \frac{w_e^p}{\|w_e^p\|} \quad \text{and} \quad e_o := e_r \times e_w,$$

so that (e_w, e_o, e_r) forms an orthogonal right-handed coordinate basis. Note that in this basis the effective wind w_e has no component in the e_o direction, as $w_e = \|w_e^p\| e_w + (w_e \cdot e_r) e_r$. We will show that the definition

$$e_t := e_w(-\cos(\psi)\sin(\eta)) + e_o(\cos(\psi)\cos(\eta)) + e_r \sin(\psi) \tag{6.7}$$

with

$$\eta := \arcsin\left(\frac{w_e \cdot e_r}{\|w_e^p\|} \tan(\psi)\right)$$

satisfies the requirements (6.4)–(6.6).[4] Equation (6.4) can be verified by substitution of the definition of η into

$$e_t \cdot w_e = -\cos(\psi)\sin(\eta)\|w_e^p\| + \sin(\psi)(w_e \cdot e_r) = 0.$$

Equation (6.5) is trivially satisfied, and (6.6) can be verified by calculation of

$$\begin{aligned}(w_e \times e_t) \cdot e_r &= (w_e \cdot e_w)\cos(\psi)\cos(\eta) - (w_e \cdot e_o)(-\cos(\psi)\sin(\eta)) \\ &= \|w_e^p\| \cos(\psi)\cos(\eta)\end{aligned}$$

(where we used the fact that $w_e \cdot e_o = 0$). For angles ψ and η in the range from $-\pi/2$ to $\pi/2$ this expression is always positive. The above considerations allow us to determine the orientation of the kite depending on the control ψ and the effective wind w_e only. Note that the considerations would break down if the projection of the effective wind w_e^p would be equal to zero, if $|\psi| \geq \frac{\pi}{2}$, or if

$$\left|\frac{w_e \cdot e_r}{\|w_e^p\|} \tan(\psi)\right| > 1.$$

The two vectors $e_n := e_l \times e_t$ and e_l are the directions of aerodynamic lift and drag, respectively. To compute the magnitudes F_{L} and F_{D} of lift and drag we assume that the lift and drag coefficients C_{L} and C_{D} are constant, so that we have

$$F_{\mathrm{L}} = \frac{1}{2}\rho\|w_e\|^2 A C_{\mathrm{L}} \quad \text{and} \quad F_{\mathrm{D}} = \frac{1}{2}\rho\|w_e\|^2 A C_{\mathrm{D}},$$

with ρ being the density of air and A being the characteristic area of the kite. Given the directions and magnitudes of lift and drag, we can compute F^{aer} as their sum, yielding $F^{\mathrm{aer}} = F_{\mathrm{L}} e_n + F_{\mathrm{D}} e_l$, or, in the local coordinate system,

$$F_\theta^{\mathrm{aer}} = F_{\mathrm{L}}(e_n \cdot e_\theta) + F_{\mathrm{D}}(e_l \cdot e_\theta) \;\text{ and }\; F_\phi^{\mathrm{aer}} = F_{\mathrm{L}}(e_n \cdot e_\phi) + F_{\mathrm{D}}(e_l \cdot e_\phi).$$

The system parameters that have been chosen for the simulation model are listed in Table 6.1. Defining the system state $\xi := (\theta, \phi, \dot\theta, \dot\phi)^T$ and the control $u := \psi$, we can summarize the four system equations, i.e., (6.2)–(6.3) and the trivial equations $\frac{\partial\theta}{\partial t} = \dot\theta$, $\frac{\partial\phi}{\partial t} = \dot\phi$, in the short form

$$\dot\xi = \hat f(\xi, u).$$

[4] It is interesting to note that the assignment of e_t can be made more transparent by considering a rotation from the (e_w, e_o, e_r) tangential frame to the body frame (e_l, e_t, e_n), with $e_n := e_l \times e_t$. Specifically, starting from (e_w, e_o, e_r), we rotate about the e_r-axis through the *yaw* angle $-\eta$ and then *roll* through the angle ψ about the e_l-axis. From this we find that e_t as the second basis vector of the body frame is represented by (6.7) in the tangential frame (e_w, e_o, e_r).

TABLE 6.1
The kite parameters.

Name	Symbol	Value
Line length	r	50 m
Kite mass	m	1 kg
Wind velocity	v_w	6 m/s
Density of air	ρ	1.2 kg/m^3
Characteristic area	A	0.5 m^2
Lift coefficient	C_{L}	1.5
Drag coefficient	C_{D}	0.29

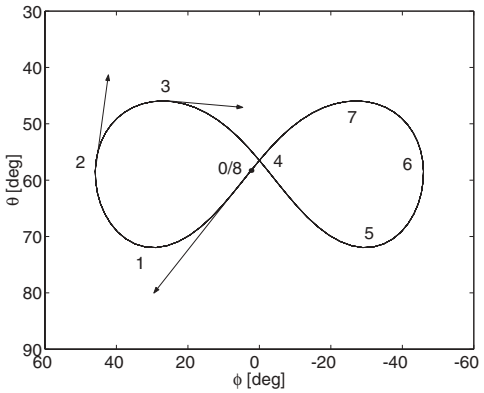


FIG. 6.3. *Periodic orbit plotted in the (ϕ, θ)-plane, as seen by the kite pilot. The dots separate intervals of one second.*

6.2. A periodic orbit. Using the above system model, a periodic orbit was determined that can be characterized as a "lying eight" and which is depicted as a (ϕ, θ)-plot in Figure 6.3. The wind is assumed to blow in the direction of the p_1-axis ($\theta = 90^o$ and $\phi = 0^o$). The periodic solution was computed using an off-line variant of the direct multiple shooting method (MUSCOD-II, due to Leineweber [31, 22, 23]), imposing periodicity conditions with period $T = 8$ seconds and suitable state bounds and a suitable objective function in order to yield a solution that was considered to be a meaningful reference orbit. We will denote the periodic reference solution by $\xi_r(t)$ and $u_r(t)$. This solution is defined for all $t \in (-\infty, \infty)$ and satisfies the periodicity condition $\xi_r(t+T) = \xi_r(t)$ and $u_r(t+T) = u_r(t)$. It is interesting to note that small errors accumulate very quickly so that the uncontrolled system will not stay in the periodic orbit very long during a numerical simulation; this observation can be confirmed by investigating the asymptotic stability properties of the periodic orbit [16], which shows that local errors are amplified by a factor of more than 5 during each period. Thus, the open-loop system is highly unstable in the periodic orbit.

We want to mention that the kite model and the periodic orbit may serve as a challenging benchmark problem for nonlinear periodic control and are available in MATLAB/SIMULINK format [17].

6.3. The optimal control problem. Given an arbitrary initial state x_0 (that we do not know in advance) we want the kite to fly three times the figure of Figure 6.3, on a time horizon of $3T = 24$ seconds. By using the figure as a reference orbit, we

formulate an optimal control problem which has the objective of bringing the system close to the reference orbit. For this aim we define a Lagrange term of least squares type

$$L(\xi, u, t) := \frac{1}{2}(\xi - \xi_r(t))^T Q(\xi - \xi_r(t)) + \frac{1}{2}(u - u_r(t))^T R(u - u_r(t))$$

with diagonal weighting matrices

$$Q := \mathrm{diag}(0.4,\ 1,\ \mathrm{s}^2,\ \mathrm{s}^2)\frac{1}{\mathrm{s}} \quad \text{and} \quad R := 1.0 \cdot 10^{-2}\mathrm{deg}^{-2}\mathrm{s}^{-1}.$$

Using these definitions, we formulate the following optimal control problem on the time horizon of interest $[0, 3T]$:

$$\begin{aligned} &\min_{u(\cdot),\xi(\cdot)} \int_0^{3T} L(\xi(t), u(t), t)\, dt \\ &\text{subject to} \\ &\qquad \dot{\xi}(t) = \hat{f}(\xi(t), u(t)) \qquad \forall t \in [0, 3T], \\ &\qquad \xi(0) = x_0. \end{aligned} \tag{6.8}$$

6.4. Direct multiple shooting formulation. In order to reformulate the above continuous optimal control problem into a discrete-time optimal control problem, we use the *direct multiple shooting* technique, originally due to Plitt and Bock [35, 12]: We divide the time horizon into $N = 24$ invervals $[t_i, t_{i+1}]$, each of one second length, and introduce a locally constant control representation $q_0, q_1, \ldots, q_{N-1}$, as well as artificial initial values $s_0, \ldots, s_N$, as depicted in Figure 1.1. On each of these intervals we solve the following initial value problem:

$$\begin{aligned} \dot{\xi}_i(t; s_i, q_i) &= \hat{f}\left(\xi_i(t; s_i, q_i), q_i\right), \quad t \in [t_i, t_{i+1}], \\ \xi_i(t_i; s_i, q_i) &= s_i, \end{aligned} \tag{6.9}$$

yielding a trajectory piece $\xi_i(t; s_i, q_i)$ for $t \in [t_i, t_{i+1}]$. This allows us to conveniently define a discrete-time system as in (1.2c) with transition function

$$f_i(s_i, q_i) := \xi_i(t_{i+1}; s_i, q_i).$$

Analogously, we define the objective contributions in (1.2a) by

$$L_i(s_i, q_i) := \int_{t_i}^{t_{i+1}} L(\xi_i(t; s_i, q_i), q_i, t)dt.$$

The main difficulty of the direct multiple shooting method lies in the efficient solution of the initial value problems (6.9) and in the sensitivity computation. For this aim we use the advanced backward differentiation formula (BDF) code DAESOL (Bauer, Bock, and Schlöder [3], Bauer [2]), which is especially suited for stiff problems, as the above kite model. It uses the principle of internal numerical differentiation (IND) as introduced by Bock [8].

Using the multiple shooting formulation, we have transformed the continuous-time optimization problem (6.8) into a nonlinear programming problem $P_0(x_0)$ of exactly the type (1.2).

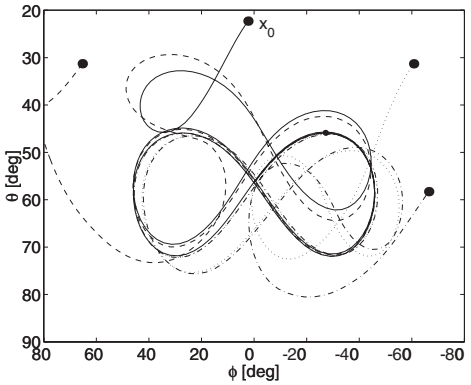


FIG. 6.4. *Closed-loop trajectories resulting from the real-time iteration scheme for different initial values x_0. The kite never crashes onto the ground ($\theta = 90$ degrees).*

6.4.1. Generation of the Gauss–Newton Hessian blocks. The efficient generation of a Gauss–Newton approximation for continuous least squares terms deserves some attention: the Hessian approximations (2.6) are determined according to

$$\begin{pmatrix} Q_i^H & M_i^H \\ (M_i^H)^T & R_i^H \end{pmatrix} := \int_{t_i}^{t_{i+1}} \left(\frac{\partial \xi_i(t; s_i, q_i)}{\partial (s_i, q_i)} \right)^T Q \; \frac{\partial \xi_i(t; s_i, q_i)}{\partial (s_i, q_i)} + \begin{pmatrix} 0 & 0 \\ 0 & R \end{pmatrix} dt.$$

These integrals are efficiently computed during the sensitivity computation using a specially adapted version of the integrator DAESOL [16, 25].

6.5. A real-time scenario. In the following real-time scenario we assume that the Newton-type optimizer is initialized with the reference trajectory itself, i.e., $y^0 := (\lambda_0^0, s_0^0, q_0^0, \ldots, \lambda_N^0, s_N^0)$, where $\lambda_i^0 := 0$, and $s_i^0 := \xi_r(t_i)$ and $q_i^0 := \frac{1}{t_{i+1}-t_i} \int_{t_i}^{t_{i+1}} u_r(t)\, dt$ are the corresponding values of the periodic reference solution. This y^0 is (nearly) identical to the solution of the problem $P_0(\xi_r(t_0))$. At the time $t_0 = 0$, when the actual value of x_0 is known, we start the iterations as described in section 3.1 by solving the first prepared linear system $\Delta y^0 = -J^0(y^0)^{-1} \nabla_y \mathcal{L}^0(y^0)$ (step 2) and give the first control $u_0 := q_0^0 + \Delta q_0^0$ immediately to the system. Then we shrink the problem (step 3) and prepare the iteration for the following one (step 1). As we assume no further disturbances, the new initial value is $x_1 = f_0(x_0, u_0) = \xi_0(t_1; x_0, u_0)$ resulting from the (continuous) system dynamics. This cycle is repeated until the $N = 24$ intervals are over.

The corresponding trajectory resulting from the real-time iteration scheme for different initial values x_0 is shown in Figure 6.4 as (ϕ, θ)-plots. For all scenarios, the third loop is already close to the reference trajectory.

In Figure 6.5 we compare the result of the real-time iteration scheme with the open-loop system dynamics without feedback (dash-dotted line) and with a hypothetical optimal feedback control (dashed line) for one initial value x_0. The open-loop system, where the controls are simply taken from the reference (and initialization) trajectory ($u_k := q_k^0$), crashes after seven seconds onto the ground.

Note that the computation of the hypothetical optimal feedback control needs about eight seconds on a Compaq Alpha XP1000 workstation. This delay means that

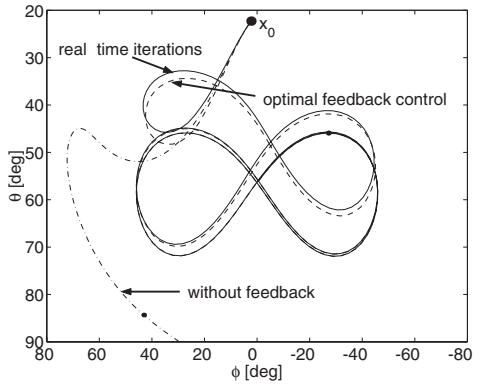


FIG. 6.5. *Comparison of trajectories resulting from the real-time iteration scheme (solid line), the open-loop controls without feedback (dash-dotted line, crashing onto the ground), and a hypothetical optimal feedback control (dashed line).*

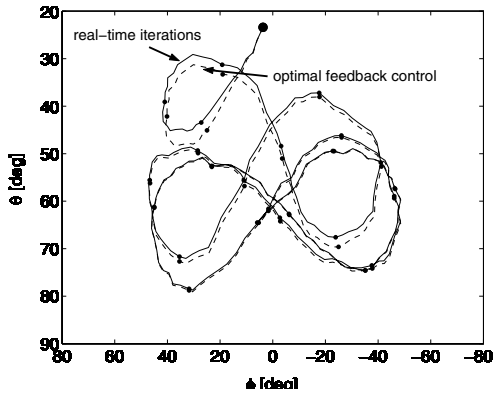


FIG. 6.6. *Disturbance scenario: Closed-loop response resulting from real-time iteration scheme (solid line) and a hypothetical optimal feedback control (dashed line). After a large initial disturbance, the state is additionally disturbed each* 0.1 *second by independent Gaussian noise. The disturbance sequence for both trajectories is identical.*

no feedback can be applied in the meantime, so the kite would have crashed onto the ground before the first response would have been computed.

In contrast to this, the first feedback control u_0 of the real-time iteration scheme was available within only 0.05 seconds delay after time t_0 (for the computations of step 2). The sampling time of one second until the next feedback can be applied was necessary to prepare the following real-time iteration (to be exact, step 1 always needed under 0.8 seconds). The comparison with the hypothetical optimal feedback control shows that the real-time iteration scheme delivers a quite good approximation even for this challenging nonlinear and unstable test example with largely disturbed initial values.

6.5.1. High frequency disturbances. In a third feedback simulation scenario shown in Figure 6.6 we test the performance of the real-time iteration scheme in the

presence of random disturbances with a frequency higher than the sampling time: each tenth of a second the state $(\theta, \phi, \dot{\theta}, \dot{\phi})$ is randomly disturbed by independent Gaussian noise of standard deviation $0.01 \cdot (1, 1, \mathrm{s}^{-1}, \mathrm{s}^{-1})$. Because feedback is provided only once a second, the kite flies open-loop during one second before feedback can be provided to the accumulated result of the disturbances. The initial state was much more strongly disturbed, in the same way as in the scenario of Figure 6.5. Despite these combined disturbances the scheme is able to lead the kite efficiently into the reference orbit. Again, it compares well with optimal feedback control. Note, however, that the results of Theorems 4.1 and 5.1 are not directly applicable to this third scenario as these theorems assume undisturbed system behavior after one initial disturbance.

For feedback control simulations of the kite using a moving horizon framework including also state constraints, we refer to [16, 24].

7. Conclusions. We have presented a recently developed Newton-type method for the real-time optimization of nonlinear processes and have given a new contractivity result and a bound on the loss of optimality when compared to optimal feedback control. In a numerical case study, the real-time control of an airborne kite, we have demonstrated the practical applicability of the method for a challenging nonlinear control example.

The "real-time iteration" scheme is based on the direct multiple shooting method, which offers several advantages in the context of real-time optimal control, among them the ability to efficiently initialize subsequent optimization problems, to treat highly nonlinear and unstable systems, and to deal efficiently with path constraints. The most important feature of the real-time iteration scheme is a dovetailing of the solution iterations with the process development which allows us to reduce sampling times to a minimum but maintains all advantages of a fully nonlinear treatment of the optimization problems. A separation of the computations in each real-time iteration into a *preparation phase* and a *feedback response phase* can be realized. The feedback phase is typically orders of magnitude shorter than the preparation phase and allows us to obtain an immediate feedback that takes all linearized constraints into account.

The contractivity of the scheme is proven under mild conditions that are nearly identical to the sufficient conditions for convergence of off-line Newton-type methods. Iterates on different horizon lengths have to be compared. The result is that the real-time iterates, after an initial disturbance, geometrically approach the exact optimal solutions during the runtime of the process. When the resulting closed-loop trajectory is compared to optimal feedback control, a bound on the loss of optimality has been established, which is of fourth order in the initial disturbance if an exact Newton–Raphson method is used.

A newly developed kite model is presented. The control aim is, starting from an arbitrary initial state, to steer the kite into a periodic orbit, a "lying eight" with a period duration of eight seconds. We consider a time horizon of 24 seconds. The initial state is only known at the moment that the first control needs already to be applied; the real-time iteration scheme delivers linearized feedback nearly without delay and provides a newly linearized feedback after each sampling time of one second, leading to a fully nonlinear optimization, and is always prepared to react to further disturbances. The scheme shows an excellent closed-loop performance for this highly nonlinear and unstable system and compares well to a hypothetical exact optimal feedback control.

The real-time iteration scheme has also been applied for NMPC of a real pilot plant distillation column described by a stiff DAE model with over 200 system states, allowing feedback sampling times of only 20 seconds [21].

Acknowledgments. The first author wants to thank F. Bonnans and M. Wright for encouraging publication of this work and also thanks the anonymous referees whose valuable comments helped to improve the paper.

REFERENCES

[1] F. Allgöwer, T. A. Badgwell, J. S. Qin, J. B. Rawlings, and S. J. Wright, *Nonlinear predictive control and moving horizon estimation: An introductory overview*, in Advances in Control, Highlights of ECC'99, P. M. Frank, ed., Springer-Verlag, New York, 1999, pp. 391–449.

[2] I. Bauer, *Numerische Verfahren zur Lösung von Anfangswertaufgaben und zur Generierung von ersten und zweiten Ableitungen mit Anwendungen bei Optimierungsaufgaben in Chemie und Verfahrenstechnik*, Ph.D. thesis, University of Heidelberg, Heidelberg, Germany, 1999, http://www.ub.uni-heidelberg.de/archiv/1513.

[3] I. Bauer, H. G. Bock, and J. P. Schlöder, *DAESOL: A BDF-Code for the Numerical Solution of Differential Algebraic Equations*, Internal report, IWR, SFB 359, University of Heidelberg, Heidelberg, Germany, 1999.

[4] L. Biegler and J. Rawlings, *Optimization approaches to nonlinear model predictive control*, in Proceedings of the 4th International Conference on Chemical Process Control, W. Ray and Y. Arkun, eds., AIChE, CACHE, Padre Island, TX, 1991, pp. 543–571.

[5] L. T. Biegler, *Efficient solution of dynamic optimization and NMPC problems*, in Nonlinear Model Predictive Control, F. Allgöwer and A. Zheng, eds., Progr. Systems Control Theory 26, Birkhäuser, Basel, 2000, pp. 219–244.

[6] T. Binder, L. Blank, H. G. Bock, R. Bulirsch, W. Dahmen, M. Diehl, T. Kronseder, W. Marquardt, J. P. Schlöder, and O. v. Stryk, *Introduction to model based optimization of chemical processes on moving horizons*, in Online Optimization of Large Scale Systems: State of the Art, M. Grötschel, S. O. Krumke, and J. Rambau, eds., Springer-Verlag, New York, 2001, pp. 295–340. also available online from http://www.zib.de/dfg-echtzeit/Publikationen/Preprints/Preprint-01-15.html.

[7] H. Bock, M. Diehl, D. B. Leineweber, and J. Schlöder, *Efficient direct multiple shooting in nonlinear model predictive control*, in Scientific Computing in Chemical Engineering II, F. Keil, W. Mackens, H. Voß, and J. Werther, eds., Springer-Verlag, Berlin, 1999, pp. 218–227.

[8] H. G. Bock, *Numerical treatment of inverse problems in chemical reaction kinetics*, in Modelling of Chemical Reaction Systems, K. H. Ebert, P. Deuflhard, and W. Jäger, eds., Springer Ser. Chem. Phys. 18, Springer-Verlag, Heidelberg, 1981, pp. 102–125.

[9] H. G. Bock, *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*, Bonner Math. Schriften 183, University of Bonn, Bonn, 1987.

[10] H. G. Bock, M. Diehl, D. B. Leineweber, and J. P. Schlöder, *A direct multiple shooting method for real-time optimization of nonlinear DAE processes*, in Nonlinear Model Predictive Control, F. Allgöwer and A. Zheng, eds., Progr. Systems Control Theory 26, Birkhäuser, Basel, 2000, pp. 246–267.

[11] H. G. Bock, M. Diehl, J. P. Schlöder, F. Allgöwer, R. Findeisen, and Z. Nagy, *Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations*, in ADCHEM2000, Proceedings of the International Symposium on Advanced Control of Chemical Processes, Vol. 2, Pisa, Italy, 2000, pp. 695–703.

[12] H. G. Bock and K. J. Plitt, *A multiple shooting algorithm for direct solution of optimal control problems*, in Proceedings of the 9th IFAC World Congress Budapest, Pergamon Press, Oxford, UK, 1984, pp. 243–247.

[13] H. Choi, M. Hinze, and K. Kunisch, *Instantaneous control of backward-facing-step flows*, Appl. Numer. Math., 31 (1999), pp. 133–158.

[14] H. Choi, R. Temam, P. Moin, and J. Kim, *Feedback control for unsteady flow and its application to the stochastic Burgers equation*, J. Fluid Mech., 253 (1993), pp. 509–543.

[15] N. de Oliveira and L. Biegler, *An extension of Newton-type algorithms for nonlinear process control*, Automatica J. IFAC, 31 (1995), pp. 281–286.

[16] M. Diehl, *Real-Time Optimization for Large Scale Nonlinear Processes*, Fortschr.-Ber. VDI Reihe 8, Meß, Steuerungs- und Regelungstechnik 920, VDI Verlag, Düsseldorf, 2002; also available online from http://www.ub.uni-heidelberg.de/archiv/1659/.

[17] M. Diehl, *The Kite Benchmark Problem Homepage*, http://www.iwr.uni-heidelberg.de/~Moritz.Diehl/KITE/kite.html, 2003.

[18] M. Diehl, H. G. Bock, and J. P. Schlöder, *Newton-type methods for the approximate solution of nonlinear programming problems in real time*, in High Performance Algorithms and Software for Nonlinear Optimization, G. D. Pillo and A. Murli, eds., Kluwer Academic Publishers B.V., Dordrecht, The Netherlands, 2002, pp. 177–200.

[19] M. Diehl, H. G. Bock, J. P. Schlöder, R. Findeisen, Z. Nagy, and F. Allgöwer, *Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations*, J. Proc. Contr., 12 (2002), pp. 577–585.

[20] M. Diehl, R. Findeisen, S. Schwarzkopf, I. Uslu, F. Allgöwer, H. G. Bock, E. D. Gilles, and J. P. Schlöder, *An efficient algorithm for nonlinear model predictive control of large-scale systems. Part* I: *Description of the method*, Automatisierungstechnik, 50 (2002), pp. 557–567.

[21] M. Diehl, R. Findeisen, S. Schwarzkopf, I. Uslu, F. Allgöwer, H. G. Bock, E. D. Gilles, and J. P. Schlöder, *An efficient algorithm for nonlinear model predictive control of large-scale systems. Part* II: *Application to a distillation column*, Automatisierungstechnik, 51 (2003), pp. 22–29.

[22] D. B. Leineweber, I. Bauer, H. G. Bock, and J. P. Schlöder, *An effficient multiple shooting based reduced SQP strategy for large-scale dyamic process optimization. Part* I: *Theoretical aspects*, Comp. Chem. Engrg., 27 (2003), pp. 157–166.

[23] D. B. Leineweber, A. Schäfer, H. G. Bock, and J. P. Schlöder, *An effficient multiple shooting based reduced SQP strategy for large-scale dyamic process optimization. Part* II: *Software aspects and applications*, Comp. Chem. Engrg., 27 (2003), pp. 167–174.

[24] M. Diehl, L. Magni, and G. De Nicolao, *Online NMPC of unstable periodic systems using approximate infinite horizon closed loop costing*, Annual Reviews in Control, 28 (2004), pp. 37–45.

[25] M. Diehl, I. Uslu, R. Findeisen, S. Schwarzkopf, F. Allgöwer, H. G. Bock, T. Bürner, E. D. Gilles, A. Kienle, J. P. Schlöder, and E. Stein, *Real-time optimization for large scale processes: Nonlinear model predictive control of a high purity distillation column*, in Online Optimization of Large Scale Systems: State of the Art, M. Grötschel, S. O. Krumke, and J. Rambau, eds., Springer-Verlag, New York, 2001, pp. 363–384; also available online from http://www.zib.de/dfg-echtzeit/Publikationen/Preprints/Preprint-01-16.html.

[26] J. C. Dunn and D. P. Bertsekas, *Efficient dynamic programming implementations of Newton's method for unconstrained optimal control problems*, J. Optim. Theory Appl., 63 (1989), pp. 23–38.

[27] A. Helbig, O. Abel, and W. Marquardt, *Model predictive control for on-line optimization of semi-batch reactors*, in Proceedings of the American Control Conference, Philadelphia, PA, 1998, pp. 1695–1699.

[28] B. Kouvaritakis and M. Cannon, eds., *Non-Linear Predictive Control: Theory and Practice*, IEE Publishing, London,, 2001.

[29] P. Krämer-Eis and H. Bock, *Numerical treatment of state and control constraints in the computation of feedback laws for nonlinear control problems*, in Large Scale Scientific Computing, P. D. et al., ed., Birkhäuser, Basel, 1987, pp. 287–306.

[30] B. Kugelmann and H. Pesch, *New general guidance method in constrained optimal control, part* 1: *Numerical method*, J. Optim. Theory Appl., 67 (1990), pp. 421–435.

[31] D. B. Leineweber, *Efficient Reduced SQP Methods for the Optimization of Chemical Processes Described by Large Sparse DAE Models*, Fortschr.-Ber. VDI Reihe 3, Verfahrenstechnik 613, VDI Verlag, Düsseldorf, 1999.

[32] W. Li and L. Biegler, *Multistep, Newton-type control strategies for constrained nonlinear processes*, Chem. Eng. Res. Des., 67 (1989), pp. 562–577.

[33] Z. Nagy, R. Findeisen, M. Diehl, F. Allgöwer, H. G. Bock, S. Agachi, J. P. Schlöder, and D. B. Leineweber, *Real-time feasibility of nonlinear predictive control for large scale processes: A case study*, in Proceedings of the American Control Conference, Chicago, IL, 2000, pp. 4249–4254.

[34] J. O. Pantoja, *Differential dynamic programming and Newton's method*, Internat. J. Control, 47 (1988), pp. 1539–1553.

[35] K. J. Plitt, *Ein superlinear konvergentes Mehrzielverfahren zur direkten Berechnung beschränkter optimaler Steuerungen*, Master's thesis, University of Bonn, Bonn, Germany, 1981.

[36] C. Rao, S. Wright, and J. Rawlings, *Application of interior-point methods to model predictive control*, J. Optim Theory Appl., 99 (1998), pp. 723–757.

[37] L. Santos, *Multivariable Predictive Control of Nonlinear Chemical Processes*, Ph.D. thesis, Universidade de Coimbra, Coimbra, Portugal, 2000.

[38] M. J. Tenny, S. J. Wright, and J. B. Rawlings, *Nonlinear model predictive control via feasibility-perturbed sequential quadratic programming*, Comput. Optim. Appl., 28 (2004), pp. 87–121.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1737–1756

STOCHASTIC DIFFERENTIAL GAMES IN A NON-MARKOVIAN SETTING*

ERHAN BAYRAKTAR† AND H. VINCENT POOR‡

Abstract. Stochastic differential games are considered in a non-Markovian setting. Typically, in stochastic differential games the modulating process of the diffusion equation describing the state flow is taken to be Markovian. Then Nash equilibria or other types of solutions such as Pareto equilibria are constructed using Hamilton–Jacobi–Bellman (HJB) equations. But in a non-Markovian setting the HJB method is not applicable. To examine the non-Markovian case, this paper considers the situation in which the modulating process is a fractional Brownian motion. Fractional noise calculus is used for such models to find the Nash equilibria explicitly. Although fractional Brownian motion is taken as the modulating process because of its versatility in modeling in the fields of finance and networks, the approach in this paper has the merit of being applicable to more general Gaussian stochastic differential games with only slight conceptual modifications. This work has applications in finance to stock price modeling which incorporates the effect of institutional investors, and to stochastic differential portfolio games in markets in which the stock prices follow diffusions modulated with fractional Brownian motion.

Key words. stochastic differential games, non-Markovian games, fractional Brownian motion, fractional noise theory

AMS subject classifications. 91A15, 91A23, 60G15, 60G18, 60H40

DOI. 10.1137/S0363012902417632

1. Introduction.

The study of stochastic differential games with controls is a part of game theory that is relatively unknown, even though it has significant potential for application as noted by Øksendal and Reikvam [27]. Prior work in this area has focused on the examination of such games in a Markovian setting (see below). In this paper we will study a type of non-Markovian stochastic differential game. In particular, we will consider a game in which the one-dimensional state X_t follows the following stochastic differential equation:

$$dX_t = \mu(t, X_t, \upsilon_1, \dots, \upsilon_N)dt + \sigma(t, X_t, \upsilon_1, \dots, \upsilon_N)dB_t^{(H)}, \tag{1}$$

$$\mu : [0,T] \times \mathbb{R} \times \Upsilon_1 \times \cdots \times \Upsilon_N \to \mathbb{R}, \tag{2}$$

$$\sigma : [0,T] \times \mathbb{R} \times \Upsilon_1 \times \cdots \times \Upsilon_N \to \mathbb{R}, \tag{3}$$

where $\upsilon_i \in \Upsilon_i \subset \mathbb{R}^{\nu_i}$ is the control of the ith player over the state and is adapted to the natural filtration of B^H. Here T is the expiration date of the game, and B^H is a fractional Brownian motion (fBm) with Hurst parameter $H \in (\frac{1}{2}, 1)$. B^H is

*Received by the editors November 9, 2002; accepted for publication (in revised form) June 16, 2004; published electronically March 22, 2005. This research was supported by U.S. Office of Naval Research grant N00014-03-1-0102.
http://www.siam.org/journals/sicon/43-5/41763.html
†Department of Mathematics, 2074 East Hall, University of Michigan, Ann Arbor, MI 48109-1109 (erhan@umich.edu).
‡Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 (poor@princeton.edu).

defined as an almost surely (a.s.) continuous zero-mean Gaussian process having the autocorrelation structure given by

$$E\left\{B_t^H B_s^H\right\} = \frac{1}{2}\left\{|s|^{2H} + |t|^{2H} - |t-s|^{2H}\right\}. \tag{4}$$

Each agent wants to maximize its own pay-off (with this feature the problem differs from the usual optimal control problem):

$$J_i(x) = E^x\left\{\int_0^T f_i(t, X_t, \upsilon_t)dt + K_i(T, X_T)\right\}, \tag{5}$$

where $E^x\{\cdot\}$ denotes conditional expectation given $X_0 = x$. In this paper we consider only the case in which the state and the source of randomness are one-dimensional. The results can be extended to the case in which there are multiple sources of randomness and multiple controlled states (see [4]).

Typically in this type of setting the modulating process in (1) is taken to be Brownian motion, i.e., $H = 1/2$, and the controls of the players are Markovian. Then Nash equilibria or other types of solutions such as Pareto equilibria are constructed using the Hamilton–Jacobi–Bellman (HJB) equations (see, e.g., Friedman [9], Gaidov [11], [12], [10], Nilakantan [23], Øksendal and Reikvam [27], and Varaiya [32]). However, fBm is not a Markov process for any H other than $\frac{1}{2}$, and therefore this approach does not work for the general case of (1). Here we will develop a quasi-martingale approach to the solution of this problem using the fractional noise calculus developed by Duncan, Hu, and Pasik-Duncan [7], and the work of Hu and Øksendal [16], which generalizes white noise calculus (see [1]) to develop an integration theory with respect to fBm. The key to our solution will be the fractional Clark–Ocone formula developed by Hu and Øksendal [16]. The integrals in (1) are Wick type integrals (see Definition 3.3) rather than Stieltjes integrals (defined pathwise; see, e.g., [33]). The motivation for using Wick type integrals is as follows: The pathwise integral $\int_0^t f_s \delta B_s^H$ with respect to fBm does not in general have a zero mean; i.e., $E\{\int_0^t f_s \delta B_s^H\} \neq 0$. However, the Wick type integral $\int_0^t f_s dB_s^H$ has zero mean; i.e., $E\{\int_0^t f_s dB_s^H\} = 0$. Therefore in a stochastic differential equation (SDE) of the form

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t^H, \tag{6}$$

the volatility term $\sigma(X_t)dB_t^H$ does not contribute to the mean rate of change, as it does with SDEs with pathwise integrals. Since separating the random fluctuations from the mean rate of change is desirable for our purposes, we prefer to use Wick type integrals for defining the integrals with respect to fBm (see [7]). (Note also that only in the Wick type calculus are the standard tools of Itô calculus such as an Itô representation theorem available.) See [5] for applications of Wick calculus to pricing weather derivatives and [8] and [16] for further applications of Wick type calculus, particularly in finance.

Fractional noise calculus reduces to white noise calculus when H is replaced by $\frac{1}{2}$. Moreover the integrals of adapted processes in this framework are equal to the Itô integrals of these processes with respect to Brownian motion. Hence our results hold in particular for the standard framework, i.e., when the modulator is a Brownian motion, and the integrals in (1) are taken to be Itô integrals.

This work has an immediate application in finance, to stock price modeling when long-range dependence is accounted for in the model. The stock prices are considered

to be states in this setting, while the agents are *not price takers*; i.e., their trading changes the price level. This models a market with institutional investors who make large transactions and therefore influence the prices of the stocks. These investors find themselves in a random environment due to the existence of small investors. The small investors are typically inert; i.e., they do not trade for long time intervals. A microstructure model taking the inertness of the agents into account is constructed in [3]. It is shown that the prices arising from the interaction of the small agents can be approximated by geometric fractional Brownian motion. The game theoretic setting in this paper is an extension to the results of [3] in the sense that we start by assuming that the noise in the environment can be modeled by an fBm differential in the controlled stochastic differential equation (1), which is the noise due to the trades of the small investors.

Another possible application is in stochastic differential portfolio games, which are studied by Browne in a Brownian motion setting in [6]. This formulation is applicable to the analysis of traders who are competing for a bonus, or to fund managers whose funds are invested in different markets and who achieve rewards based on the relative performance of their funds. Yet another possible application is in stochastic goodwill problems (finding the optimal advertising policy for the maximization of the product image) in advertising when there are more than one good of the same kind in competition. (See [22] for stochastic goodwill problems in a stochastic optimal control setting.)

By adapting the fractional noise machinery, we will see that our results can be shown to hold for more general Gaussian modulators in the state flow dynamics. However, we state the results in terms of fBm to emphasize the fact that the game under consideration becomes non-Markovian and also because this case admits an explicit equilibrium. Another motivation for this model is the fact that fBm is frequently used in modeling in various areas of research (see [2], [31], [24], [25], and [26] for applications other than finance).

The rest of the paper is organized as follows. In section 2 we give a Nash equilibrium theorem. In section 3 we introduce the necessary tools from the fractional noise calculus that we use in the proof of the equilibrium theorem that is carried out in section 4. Finally in section 5 we give a sketch of how to extend the fractional noise machinery to more general Gaussian modulation processes.

2. Nash equilibrium in a linear game of N players. For ease of exposition we consider first a one-dimensional state equation, with the drift and diffusion coefficients controlled linearly by the players:

$$dX_t = rX_t dt + \sum_{i=1}^{N} \alpha_i(t)u_i(t)dt + C\sum_{i=1}^{N} \beta_i(t)v_i(t)dt + \sum_{i=1}^{N} \beta_i(t)v_i(t)dB^H(t), \tag{7}$$

where B^H denotes the one-dimensional version of $B^{(H)}$ with $H_1 = H$. The initial state will be denoted by $X_0 = x$. The pay-off function of player i will be of the form

$$J_i(x) = E_\mu^x\left\{ \int_0^T \frac{c_i u_i^{\gamma_i}(t)}{\gamma_i} dt + \frac{b_i X_T^{\gamma_i'}}{\gamma_i'} \right\}; \tag{8}$$

that is, players are constant relative risk averse (CRRA). Here μ is the measure on the sample space under which the canonical process is an fBm. Player i controls the state by its choice of actions (u_i, v_i). We assume that $\alpha_i : [0,T] \to \mathbb{R}$ is bounded

for each $i \in \{1, \dots, N\}$. The coefficient functions $\beta_i : [0,T] \to \mathbb{R}$ will appear in the definition of admissible strategies.

Since this game is in a non-Markovian setting, it cannot be solved via the HJB method. Instead we will employ the recently developed fractional Wick calculus, which we describe briefly in section 3, and the fractional Clark–Ocone formula (which is given along with the proof of the equilibrium theorem) to find Nash equilibria for this game. Observe that u_i affects the drift of the state and also appears in the pay-off function. It can be interpreted as a cost for the player; i.e., for gaining a certain amount of riskless increase the player pays an associated cost. Whereas by choice of v_i, the player does not have to pay a cost for an associated gain (since this action does not appear in the pay-off function), but is taking some risk (since v_i affects the diffusion coefficient in addition to the drift).

Let us introduce the following notation that is necessary to define the admissible strategies and for the statement of the theorem.

Define K as

$$K(t) = \frac{C(Tt - t^2)^{\frac{1}{2}-H}}{2H(2H-1)\Gamma(2H-1)\Gamma(2-2H)\cos(\pi(H-\frac{1}{2}))} \quad \text{for } t \in [0,T], \tag{9}$$

and define ζ by

$$\begin{aligned} \big((-\Delta)^{-(H-1/2)}\zeta_t\big)(s) &= \big((-\Delta)^{-(H-1/2)}K\big)(s), \quad 0 \le s \le t, \\ \zeta_t(s) &= 0, \quad s < 0 \quad \text{or} \quad s > t, \end{aligned} \tag{10}$$

where the action of the operator $(-\Delta)^{-(H-1/2)}$ on a test function f is defined by

$$\big((-\Delta)^{-(H-1/2)}f\big)(x) = \frac{1}{2\Gamma(2H-1)\cos(\pi(H-1/2))}\int_{-\infty}^{\infty} |x-t|^{2H-2} f(t)dt, \tag{11}$$

where Γ is the Gamma function and is given by $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ for $x > 0$. The existence of such ζ is guaranteed by [15].

Define $\hat{\mu}$ and η by

$$\eta(T) := \frac{d\hat{\mu}}{d\mu} := \exp\left(-\int_0^T K(s)dB_s^H - \frac{1}{2}|K|_\phi^2\right), \tag{12}$$

with $|K|_\phi^2$ given by $|K|_\phi^2 = \int_{\mathbb{R}_+^2} K(s)K(t)\phi(s,t)dsdt$, where

$$\phi(s,t) = H(2H-1)|s-t|^{2H-2}; \quad s,t \in \mathbb{R}_+. \tag{13}$$

Here μ denotes the probability measure under which B^H is an fBm with Hurst parameter H. Note that integrals of deterministic functions w.r.t. fBm are well defined, as will become clear in section 3.

And finally define ρ by $\rho(t,w) := E_\mu\big\{\eta(T)\big|\mathcal{F}_t\big\}$, where $\mathcal{F}_t$ is the σ-algebra generated by $\{B_s^H, s \le t\}$.

Now we will introduce the solution concept of Nash equilibrium in our context. We consider a set $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$ of admissible strategies for which the admissibility conditions are adaptedness w.r.t. the filtration generated by fBm and the following integrability condition:

$$\beta_i v_i \in \mathcal{L}_\phi^{1,2}(\hat{\mu}), \tag{14}$$

where $\mathcal{L}_\phi^{1,2}(\hat{\mu})$ denotes the completion of the set of all $\mathcal{F}_t$ adapted processes f such that

$$\| f \|_{\mathcal{L}_\phi^{1,2}(\hat{\mu})} := E_{\hat{\mu}}\left\{\int_{\mathbb{R}}\int_{\mathbb{R}} f(s)f(t)\phi(s,t)dsdt\right\} + E_{\hat{\mu}}\left\{\left(\int_{\mathbb{R}} D_s^\phi f(s)ds\right)^2\right\} < \infty. \tag{15}$$

Here $D_s^\phi F = \int_{\mathbb{R}} \phi(s,t) D_t F dt$, where $D_s F$ denotes the Hida–Malliavin derivative of F, which will be introduced in section 3.

DEFINITION 2.1. *The strategy $z^e = (u^e, v^e) \in \mathcal{A}$ is called a Nash equilibrium strategy if, for each i, player i's action $z_i^e = (u_i, v_i) \in \mathcal{A}_i$ is a best response to its opponents'; i.e.,*

$$J_i^x(z_1^e, \dots, z_{i-1}^e, z_i, z_{i+1}^e, \dots, z_N^e) \le J_i^x(z^e) \tag{16}$$

for all x, for each player i, and for all $z_i \in \mathcal{A}_i$.

Then we have the following Nash equilibrium theorem.

THEOREM 2.2. *Consider the game given by* (7) *and* (8). *Then conditions* (17) and (18) *are necessary and sufficient for the existence of a Nash equilibrium:*

$$\gamma_i' = \gamma' \quad \textit{for } i = 1, \dots, N, \quad \textit{and} \tag{17}$$

$$\begin{aligned}\sum_{i=1}^N \int_0^T m^{\frac{1}{\gamma_i - 1}} b_i^{\frac{1}{\gamma_i - 1}} \alpha_i(t)^{\frac{\gamma_i}{\gamma_i - 1}} e^{-rt\frac{\gamma_i}{\gamma_i - 1}} \exp\left(\frac{\gamma_i}{2(1-\gamma_i)^2}|\zeta_t|_\phi\right)dt \\ + m^{\frac{1}{\gamma' - 1}} e^{-rT\frac{\gamma'}{\gamma' - 1}} \exp\left(\frac{2\gamma'|K|_\phi}{2(1-\gamma')^2}\right) = x\end{aligned} \tag{18}$$

has a solution $m^* \in \mathbb{R}$.

Let $\big((u_1^e, v_1^e), \dots, (u_N^e, v_N^e)\big)$ denote the agents' Nash equilibrium strategies. The first components of the equilibrium strategies are uniquely determined by

$$u_i^e(t) = \left(\frac{m^* b_i \alpha_i(t)}{c_i} e^{-rt}\rho(t,w)\right)^{\frac{1}{\gamma_i - 1}} \quad \textit{for } i = 1, \dots, N, \tag{19}$$

while the second component of the players' strategies will be any adapted (to the filtration of fBm) processes satisfying the following constraint:

$$\begin{aligned} e^{-rt}\sum_{i=1}^N \beta_i v_i^e(t) = (m^*)^{\frac{1}{\gamma' - 1}} \frac{K(t)}{1-\gamma'} \exp\left(\frac{1}{1-\gamma'}\int_0^t K(s)dB_s^H - \frac{C}{1-\gamma'}\int_t^T K(s)ds\right. \\ \left. + \frac{2-\gamma}{2(1-\gamma^2)}|K|_\phi^2 - \frac{1}{1-\gamma}|K1_{[0,t]}|_\phi^2 - rT\frac{\gamma'}{\gamma' - 1}\right) \\ - \int_0^T \sum_{i=1}^N \alpha_i(u)^{\frac{\gamma_i}{\gamma_i - 1}} \big(\frac{m^* b_i}{c_i}\big)^{\frac{1}{\gamma_i - 1}} \times \frac{\zeta_u(t)}{1-\gamma_i} e^{-ru\frac{\gamma_i}{\gamma_i - 1}} \\ \times \exp\left(\frac{1}{1-\gamma_i}\int_0^t \zeta_u(s)dB_s^H - \frac{C}{1-\gamma_i}\int_t^u \zeta_u(s)ds\right. \\ \left. + \frac{2-\gamma_i}{2(1-\gamma_i)^2}|\zeta_u|_\phi^2 - \frac{1}{1-\gamma_i}|\zeta_u 1_{[0,t]}|_\phi^2\right)du, \end{aligned} \tag{20}$$

where 1 *stands for the indicator function. Finally, the state at time* T *at Nash equilibrium is given by*

$$F^e = (m^*)^{\frac{1}{\gamma'-1}} \eta(T)^{\frac{1}{\gamma'-1}} e^{\frac{-rT}{\gamma'-1}}. \tag{21}$$

These results can be extended to games with multiple controlled states and multiple sources of randomness (see [4]).

Before proving Theorem 2.2, in the next section we will give a brief review of the facts from fractional noise calculus, largely following the treatment by Hu and Øksendal [16]. (An extended version of the following section can be found in [4].)[1]

3. Fractional noise calculus. We start this section by introducing the necessary ingredients for the definition of the stochastic integral in (1). In what follows $L^2_\phi(\mathbb{R})$ will denote the completion of the set of measurable functions satisfying

$$|f|^2_\phi := \int_{\mathbb{R}^2} f(s)f(t)\phi(s,t)dsdt < \infty. \tag{22}$$

Remark. It is shown by Pipiras and Taqqu [28] that the set of functions satisfying (22) is not a complete space.

The stochastic integral of deterministic functions in $L^2_\phi(\mathbb{R})$ w.r.t. fBm is well defined (see [13]). For $f \in L^2_\phi(\mathbb{R})$, we will denote its integral w.r.t. fBm by $\langle w, f\rangle := \int_{\mathbb{R}} f(t)dB^H_t$.

The probability space in our game will be $\Omega = \mathcal{S}'(R)$, the space of tempered distributions (the dual space of $\mathcal{S}(\mathbb{R})$, the Schwartz space of rapidly decreasing functions) equipped with the weak-star topology. And we will take the events to be Borel subsets of $\mathcal{S}'(R)$. By the Bochner–Minlos theorem there exists a probability measure μ on Ω such that $\langle \cdot, f\rangle \colon \Omega \to \mathbb{R}$ is a Gaussian random variable with mean 0 and variance $|f|^2_\phi$ (see [16]).

We will now introduce the Wiener chaos expansion of random variables in $L^2(\mu)$. One first has to find the orthonormal basis for $L^2_\phi(\mathbb{R})$. Recall that the Hermite functions (see, e.g., Appendix C of [14]), which we will denote by (z_n), form an orthonormal basis for $L^2(\mathbb{R})$.

Let us define the map from the space of functions satisfying (22) into $L^2(\mathbb{R})$ by

$$(I_\phi f)(u) = c_H \int_u^\infty (t-u)^{H-\frac{3}{2}} f(t)dt, \tag{23}$$

where $c_H = \sqrt{\frac{H(2H-1)\Gamma(\frac{3}{2}-H)}{\Gamma(H-\frac{1}{2})\Gamma(2-2H)}}$ (here Γ denotes the Gamma function). This map preserves the inner product, and the Hermite functions are in the range of this map. Let I^{-1}_ϕ denote the inverse map of I_ϕ. (For summable functions this inverse exists and is proportional to the Liouville differential of order $H - \frac{1}{2}$ [30], since $I_\phi(f)$ is proportional to the fractional integral of f of order $H - \frac{1}{2}$.) Now we see that the set $(e_n = I^{-1}_\phi(z_n))_{n=1,2,\ldots}$ constitutes an orthonormal basis for $L^2_\phi(\mathbb{R})$.

[1] Although the original name given by Hu and Øksendal in [16] to this kind of calculus was fractional white noise calculus, we prefer to omit "white" from the name since white suggests independence at each point in time, and here the noise considered is far from it.

Let $\mathcal{J}$ denote the set of all (finite) multi-indices of nonnegative integers. Then for $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m) \in \mathcal{J}$, define

$$H_\alpha(w) := h_{\alpha_1}(\langle w, e_1\rangle) \cdots h_{\alpha_m}(\langle w, e_m\rangle). \tag{24}$$

Note that $E_\mu\{H_\alpha H_\beta\} = 0$ if $\alpha \neq \beta$, and $E_\mu\{H_\alpha^2\} = \alpha!$. Now we can state what is known as the chaos decomposition for the elements of $L^2(\mu)$ (see [16]): Every $F \in L^2(\mu)$ can be decomposed uniquely as $F(w) = \sum_{\alpha\in\mathcal{J}} c_\alpha H_\alpha(w)$, where $c_\alpha \in \mathbb{R}$ for all $\alpha \in \mathcal{J}$.

For defining the integration w.r.t. fBm of random functions we will make use of the Hida test function space (a subspace of $L^2(\mu)$) and Hida distribution space (a superset of $L^2(\mu)$), which we denote by $(\mathcal{S})_H$ and $(\mathcal{S})_H^*$, respectively (see [34] for their definitions). Let $\psi(w) = \sum_{\alpha\in\mathcal{J}} a_\alpha H_\alpha(w) \in (\mathcal{S})_H$ and let $G(w) = \sum_{\beta\in\mathcal{J}} b_\beta H_\beta(w)$; then we denote the action of G on ψ by $\langle\langle G, \psi\rangle\rangle := \sum_{\alpha\in\mathcal{J}} \alpha! a_\alpha b_\alpha$.

For defining the integral w.r.t. fBm it is necessary to define $(\mathcal{S})_H^*$-valued Pettis integrals as follows.

DEFINITION 3.1. *A function $Z : \mathbb{R} \to (\mathcal{S})_H^*$ is $(\mathcal{S})_H^*$ integrable if $\langle\langle Z(t), \psi\rangle\rangle \in L^1(\mathbb{R})$ for all $\psi \in (\mathcal{S})_H$. Then the $(\mathcal{S})_H^*$ integral of Z denoted by $\int_{\mathbb{R}} Z(t)dt$ is the unique element in $(\mathcal{S})_H^*$ such that*

$$\left\langle\left\langle \int_{\mathbb{R}} Z(t)dt, \psi \right\rangle\right\rangle = \int_{\mathbb{R}} \langle\langle Z(t), \psi\rangle\rangle \, dt \quad \text{for all } \psi \in (\mathcal{S})_H. \tag{25}$$

Remark. $t \to B_t^H$ is differentiable in $(\mathcal{S})_H^*$; i.e., fractional noise is a well-defined object and we denote it by (W_t^H).

Below we describe the Wick product which is the last ingredient necessary for describing the integration w.r.t. fBm.

DEFINITION 3.2. *Suppose $F, G \in (\mathcal{S})_H^*$ are given by*

$$F(w) = \sum_{\alpha\in\mathcal{J}} a_\alpha H_\alpha(w) \quad \text{and} \quad G(w) = \sum_{\beta\in\mathcal{J}} b_\beta H_\beta(w). \tag{26}$$

Then the Wick product $F \diamond G$ of F and G is defined as

$$(F \diamond G)(w) = \sum_{\alpha,\beta\in\mathcal{J}} a_\alpha b_\beta H_{\alpha+\beta}(w). \tag{27}$$

Remark. $(\mathcal{S})_H$ and $(\mathcal{S})_H^*$ are closed under the Wick product.

The Wick exponential $\exp^\diamond$ is defined as $\exp^\diamond(X) = \sum_{n=0}^\infty \frac{X^{\diamond n}}{n!}$, provided the series converges in $(\mathcal{S})_H^*$, where $X^{\diamond n} = X \diamond \cdots \diamond X$ (n factors). And we have that $\exp^\diamond(\langle w, f\rangle) = \exp(\langle w, f\rangle - \frac{1}{2}|f|_\phi^2)$ for $f \in L_\phi^2(\mathbb{R})$ (see [16]).

DEFINITION 3.3. *Suppose $Y : \mathbb{R} \to (\mathcal{S})_H^*$ is such that $Y(t) \diamond W_t^H$ is integrable in $(\mathcal{S})_H^*$. Then $\int_{\mathbb{R}} Y(t)dB_t^H$ is defined by*

$$\int_{\mathbb{R}} Y(t)dB_t^H := \int_{\mathbb{R}} Y(t) \diamond W_t^H dt. \tag{28}$$

LEMMA 3.4. *Let $\mathcal{L}_\phi^{1,2}(\mu)$ be as in (15). If $Y \in \mathcal{L}_\phi^{1,2}(\mu)$, then $\int_{\mathbb{R}} Y_t dB_t^H$ exists as an element of $L^2(\mu)$ and its norm is given by $\| Y \|_{\mathcal{L}_\phi^{1,2}(\mu)}$.*

For finding the equilibrium strategies we also make use of the Hida derivative (which is called the Malliavin derivative in the context of Wiener space), which we will define below. We first define the directional derivative.

DEFINITION 3.5. *Suppose that $F : \mathcal{S}' \to \mathbb{R}$ and $\gamma \in \mathcal{S}'$. Then the directional (Gateaux) derivative of F in the direction of γ is given by*

$$D_\gamma F(w) := \lim_{\epsilon \to 0} \frac{F(w + \epsilon\gamma) - F(w)}{\epsilon} \quad \text{if it exists in } (\mathcal{S})_H^*. \tag{29}$$

DEFINITION 3.6. *$F : \mathcal{S}' \to \mathbb{R}$ is said to be differentiable if there is a map $K : \mathbb{R} \to (\mathcal{S})_H^*$ such that*

$$\begin{aligned} &K(t,w)\gamma(t) \text{ is } (\mathcal{S})_H^* \text{ integrable} \\ &\text{and} \quad D_\gamma F(w) = \int_{\mathbb{R}} K(t,w)\gamma(t)dt \text{ for all } \gamma \in L^2(\mathbb{R}). \end{aligned} \tag{30}$$

Then $D_t F(w) := K(t,w)$ is said to be the Hida derivative of F.

We will make use of the Pothoff–Timpel test functions and distributions (see [29] for the definitions) to define quasi-conditional expectation in the following sections. We denote these spaces by $\mathcal{G}$ and $\mathcal{G}^*$, respectively. The Hida derivative of the random variables in $\mathcal{G}^*$ exists.

Let $F = \sum_\alpha c_\alpha H_\alpha(w) \in \mathcal{G}^*$. Then the Hida derivative exists and is given by

$$D_t F(w) = \sum_\alpha c_\alpha \sum_i \alpha_i H_{\alpha - \varepsilon^i}(w) e_i(t), \tag{31}$$

where $\varepsilon^i = (0, \dots, 0, 1, 0, \dots, 0)$ with the 1 in the ith component.

We proceed by defining the quasi-conditional expectation and then introducing the fractional Clark–Ocone theorem which will be crucial in reducing the dynamic optimization problems of the next section into static optimization problems.

DEFINITION 3.7 (see [16]). *If $F \in \mathcal{G}^*(\mu)$ has the expansion*

$$F(w) = \sum_{n=0}^{\infty} \int_{[0,T]^n} f_n (dB^H)^{\otimes n}, \tag{32}$$

then its quasi-conditional expectation is given by

$$\tilde{E}_\mu\{F|\mathcal{F}_t\} = \sum_{n=0}^{\infty} \int_{[0,t]^n} f_n (dB^H)^{\otimes n}. \tag{33}$$

Note that $\tilde{E}_\mu\{F|\mathcal{F}_t\} \neq E_\mu\{F|\mathcal{F}_t\}$ in general. (Only for $H = \frac{1}{2}$ is the quasi-conditional expectation operator the same as the conditional expectation operator on $L^2(\mu)$.) However, the following holds: $\tilde{E}\{F|\mathcal{F}_t\} = F$ a.s. $\Leftrightarrow F$ is $\mathcal{F}_t$ measurable.

The following feature of the quasi-conditional expectation will be helpful in the computations in the next section:

$$\tilde{E}\{F \diamond G|\mathcal{F}_t\} = \tilde{E}\{F|\mathcal{F}_t\} \diamond \tilde{E}\{G|\mathcal{F}_t\} \quad \text{for } F, G \in \mathcal{G}^*. \tag{34}$$

We will also need the notion of a quasi martingale which is defined as follows.

DEFINITION 3.8. *Suppose M_t is an $(\mathcal{F}_t)$ adapted process in $\mathcal{G}^*$. It is called a quasi martingale if*

$$\tilde{E}\{M_t|\mathcal{F}_s\} = M_s \quad \text{for all } t \geq s. \tag{35}$$

LEMMA 3.9 (see [17]). *Let $F \in \mathcal{L}_\phi^{1,2}(\mu)$. Then $M_t = \int_0^t F_s dB_s^H$ is a quasi martingale.*

Now we will state the fractional Clark–Ocone theorem.

THEOREM 3.10 (see [16]). *Suppose $G(w) \in L^2(\mu)$ is $\mathcal{F}_T$ measurable. Define $\psi(t,w) = \tilde{E}_\mu\{D_t G | \mathcal{F}_t\}$, where $D_t G$ is the Hida derivative of G at t, which exists as an element of $\mathcal{G}^*(\mu)$. Then $\psi \in \mathcal{L}_\phi^{1,2}(\mu)$ and*

$$G(w) = E_\mu\{G\} + \int_0^T \psi(t,w) dB_t^H. \tag{36}$$

4. Proof of Theorem 2.2. Recall that in Theorem 2.2, we consider the one-dimensional state equation (7) where the pay-off function of player i is of the form (8).

As noted previously, we will employ the fractional Clark–Ocone formula and the Wick calculus introduced in section 3 to find Nash equilibria for this type of game. We begin this development by stating a fractional version of the Girsanov theorem, which is given by [16].

THEOREM 4.1 (see [16]). *Suppose $T > 0$ and $u : [0,T] \to \mathbb{R}$ is continuous. Suppose further that $K : [0,T] \to \mathbb{R}$ satisfies the equation*

$$\int_0^T K(s)\phi(s,t) ds = u(t), \quad 0 \le t \le T, \tag{37}$$

where ϕ is given by (13). *Extend K to $\mathbb{R}$ by putting $K(s) = 0$ outside $[0,T]$. Define the probability measure $\hat{\mu}$ on F_T by*

$$\frac{d\hat{\mu}(w)}{d\mu(w)} = \exp\left(- \int_0^T K(s) dB_s^H - \frac{1}{2}|K|_\phi^2 \right). \tag{38}$$

Then $\hat{B}_t^H = \int_0^t u(s) ds + B_t^H$ is an fBm with respect to $\hat{\mu}$.

The dynamics of the state (7) can be written as

$$d(e^{-rt} X_t) - e^{-rt} \sum_{i=1}^N \alpha_i(t) u_i(t) dt = e^{-rt} \sum_{i=1}^N \beta_i(t) v_i(t) (C dt + dB_t^H). \tag{39}$$

Let η and $\hat{\mu}$ be defined as in (12); i.e.,

$$\begin{aligned} \eta(T) = \frac{d\hat{\mu}}{d\mu} &= \exp\left(- \int_0^T K(s) dB_s^H - \frac{1}{2}|K|_\phi^2 \right) \\ &= \exp^\diamond\left(- \int_0^T K(s) dB_s^H \right), \end{aligned} \tag{40}$$

where K is from (9). Then since K solves (37) for $u(t) = C$ (see Lemma 7.1) and by the fractional Girsanov formula, the process

$$\hat{B}_t^H = Ct + B_t^H \tag{41}$$

is an fBm with respect to $\hat{\mu}$ having the same Hurst parameter as the modulating process in (7). Thus, the differential equation describing the flow of the state is

given in terms of $\hat{B}^H$ as

$$e^{-rt}X_t - \int_0^t e^{-rs}\sum_{i=1}^N \alpha_i(s)u_i(s)dt = x + \int_0^t e^{-rs}\sum_{i=1}^N \beta_i(s)v_i(s)d\hat{B}_s^H. \tag{42}$$

To be able to find a Nash equilibrium, we will use the quasi-martingale approach to stochastic control in the proof. (For another application of this approach see [17].) We first find the best response of a player to the given strategies of other players and for that we will use the fractional Clark–Ocone theorem (Theorem 3.10).

By (14) we have that $e^{-rt}X_t - \int_0^t e^{-rs}\sum_{i=1}^N \alpha_i(s)u_i(s)dt \in L^2(\hat{\mu})$. And note that by Lemma 3.9 and (14), $\int_0^t e^{-rs}\sum_{i=1}^N \beta_i(s)v_i(s)d\hat{B}_s^H$ is a quasi martingale. Therefore we have

$$E_{\hat{\mu}}\left\{e^{-rt}X_t - \int_0^t e^{-rs}\sum_{i=1}^N \alpha_i(s)u_i(s)dt\right\} = x.$$

Now let G be given by

$$G = e^{-rT}F_i - \int_0^T e^{-rs}\sum_{j=1}^N \alpha_j(s)u_j(s)ds. \tag{43}$$

Assume $G \in L^2(\hat{\mu})$. Then if

$$E_{\hat{\mu}}\{G\} = x, \tag{44}$$

by the fractional Clark–Ocone formula (36) we have

$$G = x + \int_0^T \tilde{E}_{\hat{\mu}}\left\{D_sG\middle|\mathcal{F}_s\right\}d\hat{B}_s^H. \tag{45}$$

If we choose v_i in (42) such that

$$v_i(s) = \frac{-\sum_{j\neq i}\beta_j(s)v_j(s) + e^{rs}\tilde{E}_{\hat{\mu}}\left\{D_sG\middle|\mathcal{F}_s\right\}}{\beta_i(s)}, \tag{46}$$

then from (45) we see that $X_T = F_i$.

By the above argument we can change the dynamic optimization problem of maximizing (8) under the dynamics (7) into a static optimization problem. In particular, given the other players' strategies, player i wishes to solve the following maximization problem:

$$\begin{aligned} K_i(x) = \sup_{u_i,F_i}\Bigg\{&E_\mu\left\{\int_0^T \frac{c_iu_i(t)^{\gamma_i}}{\gamma_i}dt + \frac{b_iF_i^{\gamma_i'}}{\gamma_i'}\right\};\ \text{given that} \\ &E_{\hat{\mu}}\left\{-\int_0^T e^{-rs}\sum_{j=1}^N \alpha_j(s)u_j(s)ds + e^{-rT}F_i\right\} = x\Bigg\}, \end{aligned} \tag{47}$$

where the supremum is taken over F_i and (u_i) such that

$$e^{-rT}F_i - \int_0^T e^{-rs}\sum_{j=1}^N \alpha_j(s)u_j(s)ds \in L^2(\hat{\mu}). \tag{48}$$

This optimization problem can be solved by first considering for each $\lambda_i > 0$ the unconstrained problem

$$C_i(x,\lambda) = \sup_{u_i,F_i} \left\{ E_\mu \left\{ \int_0^T \frac{c_i u_i(t)^{\gamma_i}}{\gamma_i} dt + \frac{b_i F_i^{\gamma_i'}}{\gamma_i'} \right\} \right.$$

(49)

$$\left. + \lambda_i E_{\hat{\mu}} \left\{ -\int_0^T e^{-rs} \sum_{j=1}^N \alpha_j(s) u_j(s) ds + e^{-rT} F_i \right\} \right\},$$

and then solving for λ_i from the slackness condition

$$E_{\hat{\mu}} \left\{ -\int_0^T e^{-rs} \sum_{j=1}^N \alpha_j(s) u_j(s) ds + e^{-rT} F_i \right\} = x. \tag{50}$$

Let us define, as before, the following random variable:

$$\rho(t,w) = E_\mu \big\{ \eta(T) \big| \mathcal{F}_t \big\}, \tag{51}$$

where η is from (40). Using the fact that

$$E_\mu\{\eta(T) u_i(t)\} = E_\mu\{\rho(t) u_i(t)\}, \tag{52}$$

we can solve (4) by maximizing pointwise, i.e., for each t and w, the functions

$$g_i(u_i) = \frac{c_i u_i^{\gamma_i}}{\gamma_i} - \lambda_i \rho(t,w) e^{-rt} \sum_{j=1}^N \alpha_j(t) u_j \tag{53}$$

$$\text{and} \quad h_i(F_i) = \frac{b_i F_i^{\gamma_i'}}{\gamma_i'} - \lambda_i \eta(T,w) e^{-rT} F_i. \tag{54}$$

Since $0 < \gamma_i < 1$, these functions are concave, and therefore we can solve $g_i'(u_i) = 0$ and $h_i'(F_i) = 0$ to find the maximum points, which are given by

$$u_i(t) = \left(\frac{\lambda_i \rho(t,w) e^{-rt} \alpha_i(t)}{c_i} \right)^{\frac{1}{\gamma_i - 1}}, \tag{55}$$

$$\text{and} \quad F_i = \left(\frac{\lambda_i \eta(T,w) e^{-rT}}{b_i} \right)^{\frac{1}{\gamma_i' - 1}}. \tag{56}$$

Since $\alpha_i(t)$ is bounded by assumption, (48) is satisfied. Note that at the Nash equilibrium F_i is independent from the player index i, i.e., $F_i = F^e$ for all i. We will use this condition to show that the Lagrange multipliers at the equilibrium are necessarily linear in b_i and then use the slackness condition to actually find their value. First we will find $E_\mu\{\eta(T)^{\frac{1}{\gamma_i' - 1}}\}$. Note that

(57)

$$\begin{aligned} \eta(T)^{\frac{1}{\gamma_i' - 1}} &= \exp\left(\frac{1}{1-\gamma_i'} \int_0^T K(s) dB_s^H + \frac{1}{2(1-\gamma_i')} |K|_\phi^2 \right) \\ &= \exp\left(\frac{1}{1-\gamma_i'} \int_0^T K(s) dB_s^H - \frac{1}{2(1-\gamma_i')^2} |K|_\phi^2 \right) \exp\left(\frac{2-\gamma_i'}{2(1-\gamma_i')^2} |K|_\phi^2 \right). \end{aligned}$$

$$\text{Since} \quad E\left\{\exp^{\diamond}\left(\int_0^T f(s)dB_s^H\right)\right\} = 1, \tag{58}$$

for all $f \in L^2(\mu)$, we have

$$E\left\{\eta(T)^{\frac{1}{\gamma_i'-1}}\right\} = \exp\left(\frac{2-\gamma_i'}{2(1-\gamma_i')^2}|K|_\phi^2\right). \tag{59}$$

Therefore using (56) we obtain

$$EF_i = \left(\frac{\lambda_i}{b_i}\right)^{\frac{1}{\gamma_i'-1}} \exp\left(\frac{2-\gamma_i'}{2(1-\gamma_i')^2}|K|_\phi^2 - \frac{rT}{\gamma_i'-1}\right). \tag{60}$$

It follows that

$$F_i = E\{F_i\}\exp^{\diamond}\left(\frac{1}{1-\gamma_i'}\int_0^T K(s)dB_s^H\right). \tag{61}$$

From (61) we see that for a Nash equilibrium to exist, it is necessary that we have $\gamma_i' = \gamma'$ and $\lambda_i^e = mb_i$. From (4) we see that m is to be found from the slackness condition:

$$\begin{aligned} E_\mu\Bigg\{ &\int_0^T e^{-rt}\rho(t)\left(\sum_{i=1}^N \alpha_i(t)\left(\frac{\lambda_i^e\rho(t)e^{-rt}\alpha_i(t)}{c_i}\right)^{\frac{1}{\gamma_i-1}}\right)dt \\ &+ e^{-rT}\eta(T)\left(\frac{\lambda_i^e\eta(T)e^{-rT}}{b_i}\right)^{\frac{1}{\gamma_i'-1}}\Bigg\} = x. \end{aligned} \tag{62}$$

Note that by (40) we have the following:

$$\begin{aligned} \eta(T)^{\frac{\gamma'}{\gamma'-1}} &= \exp\left(\frac{\gamma'}{1-\gamma'}\int_0^T K(s)dB_s^H + \frac{\gamma'}{2(1-\gamma')}|K|_\phi^2\right) \\ &= \exp^{\diamond}\left(\frac{\gamma'}{1-\gamma'}\int_0^T K(s)dB_s^H\right)\exp\left(\frac{\gamma'}{2(1-\gamma')^2}|K|_\phi^2\right), \end{aligned} \tag{63}$$

$$E\left\{\eta(T)^{\frac{\gamma'}{\gamma'-1}}\right\} = \exp\left(\frac{\gamma'}{2(\gamma'-1)^2}|K|_\phi^2\right). \tag{64}$$

Using Theorem 3.2 of [15], $\rho(t,w)$ can be written as

$$\rho(t,w) = \exp\left(-\int_0^t \zeta_t(s)dB_s^H - \frac{1}{2}|\zeta_t|_\phi^2\right), \tag{65}$$

where ζ_t is given by

$$\begin{aligned} \big((-\Delta)^{-(H-1/2)}\zeta_t\big)(s) &= \big((-\Delta)^{-(H-1/2)}K\big)(s), \quad 0 \le s \le t, \\ \zeta_t(s) &= 0, \quad s < 0 \quad \text{or} \quad s > t, \end{aligned} \tag{66}$$

with the operator $(-\Delta)^{-(H-1/2)}$ on $L^2(\mu)$ defined by (11).

Thus

$$E\left\{\rho_t^{\frac{\gamma_i}{\gamma_i-1}}\right\} = E\Bigg\{\exp\left(\frac{\gamma_i}{1-\gamma_i}\int_0^T \zeta_t(s)dB_s^H - \frac{\gamma_i^2}{2(1-\gamma_i)^2}|\zeta_t|_\phi^2 + \frac{\gamma_i}{2(1-\gamma_i)}|\zeta_t|_\phi^2 + \frac{\gamma_i^2}{2(1-\gamma_i)^2}|\zeta_t|_\phi^2\right)\Bigg\}, \tag{67}$$

from which we conclude that

$$E\left\{\rho(t)^{\frac{\gamma_i}{\gamma_i-1}}\right\} = \exp\left(\frac{\gamma_i}{2(1-\gamma_i)^2}|\zeta_t|_\phi\right), \tag{68}$$

so that m can be solved from (62), which leads to the following equation:

$$\sum_{i=1}^N \int_0^T m^{\frac{1}{\gamma_i-1}} b_i^{\frac{1}{\gamma_i-1}} \alpha_i(t)^{\frac{\gamma_i}{\gamma_i-1}} e^{-rt\frac{\gamma_i}{\gamma_i-1}} \exp\left(\frac{\gamma_i}{2(1-\gamma_i)^2}|\zeta_t|_\phi\right) dt + m^{\frac{1}{\gamma'-1}} e^{-rT\frac{\gamma'}{\gamma'-1}} \exp\left(\frac{2\gamma'|K|_\phi}{2(1-\gamma')^2}\right) = x. \tag{69}$$

After solving for m using (69), then by (55) and (56) we have the final state at the equilibrium and strategy u_i for player i leading to that state, given, respectively, by

$$F^e = m^{\frac{1}{\gamma'-1}}\eta(T)^{\frac{1}{\gamma'-1}} e^{\frac{-rT}{\gamma'-1}} \tag{70}$$

and

$$u_i^e(t) = \left(\frac{mb_i\alpha_i(t)}{c_i}e^{-rt}\rho(t,w)\right)^{\frac{1}{\gamma_i-1}}. \tag{71}$$

Observe that these controls are not Markovian. (In a Brownian motion setting the controls were assumed to be Markovian at the outset so that the HJB equations for an equilibrium solution can be developed [6], [11], and [27].)

Now we will proceed to find (v_i) at the equilibrium, which is the second component of the players' strategies. For this we will again make use of the fractional Clark–Ocone formula.

Suppose G^e is given by

$$G^e = e^{-rT}F^e - \int_0^T e^{-rs}\sum_{i=1}^N \alpha_i u_i^e(s)ds. \tag{72}$$

Since there is a unique adapted process $\psi(t,w)$ such that

$$G^e = E_\mu\{G^e\} + \int_0^T \psi(t,w)d\hat{B}_t^H, \tag{73}$$

which, from the Clark–Ocone formula, is given by

$$\psi(t,w) = \tilde{E}_{\hat{\mu}}\big\{D_tG^e\big|\mathcal{F}_t\big\}, \tag{74}$$

it can now be seen immediately that any adapted (v_i^e) that satisfies

$$\tilde{E}_\mu\{D_t G^e|\mathcal{F}_t\} = e^{-rt}\sum_{i=1}^N \beta_i v_i^e(t) \tag{75}$$

is an equilibrium strategy.

To obtain a more explicit expression, we will compute $\tilde{E}_{\hat{\mu}}\{D_t G^e|\mathcal{F}_t\}$. Using (70) and (71), G^e is given by

$$\begin{aligned} G^e(T,w) &= m^{\frac{1}{\gamma'-1}} e^{-rT\frac{\gamma'}{\gamma'-1}} \eta(T,w)^{\frac{1}{\gamma'-1}} \\ &\quad - \int_0^T \sum_{i=1}^N \alpha_i(t)^{\frac{\gamma_i}{\gamma_i-1}} \left(\frac{mb_i}{c_i}\right)^{\frac{1}{\gamma_i-1}} e^{-rt\frac{\gamma_i}{\gamma_i-1}} \rho(t,w)^{\frac{1}{\gamma_i-1}} dt. \end{aligned} \tag{76}$$

To calculate the quasi-conditional expectation of the Hida derivative of G^e we will first find it for the stochastic part of the first term on the right-hand side of (76). Define R as

$$R = \exp\left(\frac{2-\gamma'}{(1-\gamma')^2}|K|_\phi^2 - \frac{C}{1-\gamma'}\int_0^T K(s)ds\right). \tag{77}$$

Using the chain rule, (41), and (33), we have

$$\begin{aligned} \tilde{E}_{\hat{\mu}}\left\{D_t\eta(T)^{\frac{1}{\gamma'-1}}|\mathcal{F}_t\right\} &= \tilde{E}_{\hat{\mu}}\left\{\frac{K(t)}{1-\gamma'}\eta(T)^{\frac{1}{\gamma'-1}}|\mathcal{F}_t\right\} \\ &= \frac{K(t)}{1-\gamma'}R\tilde{E}_{\hat{\mu}}\left\{\exp^\diamond\left(\frac{1}{1-\gamma'}\int_0^T K(s)d\hat{B}_s^H\right)|\mathcal{F}_t\right\} \\ &= \frac{K(t)}{1-\gamma'}R\exp^\diamond\left(\frac{1}{1-\gamma'}\int_0^t K(s)d\hat{B}_s^H\right) \\ &= \frac{K(t)}{1-\gamma'}\exp\left(\frac{1}{1-\gamma'}\int_0^T K(s)dB_s^H - \frac{C}{1-\gamma'}\int_t^T K(s)ds\right. \\ &\qquad \left. + \frac{2-\gamma}{2(1-\gamma^2)}|K|_\phi^2 - \frac{1}{1-\gamma}|K1_{[0,t]}|_\phi^2\right). \end{aligned} \tag{78}$$

Now we will find the quasi-conditional expectation of the Hida derivative of the stochastic part of second term on the right-hand side of (76) using (65), i.e.,

$$\begin{aligned} &\tilde{E}_{\hat{\mu}}\left\{D_t\left(e^{-ru\frac{\gamma_i}{\gamma_i-1}}\rho(u)^{\frac{1}{\gamma_i-1}}\right)|\mathcal{F}_t\right\} = \\ &\frac{\zeta_u(t)}{1-\gamma_i}e^{-ru\frac{\gamma_i}{\gamma_i-1}}\exp\left(\frac{1}{1-\gamma_i}\int_0^t \zeta_u(s)dB_s^H - \frac{C}{1-\gamma_i}\int_t^u \zeta_u(s)ds\right. \\ &\qquad \left. + \frac{2-\gamma_i}{2(1-\gamma_i)^2}|\zeta_u|_\phi^2 - \frac{1}{1-\gamma_i}|\zeta_u 1_{[0,t]}|_\phi^2\right). \end{aligned} \tag{79}$$

Using (75) and (76), we have the result for the second component for the players' equilibrium strategies, and this concludes the proof of Theorem 2.2. □

5. Extension of the Wick calculus to arbitrary Gaussian processes. Although the results of the preceding sections have considered the explicit case in which the modulator in (1) is fBm, these results can be extended to the situation in which the modulator is a more general Gaussian process within sufficient regularity. This requires the extension of the Wick calculus to more general Gaussian processes. In this section, we sketch how this extension can be accomplished. The first step in extending the fractional noise machinery introduced in section 3 to more general Gaussian processes is the following theorem due to Loève [21] for integrating deterministic functions with respect to second-order processes.

THEOREM 5.1 (see [21]). *Suppose that X is a zero-mean process such that $E\{X_t^2\} < \infty$ for all t, and denote its covariance function by R. Then, for $-\infty<a<b<\infty$,*

$$\int_a^b f(t)dX_t \tag{80}$$

exists as the L^2-limit of Riemann sums if and only if

$$|f|_R^2 := \int_a^b \int_a^b f(t)f(s)d^2R(s,t) < \infty. \tag{81}$$

Henceforth X will denote a Gaussian process. By the Bochner–Minlos theorem, there exists a unique probability measure on the space of tempered distributions such that $\langle \cdot, f\rangle : \Omega \to \mathbb{R}$ is a Gaussian random variable with mean 0 and variance $|f|_R^2$.

We will denote $L^2(\mu)$ by $L^2(X)$, and $H(X)$ will denote the linear space of X, i.e., the closed subspace of $L^2(X)$ spanned by X_t for all $t \in [a,b]$ (i.e., the first Wiener chaos). As in [18] we construct $\Lambda(R)$, a Hilbert space of deterministic integrable "functions" isomorphic to $H(X)$ by completing the pre-Hilbert space of step functions $\mathbb{S}$ with the inner product

$$\langle f, g\rangle_{\mathbb{S}} = \int\int f(t)g(s)d^2R(t,s) \tag{82}$$

for any $f, g \in \mathbb{S}$. Then the integration operator defined on the set of step functions (the integration with respect to X) can be extended to an isomorphism between $H(X)$ and $\Lambda(R)$. (The elements of $\Lambda(R)$ are generalized functions, i.e., distributions [28].)

As a second step we will define the Wick-integrability of a random process with respect to a Gaussian process. This is done by utilizing the tensor product structure of the space $L^2(X)$. Let us define the tensor product of Hilbert spaces.

DEFINITION 5.2. *The algebraic tensor product $H_1 \otimes H_2$ of Hilbert spaces H_1 and H_2 is a pre-Hilbert space with the following inner product:*

$$\langle h_1 \otimes h_2, g_1 \otimes g_2\rangle_{H_1\otimes H_2} := \langle h_1, g_1\rangle_{H_1}\langle h_2, g_2\rangle_{H_2} \tag{83}$$

for $g_i, h_i \in H_i$ and $i = 1, 2$. The closure of this pre-Hilbert space is the tensor product of Hilbert spaces, which will still be denoted by $H_1 \otimes H_2$. $H_1 \tilde{\otimes} H_2$ will denote the symmetrized tensor product.

Then we have the following Wiener chaos isomorphism theorem.

THEOREM 5.3 (see [19]). *$\oplus_{p\geq 0} H^{\tilde{\otimes} p}(X)$ is isomorphic to $L^2(X)$ with the unique isomorphism Φ defined by*

$$\Phi(\xi_1^{\tilde{\otimes}\alpha_1} \tilde{\otimes} \cdots \tilde{\otimes} \xi_k^{\tilde{\otimes}\alpha_k}) = \frac{1}{\sqrt{p!}} \prod_{j=1}^{k} h_{\alpha_j}(\xi_j), \tag{84}$$

where $\xi_i \in H(X)$ for all i are orthonormal; $p = |\alpha| = \alpha_1 + \cdots + \alpha_k$. Here h_n is the Hermite polynomial of degree n (see Appendix C of [14]*).*

Note that for a random variable $\xi \in H(X)$ with unit variance we have

$$\Phi(e^{\tilde{\otimes}\xi}) = \exp\left(\xi - \frac{1}{2}\right), \tag{85}$$

where the exponential is defined by $e^{\tilde{\otimes}\xi} = \sum_{p\geq 0} \frac{\xi^{\tilde{\otimes}p}}{\sqrt{p!}}$.

We proceed as in [18] and in order to define the integral of a stochastic process with respect to X, we first define a tensor product integral, denoted by $\int F_t \otimes dX_t$, and its domain, denoted by $\Lambda(R)_{L^2(X)}$.

Suppose $\mathbb{S}_{L^2(X)}$ is the pre-Hilbert space of the $L^2(X)$-valued step functions F_t,

$$F_t = \sum_{i=1}^{N} F_i 1_{(t_i, t_{i+1}]}, \tag{86}$$

for $(t_i, t_{i+1}] \in [a, b]$, and $F_i \in L^2(X)$, equipped with the inner product

$$\langle F, G\rangle = \int\int \langle F_t, G_s\rangle_{L^2(X)} d^2R(t, s). \tag{87}$$

Let $\Lambda(R)_{L^2(X)}$ denote the completion of $\mathbb{S}_{L^2(X)}$. For the $F \in \mathbb{S}_{L^2(X)}$ given in (86) define the integral $I_\otimes$ as

$$\int F_t \otimes dX_t = \sum_{i=1}^{N} F_{t_i} \otimes (X_{t_{i+1}} - X_{t_i}). \tag{88}$$

Since this integral is a norm-preserving linear map, it has a unique extension to an isomorphism from $\Lambda(R)_{L^2(X)}$ into $L^2(X) \otimes H(X)$. We will construct a map Ψ from $L^2(X) \otimes H(X)$ into $L^2(X)$ and call the composition of the two maps, $\Psi(I_\otimes)$, the stochastic integral. We start by defining the linear map

$$\Psi_p : H^{\tilde{\otimes}p}(X) \otimes H(X) \to H^{\tilde{\otimes}p+1}(X) \tag{89}$$

by

$$\Psi_p\left(\left(\xi_1^{\tilde{\otimes}\alpha_1}\tilde{\otimes}\cdots\tilde{\otimes}\xi_k^{\tilde{\otimes}\alpha_k}\right)\otimes\xi_l\right) = (p+1)^{\frac{1}{2}}\xi_1^{\tilde{\otimes}\alpha_1}\tilde{\otimes}\cdots\tilde{\otimes}\xi_k^{\tilde{\otimes}\alpha_k}\tilde{\otimes}\xi_l, \tag{90}$$

where $(\xi_i) \in H(X)$ is an orthonormal set of random variables, $\alpha_1 + \cdots + \alpha_k = p$. Ψ_p can be extended uniquely to a bounded linear map with norm $(p+1)^{1/2}$ from $H^{\tilde{\otimes}p}(X) \otimes H(X)$ onto $H^{\tilde{\otimes}p+1}(X)$.

Now define Ψ^* as the map from $\oplus_{p\geq 0} H^{\tilde{\otimes}p}(X) \otimes H(x)$ onto $\oplus_{p\geq 1} H^{\tilde{\otimes}p}(X)$, by $\Psi^* = \oplus_{p\geq 0}\Psi_p$, viz., the restriction of Ψ^* to $H^{\tilde{\otimes}p}(X) \otimes H(X)$ is Ψ_p. The domain of the operator Ψ^* is given by

$$\mathcal{D}^* = \left\{\eta \in \left(H^{\tilde{\otimes}\alpha_1}(X) \oplus \cdots \oplus H^{\tilde{\otimes}\alpha_m}(X)\right) \otimes H(X) : \alpha_1 + \cdots + \alpha_m < \infty\right\}, \tag{91}$$

so that $\sum_{p\geq 0} \|\Psi_p(\eta_p)\|^2 < \infty$, where η_p is the projection of η on $H^{\tilde{\otimes}p}(X) \otimes H(X)$.

By Theorem 5.3, $\oplus_{p\geq 0}H^{\tilde{\otimes}p}(X)$ is isomorphic to $L^2(X)$. Therefore $(\oplus_{p\geq 0}H^{\tilde{\otimes}p}(X))\otimes H(X)$ is isomorphic to $L^2(X)\otimes H(X)$. Denote this isomorphism by Φ_0. Let $\mathcal{D}=\Phi_0(\mathcal{D}^*)$, which is a proper subset of $L^2(X)\otimes H(X)$. Then define Ψ by

$$\Psi=\Phi\circ\Psi^*\circ\Phi_0^{-1}. \tag{92}$$

We define the Wick product of $V\in L^2(X)$ and $W\in H(X)$ as

$$V\diamond W:=\Psi(V\otimes W). \tag{93}$$

Note that $V\diamond W$ is in $L^2(X)$ if and only if $V\otimes W\in\mathcal{D}\otimes H(X)$.

The integral $\int F_t\diamond dX_t$ is then defined by

$$\int_a^b F_t\diamond dX_t=\Psi\circ I_{\otimes}(F) \tag{94}$$

for all F such that $I_{\otimes}(F)=\int F_t\otimes dX_t\in\mathcal{D}$. The set of all F's in the domain of integration is denoted by $\Lambda(R)^*_{L^2(X)}$. Then we have the Itô representation formula as a result of the multiple Wiener integral (MWI) representation of the random variables in $L^2(X)$ [18] and the fact that each MWI can be written as an iterated integral.

THEOREM 5.4 (see [18]). *Every $\theta\in L^2(X)$ has the representation*

$$\theta=E\{\theta\}+\int_a^b F_t\diamond dX_t \tag{95}$$

for an $F\in\Lambda(R)^*_{L^2(X)}$ that is adapted to the filtration generated by X.

Now let us define the Wick product of two elements in $L^2(X)$. As a first step we define $\Upsilon_{p,q}$,

$$\Upsilon_{p,q}:H^{\tilde{\otimes}p}(X)\otimes H^{\tilde{\otimes}q}(X)\to H^{\tilde{\otimes}p+q}(X), \tag{96}$$

as

$$\begin{aligned}&\Upsilon_{p,q}\left(\left(\xi_{\gamma_1}^{\tilde{\otimes}\alpha_1}\tilde{\otimes}\cdots\tilde{\otimes}\xi_{\gamma_k}^{\tilde{\otimes}\alpha_k}\right)\otimes\left(\xi_{\lambda_1}^{\tilde{\otimes}\beta_1}\tilde{\otimes}\cdots\tilde{\otimes}\xi_{\lambda_l}^{\tilde{\otimes}\beta_l}\right)\right)\\&\quad=\sqrt{\frac{(p+q)!}{p!q!}}\xi_{\gamma_1}^{\tilde{\otimes}\alpha_1}\tilde{\otimes}\cdots\tilde{\otimes}\xi_{\gamma_k}^{\tilde{\otimes}\alpha_k}\tilde{\otimes}\xi_{\lambda_1}^{\tilde{\otimes}\beta_1}\tilde{\otimes}\cdots\tilde{\otimes}\xi_{\lambda_l}^{\tilde{\otimes}\beta_l}\end{aligned} \tag{97}$$

for any (ξ_γ) that is an orthonormal set in $H(X)$.

Let $\Upsilon=\oplus_{p\geq 0}\oplus_{q\geq 0}\Upsilon_{p,q}$; then we define the Wick product of the $W,V\in L^2(X)$ as

$$W\diamond V:=\Phi\left(\Upsilon\left(\Phi^{-1}(W)\otimes\Phi^{-1}(V)\right)\right). \tag{98}$$

Note that $L^2(X)$ is not closed under $\diamond$, since the tensor product of the random variables may not be in the domain of Υ. Then one can define the Hida distribution space, use (98) as the definition of the Wick product over this space, and see that the Wick product so defined is closed over these spaces.

The main machinery we use to develop strategies leading to a Nash equilibrium are the Girsanov formula (the absolute continuity of the translated measure w.r.t. the original measure) and the Clark–Ocone formula. These can be extended to a more general Gaussian modulator with regularity. The Girsanov theorem, Theorem 4.1,

can be stated for sufficiently regular Gaussian processes. (The proof of the Girsanov theorem in [16] does not make use of the explicit expression for ϕ.) The derivation of the Clark–Ocone theorem (Theorem 3.10) is done by using only the tensor product structure of the space L^2 (Theorem 5.3) and the spaces of generalized random variables. Defining the Hida derivative and the quasi-conditional expectation operator (w.r.t. which X is a quasi martingale) for the Gaussian process X, we can restate the Clark–Ocone theorem. Now replacing ζ_t in Theorem 2.2 by ϑ_t such that

$$E\left\{\int_0^T K(s)dX_s\Big|\mathcal{F}_t\right\} = \int_0^t \vartheta_t(s)dX_s, \tag{99}$$

we have a Nash equilibrium theorem for a general Gaussian process. Note that unlike the case of fBm, we cannot in general write ϑ explicitly in terms of K.

Hence, we cannot give an explicit solution for the Nash equilibrium. A general multidimensional theorem can also be restated for a multidimensional Gaussian process with independent components (the components do not have to be identical) by making the conceptual modifications as in the one-dimensional case.

6. Conclusion. In this paper we have explicitly found Nash equilibria for a stochastic differential game in a non-Markovian setting. In this game, the agents modify the dynamics of a common observable state by modifying its drift and volatility. The agents are heterogeneous in their controls and utility functions. We have taken the modulating process to be fractional Brownian motion, since fBm is versatile in modeling long-range dependence phenomena in finance and networks.

Since the diffusion is modulated by a non-Markovian process, the usual way of finding Nash equilibria via HJB equations is not available. Therefore we have made use of the fractional noise calculus to calculate the agents' Nash equilibrium strategies. Although we have taken the modulating process of the diffusions to be fBm, our results hold for more general Gaussian modulating processes with only slight modifications to the white noise machinery.

Our results are applicable to financial markets in which stock price dynamics are modulated with fractional Brownian motion. One of the candidate applications is stock price modeling when each agent's activities in the market affect the price flow (institutional investors are such examples) or if there are transaction costs. This work is also applicable to stochastic portfolio games, in which agents compete for a bonus.

7. Appendix.

LEMMA 7.1 (see [20]). *Let $f : [0,T] \to \mathbb{R}$ be a continuous function and introduce the following integral equation:*

$$\int_0^T \hat{f}(s)\phi(s,t)ds = f(t) \quad \textit{for } t \in [0,T], \tag{100}$$

where ϕ is given by (13)*. The solution to this equation is given by*

(101)
$$\hat{f}(t) = -\frac{1}{d_H} t^{\frac{1}{2}-H} \frac{d}{ds} \int_t^T dw w^{2H-1}(w-t)^{\frac{1}{2}-H} \frac{d}{dw} \int_0^w dz z^{\frac{1}{2}-H}(w-z)^{\frac{1}{2}-H} f(z),$$

where

$$d_H = 2H(2H-1)\left(\Gamma\left(\frac{3}{2}-H\right)\right)^2 \Gamma(2H-1)\cos\left(\pi\left(H-\frac{1}{2}\right)\right). \tag{102}$$

COROLLARY 7.2. *If we take* $f(t) = C$ *on* $[0, T]$ *in the integral equation given by* (100)*, then the solution* $\hat{f}(t)$ *is given by*

$$\hat{f}(t) = \frac{C}{k_H} t^{\frac{1}{2}-H} (T-t)^{\frac{1}{2}-H}, \tag{103}$$

where

$$k_H = 2H(2H-1)\Gamma(2-2H)\Gamma(2H-1)\cos\left(\pi\left(H-\frac{1}{2}\right)\right). \tag{104}$$

Proof. The proof can be found in [16], but we present it here for the sake of completeness:

$$\hat{f}(t) = -\frac{1}{d_H} C t^{\frac{1}{2}-H} \frac{d}{ds} \int_t^T dw w^{2H-1} (w-t)^{\frac{1}{2}-H} \frac{d}{dw} \int_0^w dz z^{\frac{1}{2}-H} (w-z)^{\frac{1}{2}-H}. \tag{105}$$

Note that

$$\frac{\int_0^w z^{\frac{1}{2}-H}(w-z)^{\frac{1}{2}-H} dz}{w^{2-2H}} = B\left(\frac{3}{2}, \frac{3}{2}\right) = \frac{\Gamma\left(\frac{3}{2}-H\right)^2}{\Gamma(3-H)}, \tag{106}$$

where $B(\cdot,\cdot)$ is the beta function given by

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt. \tag{107}$$

Hence

$$\frac{d}{dw} \int_0^w z^{\frac{1}{2}-H}(w-z)^{\frac{1}{2}-H} dz = \frac{\Gamma\left(\frac{3}{2}-H\right)^2}{\Gamma(2-H)} w^{1-2H}. \tag{108}$$

Using (108) it is not hard to evaluate (105) to get (103). □

REFERENCES

[1] K. AASE, B. ØKSENDAL, AND J. UBØE, *White noise generalizations of the Clark-Haussmann-Ocone theorem with application to mathematical finance*, Finance Stoch., 4 (2000), pp. 465–496.

[2] R. J. BARTON AND H. V. POOR, *Signal detection in fractional Gaussian noise*, IEEE Trans. Inform. Theory, 34 (1988), pp. 943–959.

[3] E. BAYRAKTAR, U. HORST, AND R. SIRCAR, *A Limit Theorem for Financial Markets with Inert Investors*, Princeton University, Princeton, NJ, 2003, preprint.

[4] E. BAYRAKTAR AND H. V. POOR, *Stochastic differential games in a non-Markovian setting* (longer version), 2004, available at http://www.math.lsa.umich.edu/∼erhan/.

[5] D. C. BRODY, J. SYROKA, AND M. ZERVOS, *Dynamical pricing of weather derivatives*, Quant. Finance, 2 (2002), pp. 189–198.

[6] S. BROWNE, *Stochastic differential portfolio games*, J. Appl. Probab., 37 (2000), pp. 126–147.

[7] T. E. DUNCAN, Y. HU, AND B. PASIK-DUNCAN, *Stochastic calculus for fractional Brownian motion* I. Theory, SIAM J. Control Optim., 38 (2000), pp. 582–612.

[8] R. J. ELLIOTT AND J. VAN DER HOEK, *A general fractional white noise theory and applications to finance*, Math. Finance, 13 (2003), pp. 301–330.

[9] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, Academic Press, New York, 1976.

[10] S. D. Gaidov, *Pareto-optimality in stochastic differential games*, Problems Control Inform. Theory, 15 (1986), pp. 439–450.
[11] S. D. Gaidov, *Nash equilibrium in stochastic differential games*, Computers Math. Appl., 12A (1986), pp. 761–768.
[12] S. D. Gaidov, *Z-equilibria in many-player stochastic differential games*, Arch. Math. (Brno), 29 (1993), pp. 123–133.
[13] G. Gripenberg and I. Norros, *On the prediction of fractional Brownian motion*, J. Appl. Probab., 33 (1996), pp. 400–410.
[14] H. Holden, B. Øksendal, J. Ubøe, and T. Zhang, *Stochastic Partial Differential Equations*, Probab. Appl., Birkhäuser Boston, Boston, 1996.
[15] Y. Hu, *Prediction and translation of fractional Brownian motions*, in Stochastics in Finite and Infinite Dimensions, Trends Math., T. Hida et al., eds., Birkhäuser Boston, Boston, 2001, pp. 153–171.
[16] Y. Hu and B. Øksendal, *Fractional white noise calculus and applications to finance*, Infin. Dimens. Anal. Quantum Probab. Relat. Top., 6 (2003), pp. 1–32.
[17] Y. Hu, B. Øksendal, and A. Sulem, *Optimal consumption and portfolio in a Black-Scholes market driven by fractional Brownian motion*, Infin. Dimens. Anal. Quantum Probab. Relat. Top., 6 (2003), pp. 519–536.
[18] S. T. Huang and S. Cambanis, *Stochastic and multiple Wiener integrals for Gaussian processes*, Ann. Probab., 6 (1978), pp. 585–614.
[19] G. Kallianpur, *The role of reproducing kernel Hilbert spaces in the study of Gaussian processes*, in Advances in Probability and Related Topics, Vol. 2, Dekker, NY, 1970, pp. 49–83.
[20] A. Le Breton, *Filtering and parameter estimation in a simple linear system driven by fractional Brownian motion*, Statist. Probab. Lett., 39 (1998), pp. 263–274.
[21] M. Loève, *Probability Theory*, Van Nostrand, New York, 1955.
[22] C. Marinelli, *The Stochastic Goodwill Problem*, preprint, http://www.arxiv.org/abs/math.OC/0310316 (2003).
[23] L. Nilakantan, *Continuous Time Stochastic Games*, Ph.D. dissertation, School of Statistics, University of Minnesota, Minneapolis, MN, 1993.
[24] I. Norros, *On the use of fractional Brownian motion in the theory of connectionless networks*, IEEE J. on Selected Areas in Communications, 13 (1995), pp. 953–962.
[25] C. J. Nuzman and H. V. Poor, *Linear estimation of self-similar processes via Lamberti's transformation*, J. Appl. Probab., 37 (2000), pp. 429–452.
[26] C. J. Nuzman and H. V. Poor, *Reproducing kernel Hilbert space methods for wide-sense self-similar processes*, Ann. Appl. Probab., 11 (2001), pp. 1199–1219.
[27] B. Øksendal and K. Reikvam, *Stochastic differential games with controls - discussion of a specific example*, in Proceedings of the First Symposium on Mathematical Finance, Gaborone, Botswana, 1997, E. Lungu, ed., pp. 74–82; also available at http://www.math.uio.no/eprint/pure_math/1998/pure_1998.html.
[28] V. Pipiras and M. S. Taqqu, *Integration questions related to fractional Brownian motion*, Probab. Theory Related Fields, 118 (2000), pp. 251–291.
[29] J. Pothoff and M. Timpel, *On a dual pair of smooth and generalized random variables*, Potential Anal., 4 (1995), pp. 637–654.
[30] S. G. Samko, A. A. Kilbas, and O. I. Marichev, *Fractional Integrals and Derivatives*, Gordon and Breach, Yverdon, Switzerland, 1993.
[31] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes*, Chapman and Hall, London, 1994.
[32] P. Varaiya, *N-player stochastic differential games*, SIAM J. Control Optim., 14 (1976), pp. 538–545.
[33] M. Zähle, *Integration with respect to fractal functions and stochastic calculus*, Probab. Theory Related Fields, 21 (1998), pp. 333–374.
[34] T. S. Zhang, *Characterizations of the white noise test functionals and Hida distributions*, Stochastics, Stochastic Rep. 41 (1992), pp. 71–87.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1757–1786

THE INSTABILITY OF OPTIMAL CONTROL PROBLEMS TO TIME DELAY*

ALEXEY S. MATVEEV†

Abstract. This paper considers a general optimal control problem with a small time delay Δ in both the state and control. The objective is to find a correct way to neglect the delay or, in other words, to construct a well-posed instantaneous approximation of the delayed problem. The natural way to do so is by merely ignoring the delay in the model. It is shown that this way is almost always incorrect: it gives rise to an error that does not vanish as $\Delta \to +0$, and so results in an ill-posed model. A proper approximation of the delayed problem is offered.

Key words. optimal control, time delay, stability, well-posedness

AMS subject classifications. 49K40, 49K45, 49K25

DOI. 10.1137/S0363012903423211

1. Introduction.

The time delay optimal control problem is a good model of a wide range of real-life phenomena. Unfortunately, the solution of such problems encounters considerable troubles in even the simplest cases. This impels one to employ instantaneous models whenever the delay is small. The conventional way to construct such a model is via merely ignoring the delay. Is this correct? To get an answer, we consider infinitely small delays and analyze the error that accrues from neglecting them. At first sight, the outcome of such an analysis can be easily foreseen and comes to the platitude: the error must be infinitely small as well. This is true indeed if the delay occurs only in the state. However, in the case of the delayed control, this platitudinous property fails to be true not only in general but also "almost always." Therefore the natural way to ignore the delay results in an ill-posed model and so is not acceptable. The objective of this research is to justify these claims and offer well-posed instantaneous approximations of optimization problems with small delays.

The motivation of the issues addressed in this paper comes from many sources. For example, due to the recent growth in communication technology, it is becoming more common to employ digital finite capacity channels for communicating information from controllers to actuators. Not only quantization but also time delay is inherent in transmission via such channels. Moreover, serial digital networks have become popular to serve complex systems with spatially distributed components like advanced aircraft, spacecraft, automotive, industrial and defense systems, arrays of microactuators, power control in mobile communication, multiagent mobile robots, etc. The delay effects are typically enhanced in networks. Indeed, then transmission delays are combined with those induced by the protocol to resolve conflicts between several data sources sharing a common channel [11]. These communication delays are usually small. So as a rule, they are ignored at the theoretical stage of controller design. There are two more reasons to do so. First, in many cases the communication delays are irregular and a priori unknown [11]. Second, ironically, the smaller the delay the harder the delayed optimization from the computational point of view.

*Received by the editors February 15, 2003; accepted for publication (in revised form) May 3, 2004; published electronically March 22, 2005.
http://www.siam.org/journals/sicon/43-5/42321.html
†Department of Mathematics and Mechanics, Saint Petersburg University, Universitetskii pr. 28, Petrodvoretz, St. Petersburg, 198904, Russia (almat@am1540.spb.edu).

In this paper, we consider an optimal control problem in the situation where controls are transmitted to the actuators over two channels. One of them is instantaneous whereas the other provides a constant delay $\Delta \approx 0$. Modulo a shift of the time, this also includes the case where the delays in both channels are nonzero and close to each other. All measurable controls are admissible. This situation is the simplest among those worth studying. However, it is enough to reveal all major points and so is a good option to start with.

We compare the delayed problem with its *natural limit model* (*problem*) that results from putting $\Delta := 0$. We show that in general this problem is not a proper (well-posed) approximation of the delayed one: the optimal value of the cost functional in the latter does not converge to that in the former as $\Delta \to +0$. Moreover, this phenomenon is robust and occurs for almost all delayed problems; the corresponding problems constitute an open and dense subset in the space of all problems. (Any problem is identified with its data.)

Thus the natural way to neglect the delay is almost always incorrect. What is the correct one? As an answer to this question, a proper "limit" (as $\Delta \to 0$) of the delayed problem is found. In brief, this is an instantaneous ($\Delta = 0$) optimal control problem well posed by perturbation [4, 24] of the delay. Here the perturbed problem is the original Δ-delayed one. More specifically, the minimum values of the cost functionals in the limit and Δ-delayed problems, respectively, are close: the latter converges to the former as $\Delta \to 0$. Moreover, solution of the limit problem gives an exhaustive description of the asymptotically optimal Δ-parametric control sequences. For the Δ-delayed problem, the control from such a sequence provides the level of performance, which differs from the optimal one by a vanishing (as $\Delta \to 0$) quantity. The above description is given by the fact that a sequence is asymptotically optimal if and only if it converges to the set of all optimal controls in the limit problem.

The proper limit problem is not standard, in contrast with the natural limit model. This promotes the interest to the cases where the natural model is yet well-posed. Such a case occurs if and only if the optimal values of the cost functionals in the two concerned problems coincide. This criterion is exhaustive but not relevant in the current context since it employs the proper limit problem. We show that the wellposedness of the natural model is not recognizable if the knowledge about the system is confined to only this model. Moreover, then it is impossible to estimate to which extent this model may be ill-posed. Thus information is required about the delayed model, though not the values of the delays. We also offer a simple sufficient criterion. It is far from being exhaustive but deals with a popular situation where the system's equations and the cost functional are linear and convex, respectively, in controls, and the domain of admissible controls is also convex.

As will be shown, the defect of the natural limit model is that it conceals certain possibilities to improve the performance. What is their source? We show that they are due to chattering controls. To this end, we examine the level of performance attainable under constraints on the rate of chattering. More precisely, we consider optimization in a class of sampled controls. The above constraints are given by the lower bound q of the sample period, which is commensurable with the delay: $q/\Delta \to N \in [0, \infty]$ as $\Delta \to +0$.

Optimization over sampled controls is of interest in its own right. For example, sampling of controls is typical for computer controlled systems. The sample period is usually commensurable with transmission delays in digital channels communicating information from controllers to actuators. Then the class of all measurable controls is

not relevant; it can be accepted only if the delays are much greater than the sample period.

We show that the harder the constraint on the rate of chattering (i.e., the larger the N), the worse the limit (as $\Delta \to 0$) optimal performance, and so the smaller the "gap" between the delayed problem and its natural limit model. For $N = \infty$ (the sample period is much greater than the delay), the gap disappears; i.e., the natural limit model becomes proper in the class of sampled controls. Typically, this is the only case where this model is proper. The limit problem that is proper in the class of all measurable controls remains proper for sampled ones if and only if the lower bound of the sample period does not exceed the delay $N \leq 1$. At the same time, neither of these two problems is proper if $1 < N < \infty$. Moreover, no pair $N_1 \neq N_2$ can be served by a common proper limit problem. So the parameter N is essential for construction of a well-posed limit problem. This problem is also explicitly described.

Up to now, not much attention was given to stability (wellposedness) of optimal control problems with respect to perturbation of the delay. The research activity was focused on the case where the delay occurs in the state alone. It was shown in [3, 20] that the trajectory of an uncontrolled delayed equation depends on the delay continuously. Conditions under which such an equation exhibits the same asymptotic behavior as that without delay were offered in [8]. An optimal control problem with time delay in the state was studied in [7]. It was shown that a small perturbation of the delay causes a small disturbance of the optimum. The bounds for certain Dini derivatives of the cost functional optimal value with respect to the delay were also established. By the author's knowledge, the current paper pioneers an analysis of stability with respect to the delay in control.

An introduction to delayed optimal control problems can be found in [16, 17, 22]. The study of the maximum principle for such problems was initiated in [12] and continued in [1, 2, 5, 6, 13, 14, 15, 22]. Relaxation issues were addressed in [18, 19, 22, 23].

The paper is organized as follows. Sections 2 and 3 offer the statement of the optimization problem and a simple example illustrating illposedness of the natural limit problem, respectively. Section 4 shows that this problem is almost always ill-posed. In section 6, we discuss when this problem is yet proper. The well-posed limit problem is presented in section 5. Section 7 discusses optimization in the class of sampled controls. The remainder of the paper is devoted to proofs. They employ certain limits of the set of admissible controls, which are introduced in section 8. Section 9 reveals their general properties. Computation of these limits in section 11 is based upon calculation of extremal points of a polygon in the space of matrices, which is done in section 10. Finally, sections 12, 13, and 14 contain the proofs of the results stated in sections 5, 4, and 6, respectively.

2. The problem statement and assumptions. We consider time delay optimal control problems of the form

$$\text{minimize} \quad \int_T \varphi\left[t, x(t), x(t-\Delta), u(t), u(t-\Delta)\right] \, dt \qquad \text{subject to} \tag{2.1}$$

$$\dot{x}(t) = f\left[t, x(t), x(t-\Delta), u(t), u(t-\Delta)\right] \quad \text{a.a.} \quad t \in T := [t_0, t_1]\,, \tag{2.2}$$

$$x(t) = a(t) \quad \text{for} \quad t \in [t_0 - \Delta, t_0]\,, \tag{2.3}$$

$$u(t) \in \Omega \quad \text{a.a.} \quad t \in [t_0 - \Delta, t_1]\,. \tag{2.4}$$

Here "a.a." means "for almost all." The state $x = x(t) \in \mathbb{R}^n$ and the control $u = u(t) \in \mathbb{R}^m$ are absolutely continuous and measurable functions of time $t \in [t_0 - \Delta, t_1]$, respectively. We interpret the delay $\Delta > 0$ as the parameter of the problem, which is denoted by $\mathfrak{P}(\Delta)$.

DEFINITION 2.1. *The* natural limit problem *is the problem* $\mathfrak{P}(0)$ *obtained by ignoring the delay in* (2.1)–(2.4)

$$\text{minimize} \quad \int_{t_0}^{t_1} \overline{\varphi}\,[t, x(t), u(t)]\,dt \quad \text{subject to} \tag{2.5}$$

$$\dot{x}(t) = \overline{f}\,[t, x(t), u(t)]\,, \quad u(t) \in \Omega \quad \text{a.a.} \quad t \in T, \quad x(t_0) = a_0, \quad \text{where} \tag{2.6}$$

$$\overline{\varphi}(t, x, u) := \varphi(t, x, x, u, u), \quad \overline{f}(t, x, u) := f(t, x, x, u, u), \quad a_0 := a(t_0). \tag{2.7}$$

Can this problem be considered as a proper (well-posed) instantaneous approximation of the delayed one $\mathfrak{P}(\Delta)$ with $\Delta \approx 0$? The answer is in the affirmative only if

$$\lim_{\Delta \to +0} \mathfrak{I}_{opt}(\Delta) = \mathfrak{I}_{opt}(0), \tag{2.8}$$

where $\mathfrak{I}_{opt}(\Delta)$ is the infimum of the cost functional in the problem $\mathfrak{P}(\Delta)$. Our first objective is to check whether relation (2.8) is true or not.

For definiteness, we consider delays $\Delta \in (0, 1)$. Suppose that in (2.1)–(2.4),

(i) the set $\Omega \subset \mathbb{R}^m$ is compact and contains at least two points;

(ii) the functions $f(t, x_1, x_2, u_1, u_2) \in \mathbb{R}^n$ and $\varphi(t, x_1, x_2, u_1, u_2) \in \mathbb{R}$ of the variables $t \in T = [t_0, t_1]$, $x_1, x_2 \in \mathbb{R}^n$, and $u_1, u_2 \in \Omega$ are continuous and continuously differentiable with respect to x_1 and x_2;

(iii) the function $a(\cdot) : [t_0 - 1, t_0] \to \mathbb{R}^n$ is absolutely continuous;

(iv) for any $\Delta \in [0, 1)$ and any measurable control $u(\cdot) : [t_0 - \Delta, t_1] \to \Omega$, the Cauchy problem (2.2), (2.3) has a solution $x(\cdot) : [t_0 - \Delta, t_1] \to \mathbb{R}^n$ such that $|x(t)| \le k < \infty$, where the constant k does not depend on $t \in [t_0 - \Delta, t_1]$, $u(\cdot)$, or $\Delta \in [0, 1)$.

It follows from (i)–(iv) that $\mathfrak{I}_{opt}(\Delta) > -\infty$.

3. Example. We start with a simple example of the situation where the natural limit model is not a proper approximation of the delayed problem. This example is as follows:

$$\text{minimize} \quad I := x_1(1)^2 + x_2(1)^2 \quad \text{subject to} \quad x_1(0) = x_2(0) = 0, \tag{3.1}$$
$$\begin{matrix} \dot{x}_1(t) = \min\{u(t), u(t-\Delta)\} \\ \dot{x}_2(t) = 1 - \max\{u(t), u(t-\Delta)\} \end{matrix}, \ 0 \le t \le 1, \quad 0 \le u(t) \le 1, \ -\Delta \le t \le 1.$$

The natural limit problem results from putting $\Delta := 0$ here is

$$\text{minimize} \quad x_1(1)^2 + x_2(1)^2 \quad \text{subject to} \tag{3.2}$$
$$\dot{x}_1 = u, \dot{x}_2 = 1 - u, \quad 0 \le u = u(t) \le 1, \quad t \in [0, 1], \quad x_1(0) = x_2(0) = 0.$$

In this problem, the trajectories obey the relation $x_1(t) + x_2(t) = t$. So it is easy to see that the states attainable at time $t = 1$ constitute the segment $S = \{(x_1, x_2) = (\theta, 1 - \theta) : 0 \le \theta \le 1\}$. Hence *the minimum of the cost functional in the natural limit problem equals* $1/2$.

At the same time, *the minimum of the cost functional in the original* Δ*-delayed problem is* 0 *irrespective of the delay value* $\Delta > 0$. This minimum is attained at the control

$$u(t) = \begin{cases} 0 & \text{for } t \in [(i-1)\Delta, i\Delta) \cap [0,1], \quad i = 0,2,4,\dots, \\ 1 & \text{for } t \in [i\Delta, (i+1)\Delta) \cap [0,1], \quad i = 0,2,4,\dots. \end{cases} \tag{3.3}$$

Indeed this control is such that for almost all t, one of the numbers $u(t)$, $u(t-\Delta)$ equals 0 whereas the other is 1. Hence $\dot{x}_1(t) = \dot{x}_2(t) = 0 \Rightarrow x_1(\cdot) \equiv x_2(\cdot) \equiv 0$ and so $I = 0$.

Thus in this example, *the natural limit problem is not a proper approximation of the delayed problem with* $\Delta \approx 0$. The "gap" (i.e., the discrepancy between the optimal values of the cost functionals) between these problems amounts to 1/2 and does not depend on Δ.

Formula (3.3) resembles that giving the standard chattering control sequence [18, 19, 22], i.e., a sequence of ordinary controls approximating a randomized one. Such sequences are used in studies of optimal control existence. As is known, there may be no ordinary optimal control. However, for a wide class of problems, the optimum is necessarily attained at some randomized control. Approximation of this control by the above sequence yields a minimizing sequence of ordinary controls, i.e., an asymptotic solution of the original problem.

It should be stressed that in the current context, the situation is different. Formula (3.3) associates not a sequence but a single control with a given Δ-delayed problem. There is no problem with existence of optimal controls; they exist in problems (3.1) and (3.2).

Illposedness of the problem (3.2) is due to the fact that even if $\Delta \approx 0$, the set $\mathcal{X}(\Delta)$ of trajectories $\mathfrak{x} = [x_1(\cdot), x_2(\cdot)]$ generated in the delayed model (3.1) sufficiently differs from that $\mathcal{X}(0)$ related to the natural limit problem (3.2). Indeed, it is easy to see that

$$\mathcal{X}(0) = \Big\{\mathfrak{x} : \dot{x}_1(t) + \dot{x}_2(t) = 1 \quad 0 \le \dot{x}_i(t) \le 1 \quad \text{a.a. } t, \quad x_i(0) = 0,\ i = 1,2\Big\}.$$

At the same time, $\mathcal{X}(\Delta)$ contains the zero trajectory $x_1(\cdot) \equiv x_2(\cdot) \equiv 0$, which does not belong to the closed convex set $\mathcal{X}(0)$. Thus $\mathcal{X}(\Delta) \not\to \mathcal{X}(0)$ (in any reasonable sense) as $\Delta \to 0$. It can be shown that in fact $\mathcal{X}(\Delta) \approx \mathcal{X}_{\lim}$ for $\Delta \approx 0$, where $\mathcal{X}_{\lim}$ results from putting $\dot{x}_1(t) + \dot{x}_2(t) \le 1$ in place of $\dot{x}_1(t) + \dot{x}_2(t) = 1$ in the formulas for $\mathcal{X}(0)$. (We do not prove this fact since it is not used further. It easily follows from the results of sections 8–11.) Thus $\mathcal{X}(\Delta)$ with $\Delta \approx 0$ is sufficiently larger than $\mathcal{X}(0)$.

Unlike (2.1), we considered a terminal cost functional in this section to underscore that the phenomenon of illposedness is not caused by the form of this functional.

4. Illposedness of the natural limit model. Now we revert to the problem (2.1)–(2.4).

LEMMA 4.1. *The limit* $\lim_{\Delta\to+0} \mathfrak{I}_{opt}(\Delta)$ *in* (2.8) *exists and is finite.*

The proofs of the results presented in this section will be given in section 13.

Our first theorem shows that in general the natural limit problem is ill-posed.

THEOREM 4.2. *Given an interval* $T = [t_0, t_1]$ *and a set* Ω *satisfying* (i), *there exists problem* (2.1)–(2.4) *for which the properties* (ii)–(iv) *hold and relation* (2.8) *violates*

$$\lim_{\Delta\to+0} \mathfrak{I}_{opt}(\Delta) < \mathfrak{I}_{opt}(0). \tag{4.1}$$

This inequality means that the natural limit problem conceals certain possibilities of improving the performance.

Now we are going to analyze whether the property (4.1) is typical or not. In doing so, it is natural to compare the variety of all of problem (2.1)–(2.4) with the set of those for which (4.1) holds. To this end, we fix the interval T and the set Ω. We also identify the problem (2.1)–(2.4) with the triplet of functions $\xi := [f(\cdot), \varphi(\cdot), a(\cdot)]$. The linear space Ξ of all such triplets ξ satisfying (ii) and (iii) is equipped with the locally convex topology generated by the following c-parametric ($c > 0$) family of hemi-norms:

(4.2)
$$\mathfrak{H}_c(\xi) := \sum_{r=f,\varphi,\partial f/\partial x_i,\partial\varphi/\partial x_i,i=1,2} \max |r(t,x_1,x_2,u_1,u_2)| + \max_{t\in T^0} |a(t)| + |\dot a(\cdot)|_1,$$

where the first maximum is over $t \in T, |x_i| \le c, u_i \in \Omega$, and $T^0 := [t_0 - 1, t_0], |\dot a(\cdot)|_1 := \int_{T^0} |\dot a(t)|\, dt$. As can be shown, the set $\Xi_{(iv)} \subset \Xi$ of all triplets $\xi \in \Xi$ satisfying (iv) is open.

THEOREM 4.3. *Given an interval $T = [t_0, t_1]$ and a set Ω satisfying* (i)*, consider the set $\Xi_{(4.1)}$ of all triplets $\xi \in \Xi_{(iv)}$ for which* (4.1) *is valid. The set $\Xi_{(4.1)}$ is nonempty and open. If the domain Ω is connected, then the set $\Xi_{(4.1)}$ is dense in $\Xi_{(iv)}$.*

The second claim means that the phenomenon (4.1) is typical.

5. The proper instantaneous (limit) model of the delayed optimal control problem. We recall that this is an instantaneous (undelayed) optimal control problem $\mathcal{P}$ such that first, the minimum of the cost functional in $\mathcal{P}$ equals the limit from (4.1) and second, the solution of this problem provides a complete characterization of *asymptotically optimal sequences* of controls in the original Δ-parametric family of problem (2.1)–(2.4). These sequences $\{u_\Delta(\cdot)\}$ are those for which $u_\Delta(\cdot)$ is "almost optimal" in $\mathfrak{P}(\Delta)$ for $\Delta \approx 0$,

$$I_\Delta\,[u_\Delta(\cdot)] - \mathfrak{I}_{opt}(\Delta) \to 0 \quad \text{as} \quad \Delta \to +0. \tag{5.1}$$

Here $I_\Delta\,[u_\Delta(\cdot)]$ is the value of the cost functional from (2.1). More specifically, the sequence is asymptotically optimal if and only if it converges to the optimal control in the proper limit problem $\mathcal{P}$ if this control is unique, and to the set of all such controls otherwise.

To introduce $\mathcal{P}$, we need some notation. Following [22], we denote by $\mathbf{frm}(\Omega^2)$ the collection of all Radon (finite regular Borel) measures in $\Omega^2 := \Omega \times \Omega$ and put

$$\mathbf{srpm}(\Omega^2) := \Big\{\eta \in \mathbf{frm}\,(\Omega^2) : \eta(du_1, du_2) \ge 0, \quad \eta(\Omega^2) = 1, \\ \text{and the measure } \eta \text{ is symmetric:} \quad \eta\,(du, \Omega) = \eta\,(\Omega, du)\Big\}. \tag{5.2}$$

(The last relation means that $\eta(E \times \Omega) = \eta(\Omega \times E)$ for any Borel set $E \subset \Omega$.) By the Riesz theorem, $\mathbf{frm}(\Omega^2)$ is the topological dual to the space $\mathbb{C}(\Omega^2)$ of continuous functions $g : \Omega^2 \to \mathbb{R}$. We endow $\mathbf{frm}(\Omega^2)$ and, thereby, $\mathbf{srpm}(\Omega^2)$ with the corresponding weak star topology. Denote by $\mathfrak{G}^s$ the set of all measurable maps $\nu : T \to \mathbf{srpm}(\Omega^2)$. They are controls in the *proper limit problem*, which is as follows:

$$\text{minimize} \quad \int_T dt \int_{\Omega^2} \varphi\,[t, x(t), x(t), w]\,\nu(t, dw) \quad \text{subject to} \quad \nu \in \mathfrak{G}^s, \tag{5.3}$$

$$\dot x(t) = \int_{\Omega^2} f\,[t, x(t), x(t), w]\,\nu(t, dw), \qquad x(t_0) = a_0 := a(t_0). \tag{5.4}$$

THEOREM 5.1. *Let the hypotheses* (i)–(iv) *be fulfilled. In the problem* (5.3), (5.4), *the minimum of the cost functional is archived and equals* $\lim_{\Delta\to+0}\mathfrak{I}_{opt}(\Delta)$.

The proofs of the results presented in this section will be given in section 12.

The problem (5.3), (5.4) has several advantages over (2.1)–(2.4). For example, along the lines of the theory of functional differential equations [10], the state space in the latter should be viewed as infinite dimensional in contrast with the former. Though another approach [21] reduces the problem (2.1)–(2.4) to one without delay and with the state space of finite dimension k, this dimension is much greater $k \geq n(t_1-t_0)/\Delta \ggg n$ than that in (5.3), (5.4) if $\Delta \approx 0$. Both interpretations reflect the point of the matter and affect the formulations of basic optimization results such as the maximum or dynamic programming principles [22].

The problem (5.3), (5.4) has deeper relations to the primal one, (2.1)–(2.4), as compared with those indicated in Theorem 5.1. For example, the set $\mathfrak{R}_\Delta$ of admissible controls $u:[t_0-\Delta,t_1]\to\Omega$ in the problem (2.1)–(2.4) converges to the set $\mathfrak{G}^s$ from (5.3) as $\Delta\to 0$. To elucidate this claim, we need to embed both $\mathfrak{G}^s$ and $\mathfrak{R}_\Delta$ into a common enveloping space. To this end, we denote by $\mathbb{L}_1\big[T,\mathbb{C}(K)\big]$ (where K is a compact topological space) the Banach space of the equivalence classes of maps $g:T\times K\to\mathbb{R}$ such that the function $g(\cdot,\varkappa)$ is measurable for all $\varkappa\in K$, the function $g(t,\cdot)$ is continuous for almost all $t\in T$, and

$$\|g\| := \int_T \max_{\varkappa\in K}|g(t,\varkappa)|\,dt < \infty. \tag{5.5}$$

The above equivalence $g_1 \backsim g_2$ holds if and only if $g_1(t,\cdot)\equiv g_2(t,\cdot)$ for almost all $t\in T$. The norm in $\mathbb{L}_1\big[T,\mathbb{C}(K)\big]$ is given by (5.5). We also denote by $\mathbf{frm}(K)$ and $\mathbf{rpm}(K)$ the collections of all the Radon and Radon probability measures in K, respectively, and put

$$\|\nu\| := \sup\int g(\varkappa)\nu(d\varkappa) \qquad \forall\,\nu\in\mathbf{frm}(K),$$

where sup is over $g(\cdot)\in\mathbb{C}(K)$ with $\max_\varkappa|g(\varkappa)|\leq 1$. Denote by $\boldsymbol{\mathcal{N}}(T,K)$ the space of equivalence classes of measurable functions $\nu:T\to\mathbf{frm}(K)$ such that $\operatorname{ess\,sup}_t\|\nu(t)\|<\infty$. This space can be put in duality to $\mathbb{L}_1\big[T,\mathbb{C}(K)\big]=\{g(\cdot)\}$ by setting

$$\langle\nu,g\rangle := \int_T dt\int_K g(t,\varkappa)\nu(t,d\varkappa) \tag{5.6}$$

(see [22]). By the Bishop and Dunford–Pettis theorems, the relativization to $\mathfrak{G}(T,K) := \{\nu(\cdot)\in\boldsymbol{\mathcal{N}}(T,K):\nu(t)\in\mathbf{rpm}(K)\ \text{a.a.}\ t\in T\}$ of the weak topology generated on $\boldsymbol{\mathcal{N}}(T,K)$ by the above duality is induced on $\mathfrak{G}(T,K)$ by a certain norm $\|\cdot\|_w$ in $\boldsymbol{\mathcal{N}}(T,K)$ (weak norm). The space $\mathfrak{G}:=\mathfrak{G}(T,\Omega^2)$ will be of further special interest. We denote by $\operatorname{dist}(A,B)$ the Hausdorff distance between the sets $A,B\subset\mathfrak{G}$, i.e.,

$$\operatorname{dist}(A,B) := \max\left\{\sup_{\nu\in A}\inf_{\mu\in B}\|\nu-\mu\|_w;\sup_{\mu\in B}\inf_{\nu\in A}\|\nu-\mu\|_w\right\}. \tag{5.7}$$

Given $\Delta\in(0,1)$, we embed $\mathfrak{R}_\Delta$ into $\mathfrak{G}$ by identifying $u(\cdot)\in\mathfrak{R}_\Delta$ with the function $\nu:T\to\mathbf{rpm}(\Omega^2)$, where $\nu(t)$ is the Dirac measure at the point $[u(t),u(t-\Delta)]\in\Omega^2$, i.e.,

$$\int r(u',u'')\nu(t,du',du'') := r[u(t),u(t-\Delta)] \tag{5.8}$$

for any $r(\cdot)\in\mathbb{C}(\Omega^2)$. Thus $\mathfrak{R}_\Delta\subset\mathfrak{G}$. The inclusion $\mathfrak{G}^s\subset\mathfrak{G}$ is obvious.

LEMMA 5.2. *The Hausdorff distance between* $\mathcal{G}^s$ *and* $\mathcal{R}_\Delta$ *converges to* 0 *as* $\Delta \to +0$.

It readily follows from this lemma that any $\nu^0(\cdot) \in \mathcal{G}^s$ can be approximated

$$u_\Delta(\cdot) \to \nu^0(\cdot) \quad \text{as} \quad \Delta \to +0 \tag{5.9}$$

by "regular" controls $u_\Delta(\cdot) \in \mathcal{R}_\Delta$. A number of algorithms can be proposed to construct such a sequence $\{u_\Delta(\cdot)\}$ explicitly. Its importance is demonstrated by the following theorem.

THEOREM 5.3. *Let the hypotheses* (i)–(iv) *be fulfilled and* $\nu^0(\cdot)$ *be an optimal control in the problem* (5.3), (5.4). *Consider a sequence* $\{u_\Delta(\cdot)\}_{\Delta\in(0,1)}$ *of "regular" controls* $u_\Delta(\cdot) \in \mathcal{R}_\Delta$, *the solution* $x_\Delta(\cdot)$ *of the Cauchy problem* (2.2), (2.3) *for* $u(\cdot) := u_\Delta(\cdot)$, *and the trajectory* $x^0(\cdot)$ *of the system* (5.4) *related to the control* $\nu^0(\cdot)$.

If (5.9) *is true, the sequence* $\{u_\Delta(\cdot)\}$ *is asymptotically optimal, i.e.,* (5.1) *holds, and*

$$\max_{t\in T} \left|x_\Delta(t) - x^0(t)\right| \to 0 \quad \textit{as} \quad \Delta \to 0.$$

The last theorem shows that a sequence is asymptotically optimal only if (5.9) holds.

THEOREM 5.4. *Let the optimal control* $\nu^0(\cdot)$ *in the problem* (5.3), (5.4) *be unique. A sequence* $\{u_\Delta(\cdot)\}_{\Delta\in(0,1)}$ *of "regular" controls* $u_\Delta(\cdot) \in \mathcal{R}_\Delta$ *is asymptotically optimal if and only if it converges to* $\nu^0(\cdot)$, *i.e.,* (5.9) *holds.*

In general, such a sequence is asymptotically optimal if and only if it converges to the set $\mathfrak{OC}$ *of all optimal controls* $\nu^0(\cdot)$ *in the problem* (5.3), (5.4), *i.e.,*

$$\textit{dist}\,[u_\Delta(\cdot), \mathfrak{OC}] := \inf_{\nu^0(\cdot)\in\mathfrak{OC}} \|u_\Delta(\cdot) - \nu^0(\cdot)\|_w \to 0 \quad \textit{as} \quad \Delta \to +0. \tag{5.10}$$

Theorems 5.1 and 5.4 mean that the problem $\mathcal{P}$ is well posed by perturbation of the delay [4, 24], provided (2.1)–(2.4) is interpreted as the perturbed problem.

6. Problems for which the natural limit model is proper. In this section, we discuss how to recognize problems for which (2.8) is true. This remains of interest despite the fact that by Theorem 4.3, relation (2.8) is likely to be violated because of errors in our knowledge of ξ. However, it can be shown that both quantities in (2.8) depend on ξ continuously. So whenever (2.8) holds for the nominal model, small errors in the knowledge of ξ cannot spoil (2.8) much. Furthermore, the following lemma shows that in the case (2.8), the control optimal for the natural limit problem is suboptimal for the delayed one.

LEMMA 6.1. *Let the hypotheses* (i)–(iv) *be fulfilled and a measurable function* $u(\cdot) : [t_0 - 1, t_1] \to \Omega$ *be given. If the control* $u|_{[t_0,t_1]}$ *is optimal in the natural limit problem* $\mathfrak{P}(0)$, *the control* $u|_{[t_0-\Delta,t_1]}$ *is* $\mu(\Delta)$*-suboptimal in the primal one* (2.1)–(2.4). *More precisely, the corresponding value of the cost functional does not exceed* $\mathfrak{I}_{opt}(\Delta) + \mu(\Delta)$, *where*

$$\mu(\Delta) \le \left|\lim_{\Delta\to+0} \mathfrak{I}_{opt}(\Delta) - \mathfrak{I}_{opt}(0)\right| + \varepsilon(\Delta) \quad \textit{and} \quad \varepsilon(\Delta) \to 0 \qquad \textit{as} \quad \Delta \to +0. \tag{6.1}$$

The proofs of this lemma and Theorem 6.3 (stated below) will be given in section 14.

DEFINITION 6.2. *If* (2.8) *is true, the problem* (2.1)–(2.4) *is said to be* stable to the delay.

Now we show that it is impossible to recognize such a stability and estimate the discrepancy between the left- and right-hand sides of (2.8) if only the natural limit model is known.

THEOREM 6.3. *Consider the problem* (2.5), (2.6)*, where the vector* $a_0 \in \mathbb{R}^n$*, the set* $\Omega \subset \mathbb{R}^m$*, and the functions* $\overline{\varphi}(\cdot), \overline{f}(\cdot)$ *are given and satisfy* (i), (ii), (iv) *(where* $f(t, x_1, x_2, u_1, u_2) := \overline{f}(t, x_1, u_1)$ *and* $\varphi(t, x_1, x_2, u_1, u_2) := \overline{\varphi}(t, x_1, u_1)$*).*

Given $A > 0$*, there exists a problem* (2.1)–(2.4) *satisfying* (i)–(iv) *and such that the original problem* (2.5), (2.6) *is its natural limit case (i.e.,* (2.7) *is true) and*

$$\left| \lim_{\Delta \to +0} \mathfrak{I}_{opt}(\Delta) - \mathfrak{I}_{opt}(0) \right| > A. \tag{6.2}$$

An exhaustive criterion of stability to the delay is immediate from Theorem 5.1.

COROLLARY 6.4. *The problem* (2.1)–(2.4) *is stable to the delay if and only if the infimum value of the cost functional in the problem* (2.5), (2.6) *equals that in* (5.3), (5.4).

This criterion is technical and rather complicated. The next theorem offers particular and simple sufficient conditions for stability to the delay, which readily follow from Corollary 6.4.

THEOREM 6.5. *Let* (i)–(iv) *from section* 2 *be valid. Suppose also that the set* Ω *is convex, the function* $\varphi(t, x_1, x_2, u_1, u_2)$ *is convex in* (u_1, u_2)*, and the function* $f(\cdot)$ *is linear* $f_i(t, x_1, x_2, u_1, u_2) = A_1(t, x_1, x_2)u_1 + A_2(t, x_1, x_2)u_2$ *in controls. Then the problem* (2.1)–(2.4) *is stable to the delay.*

Proof. Let $[x(\cdot), \nu]$ be a process in the problem (5.3), (5.4). Thanks to the symmetry condition from (5.2),

$$u(t) := \int_\Omega u\, \nu(t, du, \Omega) = \int_\Omega u\, \nu(t, \Omega, du).$$

Since the set Ω is convex and compact, $u(t) \in \Omega$. Furthermore, (5.4) implies that

$$\begin{aligned} \dot{x}(t) &= \int_{\Omega^2} \Big\{ A_1[t, x(t), x(t)]u' + A_2[t, x(t), x(t)]u'' \Big\}\, \nu(t, du', du'') \\ &= \sum_{i=1}^{2} A_i[t, x(t), x(t)]u(t) = f[t, x(t), x(t), u(t), u(t)] \overset{(2.7)}{=\!=} \overline{f}[t, x(t), u(t)], \end{aligned}$$

and $x(0) = a(0)$. Thus (2.6) holds, and $[x(\cdot), u(\cdot)]$ is a process in the problem (2.5), (2.6). So far as the function $\varphi(\cdot)$ is convex in controls, the following relations hold (see, e.g., [9, p. 402]):

$$\begin{aligned} \int_T dt \int_{\Omega^2} &\varphi[t, x(t), x(t), u', u'']\, \nu(t, du', du'') \\ &\geq \int_T \varphi \left\{ t, x(t), x(t), \int_{\Omega^2} [u', u'']\ \nu(t, du', du'') \right\} dt \\ &= \int_T \varphi[t, x(t), x(t), u(t), u(t)]\, dt \overset{(2.7)}{=\!=} \int_T \overline{\varphi}[t, x(t), u(t)]\, dt. \end{aligned}$$

Summarizing, we see that the infimum of the cost functional in the problem (5.3), (5.4) is not less than that in the problem (2.5), (2.6). At the same time, putting the elements ν of the form (5.8) with $\Delta := 0$ in (5.3), (5.4) assures that the first infimum does not exceed the second one. Thus these infima coincide. Corollary 6.4 completes the proof. □

7. Optimization in the class of sampled controls. By (4.1), the optimal value of the cost functional in the Δ-delayed problem with $\Delta \approx 0$ is typically less than that in the natural limit problem. In this section, we show that this is due to chattering controls. We also estimate the rate of chattering that is required for the above phenomenon to take place. To this end, we focus attention on sampled controls modelled as piecewise constant functions. Then the available rate of chattering is given by the lower bound on the sample period.

Now we consider the following class of admissible controls:

(7.1)
$$u(\cdot) \in \mathfrak{R}^{\rm sam}_\Delta := \bigcup_{q \geq q(\Delta)} \mathfrak{R}^q_\Delta\,, \text{ where } \mathfrak{R}^q_\Delta := \Big\{u(\cdot) : u(t) \in \Omega \text{ a.a. } t \in T_\Delta := [t_0 - \Delta, t_1]$$
$$\text{and the control } u = u(t) \in \mathbb{R}^m \text{ is constant on any time interval}$$
$$\text{of the form } \big[t_0 - \Delta + iq, t_0 - \Delta + (i+1)q\big) \cap T_\Delta, i = 0, 1, 2, \dots\Big\}.$$

Here the lower bound $q(\Delta) > 0$ on the sample period q is a given function of the delay. In this section, we study the optimization problem $\mathfrak{P}^{\rm sam}(\Delta)$ described by (2.1)–(2.4) and (7.1). Note that in this problem, the sample period is optimized above the bound $q(\Delta)$. Many of the results to follow remain true if this period is given a priori, i.e., $\mathfrak{R}^{\rm sam}_\Delta := \mathfrak{R}^{q(\Delta)}_\Delta$.

The goal is to find a proper limit model $\mathcal{P}^{\rm sam}$ of $\mathfrak{P}^{\rm sam}(\Delta)$ as $\Delta \to 0$. In this model, the set of admissible controls $\mathcal{G}^s_N$ depends on the parameter $N = \lim q/\Delta \in [0, +\infty]$ and is the set of all measurable maps $\nu : T \to \mathbf{srpm}_N(\Omega^2)$. The set $\mathbf{srpm}_N(\Omega^2)$ of controls admissible at a given time is obtained by restricting the similar set (5.2) serving the problem (5.3), (5.4),

$$\mathbf{srpm}_N(\Omega^2) := \Big\{\eta \in \mathbf{srpm}(\Omega^2) : (N-1)\,[\eta(du, \Omega) - \eta_d(du)] \leq \eta_d(du)\Big\}. \tag{7.2}$$

Here $\eta_d(du)$ is the component of η concentrated on the diagonal $D := \{(u, u) : u \in \Omega\}$,

$$\eta_d(E) := \eta(E_d), \quad \text{where} \quad E_d := \{w = (u, u) : u \in E\}. \tag{7.3}$$

Remark 7.1. In (7.2), $\eta_-(du) := \eta(du, \Omega) - \eta_d(du) \geq 0$.

Indeed in (7.3), $E_d \subset E \times \Omega$ and so $\eta_-(E) = \eta(E \times \Omega) - \eta(E_d) \geq 0$ for any Borel set E.

Hence for $N \in [0, 1]$, the inequality from (7.2) is true and $\mathcal{G}^s_N$ equals the set $\mathcal{G}^s$ of controls in the limit problem (5.3), (5.4). For $N \in (1, \infty)$, the inequality means that η is concentrated mainly on the diagonal D. The set $\mathcal{G}^s_\infty$ consists of all probability measures η concentrated on the diagonal since $\infty - 1 := \infty$ and $\infty \cdot b := \infty\ \forall b > 0$. It is clear that

$$\mathcal{G}^s_1 \supset \mathcal{G}^s_{N_1} \supset \mathcal{G}^s_{N_2} \supset \bigcap_{N>1} \mathcal{G}^s_N = \mathcal{G}^s_\infty \quad \text{if} \quad 1 < N_1 < N_2 < \infty. \tag{7.4}$$

Now we are in a position to state the main result of this section.

THEOREM 7.2. *Let the hypotheses* (i)–(iv) *from section* 2 *be fulfilled. Suppose also that*

$$\textit{either} \quad \frac{q(\Delta)}{\Delta} \to N \in [1, \infty] \textit{ as } \Delta \to +0 \qquad \textit{or} \quad \varlimsup_{\Delta \to +0} \frac{q(\Delta)}{\Delta} \leq 1\ (=: N), \tag{7.5}$$

and $q(\Delta) \to 0$ *as* $\Delta \to +0$ *if* $N = \infty$. *Then the problem* $\mathcal{P}^{sam}$ *that results from* (5.3), (5.4) *by putting* $\mathcal{G}^s_N$ *in place of* $\mathcal{G}^s$ *in* (5.3) *is the proper limit model for the* Δ*-delayed problem* $\mathfrak{P}^{sam}(\Delta)$ *given by* (2.1)–(2.4) *and* (7.1). *In other words, the following statements hold:*

(i) *In the problem* $\mathcal{P}^{sam}$, *the minimum value of the cost functional is archived and equals* $\lim_{\Delta\to+0} \mathfrak{I}^{sam}_{opt}(\Delta)$, *where* $\mathfrak{I}^{sam}_{opt}(\Delta)$ *is the infimum of the cost functional in* $\mathfrak{P}^{sam}(\Delta)$.

(ii) *A sequence* $\{u_\Delta(\cdot)\}_{\Delta\in(0,1)}, u_\Delta(\cdot) \in \mathfrak{R}^{sam}_\Delta$ *is asymptotically optimal if and only if it approaches the set* $\mathfrak{OC}$ *of all optimal controls in the problem* $\mathcal{P}^{sam}$, *i.e.,* (5.10) *holds.*

(iii) *The set* $\mathfrak{R}^{sam}_\Delta$ *of "regular" controls defined by* (7.1) *approximates the set of "generalized" ones* $\mathcal{G}^s_N$: *the distance* (5.7) *between them converges to zero as* $\Delta \to +0$.

The proof of this theorem will be given in section 12.

Remark 7.3. The problem (5.3), (5.4) with $\mathcal{G}^s := \mathcal{G}^s_\infty$ is in fact the relaxation [22] of the natural limit problem (2.5), (2.6) that results from merely ignoring the delay in (2.1)–(2.4).

Indeed this relaxation is as follows [22]:

$$\text{(7.6)}\quad \text{minimize} \quad \int_T dt \int_\Omega \overline{\varphi}\,[t, x(t), u]\,\mu(t, du) \quad \text{subject to} \quad \mu \in \mathcal{G}(T, \Omega),$$
$$\dot{x}(t) = \int_\Omega \overline{f}\,[t, x(t), u]\,\mu(t, du), \quad x(t_0) = a_0 := a(0).$$

Here $\overline{\varphi}(\cdot)$ and $\overline{f}(\cdot)$ are given by (2.7), and $\mathcal{G}(T, \Omega)$ is the set of all measurable maps from T to the space $\mathbf{rpm}(\Omega)$ of probability measures in Ω. The isomorphism $u \in \Omega \leftrightarrow (u, u) \in D$ establishes a one-to-one correspondence between $\mu \in \mathbf{rpm}(\Omega)$ and $\zeta \in \mathbf{rpm}(D)$. At the same time, the elements ζ can be identified with probability measures η in Ω^2 concentrated on the diagonal. Thus the elements ζ constitute $\mathbf{srpm}_\infty(\Omega^2)$, and there is a one-to-one correspondence $\mu \in \mathcal{G}(T, \Omega) \leftrightarrow \nu \in \mathcal{G}^s_\infty$. It is straightforward to check that whenever $\mu \leftrightarrow \nu$,

$$\int_\Omega \overline{r}\,[t, x(t), u]\,\mu(t, du) = \int_{\Omega^2} r\,[t, x(t), x(t), u_1, u_2]\,\nu(t, du_1, du_2) \quad \text{for} \quad r := f, \varphi.$$

This implies that the relaxation (7.6) of (2.5), (2.6) does reduce to (5.3), (5.4) with $\mathcal{G}^s := \mathcal{G}^s_\infty$.

Hence relations (7.4) mean that for $1 < N < \infty$, the proper limit problem (5.3), (5.4) (where $\mathcal{G}^s := \mathcal{G}^s_N$) in the class of sampled controls is intermediate between the natural limit one (2.5), (2.6) and the proper limit problem (5.3), (5.4) in the class of all measurable controls.

Now we are going to show that the natural limit problem is typically ill-posed in the class of sampled controls unless the sample period greatly exceeds the delay. We also demonstrate that the parameter $N = \lim q/\Delta$ is essential for construction of a proper limit problem $\mathcal{P}^{\mathrm{sam}}$. Strictly speaking, this does not follow from the fact that $\mathcal{P}^{\mathrm{sam}}$ formally depends on N. This will follow from the fact that the limit optimal cost $\lim_{\Delta\to+0} \mathfrak{I}^{\mathrm{sam}}_{\mathrm{opt}}(\Delta)$ strictly increases as N grows. So different N cannot be served by a common limit problem.

We recall that problem (2.1)–(2.4) is identified with the triplet $\xi := [f(\cdot), \varphi(\cdot), a(\cdot)]$; the space Ξ of all such triplets ξ satisfying (ii) and (iii) from section 2 is equipped with the topology generated by the hemi-norms (4.2), and $\Xi_{(iv)}$ denotes the set of all

$\xi \in \Xi$ satisfying (iv). The space Ξ is complete and metricizible, and the set $\Xi_{(iv)}$ is open. We recall that a subset $\Upsilon \subset \Xi_{(iv)}$ is said to be *residual* if it is representative as the intersection of a countable collection of open dense (in $\Xi_{(iv)}$) sets. For $\xi \in \Xi_{(iv)}$, (i) of Theorem 5.1 ensures that

$$\frac{q(\Delta)}{\Delta} \to N \in [1,+\infty] \text{ as } \Delta \to +0 \Big| \Rightarrow \exists \lim_{\Delta\to+0} \mathfrak{I}^{\rm sam}_{\rm opt}(\Delta) =: \boldsymbol{\mathfrak{J}}^{\rm sam}_{\rm opt}[N,\xi]. \tag{7.7}$$

THEOREM 7.4. *Let* (i) *from section* 2 *be true and the set* Ω *be connected. Then there exists a residual subset* $\Upsilon \subset \Xi_{(iv)}$ *consisting of delayed problem* (2.1)–(2.4) *for which*

(i) *the minimum of the cost functional* $\mathfrak{I}^{sam}_{opt}(\Delta)$ *obeys the bound*

$$\overline{\lim_{\Delta\to+0}}\, \mathfrak{I}^{sam}_{opt}(\Delta) < \mathfrak{I}_{opt}(0) \quad \textit{whenever} \quad \overline{\lim_{\Delta\to+0}}\, q(\Delta)/\Delta < \infty; \tag{7.8}$$

(ii) *the limit optimal cost* $\boldsymbol{\mathfrak{J}}^{sam}_{opt}[N,\xi]$ *defined in* (7.7) *is a strictly increasing function of the parameter* $N = \lim_{\Delta\to+0} q(\Delta)/\Delta \in [1,\infty]$.

Moreover, given $1 < N_- < N_+ < \infty$, *there exists an open and dense subset* $\Xi_-(N_-,N_+) \subset \Xi_{(iv)}$ *such that for* $\xi \in \Xi_-(N_-,N_+)$, *the limit optimal cost* $\boldsymbol{\mathfrak{J}}^{sam}_{opt}[N,\xi]$ *strictly increases on the interval* $[N_-,N_+]$. *If the set* Ω *is not connected, there exists an open, nonempty, but not necessarily dense such a subset* $\Xi_-(N_-,N_+)$.

The proof of this theorem will be given in section 13.

Remark 7.5. In general, the function $\boldsymbol{\mathfrak{J}}^{\rm sam}_{\rm opt}[N,\xi]$ of N does not decrease.

This assertion follows from (7.7), (i) of Theorem 5.1, and (7.4).

By the first claim of the theorem, it is typical that the natural limit problem is ill-posed in the class of sampled controls, provided the set Ω is connected and $N < \infty$.

8. Limits of the set of admissible controls. In the remainder of the paper, we prove the results stated above. The key to the proofs is studying some limits of Δ-parametric sets $\mathfrak{R}_\Delta$ of measurable controls $u(\cdot) : [t_0-\Delta, t_1] \to \Omega$ as $\Delta \to 0$. (Here and throughout, $\Delta \in (0,1)$.)

In this section, we introduce these limits and reveal their role. Furthermore, we employ the notation introduced before Lemma 5.2 and embed $\mathfrak{R}_\Delta$ into $\mathcal{G}$ by identifying $u(\cdot) \in \mathfrak{R}_\Delta$ with the function $\nu : T \to \mathbf{rpm}(\Omega^2)$ given by (5.8). We also introduce the following "limits":

$$\begin{aligned}\mathrm{LIM}_{\Delta\to0}\mathfrak{R}_\Delta := \big\{\nu \in \mathcal{G} :\ &\text{there exists a sequence } \{u_\Delta(\cdot)\} \text{ of "regular" controls}\\ &u_\Delta(\cdot) \in \mathfrak{R}_\Delta\ \forall\Delta \text{ such that } u_\Delta(\cdot) \to \nu \text{ as } \Delta \to 0\big\},\end{aligned} \tag{8.1}$$

$$\begin{aligned}\mathrm{LIM}^{\rm seq}_{\Delta\to0}\mathfrak{R}_\Delta := \big\{\nu \in \mathcal{G} :\ &\text{there exist sequences } \{\Delta_i\}_{i=1}^\infty \subset (0,1) \text{ and } \{u_i(\cdot)\}_{i=1}^\infty\\ &\text{such that}\quad u_i(\cdot) \in \boldsymbol{\mathfrak{R}}_{\Delta_i}\ \forall i \text{ and } \Delta_i \to 0, u_i(\cdot) \to \nu \text{ as } i \to \infty\big\}.\end{aligned} \tag{8.2}$$

(The space $\mathcal{G}$ is endowed with the week topology generated by the duality (5.6).)

DEFINITION 8.1. Modified *problems* (2.1)–(2.4) *and* (5.3), (5.4) *are defined to be those resulting from the original problems* (2.1)–(2.4) *and* (5.3), (5.4) *by substitution of the inclusion* $u(\cdot) \in \mathfrak{R}_\Delta$ *in place of* (2.4) *and the set* $\mathrm{LIM}_{\Delta\to0}\mathfrak{R}_\Delta$ *in place of* $\mathcal{G}^s$ *in* (5.3), *respectively.*

In this section, we show that under natural assumptions, the proper limit (as $\Delta \to 0$) of the first modified problem is the second one. To this end, we start with two technical facts.

LEMMA 8.2. *Both sets* (8.1) *and* (8.2) *are closed in* $\mathfrak{G}$.

Proof. Let $\mathfrak{S}$ denote the set (8.1) and $\nu \in \overline{\mathfrak{S}}$. Then $\|\nu - \nu_j\|_w < j^{-1}$ for some $\nu_j \in \mathfrak{S}$ for $j = 1, 2, \ldots$. By (8.1), there exist sequences $\{u_\Delta^{(j)}\}$ of controls $u_\Delta^{(j)} \in \mathfrak{R}_\Delta$ such that $u_\Delta^{(j)} \to \nu_j$ as $\Delta \to +0$ for $j = 1, 2, \ldots$. Hence $\|\nu_j - u_\Delta^{(j)}\|_w < j^{-1}$ whenever $\Delta < \Delta_j$ and $\Delta_j > 0$ is small enough. Without loss of generality, we can assume that $\Delta_{j+1} < \Delta_j \ \forall j$ and $\Delta_j \to 0$ as $j \to \infty$. Putting $u_\Delta := u_\Delta^{(j)} \in \mathfrak{R}_\Delta$ for $\Delta \in (\Delta_{j+1}, \Delta_j]$, we get

$$\|\nu - u_\Delta\|_w \le \|\nu - \nu_j\|_w + \|\nu_j - u_\Delta\|_w \le 2j^{-1} \to 0 \quad \text{as} \quad \Delta \to +0.$$

Therefore, $\nu \in \mathfrak{S}$ and $\overline{\mathfrak{S}} = \mathfrak{S}$. The closedness of the set (8.2) is established likewise. □

Remark 8.3. Since $\mathfrak{G}$ is compact, so are the sets (8.1) and (8.2) by Lemma 8.2.

LEMMA 8.4. *For any* $\nu \in \mathfrak{S} := \mathrm{LIM}_{\Delta \to 0}^{seq} \mathfrak{R}_\Delta$, *the Cauchy problem* (5.4) *has a unique solution* $x(\cdot|\nu) : T \to \mathbb{R}^n$. *This solution obeys the estimation* $|x(\cdot|\nu)| \le k \ \forall t \in T$, *where* k *is taken from* (iv) *in section* 2. *Let* $\{\Delta_i\}_{i=1}^\infty \subset (0,1)$ *and* $\{u_i\}_{i=1}^\infty$ *be two sequences such that* $u_i \in \mathfrak{R}_{\Delta_i} \ \forall i$ *and* $\Delta_i \to 0, u_i \to \nu$ *as* $i \to \infty$. *Then,*

$$\max_{t \in T} |x_i(t) - x(t|\nu)| \to 0 \quad \textit{and} \quad I_{\Delta_i}(u_i) \to I(\nu) \quad \textit{as} \quad i \to \infty. \tag{8.3}$$

Here $x_i(\cdot)$ *is the solution of the Cauchy problem* (2.2), (2.3) *with* $\Delta := \Delta_i$, $u := u_i$, *whereas* $I_\Delta(u)$ *and* $I(\nu)$ *are the values of the cost functionals generated by* $u \in \mathfrak{R}_\Delta$ *and* $\nu \in \mathfrak{S}$ *in the problems* (2.1)–(2.4) *and* (5.3), (5.4), *respectively. The map* $\nu \in \mathfrak{S} \mapsto I(\nu)$ *is continuous.*

Proof. By (8.2), any $\nu \in \mathfrak{S}$ is related to sequences $\{\Delta_i\}$ and $\{u_i\}$ with the properties given by the third sentence of the lemma. The Arzela theorem and (iv) from section 2 imply that $x_i(\cdot)$ is defined on $[t_0 - \Delta_i, t_1]$ and $|x_i(t)| \le k \ \forall t \in T, i = 1, 2, \ldots$, and also that the sequence $\{x_i|_T\}$ is precompact in $\mathbb{C}(T, \mathbb{R}^n)$. It is easy to check (see, e.g., [22, Theorem VII.1.2]) that each of its limit points satisfies (5.4). Thus relations (5.4) do have a solution such that $|x(t|\nu)| \le k \ \forall t$. Its uniqueness results from (ii) in section 2. In particular, all limit points of the precompact sequence $\{x_i(\cdot)\}$ equal $x(\cdot|\nu)$, which implies the first relation from (8.3). We recall that the control $u_i(\cdot) \in \mathfrak{R}_{\Delta_i} \subset \boldsymbol{\mathfrak{R}}_{\Delta_i}$ is identified with the element $\nu_i(t, du_1, du_2) = \delta_{[u_i(t), u_i(t-\Delta_i)]}(du_1, du_2)$ of $\mathfrak{G}$, where δ_ω is the Dirac measure at the point ω. Denote $\widetilde{\Delta}_i := \Delta_i$. By (2.1) and (5.3), the quantity $I_{\Delta_i}(u_i) - I(\nu)$ amounts to

$$\begin{aligned} \int_T dt \int_{\Omega^2} \Big\{ \varphi \left[t, x_i(t), x_i(t - \widetilde{\Delta}_i), w \right] - \varphi \left[t, x(t|\nu), x(t|\nu), w \right] \Big\} \nu_i(t, dw) \\ + \int_T dt \int_{\Omega^2} \underbrace{\varphi \left[t, x(t|\nu), x(t|\nu), w \right]}_{\eta(t,w)} \Big[\nu_i(t, dw) - \nu(t, d\omega) \Big]. \end{aligned} \tag{8.4}$$

The integrand in the first integral converges to 0 uniformly over $t \in T, w \in \Omega^2$ by the first relation from (8.3). By (5.6), the second summand equals $\langle \nu_i - \nu, \eta \rangle$. Hence (8.3) does hold.

Now let $\{\nu_i\} \subset \mathfrak{S}$ and $\nu_i \to \nu \in \mathfrak{S}$ as $i \to \infty$. Denote $x_i(\cdot) := x(\cdot|\nu_i)$. Retracing the above arguments ensures consecutively that the sequence $\{x_i(\cdot)\}$ is precompact in $\mathbb{C}(T, \mathbb{R}^n)$, each of its limit points satisfies (5.4) and so equals $x(\cdot|\nu)$, and hence $\max_{t \in T} |x_i(t) - x(t|\nu)| \to 0$ as $i \to \infty$. Then the last claim of the lemma is proved like the second relation from (8.3) since the quantity $I(\nu_i) - I(\nu)$ has the form (8.4) with $\widetilde{\Delta}_i := 0$. □

In fact, the following three theorems mean that one of the modified problems introduced by Definition 8.1 is the proper limit as $\Delta \to +0$ of the other.

THEOREM 8.5. *Suppose that* $\mathfrak{S} := \mathrm{LIM}^{seq}_{\Delta\to 0}\mathfrak{R}_\Delta = \mathrm{LIM}_{\Delta\to 0}\mathfrak{R}_\Delta$ *for some* Δ*-parametric set* $\mathfrak{R}_\Delta$ *of controls, where the limits are defined by* (8.1) *and* (8.2). *Denote by* $\mathfrak{I}^m_{opt}(\Delta)$ *the infimum of the cost functional in the modified problem* (2.1)–(2.4). *The minimum of the cost functional in the modified problem* (5.3), (5.4) *is attained and equals* $\lim_{\Delta\to+0}\mathfrak{I}^m_{opt}(\Delta)$.

Proof. By Remark 8.3, the set $\mathfrak{S}$ is compact. So thanks to the last claim from Lemma 8.4, $\inf_{\nu\in\mathfrak{S}} I(\nu)$ is attained at a control $\nu^0 \in \mathfrak{S}$. Due to (8.1), there exists a sequence $\{u_\Delta\}$ such that $u_\Delta \in \mathfrak{R}_\Delta$ and $u_\Delta \to \nu^0$ as $\Delta \to +0$. By (8.3), $\mathfrak{I}^m_{opt}(\Delta) \le I_\Delta(u_\Delta) \to I(\nu^0)$ and so

$$\overline{\lim_{\Delta\to+0}}\,\mathfrak{I}^m_{opt}(\Delta) \le I(\nu^0).$$

On the other hand, there apparently exist sequences $\{\Delta_i\}$ and $\{v_i\}$ such that $v_i \in \mathfrak{R}_{\Delta_i}\ \forall i$ and $\Delta_i \to 0, I_{\Delta_i}(v_i) \to \underline{\lim}_{\Delta\to+0}\mathfrak{I}^m_{opt}(\Delta)$ as $i \to \infty$. So far as the space $\mathfrak{G} \supset \{v_i\}$ is compact, passing to a subsequence ensures that $v_i \to \nu$ as $i \to \infty$, where $\nu \in \mathrm{LIM}^{\mathrm{seq}}_{\Delta\to+0}\mathfrak{R}_\Delta = \mathfrak{S}$ in correspondence with (8.2). By invoking (8.3), we get

$$\overline{\lim_{\Delta\to+0}}\,\mathfrak{I}^m_{opt}(\Delta) = \lim_{i\to\infty} I_{\Delta_i}(v_i) = I(\nu) \ge I(\nu^0).$$

Thus $\mathfrak{I}^m_{opt}(\Delta) \to I(\nu^0)$ as $\Delta \to +0$, which completes the proof. □

THEOREM 8.6. *Suppose that the assumption of Theorem* 8.5 *is fulfilled. The Hausdorff distance* (5.7) *between the sets* $\mathfrak{S} := \mathrm{LIM}_{\Delta\to+0}\mathfrak{R}_\Delta$ *and* $\mathfrak{R}_\Delta$ *vanishes as* $\Delta \to +0$.

Proof. Suppose the contrary. Then due to (5.7), there exist $\delta > 0$ and a sequence $\{\Delta_i\} \subset (0,1)$ such that $\Delta_i \to 0$ as $i \to \infty$ and either

(i) $\inf_{u\in\mathfrak{R}_{\Delta_i}}\|\nu_i - u\|_w \ge \delta\ \forall i$ for some sequence $\{\nu_i\} \subset \mathfrak{S}$ or

(ii) $\inf_{\mu\in\mathfrak{S}}\|\mu - u_i\|_w \ge \delta\ \forall i$ for some sequence $\{u_i\}, u_i \in \mathfrak{R}_{\Delta_i}\ \forall i$.

By Remark 8.3, both sets $\mathfrak{G}$ and $\mathfrak{S}$ are compact. So passing to subsequences ensures that $\nu_i \to \nu \in \mathfrak{S}, u_i \to \xi \in \mathfrak{G}$ as $i \to \infty$. Due to (8.2), $\xi \in \mathfrak{S}$, which implies that (ii) cannot be true and (i) thereby holds. By (8.1), $\nu = \lim_{i\to\infty} w_i$ for some $w_i \in \mathfrak{R}_{\Delta_i}, i = 1, 2, \ldots$. So

$$\inf_{u\in\mathfrak{R}_{\Delta_i}}\|\nu_i - u\|_w \le \|\nu_i - w_i\|_w \le \|\nu_i - \nu\|_w + \|\nu - w_i\|_w \to 0 \quad \text{as} \quad i \to \infty$$

in violation of (i). The contradiction obtained proves the theorem. □

THEOREM 8.7. *Let the assumption of Theorem* 8.5 *be true. Employ the notation* $\mathfrak{I}^m_{opt}(\Delta)$ *from it. A sequence* $\{u_\Delta(\cdot)\}$ *of controls* $u_\Delta(\cdot) \in \mathfrak{R}_\Delta$ *is asymptotically optimal in the modified problem* (2.1)–(2.4), *i.e.,* $I_\Delta[u_\Delta] - \mathfrak{I}^m_{opt}(\Delta) \to 0$ *as* $\Delta \to +0$, *if and only if it converges to the set* $\mathfrak{OC}$ *of all optimal controls in the modified problem* (5.3), (5.4), *i.e.,* (5.10) *holds.*

Proof. Let (5.10) be true. Suppose that the sequence $\{u_\Delta\}$ is not asymptotically optimal. With regard to Theorem 8.5, we see that there exist $\delta > 0$ and a sequence $\{\Delta_i\}$ for which $I_{\Delta_i}(u_{\Delta_i}) \ge \delta + \min_{\mu\in\mathfrak{S}} I(\mu)\ \forall i$ and $\Delta_i \to 0$ as $i \to \infty$. Since the space $\mathfrak{G} \supset \{u_{\Delta_i}\}$ is compact, passing to a subsequence ensures that $u_{\Delta_i} \to \nu$ as $i \to \infty$, where $\nu \in \mathfrak{S}$ by (8.2). In view of (5.10), $\mathrm{dist}[\nu, \mathfrak{OC}] \le \|\nu - u_{\Delta_i}\|_w + \mathrm{dist}[u_{\Delta_i}, \mathfrak{OC}] \to 0$ as $i \to \infty$, where the set $\mathfrak{OC}$ is closed due to the last claim from Lemma 8.4. Hence $\nu \in \mathfrak{OC}$. Due to (8.3), we have

$$\delta + \min_{\mu\in\mathfrak{S}} I(\mu) \le I_{\Delta_i}(u_{\Delta_i}) \to I(\nu) = \min_{\mu\in\mathfrak{S}} I(\mu) \quad \text{as} \quad i \to \infty.$$

The contradiction obtained proves that the sequence $\{u_\Delta\}$ is asymptotically optimal.

Conversely, consider such a sequence $\{u_\Delta\}$. Suppose that (5.10) is violated. Then there exist $\delta > 0$ and a sequence $\{\Delta_i\}$ such that $\operatorname{dist}[u_{\Delta_i}, \mathfrak{OC}] \geq \delta\ \forall i$ and $\Delta_i \to 0$ as $i \to \infty$. As above, it can be assumed that $u_{\Delta_i} \to \nu \in \mathfrak{S}$ as $i \to \infty$. By (8.3), $\lim_{\Delta\to+0} \mathfrak{I}^m_{opt}(\Delta) = \lim_{i\to\infty} I_{\Delta_i}(u_{\Delta_i}) = I(\nu)$. So Theorem 8.5 yields $\nu \in \mathfrak{OC}$ in violation of the above relation $\operatorname{dist}[u_{\Delta_i}, \mathfrak{OC}] \geq \delta\ \forall i$. Thus (5.10) does hold. ∎

9. General properties of the limit (8.1). The previous section shows that proofs of the results from sections 5 and 7 are reduced to calculating the limits (8.1) and (8.2) for $\mathfrak{R}_\Delta := \mathfrak{R}_\Delta, \mathfrak{R}^{\mathrm{sam}}_\Delta$. In doing so, some general properties of the limit (8.1) are helpful. In this section, we reveal them. They follow from the single property introduced by the following definition.

DEFINITION 9.1. *A subset $\mathfrak{S} \subset \mathfrak{G}$ is said to be* permutable *if for any $\nu_1, \nu_2 \in \mathfrak{S}$ and $\tau \in (t_0, t_1)$, it contains the element given by the following formula (where $E := [t_0, \tau)$):*

$$\nu_E(t, du_1, du_2) := \begin{cases} \nu_1(t, du_1, du_2) & \text{whenever } t \in E, \\ \nu_2(t, du_1, du_2) & \text{otherwise.} \end{cases} \tag{9.1}$$

The next lemma shows that this property holds in the cases of our primal interest.

LEMMA 9.2. *If either $\mathfrak{R}_\Delta := \mathfrak{R}_\Delta \forall \Delta$ or $\mathfrak{R}_\Delta := \mathfrak{R}^{sam}_\Delta\ \forall \Delta$, the set* (8.1) *is permutable.*

Proof. Let $\mathfrak{R}_\Delta := \mathfrak{R}^{\mathrm{sam}}_\Delta\ \forall \Delta$, $\nu_1, \nu_2 \in \mathfrak{S} := \mathrm{LIM}_{\Delta\to 0}\mathfrak{R}_\Delta$, and $\tau \in (t_0, t_1)$. For $i = 1, 2$, consider a sequence $\{u^{(i)}_\Delta\}$ associated with ν_i by (8.1). Let $\mathfrak{D}^{(i)}_\Delta$ denote the set of the points where $u^{(i)}_\Delta(\cdot)$ is not continuous, and $D^{(i)}_\Delta := \mathfrak{D}^{(i)}_\Delta \cup \{t_0 - \Delta, t_1\}$. We put $\tau^{(1)}_\Delta := \max\{t = s + jq(\Delta) : t < \tau, s \in D^{(1)}_\Delta, j = 0, 1, \dots\} \geq \tau - q(\Delta), \tau^{(2)}_\Delta := \min\{t = s - jq(\Delta) : t > \tau, s \in D^{(2)}_\Delta, j = 0, 1, \dots\} \leq \tau + q(\Delta)$, and $v_\Delta := u^{(2)}_\Delta(\tau^{(2)}_\Delta + 0)$. Setting $u_\Delta(t) := u^{(1)}_\Delta(t)$ for $t < \tau^{(1)}_\Delta$, $u_\Delta(t) := u^{(2)}_\Delta(t)$ for $t \geq \tau^{(2)}_\Delta$, and $u_\Delta(t) := v_\Delta$ for $t \in [\tau^{(1)}_\Delta, \tau^{(2)}_\Delta)$ gives a control $u_\Delta(\cdot) \in \mathfrak{R}^{\mathrm{sam}}_\Delta$. For $\eta \in \mathbb{L}_1\big[T, \mathbb{C}(\Omega^2)\big]$, we put $\eta_1(t, w) := \eta(t, w), \eta_2(t, w) := 0$ if $t < \tau$ and $\eta_2(t, w) := \eta(t, w), \eta_1(t, w) := 0$ if $t \geq \tau$. Then (5.6) and (5.8) yield

$$\begin{aligned}
\langle \eta; u_\Delta \rangle &= \int_T \eta\left[t, u_\Delta(t), u_\Delta(t-\Delta)\right] dt = \int_{\tau+\Delta+q(\Delta)}^{t_1} \eta\left[t, u^{(2)}_\Delta(t), u^{(2)}_\Delta(t-\Delta)\right] dt \\
&+ \underbrace{\int_{\tau-q(\Delta)}^{\tau+\Delta+q(\Delta)} \eta\left[t, u_\Delta(t), u_\Delta(t-\Delta)\right] dt}_{\alpha^{(1)}(\Delta)} + \int_{t_0}^{\tau-q(\Delta)} \eta\left[t, u^{(1)}_\Delta(t), u^{(1)}_\Delta(t-\Delta)\right] dt \\
&= \int_\tau^{t_1} \eta\left[t, u^{(2)}_\Delta(t), u^{(2)}_\Delta(t-\Delta)\right] dt - \underbrace{\int_\tau^{\tau+\Delta+q(\Delta)} \eta\left[t, u^{(2)}_\Delta(t), u^{(2)}_\Delta(t-\Delta)\right] dt}_{\alpha^{(2)}(\Delta)} + \alpha^{(1)}(\Delta) \\
&+ \int_{t_0}^{\tau} \eta\left[t, u^{(1)}_\Delta(t), u^{(1)}_\Delta(t-\Delta)\right] dt - \underbrace{\int_{\tau-q(\Delta)}^{\tau} \eta\left[t, u^{(1)}_\Delta(t), u^{(1)}_\Delta(t-\Delta)\right] dt}_{\alpha^{(3)}(\Delta)} \\
&= \left\langle \eta_2; u^{(2)}_\Delta \right\rangle + \left\langle \eta_1; u^{(1)}_\Delta \right\rangle + \alpha^{(1)}(\Delta) - \alpha^{(2)}(\Delta) - \alpha^{(3)}(\Delta).
\end{aligned}$$

In view of (5.5),

$$\left|\alpha^{(i)}(\Delta)\right| \le \int_{\tau-q(\Delta)}^{\tau+\Delta+q(\Delta)} \max_{u_1,u_2\in\Omega} \left|\eta\,[t,u_1,u_2]\right|\,dt \to 0 \quad \text{as} \quad \Delta\to 0.$$

So letting $\Delta\to 0$ gives $\langle u_\Delta;\eta\rangle \to \langle\nu_1;\eta_1\rangle + \langle\nu_2;\eta_2\rangle =: \chi$, where

$$\chi = \int_T dt \int_{\Omega^2} \eta_1(t,w)\nu_1(t,dw) + \int_T dt \int_{\Omega^2} \eta_2(t,w)\nu_2(t,dw) = \langle\nu_E;\eta\rangle$$

and $\eta\in\mathbb{L}_1\big[T,\mathbb{C}(\Omega^2)\big]$ is arbitrary. Thus $u_\Delta\to\nu_E$, and so $\nu_E\in\mathfrak{S}$ by (8.1). The case $\mathfrak{R}_\Delta := \boldsymbol{\mathcal{R}}_\Delta\ \forall\Delta$ is considered likewise. □

In the remainder of this section, we reveal consequences of permutability and closedness.

LEMMA 9.3. *Suppose that the set $\mathfrak{S}\subset\mathfrak{G}$ is closed and permutable. Given $\nu_1,\nu_2\in\mathfrak{S}$, the element* (9.1) *belongs to $\mathfrak{S}$ for any Borel set E.*

Proof. The claim is immediate from Definition 9.1 in the case where E is the union of a finite number of intervals. As is well known, any Borel set E can be approximated by such unions $E^{(1)},E^{(2)},\dots$, in the sense that $\mathrm{mes}(E^{(i)}\triangle E)\to 0$ as $i\to\infty$. Here and throughout, the symbol **mes** stands for the Lebesgue measure and $A\triangle B := \big(A\setminus B\big)\cup\big(B\setminus A\big)$. For any function $\eta\in\mathbb{L}_1\big[T,\mathbb{C}(\Omega^2)\big]$, we have due to (9.1),

$$\begin{aligned}\langle\nu_{E^{(i)}};\eta\rangle - \langle\nu_E;\eta\rangle &= \int_T dt\int_{\Omega^2}\eta(t,u_1,u_2)\,[\nu_{E^{(i)}}(t,du_1,du_2)-\nu_E(t,du_1,du_2)]\\ &= \underbrace{\int\limits_{E\setminus E^{(i)}} dt \int\limits_{\Omega^2}\eta(t,u_1,u_2)\nu_{21}(t,du_1,du_2)}_{\beta(1,i)} + \underbrace{\int\limits_{E^{(i)}\setminus E} dt\int\limits_{\Omega^2}\eta(t,u_1,u_2)\nu_{12}(t,du_1,du_2)}_{\beta(2,i)}.\end{aligned}$$

Here $\nu_{sl} := \nu_s-\nu_l$ for $s,l=1,2$ and thanks to (5.5),

$$\begin{aligned}|\beta(j,i)| &\le \int_{E^{(i)}\triangle E} dt\int_{\Omega^2}|\eta(t,u_1,u_2)|\Big[\nu_1(t,du_1,du_2)+\nu_2(t,du_1,du_2)\Big]\\ &\le 2\int_{E^{(i)}\triangle E}\max_{u_1,u_2\in\Omega}|\eta(t,u_1,u_2)|\,dt\to 0 \quad\text{as}\quad i\to\infty.\end{aligned}$$

Hence $\langle\nu_{E^{(i)}};\eta\rangle\to\langle\nu_E;\eta\rangle$ as $i\to\infty\ \forall\eta\in\mathbb{L}_1\big[T,\mathbb{C}(\Omega^2)\big]$ or in other words, $\nu_{E^{(i)}}\to\nu_E$ as $i\to\infty$. Thus, $\nu_E\in\overline{\mathfrak{S}}=\mathfrak{S}$. □

Lemma 9.3 easily implies the following claim.

COROLLARY 9.4. *Suppose that the assumptions of Lemma* 9.3 *are true, $\nu_1,\dots,\nu_s\in\mathfrak{S}$, and $E_1,\dots,E_s$ are Borel sets constituting a partition of T. Put $\nu(t,dw) := \nu_i(t,dw)$ whenever $t\in E_i$ and $i=1,\dots,s$. Then $\nu\in\mathfrak{S}$.*

Though the next fact is well known, we offer its proof for the convenience of the reader.

LEMMA 9.5. *Any closed and permutable set $\mathfrak{S}\subset\mathfrak{G}$ is convex.*

Proof. Let $\nu_1,\nu_2\in\mathfrak{S},\theta_1,\theta_2>0$, and $\theta_1+\theta_2=1$. We put $\nu_{(r)}(t,dw) := \nu_i(t,dw)$ whenever $t\in T_i(j,r)$ for some $i=1,2$ and $j=0,\dots,2^r-1$, where $r=1,2,\dots$,

$$\begin{aligned}T_1(j,r) &:= \big[t_0+j2^{-r}|T|,\,t_0+j2^{-r}|T|+\theta_1 2^{-r}|T|\big),\\ T_2(j,r) &:= \big[t_0+j2^{-r}|T|+\theta_1 2^{-r}|T|,\,t_0+(j+1)2^{-r}|T|\big),\end{aligned}$$

and $|T| := t_1 - t_0$. For any $\eta(\cdot) \in \mathbb{L}_1\big[T, \mathbb{C}(\Omega^2)\big]$, relation (5.6) yields

$$\big\langle \nu_{(r)}; \eta \big\rangle = \int_T dt \int_{\Omega^2} \eta(t,w)\nu_{(r)}(t,dw) = \mathcal{A}_r(\mathfrak{f}_1, \mathfrak{f}_2) := \sum_{j=0}^{2^r-1} \sum_{i=1,2} \int_{T_i(j,r)} \mathfrak{f}_i(t)\, dt,$$

where $\mathfrak{f}_i(t) := \int_{\Omega^2} \eta(t,w)\nu_i(t,dw)$. For any $\chi_1(\cdot), \chi_2(\cdot) \in \mathbb{L}_1(T)$,

$$|\mathcal{A}_r(\chi_1, \chi_2)| \le \sum_{j=0}^{2^r-1} \sum_{i=1,2} \int_{T_i(j,r)} |\chi_i(t)|\ dt \le \sum_{i=1,2} \int_T |\chi_i(t)|\ dt = |(\chi_1, \chi_2)|.$$

Denote by S_r the set of functions $\chi(\cdot) : T \to \mathbb{R}$ that are constant on any interval $T(j,r) := T_1(j,r) \cup T_2(j,r)$. It is easy to see that $\int_{T_i(j,r')} \chi(t)dt = \theta_i \int_{T(j,r')} \chi(t)dt$, $i = 1,2$ for $\chi(\cdot) \in S_r$ and $r' \ge r$. Hence $\mathcal{A}_r(\chi_1, \chi_2) = \mathcal{A}(\chi_1, \chi_2) := \sum_{i=1,2} \theta_i \int_T \chi_i(t)dt$ for large $r \ge \overline{r}(\chi_1, \chi_2)$ whenever $\chi_1(\cdot), \chi_2(\cdot) \in S := \bigcup_{r=1}^{\infty} S_r$. The set S is dense in $\mathbb{L}_1(T)$. So given $\varepsilon > 0$, there exist $\chi_1(\cdot), \chi_2(\cdot) \in S$ such that $|\mathfrak{f}_i - \chi_i| < \varepsilon$ for $i = 1,2$. Whence

$$\begin{aligned}
\overline{\lim_{r\to\infty}} \left|\left(\mathcal{A}_r - \mathcal{A}\right)(\mathfrak{f}_1, \mathfrak{f}_2)\right| &= \overline{\lim_{r\to\infty}} \left|\left(\mathcal{A}_r - \mathcal{A}\right)\left[(\mathfrak{f}_1, \mathfrak{f}_2) - (\chi_1, \chi_2)\right]\right| \\
&\le \overline{\lim_{r\to\infty}} \left|\mathcal{A}_r\left[(\mathfrak{f}_1, \mathfrak{f}_2) - (\chi_1, \chi_2)\right]\right| + \left|\mathcal{A}\left[(\mathfrak{f}_1, \mathfrak{f}_2) - (\chi_1, \chi_2)\right]\right| \\
&\le \left(1 + \|\mathcal{A}\|\right) \left|(\mathfrak{f}_1, \mathfrak{f}_2) - (\chi_1, \chi_2)\right| \le 2\varepsilon\left(1 + \|\mathcal{A}\|\right) \quad \forall \varepsilon > 0.
\end{aligned}$$

Thus we see that $\lim_{r\to\infty} \big\langle \nu_{(r)}; \eta \big\rangle = \mathcal{A}(\mathfrak{f}_1, \mathfrak{f}_2) = \theta_1 \int_T \mathfrak{f}_1(t)dt + \theta_2 \int_T \mathfrak{f}_2(t)dt = \int_T dt \int_{\Omega^2} \eta(t,w)[\theta_1\nu_1(t,dw) + \theta_2\nu_2(t,dw)] = \langle \theta_1\nu_1 + \theta_2\nu_2; \eta \rangle\ \forall \eta$, i.e., $\nu_{(r)} \to \theta_1\nu_1 + \theta_2\nu_2$ as $r \to \infty$. Since $\nu_{(r)} \in \mathfrak{S}\ \forall r$ by Lemma 9.3, we get $\theta_1\nu_1 + \theta_2\nu_2 \in \overline{\mathfrak{S}} = \mathfrak{S}$. □

Now we are in a position to prove the main result of this section.

THEOREM 9.6. *Let a set $\mathfrak{S} \subset \mathfrak{G}$ be closed and permutable. Whenever*

(9.2)

$$\mu_j \in \mathfrak{S}, \alpha_j(\cdot) \in \mathbb{L}_\infty(T)\ j = 1, \dots, l,\ \textit{and}\ \alpha_1(t) \ge 0, \dots, \alpha_l(t) \ge 0,\ \sum_{j=1}^{l} \alpha_j(t) = 1$$

for almost all $t \in T$, the set $\mathfrak{S}$ contains the element $\nu(t,dw) := \sum_{j=1}^{l} \alpha_j(t)\mu_j(t,dw)$.

Proof. The functions $\alpha_j(\cdot)$ can be uniformly approximated by step functions, i.e., measurable functions taking a finite number of values. These step functions can be chosen to satisfy (9.2). Since $\overline{\mathfrak{S}} = \mathfrak{S}$, this implies that it suffices to prove Lemma 9.6 in the case where $\alpha_1(t), \dots, \alpha_l(t)$ are step functions. Then there exists a Borel partition $E_1 \cup \dots \cup E_s = T$ of T such that $\alpha_j(t) = \alpha_j^{(i)}$ a.a. $t \in E_i$ and all j, i. By Lemma 9.5, $\nu_i := \sum_{j=1}^{l} \alpha_j^{(i)}\mu_j \in \mathfrak{S}\ \forall i$. It remains to note that $\nu(t,dw) = \nu_i(t,dw)$ whenever $t \in E_i$, and apply Corollary 9.4. □

Lemmas 8.2 and 9.2 and Theorem 9.6 give rise to the following claim.

COROLLARY 9.7. *The conclusion of Theorem* 9.6 *is true for the limits* (8.1) *of both $\mathfrak{R}_\Delta \equiv \mathcal{R}_\Delta$ and $\mathfrak{R}_\Delta \equiv \mathcal{R}_\Delta^{sam}$.*

10. Extremal points of a polygon in the space of matrices. Now we show that

(10.1)

$$\mathfrak{S} = \mathfrak{S}^{\mathrm{seq}} = \begin{cases} \mathfrak{G}^s & \text{if } \mathfrak{R}_\Delta \equiv \mathcal{R}_\Delta, \\ \mathfrak{G}_N^s & \text{if } \mathfrak{R}_\Delta \equiv \mathcal{R}_\Delta^{\mathrm{sam}}, \end{cases} \quad \text{where } \mathfrak{S} := \operatorname*{LIM}_{\Delta\to 0} \mathfrak{R}_\Delta,\ \mathfrak{S}^{\mathrm{seq}} := \operatorname*{LIM}_{\Delta\to 0}^{\mathrm{seq}} \mathfrak{R}_\Delta,$$

and the limits are defined by (8.1), (8.2). We recall that $\mathfrak{G}^s = \mathfrak{G}^s_1$ and $\mathfrak{G}^s_N$ stands for the set of all measurable maps $\nu : T \to \mathbf{srpm}_N(\Omega^2)$, whereas $\mathbf{srpm}_N(\Omega^2)$ is given by (5.2) and (7.2).

To show that $\nu \in \mathfrak{G}^s_N \Rightarrow \nu \in \mathfrak{S}$, we shall start with the elements ν of the form

$$\nu(t, du', du'') = \sum_{i,j=1}^{l} p_{ij}(t)\delta_{u_i}(du')\delta_{u_j}(du''), \quad \Big(p_{ij}(t)\Big)_{i,j=1}^{l} =: P(t) \in \mathbb{R}^{l\times l}, \tag{10.2}$$

where $u_1, \dots, u_l \in \Omega$, $u_i \neq u_j$ whenever $i \neq j$, and $\delta_v(du)$ is the Dirac measure at the point $v \in \Omega$. It is easy to see that $\nu \in \mathfrak{G}^s_N \Leftrightarrow P(t) \in \mathcal{M}_l(N-1)$ a.a. $t \in T$, where

$$\mathcal{M}_l(\varkappa) := \Bigg\{ P = \Big(p_{ij}\Big)_{i,j=1}^{l} : p_{i,j} \geq 0 \ \forall i,j, \quad \sum_{i,j=1}^{l} p_{i,j} = 1, \\ \sum_{j=1}^{l} p_{i,j} = \sum_{j=1}^{l} p_{j,i}, \quad \varkappa \sum_{j:j\neq i} p_{i,j} \leq p_{i,i} \ \forall i \Bigg\} \qquad (\varkappa \in [0, +\infty]). \tag{10.3}$$

(The sum over the empty set is 0 and $\infty \cdot 0 := 0$.) Since $\mathcal{M}_l(\varkappa)$ is a polygon, it is the convex hull of the set $\mathcal{M}^{\mathrm{ext}}_l(\varkappa)$ of its extremal points. By Corollary 9.7, this means that any element $\nu \in \mathfrak{G}^s_N$ of the form (10.2) belongs to $\mathfrak{S}$ if so do the elements (10.2) with the matrix $P(t)$ constant and from the set $\mathcal{M}^{\mathrm{ext}}_l(N-1)$. In this section, we compute this set.

LEMMA 10.1. *Suppose that* $\widetilde{P} = (\widetilde{p}_{ij}) \in \mathcal{M}^{ext}_l(\varkappa)$, $\varkappa \in [0, \infty]$, $P = (p_{ij}) \in \mathcal{M}_l(\varkappa)$, $\varkappa \sum_{j:j\neq i} p_{i,j} = p_{i,i}$ $\forall i$ *if* $\varkappa < \infty$, *and* $\widetilde{p}_{ij} \geq \theta p_{ij}$ $\forall i,j$, *where* $\theta > 0$. *Then* $\widetilde{P} = P$.

Proof. Reducing $\theta > 0$ yields that $\widetilde{p}_{i'j'} > \theta p_{i'j'}$ for some i', j'. By (10.3), $1 = \sum_{i,j=1}^{l} \widetilde{p}_{ij} > \theta \sum_{i,j=1}^{l} p_{ij} = \theta$, i.e., $0 < \theta < 1$. Put $p'_{ij} := (1-\theta)^{-1}(\widetilde{p}_{ij} - \theta p_{ij})$. It is easy to see that $P' := (p'_{ij})_{i,j=1}^{l} \in \mathcal{M}_l(\varkappa)$ and $\widetilde{P} = \theta P + (1-\theta)P'$. This implies $\widetilde{P} = P$ by the definition of the extremal point. □

DEFINITION 10.2. *An* $l \times l$*-matrix* P *is said to be* cyclic $\varkappa$-subdiagonal *if there exists a sequence of integers* $1 \leq i_1, i_2, \dots, i_s \leq l$ *such that* $i_h \neq i_d$ *for* $h \neq d$ *and* $s = 1$ *if* $\varkappa = \infty$,

$$\begin{array}{c} p_{hd} = 0 \quad \textit{whenever} \quad (h,d) \neq (i_\sigma, i_{\sigma-1}), (i_\sigma, i_\sigma) \quad \forall \sigma = 1, \dots, s \quad (i_0 := i_s), \\ p_{i_\sigma i_{\sigma-1}} = \frac{1}{(\varkappa+1)s},\ p_{i_\sigma i_\sigma} = \frac{\varkappa}{(\varkappa+1)s}\ \forall \sigma \quad \textit{if } s > 1, \quad \textit{otherwise} \quad p_{i_1,i_1} = 1. \end{array} \tag{10.4}$$

Any such matrix clearly belongs to $\mathcal{M}_l(\varkappa)$. Now we state the main result of this section.

LEMMA 10.3. *Any extremal point of* $\mathcal{M}_l(\varkappa)$ *is a cyclic* $\varkappa$*-subdiagonal matrix.*

Proof. Let $\widetilde{P} = (\widetilde{p}_{ij}) \in \mathcal{M}^{\mathrm{ext}}_l(\varkappa)$. Suppose first that $\widetilde{p}_{ij} = 0$ if $i \neq j$. For $d \in \Upsilon := \{d : \widetilde{p}_{dd} > 0\}$, we put $p^{\langle d\rangle}_{ij} := 1$ if $i = j = d$ and $p^{\langle d\rangle}_{ij} := 0$ otherwise. Since $P^{\langle d\rangle} := (p^{\langle d\rangle}_{ij}) \in \mathcal{M}_l(\varkappa)$, $\widetilde{P} = \sum_{d\in\Upsilon} \widetilde{p}_{dd} P^{\langle d\rangle}$, and the point $\widetilde{P}$ is extremal, the set Υ contains only one element d. So $\widetilde{P} = P^{\langle d\rangle}$, where the matrix $P^{\langle d\rangle}$ is cyclic $\varkappa$-subdiagonal.

Now suppose that $\widetilde{p}_{d_1 d_2} > 0$ for some indices $d_1 \neq d_2$. Then $\varkappa < \infty$ and $\sum_{j=1}^{l} \widetilde{p}_{jd_2} > \widetilde{p}_{d_2 d_2}$. By (10.3), $\sum_{j=1}^{l} \widetilde{p}_{d_2 j} = \sum_{j=1}^{l} \widetilde{p}_{jd_2} > \widetilde{p}_{d_2 d_2}$ and so $\widetilde{p}_{d_2 d_3} > 0$ for some $d_3 \neq d_2$. Continuing likewise, we get a sequence $d_1, d_2, d_3, \dots$, such that $d_{j+1} \neq d_j$ and $\widetilde{p}_{d_j d_{j+1}} > 0$ $\forall j$. We terminate it when some index repeats, i.e.,

$d_i \neq d_j$ whenever $i < j \leq q$ and $d_{q+1} = d_r$ for some $r = 1, \ldots, q$. Here $r \leq q-1$ since $d_q \neq d_{q+1}$. Put $s := q-r+1 \geq 2$ and $i_j := d_{s+r-j}$ for $j = 1, \ldots, s$. Then $i_j \neq i_{j'}$ whenever $j < j' \leq s$ and $\widetilde{p}_{i_j i_{j-1}} > 0$ for $j = 1, \ldots, s$, where $i_0 := d_{s+r} = d_{q+1} = d_r = i_s$. Define the cyclic $\varkappa$-subdiagonal matrix $P = (p_{ij})_{i,j=1}^{l}$ in correspondence with (10.4) and denote $\theta := s(\varkappa+1)\min_{j=1,\ldots,s} \widetilde{p}_{i_j i_{j-1}}$. With regard to (10.3), we conclude that $\widetilde{p}_{ij} \geq \theta p_{ij}$ for all i, j. Lemma 10.1 completes the proof. □

In fact, the set $\mathcal{M}_l^{\rm ext}(\varkappa)$ consists of all cyclic $\varkappa$-subdiagonal matrices. However, we do not need this strengthening of Lemma 10.3 to prove the results of this paper.

11. Proof of (10.1).

LEMMA 11.1. *For any Δ-parametric set $\mathfrak{R}_\Delta$, the following inclusions hold $\mathfrak{S} \subset \mathfrak{S}^{seq} \subset \mathfrak{G}^s (= \mathfrak{G}_1^s)$.*

Proof. The first of them is immediate from (8.1) and (8.2). To prove the second one, consider an element $\nu \in \mathfrak{S}^{\rm seq}$ and the corresponding sequences $\{\Delta_i\}$ and $\{u_i\}$ from (8.2). Given two functions $\zeta(\cdot) \in \mathbb{L}_1(T)$ and $\rho(\cdot) \in \mathbb{C}(\Omega)$, we put $\eta_j(t, u_1, u_2) := \zeta(t)\rho(u_j)$ $(j = 1, 2)$ for all $t \in T$ and $u_1, u_2 \in \Omega$. By (5.6) and (5.8),

$$\begin{aligned}\langle u_i; \eta_2\rangle - \langle u_i; \eta_1\rangle &= \int_T \rho\left[u_i(t-\Delta_i)\right]\zeta(t)\,dt - \int_T \rho\left[u_i(t)\right]\zeta(t)\,dt \\ &= \int_{t_0-\Delta_i}^{t_1-\Delta_i} \rho\left[u_i(t)\right]\zeta(t+\Delta_i)\,dt - \int_{t_0}^{t_1} \rho\left[u_i(t)\right]\zeta(t)\,dt.\end{aligned}$$

Letting $i \to \infty$ gives

$$0 = \langle \nu; \eta_2\rangle - \langle \nu; \eta_1\rangle = \int_T \zeta(t)\,dt \int_{\Omega^2} \left[\rho(u_2) - \rho(u_1)\right]\nu(t, du_1, du_2).$$

So far as the function $\zeta(\cdot) \in \mathbb{L}_1(T)$ is arbitrary, we have for almost all $t \in T$,

$$0 = \int_{\Omega^2} \left[\rho(u_2) - \rho(u_1)\right]\nu(t, du_1, du_2) = \int_\Omega \rho(u)\nu(t, du, \Omega) - \int_\Omega \rho(u)\nu(t, \Omega, du)$$

for all $\rho(\cdot) \in \mathbb{C}(\Omega)$. So $\nu(t, du, \Omega) = \nu(t, \Omega, du)$ and thus $\nu \in \mathfrak{G}^s$. □

LEMMA 11.2. *Let $\mathfrak{R}_\Delta \equiv \mathfrak{R}_\Delta^{sam}$, $q(\Delta) \to 0$, and $\frac{q(\Delta)}{\Delta} \to N \in (1, \infty]$ as $\Delta \to 0$. Then*

$$\mathfrak{S}^{seq} \subset \mathfrak{G}_N^s. \tag{11.1}$$

Proof. Consider an element $\nu \in \mathfrak{S}^{\rm seq}$ and the corresponding sequences $\{\Delta_i\}$ and $\{u_i\}$ from (8.2). Due to (7.1), the inclusion $u_i \in \mathfrak{R}_{\Delta_i}^{\rm sam}$ implies that

$$u_i(t) = u_i^{(j)}\ \forall\, t_i^{(j)} \leq t < t_i^{(j+1)},\ j = 0, \ldots, s_i, \qquad t_i^{(r+1)} - t_i^{(r)} \geq q(\Delta_i), \tag{11.2}$$

where $r = 0, \ldots, s_i - 1$, $t_i^{(0)} = t_0 - \Delta_i$, and $t_i^{(s_i+1)} = t_1$. Put

$$E_i^{[1]} := T \cap \bigcup_{j=0}^{s_i} \left[t_i^{(j)}, t_i^{(j)} + \Delta_i\right),\ E_i^{[2]} := T \setminus E_i^{[1]},\ \chi_i^{[p]}(t) := \begin{cases} 1 \text{ if } t \in E_i^{[p]} \\ 0 \text{ if } t \notin E_i^{[p]} \end{cases}. \tag{11.3}$$

Invoke the space $\boldsymbol{\mathcal{N}}(T, K)$ introduced in section 5. Passing to subsequences ensures that there exist limits in the weak topologies of $\boldsymbol{\mathcal{N}}(T, \Omega^2)$ and $\boldsymbol{\mathcal{N}}(T, \Omega)$, respectively,

$$\nu^{[1]}(t, dw) := \lim_{i\to\infty} \chi_i^{[1]}(t)\delta_{[u_i(t), u_i(t-\Delta_i)]}(dw), \mu(t, du) := \lim_{i\to\infty} \chi_i^{[2]}(t)\delta_{u_i(t)}(du). \tag{11.4}$$

We recall that δ_a is the Dirac measure at the point a. Here μ can be interpreted as an element $\nu^{[2]}$ of $\boldsymbol{\mathcal{N}}(T,\Omega^2)$ concentrated on the diagonal $D:=\{(u,u):u\in\Omega\}$. More precisely,

$$\nu^{[2]}(t,E):=\mu\big(t,E\pitchfork D\big),\quad \text{where}\quad E\pitchfork D:=\{u:(u,u)\in E\}. \tag{11.5}$$

Since $u_i(t-\Delta_i)=u_i(t)$ whenever $t\in E_i^{[2]}$, we have for any $g(\cdot)\in\mathbb{L}_1\big[T,\mathbb{C}(\Omega^2)\big]$,

$$\begin{aligned}\langle\nu,g\rangle&=\lim_{i\to\infty}\int_T g\left[t,u_i(t),u_i(t-\Delta_i)\right]\,dt\\&=\lim_{i\to\infty}\int_{E_i^{[1]}} g\left[t,u_i(t),u_i(t-\Delta_i)\right]\,dt+\lim_{i\to\infty}\int_{E_i^{[2]}} g\left[t,u_i(t),u_i(t)\right]\,dt\\&=\int_T dt\int_{\Omega^2} g(t,w)\nu^{[1]}(t,dw)+\int_T dt\int_\Omega g(t,u,u)\mu(t,du)\\&=\int_T dt\int_{\Omega^2} g(t,w)\left[\nu^{[1]}(t,dw)+\nu^{[2]}(t,dw)\right].\end{aligned}$$

Thus $\nu=\nu^{[1]}+\nu^{[2]}$. By (7.3) and (11.5), $\nu_d^{[2]}(t,du)=\mu(t,du)=\nu^{[2]}(t,du,\Omega)$. Hence

$$\nu_d(t,du)=\nu_d^{[1]}(t,du)+\mu(t,du)\geq\mu(t,du) \tag{11.6}$$

and $\nu_-(t,du):=\nu(t,du,\Omega)-\nu_d(t,du)=\nu^{[1]}(t,du,\Omega)-\nu_d^{[1]}(t,du)\leq\nu^{[1]}(t,du,\Omega)=\int_\Omega\nu^{[1]}(t,du,du_2)$. This means that for any functions $\zeta(\cdot)\in\mathbb{C}(T)$ and $\rho(\cdot)\in\mathbb{C}(\Omega)$ such that $\zeta(\cdot)\geq 0,\rho(\cdot)\geq 0$, and $\zeta(t)=0\ \forall t\in[t_1-\varepsilon,t_1]$ for some $\varepsilon\approx 0,\varepsilon>0$, we have

$$\begin{aligned}\sigma:=\int_T\zeta(t)dt\int_\Omega\rho(u_1)\nu_-(t,du_1)&\leq\int_T\zeta(t)dt\int_{\Omega^2}\rho(u_1)\nu^{[1]}(t,du_1,du_2)\\&\overset{(11.4)}{=\!=}\lim_{i\to\infty}\int_{E_i^{[1]}}\zeta(t)\rho[u_i(t)]dt=\lim_{i\to\infty}\sum_{j=0}^{\widehat{s}_i}\int_{t_i^{(j)}}^{t_i^{(j)}+\Delta_i}\zeta(t)\rho[u_i(t)]dt.\end{aligned} \tag{11.7}$$

Here $\widehat{s}_i:=\max\{j:t_i^{(j)}\leq t_1-\varepsilon\}$. For $j\leq\widehat{s}_i$ and i sufficiently large, $[t_i^{(j)}+\Delta_i,t_i^{(j)}+q(\Delta_i)]\subset T$ and $u_i(t)$ is constant on $[t_i^{(j)},t_i^{(j)}+q(\Delta_i)]$ due to (11.2). Therefore

$$\begin{aligned}&\left|\int_{t_i^{(j)}}^{t_i^{(j)}+\Delta_i}\zeta(t)\rho[u_i(t)]dt-\frac{\Delta_i}{q(\Delta_i)-\Delta_i}\int_{t_i^{(j)}+\Delta_i}^{t_i^{(j)}+q(\Delta_i)}\zeta(t)\rho[u_i(t)]dt\right|\\&=\left|\int_{t_i^{(j)}}^{t_i^{(j)}+\Delta_i}\underbrace{\Big(\zeta[t]-\zeta\left[t_i^{(j)}+\Delta_i\right]\Big)}_{\varkappa(t)}\rho[u_i(t)]dt-\frac{\Delta_i}{q(\Delta_i)-\Delta_i}\int_{t_i^{(j)}+\Delta_i}^{t_i^{(j)}+q(\Delta_i)}\varkappa(t)\rho[u_i(t)]dt\right|\\&\leq\int_{t_i^{(j)}}^{t_i^{(j)}+\Delta_i}|\varkappa(t)|\rho[u_i(t)]dt+\frac{\Delta_i}{q(\Delta_i)-\Delta_i}\int_{t_i^{(j)}+\Delta_i}^{t_i^{(j)}+q(\Delta_i)}|\varkappa(t)|\rho[u_i(t)]dt\\&\leq 2\Delta_i\Gamma_\zeta\left[q(\Delta_i)\right]c_\rho,\end{aligned}$$

where $\Gamma_\zeta(\delta):=\max_{|t_2-t_1|\leq\delta}|\zeta(t_1)-\zeta(t_2)|\to 0$ as $\delta\to 0$ and $c_\rho:=\max_{u\in\Omega}\rho(u)$.

Hence the discrepancy between the sum from (11.7) and $\frac{\Delta_i}{q(\Delta_i)-\Delta_i}\sum_{j=0}^{\widehat{s}_i}\int_{t_i^{(j)}+\Delta_i}^{t_i^{(j)}+q(\Delta_i)}\zeta(t)\rho[u_i(t)]dt$ does not exceed $2c_\rho\Gamma_\zeta\big[q(\Delta_i)\big]\sum_{j=1}^{\widehat{s}_i}\Delta_i \le 2c_\rho\Gamma_\zeta\big[q(\Delta_i)\big](t_1-t_0)$ thanks to (11.2). Thus this discrepancy converges to zero as $i\to\infty$, and (11.7) can be continued as follows:

$$\begin{aligned}\sigma &\le \lim_{i\to\infty}\sum_{j=1}^{\widehat{s}_i}\frac{\Delta_i}{q(\Delta_i)-\Delta_i}\int_{t_i^{(j)}+\Delta_i}^{t_i^{(j)}+q(\Delta_i)}\zeta(t)\rho[u_i(t)]dt\\ &\overset{(11.3)}{\le} \lim_{i\to\infty}\frac{\Delta_i}{q(\Delta_i)-\Delta_i}\int_{E_i^{[2]}}\zeta(t)\rho[u_i(t)]dt\end{aligned}$$

$$\overset{(11.3),\,(11.4)}{=\!=\!=} \frac{1}{N-1}\int_T\zeta(t)dt\int_\Omega\rho(u)\mu(t,du)\overset{(11.6)}{\le}\frac{1}{N-1}\int_T\zeta(t)dt\int_\Omega\rho(u)\nu_d(t,du).$$

(We put $1/\infty := 0$.) Since the nonnegative functions $\zeta(\cdot),\rho(\cdot)$ are arbitrary, we get $\nu(t,du,\Omega)-\nu_d(t,du)=\nu_-(t,du)\le(N-1)^{-1}\nu_d(t,du)$ a.a. $t\in T$, which completes the proof in view of (7.2) and Lemma 11.1 (and Remark 7.2 in the case $N=\infty$). □

LEMMA 11.3. *Let the assumptions of Lemma* 11.2 *be true. Suppose also that in* (10.2)*, the matrix $P(t)$ is cyclic $(N-1)$-subdiagonal and does not depend on t, i.e., $P(t)=P$ a.a. $t\in T=[t_0,t_1]$. Then the element ν given by* (10.2) *belongs to* $\mathfrak{S}=\mathrm{LIM}_{\Delta\to+0}\mathfrak{R}_\Delta^{sam}$.

Proof. Consider the sequence $1\le i_1,\dots,i_s\le l$ associated with P by (10.4) and denote

$$t_\Delta(r,j):=t_0-\Delta+srq(\Delta)+jq(\Delta),\quad T_\Delta(r,j):=\Big[t_\Delta(r,j-1)\,;t_\Delta(r,j)\Big)\cap[t_0-\Delta,t_1],$$
$$u_\Delta(t):=u_{i_j}\quad\text{whenever}\quad t\in T_\Delta(r,j)\quad\text{for some}\quad r=0,1,\dots,\ j=1,\dots,s.$$

Then $u_\Delta\in\mathfrak{R}_\Delta^{\mathrm{sam}}$. By (8.1), it suffices to show that $u_\Delta\to\nu$ as $\Delta\to+0$, i.e., $\langle u_\Delta;\eta\rangle\to\langle\nu;\eta\rangle$ as $\Delta\to+0$ for any $\eta\in\mathbb{L}_1\big[T,\mathbb{C}(\Omega^2)\big]$. By [22, Lemma IV.2.4], one can focus on the functions η of the form $\eta(t,u_1,u_2)=\chi_E(t)\rho(u_1,u_2)$, where $\rho(\cdot)\in\mathbb{C}(\Omega\times\Omega), E=[\tau_1,\tau_2], t_0<\tau_1<\tau_2<t_1$, and $\chi_E(t):=1$ for $t\in E$ whereas $\chi_E(t):=0$ if $t\notin E$. Put

$$A_\Delta:=\Big\{\alpha=0,1,\cdots:T_\Delta(\alpha):=\Big[t_\Delta(0,0)+s\alpha q(\Delta);\,t_\Delta(0,0)+s(\alpha+1)q(\Delta)\Big]\subset[\tau_1,\tau_2]\Big\},\quad \alpha_\Delta^-:=\min\{\alpha:\alpha\in A_\Delta\},\quad \alpha_\Delta^+:=\max\{\alpha:\alpha\in A_\Delta\}.$$

Then by (5.6) and (5.8),

$$\langle u_\Delta;\eta\rangle=\int_T\chi_E(t)\underbrace{\rho\,[u_\Delta(t),u_\Delta(t-\Delta)]}_{\varkappa(t)}\,dt=\int_{\tau_1}^{\tau_2}\varkappa(t)\,dt$$
$$=\underbrace{\int_{T_\Delta(\alpha_\Delta^--1)\cap[\tau_1,\tau_2]}\varkappa(t)\,dt}_{\omega^-(\Delta)}+\underbrace{\int_{T_\Delta(\alpha_\Delta^++1)\cap[\tau_1,\tau_2]}\varkappa(t)\,dt}_{\omega^+(\Delta)}+\underbrace{\sum_{r=\alpha_\Delta^-}^{\alpha_\Delta^+}\sum_{j=1}^{s}\int_{T_\Delta(r,j)}\varkappa(t)\,dt}_{\omega(\Delta)}.$$

Here $\mathrm{mes}T_\Delta(\alpha^\pm_\Delta \pm 1) = sq(\Delta)$, and so $|\omega^\pm(\Delta)| \le c_\rho sq(\Delta)$, where $c_\rho := \max_{u',u''\in\Omega}|\rho(u', u'')|$. Thus $\omega^\pm(\Delta) \to 0$ as $\Delta \to +0$. For the indices r, j involved in the last sum and $\Delta \approx 0$,

$$\int_{T_\Delta(r,j)} \varkappa(t)dt = \int_{t_\Delta(r,j-1)}^{t_\Delta(r,j-1)+\Delta} \rho\big(u_{i_j}, u_{i_{j-1}}\big)dt + \int_{t_\Delta(r,j-1)+\Delta}^{t_\Delta(r,j)} \rho\big(u_{i_j}, u_{i_j}\big)dt$$
$$= \Delta\rho\big(u_{i_j}, u_{i_{j-1}}\big) + \big[q(\Delta) - \Delta\big]\rho\big(u_{i_j}, u_{i_j}\big).$$

Note that $q(\Delta) = \mathrm{mes}T_\Delta(r, j)$. Hence

$$\omega(\Delta) = \frac{1}{s}\sum_{j=1}^{s}\left[\frac{\Delta}{q(\Delta)}\rho\big(u_{i_j}, u_{i_{j-1}}\big) + \left(1 - \frac{\Delta}{q(\Delta)}\right)\rho\big(u_{i_j}, u_{i_j}\big)\right]\sum_{r=\alpha^-_\Delta}^{\alpha^+_\Delta}\sum_{j=1}^{s}\mathrm{mes}T_\Delta(r, j).$$

The last double sum equals $\mathrm{mes}\mathcal{T}_\Delta$: $\mathcal{T}_\Delta := \bigcup_{r=\alpha^-_\Delta}^{\alpha^+_\Delta}\bigcup_{j=1}^{s} T_\Delta(r, j) \subset [\tau_1, \tau_2]$, $[\tau_1, \tau_2] \setminus \mathcal{T}_\Delta \subset \mathfrak{T}_\Delta := T_\Delta(\alpha^-_\Delta - 1) \cup T_\Delta(\alpha^+_\Delta + 1)$, and $\mathrm{mes}\mathfrak{T}_\Delta \to 0$ as $\Delta \to +0$ by the foregoing. So

$$\langle u_\Delta; \eta\rangle \to \Gamma := s^{-1}(\tau_2 - \tau_1)\sum_{j=1}^{s}\left[\frac{1}{N}\rho\big(u_{i_j}, u_{i_{j-1}}\big) + \left(1 - \frac{1}{N}\right)\rho\big(u_{i_j}, u_{i_j}\big)\right].$$

At the same time, (5.6), (10.2), and (10.4) (where $\varkappa := N - 1$) yield $\Gamma = \int_{\tau_1}^{\tau_2} dt\sum_{i,j=1}^{l} p_{ij}\rho(u_i, u_j) = \langle\nu; \eta\rangle$, which completes the proof. □

LEMMA 11.4. *Lemma* 11.3 *remains true for* $N = 1$.

Proof. Let $P \in \mathcal{M}_l(0)$ and I denote the unit $l \times l$-matrix. Given $\varkappa > 0$, it follows from (10.3) that $P_\varkappa := \frac{1}{l\varkappa+1}P + \frac{\varkappa}{(l\varkappa+1)}I \in \mathcal{M}_l(\varkappa)$. The corresponding element $\nu_\varkappa$ given by (10.2) (where $P(t) := P_\varkappa$) belongs to the set $\mathrm{LIM}_{\Delta\to+0}\mathfrak{R}^{\mathrm{sam}}_{\Delta,\varkappa}$ defined by (8.1) due to Lemma 11.3. Here the set $\mathfrak{R}^{\mathrm{sam}}_{\Delta,\varkappa}$ is defined by (7.1) with $q(\Delta) := (1 + \varkappa)\Delta$. Since $\frac{q(\Delta)}{\Delta} \to N = 1$ as $\Delta \to +0$, we have $\mathfrak{R}^{\mathrm{sam}}_{\Delta,\varkappa} \subset \mathfrak{R}^{\mathrm{sam}}_\Delta$ for $\Delta \approx 0$. So $\mathrm{LIM}_{\Delta\to+0}\mathfrak{R}^{\mathrm{sam}}_{\Delta,\varkappa} \subset \mathrm{LIM}_{\Delta\to+0}\mathfrak{R}^{\mathrm{sam}}_\Delta = \mathfrak{S}$ and thus $\nu_\varkappa \in \mathfrak{S}$. It remains to note that $\nu_\varkappa \to \nu$ as $\varkappa \to +0$ and employ Lemma 8.2. □

LEMMA 11.5. *Suppose that* $\mathfrak{R}_\Delta \equiv \mathfrak{R}^{sam}_\Delta$ *and the first relation from* (7.5) *is true. The set* $\mathfrak{S}$ *defined in* (10.1) *contains any element* (10.2) *such that* $P(t) \in \mathcal{M}_l(N - 1)$ *a.a.* $t \in T$.

Proof. Put $\mathfrak{A} := \{\alpha : \mathcal{M}_l^{\mathrm{ext}}(N - 1) \to [0, \infty) : \sum_{P\in\mathcal{M}_l^{\mathrm{ext}}(N-1)}\alpha(P) = 1\}$ and $\Lambda\big[\alpha(\cdot)\big] := \sum_{P\in\mathcal{M}_l^{\mathrm{ext}}(N-1)}\alpha(P)P$, where $\mathcal{M}_l^{\mathrm{ext}}(N - 1)$ is the set of the extremal points of $\mathcal{M}_l(N - 1)$. By the Krein–Milman theorem, $\Lambda(\mathfrak{A}) \supset \mathcal{M}_l(N - 1)$. Then Theorems I.7.6 and I.7.7 in [22] imply that $P(t) = \Lambda\big[\alpha(t, \cdot)\big]$ for a measurable function $t \mapsto \alpha(t, \cdot) \in \mathfrak{A}$. Lemmas 10.3, 11.3, and 11.4 and Corollary 9.7 complete the proof. □

Now we are in a position to prove the main result of this section.

THEOREM 11.6. *The relationships* (10.1) *are true whenever* (7.5) *is valid.*

Proof. Suppose that $\mathfrak{R}_\Delta \equiv \mathfrak{R}^{\mathrm{sam}}_\Delta$ and the first relation from (7.5) is true. Due to Lemma 11.1 and (11.1), it suffices to show that $\mathfrak{G}^s_N \subset \mathfrak{S}$. Let $\nu \in \mathfrak{G}^s_N$. Given $\delta > 0$, we introduce a Borel partition $\Omega = \Omega^\delta(1)\cup\cdots\cup\Omega^\delta(l_\delta)$ such that $\sup_{u\in\Omega^\delta(i)}|u - u^\delta_i| < \delta$ for some $u^\delta_i \in \Omega^\delta(i)$ for all i. For

$$p_{ij} := p^\delta_{ij}(t) := \nu\left[t, \Omega^\delta(i) \times \Omega^\delta(j)\right], \quad i, j = 1, \ldots, l_\delta, \qquad \text{and} \qquad l := l_\delta,$$

all relations in (10.3) except for the last one are immediate from (5.2) and (7.2). To prove the last relation, we put $\Omega_d^\delta(i) := \{w = (u,u) : u \in \Omega^\delta(i)\}$. By (7.2), we have for $N < \infty$

$$(N-1)\sum_{j=1}^{l_\delta} p_{ij} = (N-1)\nu\Big[t, \Omega^\delta(i)\times\Omega\Big] \le N\nu_d\Big[t,\Omega^\delta(i)\Big] \overset{(7.3)}{=\!=} N\nu\Big[t, \underbrace{\Omega_d^\delta(i)}_{\subset\Omega^\delta(i)\times\Omega^\delta(i)}\Big]$$
$$\le N\nu\Big[t,\Omega^\delta(i)\times\Omega^\delta(i)\Big] = Np_{ii}.$$

Hence $P(t) := \big(p_{ij}^\delta(t)\big)_{i,j=1}^{l_\delta} \in \mathcal{M}_{l_\delta}(N-1)$ a.a. t, where $\mathcal{M}_{l_\delta}(N-1)$ is given by (10.3). If $N=\infty$, then (7.4) $\Rightarrow \nu \in \mathcal{G}_\infty^s = \bigcap_{N'>1}\mathcal{G}_{N'}^s$. Thus $P(t) \in \bigcap_{N'=1,2,\ldots,} \mathcal{M}_{l_\delta}(N'-1) = \mathcal{M}_{l_\delta}(\infty) = \mathcal{M}_{l_\delta}(N-1)$ a.a. t. So Lemma 11.5 ensures that the element ν_δ defined by (10.2) (where $l := l_\delta,\ p_{ij} := p_{ij}^\delta(t),\ u_i := u_i^\delta$) belongs to $\mathfrak{S}$. For any $\rho \in \mathbb{C}(\Omega^2)$ and a.a. $t \in T$,

$$\text{(11.8)}\quad \left|\int_{\Omega^2}\rho(u',u'')\nu_\delta(t,du',du'') - \int_{\Omega^2}\rho(u',u'')\nu(t,du',du'')\right|$$
$$= \left|\sum_{i,j=1}^{l_\delta}\rho\left(u_i^\delta,u_j^\delta\right)p_{ij}^\delta(t) - \sum_{i,j=1}^{l_\delta}\int_{\Omega^\delta(i)\times\Omega^\delta(j)}\rho(u',u'')\nu(t,du',du'')\right|$$
$$\le \sum_{i,j=1}^{l_\delta}\int_{\Omega^\delta(i)\times\Omega^\delta(j)}\left|\rho\left(u_i^\delta,u_j^\delta\right)-\rho(u',u'')\right|\nu(t,du',du'') \le q_\delta[\rho(\cdot)].$$

Here $q_\delta[\rho(\cdot)] := \max|\rho(v_1,v_2)-\rho(u_1,u_2)|$, where max is over $(v_1,v_2),(u_1,u_2)\in\Omega^2$ such that $|v_1-u_1|\le\delta, |v_2-u_2|\le\delta$. Pick $\eta\in\mathbb{C}(T)$. Then

$$\left|\int_T\eta(t)\,dt\int_{\Omega^2}\rho(u',u'')\nu_\delta(t,du',du'') - \int_T\eta(t)\,dt\int_{\Omega^2}\rho(u',u'')\nu(t,du',du'')\right|$$
$$\le q_\delta[\rho(\cdot)]\int_T|\eta(t)|\,dt.$$

Let $\delta\to+0$. Since $q_\delta[\rho(\cdot)]\to 0$, then $\nu_\delta\to\nu$ by [22, Theorem IV.2.4]. As was shown, $\nu_\delta\in\mathfrak{S}$. Then Lemma 8.2 yields $\nu\in\mathfrak{S}$, $\mathcal{G}_N^s\subset\mathfrak{S}$, which implies (10.1) for the case at hand.

Now let the second relation from (7.5) hold. By Lemma 11.1, it suffices to show that $\mathcal{G}_1^s\subset\mathfrak{S}$. For the set $\mathcal{R}_\Delta^{\mathrm{sam}*}$ defined by (7.1) with $q(\Delta) := \max\{q(\Delta),\Delta\}$, the first relation from (7.5) is true with $N=1$. So $\mathcal{G}_1^s = \mathrm{LIM}_{\Delta\to0}\mathcal{R}_\Delta^{\mathrm{sam}*}$ by the foregoing. At the same time, $\mathcal{R}_\Delta^{\mathrm{sam}*}\subset\mathfrak{R}_\Delta \overset{(8.1)}{\Rightarrow} \mathrm{LIM}_{\Delta\to0}\mathcal{R}_\Delta^{\mathrm{sam}*}\subset\mathrm{LIM}_{\Delta\to0}\mathfrak{R}_\Delta = \mathfrak{S}$, which completes the proof.

For $\mathfrak{R}_\Delta\equiv\mathcal{R}_\Delta$, the proof comes to retracing the arguments from the previous paragraph for $\mathcal{R}_\Delta^{\mathrm{sam}*}$ given by (7.1) with $q(\Delta):=\Delta$. □

12. Proofs of Theorem 7.2 and the claims from section 5. Theorem 5.1 is immediate from Theorems 8.5 and 11.6. Lemma 5.2 follows from Theorems 8.6 and 11.6. Theorem 5.3 results from Theorems 8.7, 11.6, and the first relation from (8.3). Theorem 5.4 is justified by Theorems 8.7 and 11.6. Theorem 7.2 follows from Theorems 8.5, 8.6, 8.7, and 11.6.

13. Proofs of Theorem 7.4 and the claims from section 4.

Proof of Lemma 4.1. This is immediate from Theorem 5.1. □

Proof of Theorem 4.2. Put in (2.1)–(2.4) $f(\cdot)\equiv 0, a(\cdot)\equiv 0, \varphi(t,x_1,x_2,\,u_1,u_2):=-(t_1-t_0)^{-1}|u_1-u_2|^2$, where $|\cdot|$ is the Euclidean norm in $\mathbb{R}^m$. Then due to (2.7), $\overline{\varphi}(t,x,u)=0$ and so $\mathfrak{I}_{opt}(0)=0$, where $\mathfrak{I}_{opt}(0)$ is the infimum of the cost functional in the problem (2.5), (2.6). By (i) from section 4, the set Ω contains two points $u_1\neq u_2$. Put

$$\nu(t,dw):=\eta(dw):=\frac{1}{2}\Big[\delta_{(u_1,u_2)}(dw)+\delta_{(u_2,u_1)}(dw)\Big], \tag{13.1}$$

where $w=(u',u'')\in\Omega^2$ and $\delta_\omega(dw)$ is the Dirac measure at the point $\omega\in\Omega^2$. Then η satisfies the relations from (5.2) and hence $\nu\in\mathfrak{G}^s$. As a result, Theorem 5.1 yields

$$\lim_{\Delta\to+0}\mathfrak{I}_{opt}(\Delta)\leq-\int_T dt\int_{\Omega^2}\frac{|u'-u''|^2}{t_1-t_0}\nu(t,du',du'')=-|u_1-u_2|^2<0=\mathfrak{I}_{opt}(0). \quad \square$$

In the remainder of the section, we prove Theorems 4.3 and 7.4. The attention is focused on Theorem 7.4 since it and Remark 7.5 imply Theorem 4.3. We start with a technical fact.

LEMMA 13.1. *There exist $\delta=\delta_0>0, \varkappa\in(0,1)$, and a map $\chi_\delta(\cdot):\Omega\to\Omega$ such that*

$$\varkappa\delta\leq|\chi_\delta(u)-u|\leq\delta\ \forall u\in\Omega,\quad \chi_\delta(u)=\mathrm{const}=v_i\ \forall u\in W_i, i=1,\dots,k, \tag{13.2}$$

$$\text{where}\quad W_i=\begin{cases}V_1 & \text{if } i=1\\ V_i\setminus(V_{i-1}\cup\dots\cup V_1) & \text{if } i>1,\end{cases}$$

$V_1,\dots,V_k\subset\Omega$ are some open (in Ω) sets, and $V_1\cup\dots\cup V_k=\Omega$. If the set Ω is connected, the numbers δ_0 and $\varkappa$ can be chosen so that such a function $\chi_\delta(\cdot)$ exists for all $\delta\in(0,\delta_0]$.

Proof. If the set Ω is not connected, we put $\delta:=\delta_0>m_+, \varkappa:=1/2\delta^{-1}m_-$, where $m_+:=\max_{u\in\Omega}m(u)$, $m_-:=\min_{u\in\Omega}m(u)$, and $m(u):=\max_{v\in\Omega}|u-v|=|v(u)-u|$ for some $v(u)\in\Omega$. Then $\varkappa\delta<m_-\leq|v(u)-u|\leq m_+<\delta$. Hence for some $\varepsilon_\delta>0$,

$$\varkappa\delta+\varepsilon_\delta\leq|v(u)-u|\leq\delta-\varepsilon_\delta\qquad\forall u\in\Omega. \tag{13.3}$$

If the set Ω is connected, we set $\delta_0:=m_-, \varkappa:=1/2$, and $\varepsilon_\delta:=1/5\delta$ for $\delta\in(0,\delta_0]$. The connected set Ω is not covered by the disjoint sets $\{v:|v-u|\leq 7/10\delta\}$ and $\{v:|v-u|\geq 4/5\delta\}$. So $7/10\delta<|v-u|<4/5\delta$ for some $v=v(u)\in\Omega$, i.e., (13.3) still holds.

In any case for $u'\in\Omega(u):=\{u'\in\Omega:|u'-u|<\varepsilon_\delta\}$, we have $|u'-v(u)|\leq|u'-u|+|u-v(u)|<\varepsilon_\delta+\delta-\varepsilon_\delta=\delta$, $|u'-v(u)|\geq|u-v(u)|-|u'-u|>\varkappa\delta$. Since the set Ω is compact by (i) from section 2, there exist points $u_1,\dots,u_k$ such that $\Omega=V_1\cup\dots\cup V_k, V_i:=\Omega(u_i)$. It remains to define $\chi_\delta(\cdot)$ by (13.2), where $v_i:=v(u_i)$. □

From now on, we fix $\delta,\varkappa$, and $\chi_\delta(\cdot)$ from (13.2). To prove Theorem 7.4, we show that $\mathcal{J}^{\mathrm{sam}}_{\mathrm{opt}}[N-\varepsilon,\xi]<\mathcal{J}^{\mathrm{sam}}_{\mathrm{opt}}[N,\xi]$ if $\varepsilon>0,\varepsilon\approx 0$. To this end, we find a variation $\nu(\varepsilon)\in\mathfrak{G}^s_{N-\varepsilon}$ of the optimal control ν in (5.3), (5.4) (where $\mathfrak{G}^s:=\mathfrak{G}^s_N$) that decreases the cost. It is obtained by "spraying" a small portion of the diagonal component of ν outside the diagonal D of Ω^2. This is possible since, due to (7.2), the lesser the N, the lesser this component may be.

Now we introduce a "dispersion" of a measure $\eta \in \mathbf{frm}(\Omega^2)$ concentrated on D. The isomorphism $u \in \Omega \leftrightarrow (u,u) \in D$ identifies η with an element of $\mathbf{frm}(\Omega)$, which is denoted by the same symbol. For any Borel subsets $E \subset \Omega^2$ and $\mathcal{E} \subset \Omega$, we put

$$\eta^{\langle\delta\rangle}(E) := 1/2\left(\eta\left\{u : [u, \chi_\delta(u)] \in E\right\} + \eta\left\{u : [\chi_\delta(u), u] \in E\right\}\right), \tag{13.4}$$

$$\widehat{\eta}^{\langle\delta\rangle}(\mathcal{E}) := 1/2\left[\eta(\mathcal{E}) + \eta\{u : \chi_\delta(u) \in \mathcal{E}\}\right]. \tag{13.5}$$

For $\eta \in \mathbf{rpm}(\Omega^2)$, we consider the diagonal component η_d given by (7.3), and put

$$\widehat{\eta}_N^{\langle\delta,\varepsilon\rangle} := \tau(\eta_d)^{\langle\delta\rangle} + (1-\tau)\widehat{(\eta_d)}^{\langle\delta\rangle}, \quad \text{where} \quad \tau := \frac{1}{N-\varepsilon}, \tag{13.6}$$

$$\eta_N^{\langle\delta,\varepsilon\rangle} := \underbrace{\eta - \eta_d}_{=:\eta^-} + (1-\theta)\,\eta_d + \theta\widehat{\eta}_N^{\langle\delta,\varepsilon\rangle}, \quad \text{where} \quad \theta := \frac{\varepsilon}{N-1}. \tag{13.7}$$

LEMMA 13.2. *Suppose that $N \in (1,\infty), \varepsilon \in (0, N-1)$, and $\eta \in \mathbf{srpm}_N(\Omega^2)$, where $\mathbf{srpm}_N(\Omega^2)$ is given by* (7.2) *and* (5.2). *Then $\eta_N^{\langle\delta,\varepsilon\rangle} \in \mathbf{srpm}_{N-\varepsilon}(\Omega^2)$.*

Proof. The first two relations from (5.2) are clearly true for $\eta := \eta_N^{\langle\delta,\varepsilon\rangle}$. By (13.4), (13.5),

$$(\eta_d)^{\langle\delta\rangle}(du,\Omega) = \widehat{(\eta_d)}^{\langle\delta\rangle}(du) = (\eta_d)^{\langle\delta\rangle}(\Omega, du). \tag{13.8}$$

Thus $(\eta_d)^{\langle\delta\rangle}$ meets the third requirement from (5.2). So do the measures $\widehat{(\eta_d)}^{\langle\delta\rangle}$ and η_d, since they are concentrated on the diagonal D, and η since $\eta \in \mathbf{srpm}_N(\Omega^2)$. It follows that the measure (13.7) satisfies the third condition from (5.2) as well.

Due to (13.2), (13.4), (13.7), the diagonal components of $\eta^{\langle\delta\rangle}$ and η^- are zero. Hence

$$\left[\eta_N^{\langle\delta,\varepsilon\rangle}\right]_d \overset{(13.7)}{=\!=} (1-\theta)\,\eta_d + \theta\left[\widehat{\eta}_N^{\langle\delta,\varepsilon\rangle}\right]_d \overset{(13.5),\,(13.6)}{=\!=\!=} (1-\theta)\,\eta_d + \theta(1-\tau)\widehat{(\eta_d)}^{\langle\delta\rangle}. \tag{13.9}$$

Note that $\mu(du,\Omega) = \mu(du,du) = \mu(du)$ for any measure μ in Ω^2 concentrated on the diagonal D, and by (7.2), $\eta \in \mathbf{srpm}_N(\Omega^2) \Rightarrow \eta^-(du,\Omega) \le (N-1)^{-1}\eta_d(du)$. So

$$\begin{aligned}
&\eta_N^{\langle\delta,\varepsilon\rangle}(du,\Omega) \overset{(13.6),\,(13.7)}{=\!=\!=} \eta^-(du,\Omega) + (1-\theta)\eta_d(du) + \theta\left[\tau(\eta_d)^{\langle\delta\rangle}(du,\Omega) + (1-\tau)\widehat{(\eta_d)}^{\langle\delta\rangle}(du)\right] \\
&\overset{(13.8)}{=\!=} \eta^-(du,\Omega) + (1-\theta)\eta_d(du) + \theta\widehat{(\eta_d)}^{\langle\delta\rangle}(du);\ (N-\varepsilon-1)\left(\eta_N^{\langle\delta,\varepsilon\rangle}(du,\Omega) - \left[\eta_N^{\langle\delta,\varepsilon\rangle}\right]_d(du)\right) \\
&\le (N-\varepsilon-1)\left\{\left(\frac{1}{N-1} + 1 - \theta\right)\eta_d(du) + \theta\widehat{(\eta_d)}^{\langle\delta\rangle}(du) - \left[\eta_N^{\langle\delta,\varepsilon\rangle}\right]_d(du)\right\} \overset{(13.6),\,(13.7)}{=\!=\!=} \\
&(N-\varepsilon)\left[(1-\theta)\eta_d + \theta(1-\tau)\widehat{(\eta_d)}^{\langle\delta\rangle}(du)\right] - (N-\varepsilon-1)\left[\eta_N^{\langle\delta,\varepsilon\rangle}\right]_d(du) \overset{(13.9)}{=\!=} \left[\eta_N^{\langle\delta,\varepsilon\rangle}\right]_d(du).
\end{aligned}$$

Thus $\eta_N^{\langle\delta,\varepsilon\rangle} \in \mathbf{srpm}_{N-\varepsilon}(\Omega^2)$ by (7.2). □

LEMMA 13.3. *Let $\mu \in \mathfrak{N}(T,\Omega), \nu \in \mathfrak{N}(T,\Omega^2)$, and $\mu(t,du) \ge 0, \nu(t,dw) \ge 0, \mu(t,\Omega) \le 1, \nu(t,\Omega^2) \le 1$ a.a. $t \in T$. Then the maps $t \in T \mapsto \mu^{\langle\delta\rangle}(t) := [\mu(t)]^{\langle\delta\rangle} \in \mathbf{frm}(\Omega^2), t \mapsto \widehat{\mu}^{\langle\delta\rangle}(t) := \widehat{[\mu(t)]}^{\langle\delta\rangle} \in \mathbf{frm}(\Omega)$ and $t \mapsto \nu_d(t) := [\nu(t)]_d \in \mathbf{frm}(\Omega^2)$ are measurable.*

Proof. For $u\in\Omega, w\in\Omega^2, j=1,2,\dots$, we put $\omega^i_j(u):=\min\big\{j\cdot\mathrm{dist}\big(u;\Omega\setminus V_i\big);1\big\}$, $\omega_j(w):=\min\big\{[j\cdot\mathrm{dist}\big(w,D\big)]^{-1};1\big\}$, where D is the diagonal of Ω^2, $0^{-1}:=\infty$, and $V_i, i=1,\dots,k$ are the open sets from (13.2). Then $0\le\omega^i_j(u),\omega_j(w)\le 1$ for all u,w; $\omega^i_j(u)=0$ for all $u\notin V_i$, $\omega_j(w)=1$ for all $w\in D$, and $\omega^i_j(u)\to 1$ for all $u\in V_i, \omega_j(w)\to 0$ for all $w\notin D$ as $j\to\infty$. We also set $\omega^0_j(u):=0$. For any $\rho(\cdot)\in\mathbb{C}(\Omega^2), g(\cdot)\in\mathbb{C}(\Omega)$, we have by (13.4),

$$\sigma_\mu(t):=2\int_{\Omega^2}\rho(w)\mu^{\langle\delta\rangle}(t,dw)=\int_\Omega\big(\rho[u,\chi_\delta(u)]+\rho[\chi_\delta(u),u]\big)\mu(t,du)\overset{(13.2)}{=\!=}$$

$$\sum_{i=1}^k\int_{W_i}\underbrace{\big[\rho(u,v_i)+\rho(v_i,u)\big]}_{\rho_i(u)}\mu(t,du)=\sum_{i=1}^k\lim_{j\to\infty}\int_\Omega\omega^i_j(u)\Big[1-\max_{r=0,\dots,i-1}\omega^r_j(u)\Big]\rho_i(u)\mu(t,du)$$

$$\sigma_\nu(t):=\int_\Omega g(u)\nu_d(t,du)=\lim_{j\to\infty}\int_{\Omega^2}g(u_1)\omega_j(u_1,u_2)\nu(t,du_1,du_2).$$

The integrands in both final expressions are continuous. By the definition of $\mathfrak{N}(T,K)$, $K=\Omega,\Omega^2$, this means that the corresponding integrals are measurable functions of t. Hence so are their limits $\sigma_\mu(t)$ and $\sigma_\nu(t)$. This proves that the maps $\mu^{\langle\delta\rangle}$ and ν_d are measurable. The measurability of $\widehat{\mu}^{\langle\delta\rangle}$ is established likewise. □

By combining Lemmas 13.2 and 13.3 with (13.6) and (13.7), we arrive at the following corollary.

COROLLARY 13.4. *If $\nu\in\mathfrak{G}^s_N, N\in(1,\infty)$, $\varepsilon\in(0,N-1)$, then $\nu^{\langle\delta,\varepsilon\rangle}_N\in\mathfrak{G}^s_{N-\varepsilon}$.*

LEMMA 13.5. *For $\xi=[f(\cdot),\varphi(\cdot),a(\cdot)]\in\Xi_{(iv)}$, $N\in(1,\infty)$, and $\nu\in\mathfrak{G}^s_N$, we put $\widetilde{r}(t,x,w):=r(t,x,x,w)$, $r^{\blacktriangledown}(t,\xi,\nu,N,\delta):=\int_{\Omega^2}\widetilde{r}\,[t,x(t),w]\nu^{\langle\delta\rangle}_N(t,dw)$, where $r=f,\varphi$,*

$$\nu^{\langle\delta\rangle}_N:=\frac{1}{N-1}\left(\frac{1}{N}\left[(\nu_d)^{\langle\delta\rangle}+(N-1)\widehat{(\nu_d)}^{\langle\delta\rangle}\right]-\nu_d\right),\tag{13.10}$$

and $x(\cdot)$ is the solution of (5.4). *Let $J(\nu)$ denote the value of the cost functional from* (5.3). *Consider the constant k from* (iv) *in section* 2 *and the hemi-norm $\mathfrak{H}_k(\cdot)$ given by* (4.2). *Then*

$$\lim_{\varepsilon\to+0}\varepsilon^{-1}\left[J\left(\nu^{\langle\delta,\varepsilon\rangle}_N\right)-J(\nu)\right]=\sigma(\xi,\nu,N,\delta)+\int_T\varphi^{\blacktriangledown}(t,\xi,\nu,N,\delta)dt,\tag{13.11}$$

$$|\sigma(\xi,\nu,N,\delta)|\le|T|\mathfrak{H}_k(\xi)e^{\mathfrak{H}_k(\xi)|T|}\max_{t\in T}\left|\int_{t_0}^t f^{\blacktriangledown}(\theta,\xi,\nu,N,\delta)d\theta\right|,|T|:=t_1-t_0.\tag{13.12}$$

Proof. As follows from (13.6), (13.7), (13.10), $\frac{d}{d\varepsilon}\nu^{\langle\delta,\varepsilon\rangle}_N|_{\varepsilon=0}=\nu^{\langle\delta\rangle}_N$. Then (5.3), (5.4) imply via the standard arguments (see, e.g., the proof of Theorem II.4.11 in [22]) (13.11) with

$$\sigma(\xi,\nu,N,\delta):=\int_T\breve{\varphi}_\partial(t)y(t)dt,\quad\text{where}\quad y(t)=\int_{t_0}^t\breve{f}_\partial(\theta)y(\theta)d\theta+z(t),\tag{13.13}$$

$\breve{r}_\partial(t):=\int_{\Omega^2}\frac{\partial\widetilde{r}}{\partial x}[t,x(t),w]\nu(t,dw)$ for $r=\varphi,f$, and $z(t)$ is the integral embraced by $|\cdot|$ in (13.12). Due to Lemma 8.4, $|x(t)|\le k$ for all $t\in T$. So (4.2) yields $|\breve{r}_\partial(t)|\le c:=\mathfrak{H}_k(\xi), r=f,\varphi$. Thus $|y(t)|\le c\int_{t_0}^t|y(\theta)|d\theta+|z(t)|$ for all t, and Theorem II.4.4 [22] gives $|y(t)|\le|z(t)|+c\int_{t_0}^t e^{c(t-\theta)}|z(\theta)|d\theta\le e^{c(t_1-t_0)}\max_t|z(t)|$. Then (13.13) ⇒ (13.12). □

LEMMA 13.6. *We employ the notation from Lemmas* 13.1 *and* 13.5 *and put for* $r = f, \varphi$,

(13.14)
$$\alpha_r^k(\delta) := \max_{\substack{t\in T,|x|\le k,\\ u,v\in\Omega,|u-v|\le\delta}} \left\{\widetilde{r}(t,x,v,v) - \widetilde{r}(t,x,u,u)\right\}_r, \quad \{b\}_r := \begin{cases} b & \text{for } r=\varphi,\\ |b| & \text{for } r=f,\end{cases}$$

$$\beta_r^k(\delta) := \max_{t\in T,|x|\le k,u,v\in\Omega,\varkappa\delta\le|u-v|\le\delta} \max_{w=(u,v),(v,u)} \left\{\widetilde{r}(t,x,w) - \widetilde{r}(t,x,u,u)\right\}_r. \tag{13.15}$$

Suppose that $1 < N_- \le N \le N_+ < \infty$ *and denote* $\lambda_r^k(\delta, N_+) := 2\beta_r^k(\delta) + N_+\alpha_r^k(\delta)$. *Then*

$$\left\{r^{\blacktriangledown}(t,\xi,\nu,N,\delta)\right\}_r \le \lambda_r^k(\delta,N_+) \times \begin{cases} 1/2(N_--1)^{-2} & \text{if } r=f,\\ 1/2N_+^{-2} & \text{if } r=\varphi \text{ and } \lambda_\varphi^k(\delta,N_+)\le 0.\end{cases}$$

Proof. The definition of $r^{\blacktriangledown}(\cdot)$ given in Lemma 13.5 and (13.4), (13.5), and (13.10) imply that

$$\begin{aligned}\left\{r^{\blacktriangledown}(t,\xi,\nu,N,\delta)\right\}_r \le \int_\Omega \Big\{ &\frac{1}{2N(N-1)}\Big[\left\{\widetilde{r}[t,x(t),u,\chi_\delta(u)] - \widetilde{r}[t,x(t),u,u]\right\}_r\\ &+\left\{\widetilde{r}[t,x(t),\chi_\delta(u),u] - \widetilde{r}[t,x(t),u,u]\right\}_r\Big]\\ &+\frac{1}{2N}\Big[\left\{\widetilde{r}[t,x(t),\chi_\delta(u),\chi_\delta(u)] - \widetilde{r}[t,x(t),u,u]\right\}_r\Big]\Big\}\nu_d(t,du) \overset{(13.2),\,(13.14),\,(13.15)}{\le}\\ \frac{1}{2N(N-1)}\left[2\beta_r^k(\delta) + (N-1)\alpha_r^k(\delta)\right]&\int_\Omega \nu_d(t,du) \le \frac{1}{2N(N-1)}\lambda_r^k(\delta,N_+)\nu_d(t,\Omega).\end{aligned}$$

Here $\lambda_f^k(\delta,N_+) \ge 0$ by (13.14), (13.15), whereas (7.3), (7.2), and (5.2) yield $\nu_d(t,\Omega) = \nu(t,D) \le \nu(t,\Omega^2) \le 1$. Since $2N(N-1) \ge 2(N_--1)^2$, this completes the proof for $r := f$. Now let $r = \varphi$ and $\lambda_\varphi^k(\delta,N_+) \le 0$. The last relation from (7.2) gives

$$(N-1)[\nu(t,\Omega^2) - \nu_d(t,\Omega)] \le \nu_d(t,\Omega) \Rightarrow \nu_d(t,\Omega) \ge (N-1)/N\nu(t,\Omega^2) = (N-1)/N,$$
$$\left\{\varphi^{\blacktriangledown}(t,\xi,\nu,N,\delta)\right\}_r \le \frac{1}{2N^2}\lambda_\varphi^k(\delta,N_+) \le \frac{1}{2N_+^2}\lambda_\varphi^k(\delta,N_+). \qquad \square$$

LEMMA 13.7. *Given* $1 < N_- < N_+ < \infty$, *the following set is open and nonempty:*

$$\Xi[\delta,N_-,N_+] := \left\{\xi = [f(\cdot),\varphi(\cdot),a(\cdot)] \in \Xi_{(iv)} : \gamma_\xi(\delta,N_-,N_+) < 0\right\}, \quad \text{where} \tag{13.16}$$
$$\gamma_\xi(\delta,N_-,N_+) := \frac{|T|^2}{2(N_--1)^2}\lambda_f^k(\delta,N_+)\mathfrak{H}_k(\xi)e^{\mathfrak{H}_k(\xi)|T|} + \frac{|T|}{2N_+^2}\lambda_\varphi^k(\delta,N_+).$$

Whenever $\xi \in \Xi[\delta,N_-,N_+]$, *the limit optimal cost* $\mathfrak{J}_{opt}^{sam}[N,\xi]$ *defined in* (7.7) *strictly increases on the interval* $[N_-,N_+]$.

Proof. In the topology generated by the family of hemi-norms (4.2), the hemi-norm $\mathfrak{H}_k(\cdot)$ is evidently continuous. Due to (13.14), (13.15), so is the function $\xi \mapsto \lambda_r^k(\delta,N_+)$ for $r = f,\varphi$ and thus $\xi \mapsto \gamma_\xi(\delta,N_-,N_+)$. It follows that the set $\Xi[\delta,N_-,N_+]$ is open. For $f(\cdot) \equiv 0, a(\cdot) \equiv 0, \varphi(t,x_1,x_2,u,v) := -|u-v|, \xi := [f(\cdot),\varphi(\cdot),a(\cdot)]$,

we have by (13.14), (13.15), $\alpha^k_f(\delta)=\alpha^k_\varphi(\delta)=\beta^k_f(\delta)=0,\beta^k_\varphi(\delta)\le-\varkappa\delta<0$. So $\lambda^k_f(\delta,N_+)=0,\lambda^k_\varphi(\delta,N_+)<0$. Hence $\xi\in\Xi[\delta,N_-,N_+]$ by (13.16), i.e., $\Xi[\delta,N_-,N_+]\neq\emptyset$.

Now let $N\in(N_-,N_+]$, $\xi\in\Xi[\delta,N_-,N_+]$, and $J(\cdot)$ denote the functional from (5.3). Due to Theorem 7.2, there is an optimal control $\nu\in\mathfrak{G}^s_N$ in the problem (5.3), (5.4) with $\mathfrak{G}^s:=\mathfrak{G}^s_N$, and $J(\nu)=\mathcal{J}^{\rm sam}_{\rm opt}[N,\xi]$. By Corollary 13.4, $\nu^{\langle\delta,\varepsilon\rangle}_N\in\mathfrak{G}^s_{N-\varepsilon}$ for all $\varepsilon>0,\varepsilon\approx0$. So

$$\overline{\lim_{\varepsilon\to+0}}\,\varepsilon^{-1}\left[\mathcal{J}^{\rm sam}_{\rm opt}(N-\varepsilon,\xi)-\mathcal{J}^{\rm sam}_{\rm opt}(N,\xi)\right]\le\lim_{\varepsilon\to+0}\varepsilon^{-1}\left[J\left(\nu^{\langle\delta,\varepsilon\rangle}_N\right)-J(\nu)\right]. \tag{13.17}$$

In (13.14), (13.15), $\nmid b\nmid_f\ge0$. Hence $\lambda^k_f(\delta,N_+)\ge0$, and the inequality from (13.16) yields $\lambda^k_\varphi(\delta,N_+)<0$. This permits us to employ Lemma 13.6. Combining it with (13.11) and (13.12) shows that the right-hand side of (13.17) does not exceed $\gamma_\xi(\delta,N_-,N_+)$. So it is negative by (13.16). At the same time, the function $\mathcal{J}^{\rm sam}_{\rm opt}[N,\xi]$ of N does not decrease by Remark 7.5. So if it does not strictly increase on $[N_-,N_+]$, there would be a nontrivial interval $[N'_-,N'_+]\subset[N_-,N_+]$ on which this function would be constant in violation of (13.17) with $N\in(N'_-,N'_+)$. Thus we see that $\mathcal{J}^{\rm sam}_{\rm opt}[N,\xi]$ strictly increases on $[N_-,N_+]$. □

LEMMA 13.8. *If the set Ω is connected, the following set is open and dense in $\Xi_{(iv)}$:*

$$\Xi[N_-,N_+]:=\bigcup_{\delta\in(0,\delta_0]}\Xi[\delta,N_-,N_+]. \tag{13.18}$$

Proof. The set (13.18) is open as the union of the open sets. Let $\xi=[f(\cdot),\varphi(\cdot),a(\cdot)]\in\Xi_{(iv)}$. Define $\gamma(\delta)$ to be the right-hand side of the last formula from (13.16), where $\mathfrak{H}_k(\xi)+1$ is put in place of $\mathfrak{H}_k(\xi)$. Due to (13.14) and (13.15), $\gamma(\delta)\to0$ as $\delta\to+0$. For $z=(t,x_1,x_2)\in T\times\mathbb{R}^n\times\mathbb{R}^n$ and $u,v\in\Omega$, we put $\varphi^\delta(z,u,v):=\varphi(z,u,v)-|T|N^2_+[\gamma(\delta)+\delta]\min\left\{\frac{1}{\varkappa\delta}|u-v|,1\right\}$, $\xi^\delta:=[f(\cdot),\varphi^\delta(\cdot),a(\cdot)]$, where $\varkappa$ is taken from Lemma 13.1. By (13.14), (13.15), $\alpha^k_{\varphi^\delta}(\delta)=\alpha^k_\varphi(\delta)$, $\beta^k_{\varphi^\delta}(\delta)\le\beta^k_\varphi(\delta)-|T|N^2_+[\gamma(\delta)+\delta]$, and so $\lambda^k_{\varphi^\delta}(\delta,N_+)\le\lambda^k_\varphi(\delta,N_+)-2|T|N^2_+[\gamma(\delta)+\delta]$. Furthermore, $\xi^\delta\to\xi$ as $\delta\to+0$ owing to (4.2) and hence $\mathfrak{H}_k(\xi^\delta)\le\mathfrak{H}_k(\xi^\delta-\xi)+\mathfrak{H}_k(\xi)\le\mathfrak{H}_k(\xi)+1$ for $\delta\approx0$. This, the definition of $\gamma(\cdot)$, and (13.16) imply that $\gamma_{\xi^\delta}(\delta,N_-,N_+)\le\gamma(\delta)-[\gamma(\delta)+\delta]<0$. Thus $\xi^\delta\in\Xi[\delta,N_-,N_+]\subset\Xi[N_-,N_+]$ for all $\delta\approx0$. It remains to be noted that $\xi^\delta\to\xi$ as $\delta\to0$. □

Proof of Theorem 7.4. By Lemma 13.8, the set $\Upsilon:=\bigcap_{N_-<N_+}\Xi[N_-,N_+]$ is residual in $\Xi_{(iv)}$ if the set Ω is connected. For $\xi\in\Upsilon$, (ii) of Theorem 7.4 is immediate from Lemma 13.7. To prove (i), we pick N that exceeds the second $\overline{\lim}$ from (7.8) and consider the set $\mathfrak{R}^{\rm sam*}_\Delta$ defined by (7.1), where $q(\Delta)$ is replaced by $N\Delta$. Since $q(\Delta)\le N\Delta$ for all $\Delta\approx0$, then $\mathfrak{R}^{\rm sam}_\Delta\supset\mathfrak{R}^{\rm sam*}_\Delta$. So $\mathfrak{I}_{opt}(\Delta)\le\mathfrak{I}^*_{opt}(\Delta)$, where $\mathfrak{I}_{opt}(\Delta)$ and $\mathfrak{I}^*_{opt}(\Delta)$ are the infima of the cost functionals in problem (2.1)–(2.4) related to the classes $\mathfrak{R}^{\rm sam}_\Delta$ and $\mathfrak{R}^{\rm sam*}_\Delta$, respectively. With regard to (7.7) and (ii) of the theorem, letting $\Delta\to+0$ shows that the first $\overline{\lim}$ from (7.8) does not exceed $\mathcal{J}^{\rm sam}_{\rm opt}[N,\xi]$. Then (i) of Theorem 7.4 follows from (ii) since $\mathfrak{I}_{opt}(0)=\mathcal{J}^{\rm sam}_{\rm opt}[\infty,\xi]$ due to Remark 7.3 and (i) of Theorem 7.2.

The second and third claims of Theorem 7.4 follow from Lemmas 13.7 and 13.8. □

Proof of Theorem 4.3. This is stated by the results from Theorem 7.4, Remarks 7.3 and 7.5, and (i) of Theorem 7.2. □

14. Proofs of the statements from section 6.

Proof of Theorem 6.3. Consider the problem (2.1)–(2.4) with $f(t,x_1,x_2,u_1,u_2) := \overline{f}(t,x_1,u_1), a(t) := a_0, \varphi(t,x_1,x_2,u_1,u_2) := \overline{\varphi}(t,x_1,u_1) - B|u_1-u_2|$, where B will be specified later. This problem meets the requirements (i)–(iv) from section 2 and its natural "limit" case equals the original problem (2.5), (2.6) (i.e., (2.7) is true). By invoking Theorem 5.1 and employing the element (13.1), we get

$$\lim_{\Delta\to+0}\mathfrak{I}_{opt}(\Delta) \le \int_T dt \int_{\Omega\times\Omega} \left(\overline{\varphi}[t,x(t),u'] - B|u'-u''|\right)\,\nu(t,du',du'')$$
$$= \frac{1}{2}\int_T \left(\overline{\varphi}[t,x(t),u_1] + \overline{\varphi}[t,x(t),u_2]\right)\,dt - B(t_1-t_0)|u_1-u_2|.$$

Here $x(\cdot)$ is the solution of (5.4). Due to Lemma 8.4, $|x(t)| \le k$ for all $t \in [t_0,t_1]$. Hence

$$\lim_{\Delta\to+0}\mathfrak{I}_{opt}(\Delta) \le \overbrace{\max_{t\in[t_0,t_1],u\in\Omega,|x|\le k}|\overline{\varphi}(t,x,u)|}^{\varkappa} - B\overbrace{(t_1-t_0)|u_1-u_2|}^{\delta}.$$

By picking $B > \delta^{-1}(\varkappa - \mathfrak{I}_{opt}(0) + A)$, we arrive at (6.2) and complete the proof. □

The proof of Lemma 6.1 is prefaced by the following technical fact.

LEMMA 14.1. *Given $u(\cdot) \in \mathbb{L}_\infty([t_0-1,t_1] \to \Omega)$ and $\Delta \in [0,1]$, the symbol ν_Δ stands for the element given by* (5.8). *The function $\Delta \mapsto \nu_\Delta \in \mathfrak{G}$ is continuous.*

Proof. By [22, Theorem IV.2.4], it suffices to show that for any $\zeta(\cdot) \in \mathbb{C}(T), \rho(\cdot) \in \mathbb{Y} := \mathbb{C}(\Omega^2)$, and $\Delta \in [0,1]$,

$$\Lambda_{\Delta',\Delta}[\rho(\cdot)] := \int_T \zeta(t)\left\{\rho[u(t),u(t-\Delta')] - \rho[u(t),u(t-\Delta)]\right\}\,dt \to 0 \tag{14.1}$$

as $\Delta' \to \Delta$. Fix Δ and $\zeta(\cdot)$. All $\rho(\cdot) \in \mathbb{Y}$ such that (14.1) is true and constitute a linear subspace $\mathbb{Y}_\rho$ of $\mathbb{Y}$. We are going to show that it is closed. Indeed, let $\rho_*(\cdot) \in \overline{\mathbb{Y}_\rho}$. For any $\varepsilon > 0$, there exists $\rho(\cdot) \in \mathbb{Y}_\rho$ such that $\varkappa := \max_{\omega\in\Omega^2}|\rho(\omega)-\rho_*(\omega)| \le \varepsilon$. Then

$$|\Lambda_{\Delta',\Delta}[\rho_*(\cdot)]| \le |\Lambda_{\Delta',\Delta}[\rho_*(\cdot)-\rho(\cdot)]| + |\Lambda_{\Delta',\Delta}[\rho(\cdot)]| \le 2\varkappa\int_T|\zeta(t)|\,dt + |\Lambda_{\Delta',\Delta}[\rho(\cdot)]|;$$
$$\overline{\lim_{\Delta'\to\Delta}}\,|\Lambda_{\Delta',\Delta}[\rho_*(\cdot)]| \le 2\varepsilon\int_T|\zeta(t)|\,dt + \lim_{\Delta'\to\Delta}|\Lambda_{\Delta',\Delta}[\rho(\cdot)]| = 2\varepsilon\int_T|\zeta(t)|\,dt.$$

Letting $\varepsilon \to +0$ shows that $\rho_*(\cdot) \in \mathbb{Y}_\rho$, i.e., $\mathbb{Y}_\rho = \overline{\mathbb{Y}_\rho}$. So it suffices to prove (14.1) for $\rho(\cdot) \in \mathcal{E}$, where $\mathcal{E}$ is an arbitrary set whose linear hull is dense in $\mathbb{Y}$. This permits us to focus on the functions $\rho(\cdot)$ of the form $\rho(u',u'') = \rho_1(u')\rho_2(u''), \rho_1(\cdot),\rho_2(\cdot) \in \mathbb{C}(\Omega)$. For them,

$$|\Lambda_{\Delta',\Delta}[\rho(\cdot)]| = \left|\int_T \underbrace{\zeta(t)\rho_1[u(t)]}_{\chi(t)}\left\{\rho_2[u(t-\Delta')] - \rho_2[u(t-\Delta)]\right\}\,dt\right|$$
$$\le \operatorname{ess\,sup}_{t\in T}|\chi(t)|\int_T|r(t-\Delta')-r(t-\Delta)|\,dt, \to 0 \quad \text{as} \quad \Delta' \to \Delta,$$

where $r(t) := \rho_2[u(t)]$. Thus (14.1) is valid. □

Proof of Lemma 6.1. We denote $u_\Delta(\cdot) := u|_{[t_0-\Delta,t_1]}(\cdot) \in \mathfrak{R}_\Delta$, $\mu(\Delta) := I_\Delta(u_\Delta) - \mathfrak{I}_{opt}(\Delta)$, and employ the notation ν_Δ introduced in Lemma 14.1. Then $I_\Delta(u_\Delta) = \mathfrak{I}_{opt}(\Delta) + \mu(\Delta)$, i.e., the control u_Δ is $\mu(\Delta)$-suboptimal. Furthermore, thanks to Lemmas 8.4 and 14.1,

$$\lim_{\Delta\to 0}\mu(\Delta) = \lim_{\Delta\to 0} I_\Delta(u_\Delta) - \lim_{\Delta\to 0}\mathfrak{I}_{opt}(\Delta) = I(\nu_0) - \lim_{\Delta\to 0}\mathfrak{I}_{opt}(\Delta) = \mathfrak{I}_{opt}(0) - \lim_{\Delta\to 0}\mathfrak{I}_{opt}(\Delta),$$

which implies (6.1) and completes the proof. □

REFERENCES

[1] A. Arutyunov and M. Mardanov, *On the theory of the maximum principle in problems with delays*, Differential Equations, 29 (1989), pp. 2048–2058, 2205.

[2] H. Banks, *Necessary conditions for control problems with variable time lags*, SIAM J. Control, 6 (1968), pp. 9–47.

[3] R. Bellman and K. Cooke, *On the limit of solutions of differential-difference equations as the retardation approached zero*, Proc. Nat. Acad. Sci. U.S.A., 45 (1959), pp. 1026–1028.

[4] M. Bennati, *Wellposedness by perturbation in optimization problems and metric characterization*, Rend. Mat. Appl., 16 (1996), pp. 613–623.

[5] G. Bunce, *A maximum principle for time-lag control problems with bounded state*, J. Optimization Theory Appl., 22 (1977), pp. 563–606.

[6] F. Clarke and G. Watkins, *Necessary conditions, controllability and the value function for differential-difference inclusions*, Nonlinear Anal., 10 (1986), pp. 1155–1179.

[7] F. Clarke and P. Wolenski, *The sensitivity of optimal control problems to time delay*, SIAM J. Control Optim., 29 (1991), pp. 1176–1215.

[8] R. Driver, *Some harmless delays*, in Delay and Functional Differential Equations and Their Applications, K. Schmitt, ed., Academic Press, New York, 1972, pp. 103–119.

[9] R. Edwards, *Functional Analysis. Theory and Applications*, Holt, Rinehart and Winston, New York, London, 1965.

[10] J. Hale, *Theory of Functional Differential Equations*, 2nd ed., Springer-Verlag, New York, 1977.

[11] H. Ishii and B. Francis, *Limited Data Rate in Control Systems with Networks*, Lecture Notes in Control and Information Sciences 275, M. Thoma and M. Morari, eds., Springer-Verlag, Berlin, 2002.

[12] G. Kharatishvili, *Maximum principle in the theory of optimum time-delay processes*, Dokl. Akad. Nauk USSR, 136 (1961), pp. 39–42.

[13] G. Kharatishvili and T. Tadumadze, *Nonlinear optimal control systems with varying delays*, Matem. Sbornik, 107 (1973), pp. 613–633 (in Russian).

[14] A. Matveev, *Optimal control problems with delays of general form and phase constraints*, Math. USSR Izv., 33 (1989), pp. 521–552.

[15] A. Matveev, *On nonequivalence of the pointwise and integral maximum principles for systems with delays in the controls*, St. Petersburg Math. J., 4 (1993), pp. 749–775.

[16] M. Oğuztoreli, *Time-Lag Control Systems*, Academic Press, New York-London, 1966.

[17] L. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mishchenko, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[18] J. Rosenblueth, *Strongly and weakly relaxed controls for time delay systems*, SIAM J. Control Optim., 30 (1992), pp. 856–866.

[19] J. Rosenblueth and R. Vinter, *Relaxation procedures for time delay systems*, J. Math. Anal. Appl., 162 (1991), pp. 542–563.

[20] S. Sugiyama, *Continuity properties on the retardation in the theory of difference-differential equations*, Proc. Japan Acad., 37 (1961), pp. 179–182.

[21] J. Warga, *The reduction of certain control problems to an "ordinary differential" type*, SIAM Rev., 10 (1968), pp. 219–222.

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

[23] J. Warga and Q. Zhu, *A proper relaxation of shifted and delayed controls*, J. Math. Anal. Appl., 169 (1992), pp. 546–561.

[24] T. Zolezzi, *Extended well-posedness of optimization problems*, J. Optimization Theory Appl., 91 (1996), pp. 257–266.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1787–1811

CONVEX OPTIMAL CONTROL PROBLEMS WITH SMOOTH HAMILTONIANS*

RAFAL GOEBEL†

Abstract. Optimal control problems with convex costs, for which Hamiltonians have Lipschitz continuous gradients, are considered. Examples of such problems, including extensions of the linear-quadratic regulator with hard and possibly state-dependent control constraints, and piecewise linear-quadratic penalties are given. Lipschitz continuous differentiability and strong convexity of the terminal cost are shown to be inherited by the value function, leading to Lipschitz continuity of the optimal feedback. With no regularity assumptions on the limiting problem, epi-convergence of costs, which can be equivalently described by pointwise convergence of Hamiltonians, is shown to guarantee epi-convergence of value functions. Resulting schemes of approximating any concave-convex Hamiltonian by continuously differentiable ones are displayed. Auxiliary results about existence and stability of saddle points of quadratic functions over polyhedral sets are also proved. Tools used are based on duality theory of convex and saddle functions.

Key words. optimal control, differentiable Hamiltonian, convex value function, optimal feedback regularity, conjugate duality, epi-convergence, piecewise linear-quadratic function, saddle function

AMS subject classifications. 49N60, 49N10, 49M29, 90C47

DOI. 10.1137/S0363012902411581

1. Introduction.

Given a point $(\tau, \xi) \in (-\infty, T] \times \mathbb{R}^n$, a *terminal cost* $g : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ and a *Lagrangian* $L : \mathbb{R}^{2n} \mapsto \overline{\mathbb{R}}$, consider the generalized problem of Bolza:

$$\mathcal{P}(\tau, \xi): \quad \text{minimize } \int_{\tau}^{T} L(x(t), \dot{x}(t))\, dt + g(x(T)) \text{ subject to } x(\tau) = \xi, \tag{1}$$

with the minimization carried out over all absolutely continuous arcs $x : [\tau, T] \mapsto \mathbb{R}^n$. While it is well known that a smooth Lagrangian need not lead to a regular (maximized) *Hamiltonian*, which is defined by

$$H(x, y) = \sup_{v \in \mathbb{R}^n} \{y \cdot v - L(x, v)\}, \tag{2}$$

it is less appreciated that nonsmooth and infinite-valued L may give rise to a smooth H. We explore this fact here, focusing on problems with convex g and L, and with Hamiltonians for which ∇H is Lipschitz continuous.

Optimal control problems with explicit linear dynamics, hard and possibly state-dependent control constraints, and state and control penalties can be reformulated in Bolza format; see Clarke [10] or Rockafellar [18]. In section 2 we show that a broad range of optimal control problems, including various extensions of the classical linear-quadratic regulator, can lead to a smooth Hamiltonian. This makes the results of section 3 applicable to the control framework.

*Received by the editors July 18, 2002; accepted for publication (in revised form) May 31, 2004; published electronically March 22, 2005. Research carried out at the Department of Mathematics at the University of Washington, the Centre for Experimental and Constructive Mathematics at Simon Fraser University, and the Department of Mathematics at the University of British Columbia.
http://www.siam.org/journals/sicon/43-5/41158.html
†Center for Control Engineering and Computation, ECE, University of California, Santa Barbara, CA 93106-9650 (rafal@ece.ucsb.edu).

Section 3 studies regularity of the *value function* $V : (-\infty, T] \times \mathbb{R}^n \mapsto \overline{\mathbb{R}}$, defined as the optimal value in $\mathcal{P}(\tau, \xi)$ parameterized by the initial condition. Lipschitz continuity of ∇g and ∇H is shown to lead to Lipschitz ∇V; explicit bounds on the constants are given. We stress that no smoothness or even finiteness assumptions are made on L. For comparison, in a nonconvex setting, if the method of characteristics associated with the Hamilton–Jacobi equation has no shocks (in our setting, this automatically holds; see Goebel [14]), the value function inherits continuous differentiability from that of the terminal cost, under further regularity assumptions on L; see Byrnes and Frankowska [7] and also Caroff and Frankowska [8]. We note that while we work with continuously differentiable Hamiltonians, we do not require them to be C^2. This raises an obstacle to Riccati-like descriptions of V as given by Byrnes [6] and Caroff and Frankowska [9] but allows for treatment of problems discussed in section 2 (for those, hard constraints or piecewise linear-quadratic penalties exclude C^2 smoothness of the Hamiltonian).

Our interest in Lipschitz continuity of ∇V comes from the role the gradient plays in constructing optimal feedback. With the regularity of H and V as just mentioned, the adjoint variable to an optimal arc $x(t)$ is just $-\nabla V(t, x(t))$, and the resulting optimal feedback mapping is continuous and Lipschitz in the state variable. Consequently, the classical differential equation tools and existence and uniqueness results apply. This is not the case for the general convex but nonsmooth setting—there, the resulting set-valued feedback may be highly irregular, even for piecewise linear-quadratic costs; see Goebel [14].

In section 4 we show that regular Bolza problems—those with Lipschitz ∇g and ∇H—can approximate any convex problem fitting our mild growth conditions. The approximations are explicit and, together with direct proofs in section 3, they should yield insights to numerical implementation of the method of characteristics. The approximations rely on a more general result concluding the convergence of value functions, defined by any converging to g and L sequences of initial costs and Lagrangians. As the functions in question need not be finite, we rely on the concept of epi-convergence. Its extensions to infinite dimensions, where various topologies have to be considered, have been used to study control problems; see Buttazzo and Dal Maso [5] and Briani [4]. These works, while not requiring full convexity, had stricter growth assumptions and finite cost functions, and dealt, respectively, with convergence of optimal solutions and pointwise convergence of value functions. Moreover, methods used here are significantly different; we employ a dual problem leading to a dual value function, as described by Rockafellar and Wolenski [27]. The symmetry between the primal and dual problem, and the fact that epi-convergence is preserved by convex conjugacy (vaguely speaking, the "lower half" of epi-convergence dualizes to the "upper" and vice-versa), requires us to show just one side (the easier one) of epi-convergence. A similar idea was employed by Joly and Thelin [17] in the study of convex integral functionals; here we keep to a minimum the discussion of such issues, preferring to work with functions on finite-dimensional spaces.

Some of our results are most conveniently handled with the tools related to conjugacy and epi/hypo-convergence of saddle functions; see, respectively, chapters 33–37 in Rockafellar [19], Attouch and Wets [2], and Attouch, Azé, and Wets [1]. We present the necessary background in section 5. In particular, our results on finiteness and differentiability of piecewise linear-quadratic Hamiltonians are closely related to existence and uniqueness of saddle points of an auxiliary quadratic function defined on a product of polyhedral sets. Such a function also appears as a Lagrangian in ex-

tended linear-quadratic programming; see Rockafellar [25]. (In convex optimization, Lagrangians are saddle functions used, in particular, to express optimality conditions.)

2. Extended piecewise linear-quadratic optimal control. In this section we illustrate that control problems with constraints and nondifferentiable costs can possess Hamiltonians with desirable smoothness properties. Let us start with the following example.

Example 2.1 (separable smooth Hamiltonian). Suppose that $L(x,v) = k(x) + l(v)$, and l is a convex function. Then the Hamiltonian $H(x,y)$ is differentiable and ∇H is (globally) Lipschitz continuous if and only if k has this property and l is strongly convex (that is, $v \mapsto l(v) - \rho\|v\|^2$ is convex for some $\rho > 0$). Indeed, $H(x,y) = -k(x) + l^*(y)$, where $l^*(y) = \sup_v \{y \cdot v - l(v)\}$ is the convex function conjugate to l. The statement about differentiability of l^*, and the bound $(2\rho)^{-1}$ on its Lipschitz constant, can be found in [26, Proposition 12.60]. Strongly convex functions include functions of the form $l(v) = v \cdot Pv$ for $v \in C$ while $l(v) = +\infty$ for $v \notin C$, where P is a symmetric positive definite matrix and C is any convex set, but the (piecewise) quadratic structure is not necessary. For example, the "barrier function" $l(v) = -\log(1 - |v|)$ for $v \in (-1,1)$, $l(v) = +\infty$ otherwise, is strongly convex (note the nondifferentiability at the origin), we have $l^*(y) = 0$ for $y \in [-1,1]$, $l^*(y) = |y| - \log|y| - 1$ otherwise, and l^* has a Lipschitz continuous gradient.

In the remainder of this section, we discuss control problems with explicit mention of controls, dynamics, constraints, and penalties. Translating such problems to the generalized format of Bolza (1) is possible; see Clarke [10] for a general exposition or Rockafellar [18] for details in the convex case. This enables the translation of results of sections 3 and 4 to the control setting. As finiteness of the Hamiltonian and of the value function implies that an optimal arc $x(\cdot)$ has a bounded derivative—in the control setting below, $u(\cdot)$ is bounded—the potential discrepancy between minimizing over absolutely continuous arcs in $\mathcal{P}(\tau,\xi)$ and over L^2 controls in the linear-quadratic regulator is avoided in most cases under discussion.

Separable Hamiltonians of Example 2.1, and their biaffine perturbations given by $H(x,y) = y \cdot Ax - k(x) + l^*(y)$, appear, for example, in the linear-quadratic regulator with control constraints of the type $u(t) \in U$. However, state-dependent constraints $u(t) \le Cx(t) + d$ or mixed control and state penalties call for the analysis of a more general class of Hamiltonians.

Given vectors p, q; matrices A, B, C, D, P, Q; and sets U, V of appropriate dimensions, consider the following control problem $\mathcal{C}(\tau,\xi)$:

$$\begin{aligned} &\min \int_\tau^T \left[p \cdot u(t) + \frac{1}{2} u(t) \cdot Pu(t) + \rho_{V,Q}(q - Cx(t) - Du(t)) \right] dt + g(x(T)) \\ &\text{s.t.} \qquad \dot{x}(t) = Ax(t) + Bu(t), \quad u(t) \in U \text{ a.e.}, \quad x(\tau) = \xi, \end{aligned} \tag{3}$$

with the minimization carried out over all integrable controls $u : [\tau, T] \mapsto \mathbb{R}^k$. The convex and possibly infinite-valued penalty function $\rho_{V,Q}(\cdot)$ is given by

$$\rho_{V,Q}(s) = \sup_{v \in V} \left\{ s \cdot v - \frac{1}{2} v \cdot Qv \right\}. \tag{4}$$

The key assumptions, guaranteeing not only the convex structure of the problem, but also the piecewise linear-quadratic structure of the resulting Hamiltonian, is stated below. We recall that a set is polyhedral if it is the intersection of finitely many

closed half-spaces; consequently, a polyhedral set is always closed and convex (but not necessarily bounded).

$$\begin{array}{c}\text{Matrices } P \text{ and } Q \text{ are symmetric positive semidefinite.}\\ \text{Sets } U \text{ and } V \text{ are nonempty and polyhedral.}\end{array} \tag{5}$$

Such *extended piecewise linear-quadratic optimal control* format was proposed by Rockafellar [22]. Therein, optimality conditions taking advantage of duality were stated. Their minimax form (related to the structure of the Hamiltonian as outlined in Example 5.1 and the surrounding discussion) facilitates the use of various primal-dual optimization methods to discretized problems; see Rockafellar and Zhu [28], Wright [29], and Zhu [30].

Here, we begin by describing when the control problem (3) fits the convex duality framework of Rockafellar and Wolenski [27], we call upon some of their results in later sections. The Hamiltonian for $\mathcal{C}(\tau,\xi)$ (see Rockafellar [23] or apply (2) to the Lagrangian (11)) is

$$H(x,y) = y\cdot Ax + J^*(B^*y, Cx), \tag{6}$$

where the function J^*, convex in a and concave in b, is given by

$$J^*(a,b) = \sup_{u\in U}\inf_{v\in V}\left\{a\cdot u + b\cdot v - p\cdot u - \frac{1}{2}u\cdot Pu + q\cdot v + \frac{1}{2}v\cdot Qv + v\cdot Du\right\}. \tag{7}$$

Here and in what follows, B^* denotes the transpose of B. The Hamiltonian (and the Lagrangian (11)) are piecewise linear-quadratic: their effective domains are unions of finitely many polyhedral sets, relative to each of which the functions are linear-quadratic (Goebel [12]). Goebel and Rockafellar [15] showed that if a piecewise linear-quadratic Hamiltonian is finite, the control problem fits the framework of [27]. A particular consequence of such a structure of the Hamiltonian, shown in [15], is that the knowledge of $V(\overline{\tau},\cdot)$ at any particular $\overline{\tau}\in(-\infty,T]$ determines V (uniquely) for all times $\tau\in(-\infty,T]$.

In our setting, the finiteness of J^*, which implies that of the Hamiltonian, is described by the following result. For a given set S, the *recession cone* S^∞ consists of all z such that $S+z\subset S$, while for a cone K, the polar cone K^* is $\{w \mid w\cdot z\le 0 \text{ for all } z\in K\}$.

THEOREM 2.2 (finiteness of J^*). *Assume that* (5) *holds. Then, the function J^* is finite if and only if the following is satisfied:*

$$\begin{cases} U^\infty\cap\ker P\cap(-D^*V^\infty)^* = \{0\},\\ V^\infty\cap\ker Q\cap(DU^\infty)^* = \{0\}.\end{cases} \tag{8}$$

Above, $(DU^\infty)^* = \{w \mid D^*w\in U^{\infty *}\}$ and $(-D^*V^\infty) = \{z \mid -Dz\in V^{\infty *}\}$, this comes directly from the definitions. The proof of Theorem 2.2, as well as that of Theorem 2.4, requires some notions of saddle function theory. We present them and the proofs in section 5. Note that if D is the zero matrix (which excludes many modeling options), the function J^* is separable: $J^*(a,b) = \sup_{u\in U}\{a\cdot u - \frac{1}{2}u\cdot Pu\} - \sup_{v\in V}\{-b\cdot v - \frac{1}{2}v\cdot Qv\}$, and (8) reduce to known conditions on recession cones and kernels, we mention them in the discussion preceding Example 3.8.

COROLLARY 2.3. *Assume that* (5) *holds.*

(a) *If U is a bounded set, J^* is finite if and only if $V^\infty\cap\ker Q = \{0\}$ (and this holds in particular when V is bounded or Q is positive definite).*

(b) *When sets U and V are cones, J^* is finite if and only if*

$$\begin{cases} U \cap \ker P \cap (-D^*V)^* = \{0\}, \\ V \cap \ker Q \cap (DU)^* = \{0\}. \end{cases}$$

Arguments of Example 2.1 imply that in the separable case, as described before Corollary 2.3, positive definiteness of P and Q is equivalent to the differentiability of J^*. Below, we give a sufficient condition for differentiability, applicable to cases where $D \neq 0$ and not requiring the positive definiteness of P and Q. A somewhat extreme example, showing that this last property is not necessary, is as follows. For a and b one-dimensional, consider J^* with $p = q = P = Q = 0$, $D = 1$, and $U = V = \mathbb{R}$. Direct calculation shows that $J^*(a, b) = ab$.

THEOREM 2.4 (differentiability of J^*). *Assume that the following condition holds:*

$$\begin{cases} \ker P \cap [D^*(V^\infty \cap -V^\infty)]^\perp = \{0\}, \\ \ker Q \cap [D(U^\infty \cap -U^\infty)]^\perp = \{0\}. \end{cases} \tag{9}$$

Then J^* is differentiable and ∇J^* is Lipschitz continuous.

Lipschitz continuity of ∇J^*, while guaranteed by the proof, is automatic in the presence of differentiability of J^*. This is thanks to the piecewise linear-quadratic structure; if J^* is differentiable, then ∇J^* is piecewise affine (and there is finitely many pieces). The piecewise linear-quadratic structure furthermore implies that J^* is not C^2, unless it is in fact quadratic (and this excludes any hard constraints or piecewise linear-quadratic penalties in the underlying problem).

In the remainder of this section, we illustrate the modeling capabilities of the extended piecewise linear-quadratic control, and use Theorem 2.4 to conclude the differentiability of the Hamiltonian for various extensions of the linear-quadratic regulator. Computational methods for such problems in discrete time are of great interest in the engineering literature; see Bemporad et al. [3] and the references therein.

Given symmetric positive semidefinite matrices E and G and a symmetric and positive definite F, this classical problem is as follows:

$$\begin{aligned} &\min \int_\tau^T \frac{1}{2}\left(x(t) \cdot Ex(t) + u(t) \cdot Fu(t)\right) dt + \frac{1}{2} x(T) \cdot Gx(T), \\ &\text{s.t.} \qquad \dot{x}(t) = Ax(t) + Bu(t), \quad x(\tau) = \xi. \end{aligned} \tag{10}$$

Minimization is carried out over all L^2 controls $u(\cdot)$ on $[\tau, T]$ (optimal controls turn out to be bounded, and in fact continuous). The value function for (10) is $V(\tau, \xi) = \frac{1}{2}\xi \cdot S(\tau)\xi$, where the matrix $S(\cdot)$ solves the associated Riccati equation, the Hamiltonian is quadratic, and the optimal feedback is linear in the state. Results of section 3 will show that while constraints and penalties destroy the linear structure, the optimal feedback may still be Lipschitz continuous. Here, we focus on the regularity of the Hamiltonian.

The linear-quadratic regulator can of course be cast in the format (3), by taking

$$P = F,\ Q = I,\ U = \mathbb{R}^k,\ V = \mathbb{R}^n,\ C = \sqrt{E},\ D = 0,\ p = 0,\ q = 0.$$

Indeed, we obtain $\rho_{V,Q}(q - Cu - Dv) = \sup_v\{(-\sqrt{E}u) \cdot v - \frac{1}{2}v \cdot v\} = \frac{1}{2}u \cdot \sqrt{E}^* \sqrt{E}u$. It can be easily verified that conditions (8) and (9) are (obviously) satisfied.

Example 2.5 (fixed control constraints). A linear-quadratic regulator with a constraint $u(t) \in U$, for a nonempty polyhedral set U, certainly fits the format (3).

Thanks to the positive definiteness of $P = F$ and $Q = I$, conditions (8) and (9) hold, and thus the Hamiltonian is finite and differentiable (if U is bounded, the Hamiltonian remains finite but not differentiable if the matrix F is just positive-semidefinite). Direct calculation yields $J^*(a,b) = \rho_{U,F}(a) - \frac{1}{2}\|b\|^2$, and thus the Hamiltonian is $H(x,y) = y \cdot Ax - \frac{1}{2}x \cdot Ex + \rho_{U,F}(B^*y)$. Note that H is not C^2.

Example 2.6 (state-dependent inequality constraints on controls). Consider (10) with the following constraint on the control:

$$u(t) \le C_0 x(t) - q_0,$$

for some matrix C_0. Taking $U = \mathbb{R}^k,\ V = \mathbb{R}^n \times \mathbb{R}^k_+,\ P = F,\ p = 0$, and

$$Q = \begin{bmatrix} I_{n\times n} & 0_{n\times k} \\ 0_{k\times n} & 0_{k\times k} \end{bmatrix},\ q = \begin{bmatrix} 0_{n\times n} \\ q_0 \end{bmatrix},\ C = \begin{bmatrix} \sqrt{E} \\ C_0 \end{bmatrix},\ D = \begin{bmatrix} 0_{n\times k} \\ -I_{k\times k} \end{bmatrix},$$

where 0_n is a zero vector in $\mathbb{R}^n$, $0_{n\times k}$ is the zero matrix of appropriate dimension, etc., casts the problem in the framework of (3). We get, for $s = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$ with $s_1 \in \mathbb{R}^n$, $s_2 \in \mathbb{R}^k$,

$$\begin{aligned}\rho_{V,Q}(s) &= \sup_{v\in V}\left\{s\cdot v - \frac{1}{2}v\cdot Qv\right\} = \sup_{v_1\in\mathbb{R}^n, v_2\in\mathbb{R}^k_+}\left\{s_1\cdot v_1 + s_2\cdot v_2 - \frac{1}{2}v_1\cdot v_1\right\} \\ &= \sup_{v_1\in\mathbb{R}^n}\left\{s_1\cdot v_1 - \frac{1}{2}v_1\cdot v_1\right\} + \sup_{v_2\in\mathbb{R}^k_+}\{s_2\cdot v_2\} = \frac{1}{2}|s_1|^2 + \delta_{\mathbb{R}^k_-}(s_2),\end{aligned}$$

and thus, since

$$q - Cx - Du = \begin{pmatrix} -\sqrt{E}x \\ q_0 - C_0x + u \end{pmatrix},$$

expression $\rho_{V,Q}(q - Cu - Dv)$ equals

$$\frac{1}{2}x\cdot Ex + \delta_{\mathbb{R}^k_-}(q_0 - C_0x + u) = \frac{1}{2}x\cdot Ex + \begin{cases} 0 & \text{if } u \le C_0x - q_0, \\ +\infty & \text{otherwise.}\end{cases}$$

As desired, the penalty function enforces the inequality constraint.

We now check the finiteness and differentiability of the Hamiltonian. First, conditions in both (8) and (9) are satisfied since P is positive definite. We have $V^\infty = V$, $\ker Q = \{0_n\}\times\mathbb{R}^k$, and, since $U^{\infty *} = 0_n$, $(DU^\infty)^* = \{w \mid [0_{n\times n}, I_{k\times k}]w = 0\} = \mathbb{R}^n \times 0_k$; thus the second condition for finiteness is satisfied. Similarly, $[D(U^\infty \cap -U^\infty)]^\perp = \mathbb{R}^n \times 0_k$, and the Hamiltonian is differentiable.

Example 2.7 (state-dependent control constraints through quadratic penalties). Adding to the integrand in (10) the penalty function

$$\sum_{i=1}^{s}\begin{cases} 0 & \text{if } q_i - c_i\cdot x(t) - d_i\cdot u(t) \le 0, \\ \frac{1}{2}\lambda_i\left(q_i - c_i\cdot x(t) - d_i\cdot u(t)\right)^2 & \text{if } q_i - c_i\cdot x(t) - d_i\cdot u(t) > 0,\end{cases}$$

with $\lambda_i > 0$ leads to another problem in the extended piecewise linear-quadratic format. Indeed, set $U = \mathbb{R}^k,\ V = \mathbb{R}^n \times \mathbb{R}^s_+,\ P = F,\ p = 0$, and

$$Q = \begin{bmatrix} I_{n\times n} & 0_{n\times s} \\ 0_{s\times n} & \Lambda^{-1} \end{bmatrix},\ q = \begin{bmatrix} 0_{n\times n} \\ q_1 \\ \vdots \\ q_s \end{bmatrix},\ C = \begin{bmatrix} \sqrt{E} \\ c_1 \\ \vdots \\ c_s \end{bmatrix},\ D = \begin{bmatrix} 0_{n\times k} \\ d_1 \\ \cdots \\ d_s \end{bmatrix},$$

where Λ is a diagonal matrix with diagonal entries λ_i. It can be verified that the corresponding Hamiltonian function is finite and continuously differentiable (but not C^2).

We add that combining penalty functions from Example 2.7, with constraints of either Example 2.5 or 2.6, is possible in the extended piecewise linear-quadratic format. Moreover, these suggested combinations will lead to a differentiable Hamiltonian. In section 3 we will return to the examples above to describe the corresponding optimal feedback mappings.

3. Value function regularity. Techniques used in this and the following sections will rely in part on the Hamilton–Jacobi and duality theories developed for convex control problems in Rockafellar and Wolenski [27]. The required assumptions on the problem $\mathcal{P}(\tau,\xi)$ defined in (1), which we pose throughout this section, are stated below. The growth conditions in (A2), (A3) are quite mild, their detailed discussions can be found in [27] and also Goebel [14].

Assumption 3.1 (basic assumptions).

(A1) The functions $g:\mathbb{R}^n \mapsto \overline{\mathbb{R}}$ and $L:\mathbb{R}^{2n} \mapsto \overline{\mathbb{R}}$ are proper, l.s.c., and convex.

(A2) The set $F(x) = \{v \mid L(x,v) < \infty\}$ is nonempty for all x, and there is a constant ρ such that $\operatorname{dist}(0,F(x)) \le \rho(1+|x|)$ for all x.

(A3) There exist constants α, β and a coercive, proper, nondecreasing function $\theta(\cdot)$ on $[0,\infty)$ such that $L(x,v) \ge \theta(\max\{0,|v|-\alpha|x|\}) - \beta|x|$ for all x and v.

The symbol $\overline{\mathbb{R}}$ stands for the interval $[-\infty,+\infty]$, a function $f:\mathbb{R}^n \mapsto \overline{\mathbb{R}}$ is said to be proper if it does not take on the value $-\infty$, and its effective domain $\operatorname{dom} f = \{x \mid f(x) < +\infty\}$ is nonempty; a function f is called coercive if $\lim_{|x|\to+\infty} \frac{f(x)}{|x|} = +\infty$.

Example 3.2 (piecewise linear-quadratic Lagrangian). Translating the control problem $\mathcal{C}(\tau,\xi)$ discussed in section 2 to the format of Bolza (1) (see [10] or [18]) leads to the Lagrangian

$$L(x,v) = \inf_u \left\{ p\cdot u + \frac{1}{2}u\cdot Pu + \rho_{V,Q}(q - Cx - Du) \mid v = Ax + Bu, u \in U \right\}. \tag{11}$$

In particular, the value function defined by (1) with the Lagrangian (11) is the same as that defined by (3). If (5) holds and the corresponding Hamiltonian (6) is finite (as is always the case if conditions (8) are in place), then the Lagrangian above satisfies Assumption 3.1; see [15, Corollary 4.5].

A key tool for the analysis of the regularity of the value function V is the global description of the graph of $\partial_\xi V(\tau,\cdot)$ as the image of $\operatorname{gph}\partial g$ under a certain flow mapping. Here, and in what follows, $\partial_\xi V$ denotes the subdifferential in the sense of convex analysis, of the convex function $\xi \mapsto V(\tau,\xi)$; the subdifferentials ∂g and $\partial_y H$ should also be understood in the convex sense; see Rockafellar [19, section 23]. The subdifferential $\tilde{\partial}_x H(x,y)$ of the concave function $H(\cdot,y)$ equals $-\partial_x(-H(x,y))$. If any of the mentioned functions are differentiable, the subdifferential reduces to the gradient. Consider the Hamiltonian inclusion

$$-\dot{y}(t) \in \tilde{\partial}_x H(x(t),y(t)), \quad \dot{x}(t) \in \partial_y H(x(t),y(t)). \tag{12}$$

A pair of absolutely continuous arcs $(x(\cdot),y(\cdot))$ on $[a,b]$ will be called a *Hamiltonian trajectory* if it satisfies (12) for almost all $t \in [a,b]$.

THEOREM 3.3 (flow of the value function). *One has $\eta \in \partial_\xi V(\tau,\xi)$ if and only if, for some $\eta^T \in \partial g(\xi^T)$, there is a Hamiltonian trajectory on $[T-\tau,T]$ from $(\xi,-\eta)$ to $(\xi^T,-\eta^T)$.*

The above result was shown by Rockafellar and Wolenski [27], as Theorem 2.4, in the setting of control problems with an initial cost function, and for which the value function is parameterized by a terminal constraint. A change of variables in the expression for the value function yields the result as described above.

In a less convex setting, descriptions of the (appropriately understood) subdifferential of the value function in the flavor of Theorem 3.3 are possible in some local sense, as long as the image of the subdifferential of the terminal cost under the Hamiltonian flow remains a subdifferential of a function—this is the case in our convex setting for any length of the time interval $[\tau, T]$. Under stronger smoothness assumptions than used here, the Hessian of the value function may then turn out to be a solution of an appropriate matrix Riccati differential equation; see Byrnes [6] and Caroff and Frankowska [9].

To illustrate Theorem 3.3, we show that a piecewise linear-quadratic problem need not yield a piecewise linear-quadratic value function. This is in contrast to discrete time problems.

Example 3.4 (loss of piecewise linear-quadratic structure). Consider a one-dimensional problem of Bolza with the cost functions

$$L(x,v) = \frac{1}{2}v^2 + \begin{cases} 0, & x<0,\\ \frac{1}{2}x^2, & x\geq 0,\end{cases} \qquad g(x) = \frac{1}{2}(x+3)^2.$$

The corresponding Hamiltonian is piecewise linear-quadratic and differentiable, and its gradient is piecewise linear:

$$H(x,y) = \begin{cases} \frac{1}{2}y^2, & x<0,\\ -\frac{1}{2}x^2+\frac{1}{2}y^2, & x\geq 0,\end{cases} \qquad \nabla H(x,y) = \begin{cases} (0,y), & x<0,\\ (-x,y), & x\geq 0.\end{cases}$$

A Hamiltonian trajectory $(x(\cdot), y(\cdot))$ must satisfy $\dot{x}(t) = y(t) = const$ when $x(t) < 0$, and $x(t) = \alpha e^t + \beta e^{-t}$, $y(t) = \alpha e^t - \beta e^{-t}$ for suitably chosen α, β when $x(t) > 0$.

The segment between $(-2,-1)$ and $(-1,-2)$ is contained in $\operatorname{gph}(-\nabla g)$. Parameterize the segment by $(x_s(T), y_s(T)) = (s-2, -s-1)$ with $s \in [0,1]$. Hamiltonian trajectories terminating at $(x_s(T), y_s(T))$ are given by

$$(x_s(t), y_s(t)) = \begin{cases} \Big((s+1)(T-t)+s-2, -s-1\Big), & 0\leq T-t\leq \frac{2-s}{s+1},\\ (s+1)\Big(\sinh(T-t-\frac{2-s}{s+1}), -\cosh(T-t-\frac{2-s}{s+1})\Big), & T-t\geq \frac{2-s}{s+1}.\end{cases}$$

It is easy to check that for any $t < T-1$, the set $\{(x_s(t), y_s(t)), s\in[0,1]\}$ is not a straight line segment, nor is it a union of segments. But $\{(x_s(t), y_s(t)), s\in[0,1]\} \subset \operatorname{gph} -\partial_\xi V(T-t,\cdot)$, and consequently, $V(T-t,\cdot)$ is not piecewise linear-quadratic.

LEMMA 3.5. *Suppose that H is differentiable and ∇H is Lipschitz continuous with constant K. Let $g(x) = \frac{1}{2}a\|x\|^2 + b\cdot x + c$, with $a>0$. Then, for all $\tau \leq T$,*

(a) *$V(\tau,\cdot)$ is differentiable with $\nabla_\xi V(\tau,\cdot)$ Lipschitz continuous, with constant*

$$a\left[1+\left(e^{K(T-\tau)}-1\right)\sqrt{1+a^{-2}}\right]^2;$$

(b) *$V(\tau,\cdot)$ is strongly convex with constant*

$$a\left[1+\left(e^{K(T-\tau)}-1\right)\sqrt{1+a^{2}}\right]^{-2}.$$

Proof. Fix $\tau \le T$. Pick two points, $\xi_1^T \ne \xi_2^T$ in $\mathbb{R}^n$, and let $\eta_i^T = \nabla g(\xi_i^T) = a\,\xi_i^T + b$, $i = 1, 2$. Let $(x_i(\cdot), y_i(\cdot))$ be the Hamiltonian trajectory on $[\tau, T]$ with $(x_i(T), y_i(T)) = (\xi_i^T, -\eta_i^T)$ for $i = 1, 2$. As ∇H is Lipschitz continuous, the Hamiltonian trajectories and the endpoints just mentioned are well defined. To shorten the notation, let $\alpha(t) = x_1(t) - x_2(t)$, $\beta(t) = y_1(t) - y_2(t)$.

The monotone structure of ∇H implies that $\alpha(t) \cdot \beta(t)$ is a nondecreasing function of t; see Theorem 4 in [20]—this is a distinguishing feature of a convex problem. Consequently,

$$-\|\alpha(\tau)\|\,\|\beta(\tau)\| \le \alpha(\tau) \cdot \beta(\tau) \le \alpha(T) \cdot \beta(T) = -a\|\alpha(T)\|^2,$$

and thus $\|\alpha(\tau)\|\,\|\beta(\tau)\| \ge a\|\alpha(T)\|^2$. Lipschitz continuity of ∇H implies that

$$\|\beta(\tau)\| \le \|\beta(T)\| + \left(e^{K(T-\tau)} - 1\right) \|(a(T), b(T))\|. \tag{13}$$

Maximizing $\|\beta(\tau)\|/\|\alpha(\tau)\|$ subject to the last two inequalities (this is a simple two-dimensional calculus problem) yields

$$\frac{\|\beta(\tau)\|}{\|\alpha(\tau)\|} \le \frac{\|\beta(T)\| + \left(e^{K(T-\tau)} - 1\right) \|(a(T), b(T))\|}{a\|\alpha(T)\|^2 \Big/ \left[\|\beta(T)\| + \left(e^{K(T-\tau)} - 1\right) \|(a(T), b(T))\|\right]},$$

which simplifies to $\|\beta(\tau)\|/\|\alpha(\tau)\| \le a\left[1 + \left(e^{K(T-\tau)} - 1\right)\sqrt{1 + a^{-2}}\right]^2$, since $\beta(T) = -a\,\alpha(T)$. Thanks to Theorem 3.3, the last bound is in fact a bound on $\|\eta_1 - \eta_2\|/\|\xi_1 - \xi_2\|$ over all (ξ_i, η_i) such that $\eta_i \in \partial_\xi V(\tau, \xi_i)$, $i = 1, 2$. This shows (a).

A lower bound on $\|\eta_1 - \eta_2\|/\|\xi_1 - \xi_2\|$ and the relationship between strong convexity of a convex function and the Lipschitz continuity of the gradient of its conjugate [26, Proposition 12.60] yield (b); see also Example 4.2. □

THEOREM 3.6 (Lipschitz gradient). *Assume that H is differentiable and ∇H is Lipschitz continuous with constant K.*

(a) *Suppose that g is differentiable and ∇g is Lipschitz with constant γ_0. Then $V(\tau, \cdot)$ is differentiable for all $\tau < T$, and there exists a continuous function $\gamma : (-\infty, T] \mapsto \mathbb{R}$ with $\gamma(T) = \gamma_0$ such that $\nabla_\xi V(\tau, \cdot)$ is Lipschitz with constant $\gamma(\tau)$.*

(b) *Suppose that g is strongly convex with constant δ_0. Then there exists a continuous (and positive) function $\delta : (-\infty, T] \mapsto \mathbb{R}$ with $\delta(T) = \delta_0$ such that for all $\tau < T$, $V(\tau, \cdot)$ is strongly convex with constant $\delta(\tau)$.*

In fact, one can choose $\gamma(\tau) = \frac{c^2-1}{2c}$ with $c = (\gamma_0 + \sqrt{1 + \gamma_0^2})e^{2K(T-\tau)}$, and $\delta(\tau) = \frac{2d}{d^2-1}$ with $d = \left(\delta_0^{-1} + \sqrt{1 + \delta_0^{-2}}\right)e^{2K(T-\tau)}$. In particular, $\gamma(\tau) \le (\gamma_0 + \frac{1}{2})e^{2K(T-\tau)}$ and $\delta(\tau) \ge \frac{2\delta_0}{2+\delta_0}e^{-2K(T-\tau)}$.

Proof. The gradient of a differentiable convex function f is Lipschitz continuous with constant a if and only if, for all x, x',

$$f(x') \le f(x) + \nabla f(x) \cdot (x' - x) + \frac{1}{2}a\|x' - x\|^2; \tag{14}$$

see Proposition 12.60 in [26]. If g is as assumed in (a), we have for any x, x', $g(x') \le \overline{g}^x(x')$, where $\overline{g}^x(x') = g(x) + \nabla g(x) \cdot (x' - x) + \frac{1}{2}\gamma_0\|x' - x\|^2$. Then for any $\tau \le T$, $V(\tau, \xi) \le \overline{V}^x(\tau, \xi)$, where $\overline{V}^x(\tau, \cdot)$ is the value function corresponding to the initial cost g_x. The latter value function is differentiable, as shown in Lemma

3.5. Also, $V(\tau,\xi_x) = \overline{V}^x(\tau,\xi_x)$, where ξ_x is the first coordinate of the initial point of the Hamiltonian trajectory on $[\tau,T]$ terminating at $(x,-\nabla g(x))$; this follows from Theorem 3.3 and from the optimality of the first arc constituting the mentioned Hamiltonian trajectory in the definition of both value functions. Consequently, $V(\tau,\cdot)$ is also differentiable at ξ_x, and $\nabla_\xi V(\tau,\xi_x) = \nabla_\xi \overline{V}^x(\tau,\xi_x)$. Now, Lemma 3.5 implies that the gradient of $\overline{V}^x(\tau,\cdot)$ is Lipschitz continuous with constant γ' as described in (a) of the lemma. Combining this, the inequality (14), and the comparison between $V(\tau,\cdot)$ and $\overline{V}^x(\tau,\cdot)$ yields

$$V(\tau,\xi) \le V(\tau,\xi_x) + \nabla_\xi V(\tau,\xi_x) + \frac{1}{2}\gamma'\|\xi-\xi_x\|^2.$$

In light of Theorem 3.3 this bound holds for any ξ, ξ', and thus $\nabla_\xi V(\tau,\cdot)$ is Lipschitz continuous with constant γ'.

The Optimality Principle and time-invariance of the Hamiltonian allow us to derive, through arguments similar to those above, a Lipschitz constant for $\nabla_\xi V(\tau',\cdot)$ given a constant for $\nabla_\xi V(\tau,\cdot)$, with $\tau' < \tau$. Let $\gamma(t)$ denote the (smallest possible) Lipschitz constant for $\nabla_\xi V(t,\cdot)$. Then $\gamma(\tau') \le \gamma(\tau)[1+(e^{K(\tau-\tau')}-1)\sqrt{1+\gamma(\tau)^{-2}}]^2$ whenever $\gamma(\tau) > 0$; a similar bound can be obtained for the case of $\gamma(\tau) = 0$ for small values of $\tau-\tau'$ (by estimating $\|a(\tau')\|$, $\|b(\tau')\|$ from the proof of Lemma 3.5 as in (13)). Consequently, we can show that

$$\liminf_{\tau'\to\tau} \frac{\gamma(\tau')-\gamma(\tau)}{\tau'-\tau} \ge -2K\sqrt{1+\gamma^2(\tau)}.$$

Thus $\gamma(\tau) \le \phi(\tau)$, where ϕ is the solution of $\phi'(t) = -2K\sqrt{1+\phi^2(t)}$, $\phi(T) = \gamma_0$. This yields the bound at the end of Theorem 3.6 and proves (a).

A direct proof of (b) is symmetrical to the one just presented for (a), and an alternate approach is explained in Example 4.2. □

The factor 2 in the exponent in formulas for c and d at the end of Theorem 3.6 is not surprising. Consider $H(x,y) = x\cdot y$ corresponding to $L(x,v) = \delta_x(v)$. Then for any g, $V(\tau,\xi) = g(e^{T-\tau}\xi)$ and the Lipschitz constant for $\nabla V(\tau,\cdot)$ is $e^{2(T-\tau)}$ times that of ∇g.

Under the assumptions of Theorem 3.6 (a), an arc $x(\cdot)$ is optimal for $\mathcal{P}(\tau,\xi)$ in (1) if and only if

$$x(\tau) = \xi, \quad \dot{x}(t) = \nabla_y H(x(t), -\nabla_\xi V(t,x(t))) \text{ for almost all } t\in[\tau,T]. \tag{15}$$

The properties of the optimal feedback mapping $\Phi : (-\infty,T]\times\mathbb{R}^n$, defined by $\Phi(t,x) = \nabla_y H(x,-\nabla_\xi V(t,x))$, are summarized below. Continuity of ϕ in both variables follows from that of $\nabla_\xi V$, which in turn is a consequence of graphical continuity of $\nabla_\xi V(t,\cdot)$ in t, as stated in [27, Corollary 2.2]; details were worked out in Goebel [14].

Corollary 3.7 (Lipschitz optimal feedback). *Suppose that H and g are differentiable and their gradients are Lipschitz continuous. Then the optimal feedback mapping Φ is continuous on $(-\infty,T]\times\mathbb{R}^n$, and there exists a continuous function $\mu : (-\infty,T]\mapsto\mathbb{R}$ such that for all $t\le T$, $\Phi(t,\cdot)$ is Lipschitz continuous with constant $\mu(t)$.*

If the problem of Bolza $\mathcal{P}(\tau,\xi)$ represents a control problem $\mathcal{C}(\tau,\xi)$ in (3) via the transformation (11), an optimal control minimizes the right-hand side in (11). This translates (15) to necessary and sufficient optimality conditions for $\mathcal{C}(\tau,\xi)$ (general

case with no smoothness present was discussed in [14]), $x(\tau) = \xi$, $\dot{x}(t) = Ax(t)+Bu(t)$, and

$$u(t) = \nabla_1 J^*(-B^*\nabla_\xi V(t, x(t)), Cx(t)) \text{ for almost all } t \in [\tau, T]. \tag{16}$$

Under the assumptions of Theorem 2.4, conclusions similar to those in Corollary 3.7 can be made about $\phi(t, x) = \nabla_1 J^*(-B^*\nabla_\xi V(t, x), Cx)$. In particular, optimal controls turn out to be continuous. To finish this section, we calculate ϕ for some of the examples of section 2.

We will need some properties of $\rho_{V,Q}$ defined in (4) (recall that Q is positive semidefinite and V is polyhedral). The function $\rho_{V,Q}$ is proper, convex, and piecewise linear-quadratic; $\operatorname{dom} \rho_{V,Q} = (V^\infty \cap \ker Q)^*$ and, in particular, $\rho_{V,Q}$ is finite-valued if and only if $V^\infty \cap \ker Q = \{0\}$ (Theorem 2.2 generalizes this fact). If this condition holds, then

$$\partial\rho_{V,Q}(s) = \operatorname*{argmax}_{v\in V}\left\{s \cdot v - \frac{1}{2}v \cdot Qv\right\} = \{v \mid s - Qv \in N_V(v)\} = (Q + N_V)^{-1}(s),$$

where $N_V(v)$ is the normal cone to the set V at v. For details, see Example 11.18 in Rockafellar and Wets [26]. If Q is actually positive definite, and thus invertible, we have, with $\operatorname{proj}_{\sqrt{V}Q}$ being the projection onto $\sqrt{Q}V$,

$$\partial\rho_{Q,V}(s) = \left(\sqrt{Q}\right)^{-1} \operatorname{proj}_{\sqrt{Q}V}\left(\left(\sqrt{Q}\right)^{-1} s\right).$$

Indeed for any convex set C, $(\operatorname{proj}_C)^{-1} = I + N_C$. Then

$$\left[\left(\sqrt{Q}\right)^{-1} \operatorname{proj}_{\sqrt{Q}V}\left(\sqrt{Q}\right)^{-1}\right]^{-1} = \sqrt{Q}\left(\operatorname{proj}_{\sqrt{Q}V}\right)^{-1}\sqrt{Q} = \sqrt{Q}\left(I + N_{\sqrt{Q}V}\right)\sqrt{Q}.$$

The last expression equals $Q+N_V$. It follows from the fact that $\sqrt{Q}N_{\sqrt{Q}V}\sqrt{Q} = N_V$, and this can be deduced from the properties of the normal cone under a change of coordinates.

Example 3.8 (optimal controls in feedback form). The linear-quadratic regulator (10) with a constraint $u(t) \in U$ (Example 2.5) has the following feedback mapping:

$$\phi(t, x) = (F + N_U)^{-1}(-B^*\nabla_\xi V(t, x)) = \sqrt{F}^{-1} \operatorname{proj}_{\sqrt{F}U}\left(-\sqrt{F}^{-1}B^*\nabla_\xi V(t, x)\right).$$

A similar formula was obtained by Heemels, Van Eijndhoven, and Stoorvogel [16] for a conical U; our regularity results are also stronger than those therein.

Example 2.6 discussed (10) with a constraint $u(t) \le C_0x(t) - q_0$. With the matrices as defined in the mentioned example, we obtain

$$\begin{aligned}
J^*(a, b) &= \sup_{u\in U}\inf_{v\in V}\left\{a \cdot u + b \cdot v - \frac{1}{2}u \cdot Pu + \frac{1}{2}v \cdot Qv + v \cdot Du\right\} \\
&= \inf_{v\in V}\left\{b \cdot v + \frac{1}{2}v \cdot Qv + \sup_{u\in\mathbb{R}^k}\left\{(a + D^*v) \cdot u - \frac{1}{2}u \cdot Pu\right\}\right\} \\
&= \inf_{v\in V}\left\{b \cdot v + \frac{1}{2}v \cdot Qv + \frac{1}{2}(a + D^*v) \cdot P^{-1}(a + D^*v)\right\} \\
&= \frac{1}{2}a \cdot P^{-1}a - \sup_{v\in V}\left\{(-b - DP^{-1}a) \cdot v - \frac{1}{2}v \cdot (Q + DP^{-1}D^*)v\right\}.
\end{aligned}$$

The matrix $Q + DP^{-1}D^*$ equals $\left[\begin{smallmatrix} I_{n\times n} & 0_{n\times k} \\ 0_{k\times n} & F^{-1} \end{smallmatrix}\right]$, and thus the sup expression above is separable. Also, $DP^{-1} = \left[\begin{smallmatrix} 0_{k\times k} \\ -F^{-1} \end{smallmatrix}\right]$, so $(-b - DP^{-1}a)_1 = -b_1$, $(-b - DP^{-1}a)_2 = -b_2 - F^{-1}a$. Then $J^*(a,b)$ equals

$$\begin{aligned}\frac{1}{2}a\cdot F^{-1}a - \frac{1}{2}\left\|(-b-DF^{-1}a)_1\right\|^2 - \sup_{v_2\in\mathbb{R}^k_+}\left\{(-b-DF^{-1}a)_2\cdot v_2 - v_2\cdot F^{-1}v_2\right\}\\ = \frac{1}{2}a\cdot F^{-1}a - \frac{1}{2}\left\|b_1\right\|^2 - \rho_{\mathbb{R}^k_-,F^{-1}}\left(-b_2 - F^{-1}a\right)\end{aligned}$$

and thus $\nabla_1 J^*(a,b) = F^{-1}\left[a + (N_{\mathbb{R}^k_-} + F^{-1})^{-1}(-b_2 - F^{-1}a)\right]$. Since for $b = Cx$ we have $b_1 = \sqrt{E}x$, $b_2 = C_2x$, the optimal feedback map is

$$\phi(t,x) = -F^{-1}\left[B^*\nabla_\xi V(t,x) - \left(N_{\mathbb{R}^k_-} + F^{-1}\right)^{-1}\left(F^{-1}B^*\nabla_\xi V(t,x) - C_2x\right)\right].$$

4. Convergence and approximation of value functions. In this section we study the convergence of value functions defined by sequences of converging costs $\{g_i\}$ and $\{L_i\}$,

$$V_i(\tau,\xi) = \inf\left\{g_i(x(0)) + \int_0^\tau L_i(x(t),\dot{x}(t))\,dt \;\middle|\; x(\tau) = \xi\right\}, \tag{17}$$

to $V(\tau,\xi)$ defined in (1). To treat sequences of possibly infinite-valued functions we use the well appreciated in optimization notion of epi-convergence. A sequence of functions $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$, $i = 1,2,\ldots$, is said to *epi-converge* to f ($\text{e-lim}_i f_i = f$ for short) if, for every point $x \in \mathbb{R}^n$,

(a) $\liminf_i f_i(x_i) \geq f(x)$ for every sequence $x_i \to x$,

(b) $\limsup_i f_i(x_i) \leq f(x)$ for some sequence $x_i \to x$.

For details, consult Rockafellar and Wets [26, Chapter 7]. We will only need to directly show the "lower" part of epi-convergence of value functions and rely on duality results to complete the argument. Let us briefly introduce the needed concepts.

For a function $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ its convex conjugate is defined by

$$f^*(y) = \sup_{x\in\mathbb{R}^n}\{y\cdot x - f(x)\}.$$

If f is proper, l.s.c., and convex, then so is f^*, and the conjugate of f^* equals f (that is, $f(x) = \sup_{y\in\mathbb{R}^n}\{x\cdot y - f^*(y)\}$). For details, consult Rockafellar [19, section 12]. Relations of certain properties of f to some other properties of f^*, say of coercivity and finiteness, were alluded to in the previous sections; in Example 4.2 we discuss the symmetry between strong convexity of f and Lipschitz continuity of ∇f^*, and revisit Lemma 3.5 and Theorem 3.6. Epi-convergence of a sequence of convex function is equivalent to that of the sequence of conjugates; we will need the following related facts. Below, $\text{e-lim}\inf_i f_i \geq f$ means that condition (a) in the definition of epi-convergence holds. A sequence $\{f_i\}$ is said to escape epigraphically to the horizon if the epigraphical limit of f_i is equal to $+\infty$ everywhere.

LEMMA 4.1. *Suppose that functions $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ and $f_i : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$, $i = 1,2,\ldots$, are proper, l.s.c., and convex.*

(a) *If $\text{e-lim}\inf_i f_i \geq f$ and $\text{e-lim}\inf_i f_i^* \geq f^*$ and neither sequence escapes epigraphically to the horizon, then $\text{e-lim}_i f_i = f$ and $\text{e-lim}_i f_i^* = f^*$.*

(b) *Neither of the sequences f_i, f_i^* escapes epigraphically to the horizon provided there exists a constant $\rho > 0$ such that $f_i(x) \geq -\rho(\|x\| + 1)$ and $f_i^*(x) \geq -\rho(\|x\| + 1)$ for all x and $i = 1, 2, \ldots$.*

Proof. Statement (a) essentially follows from the statement and proof of Theorem 11.34 in [26]. We show (b). An application of a separation principle (for example, Theorem 11.3 in [19]) implies that for every $i = 1, 2, \ldots$, there exist $\alpha_i \in \mathbb{R}^n$, $\beta_i \in \mathbb{R}$ such that $f_i(x) \geq \alpha_i \cdot x + \beta_i \geq -\rho(\|x\| + 1)$ for every $x \in \mathbb{R}^n$. It must be the case that $\|\alpha_i\| \leq \rho$ while $\beta_i \geq -\rho$. We then obtain

$$f_i^*(\alpha_i) = \sup_x\{\alpha_i \cdot x - f(x)\} \leq \sup_x\{\alpha_i \cdot x - \alpha_i \cdot x - \beta_i\} = -\beta_i \leq \rho.$$

Thus $f_i^*(\alpha_i) \leq \rho$ while by assumption, $f_i^*(\alpha_i) \geq -\rho(\|\alpha_i\| + 1)$. As $\|\alpha_i\| \leq \rho$, there exists a convergent subsequence of $(\alpha_i, f_i^*(\alpha_i))$, and, consequently, the sequence f_i^* cannot escape to the horizon. A symmetric argument shows the corresponding fact for the sequence f_i. ∎

For a given initial cost g and Lagrangian L, the *dual value function* $\widetilde{V} : (-\infty, T] \times \mathbb{R}^n \mapsto \mathbb{R}$ is defined in a fashion similar to V,

$$\widetilde{V}(\tau, \eta) = \inf\left\{g^*(y(0)) + \int_0^\tau \widetilde{L}(y(t), \dot{y}(t))\, dt \;\middle|\; y(\tau) = \eta\right\}, \tag{18}$$

where the dual Lagrangian is

$$\widetilde{L}(y, w) = L^*(w, y) = \sup_{(x,v)\in\mathbb{R}^{2n}} \{w \cdot x + y \cdot v - L(x, v)\}. \tag{19}$$

If L satisfies Assumption 3.1, then so does $\widetilde{L}$ (and consequently $\widetilde{V}(\tau, \cdot)$ is proper, l.s.c., and convex for every $\tau \leq T$), and in fact for any $\tau \leq T$, the functions $V(\tau, \cdot)$ and $\widetilde{V}(\tau, \cdot)$ are conjugate to each other:

$$\widetilde{V}(\tau, \eta) = \sup_{\xi\in\mathbb{R}^n} \big\{\eta \cdot \xi - V(\tau, \xi)\big\}, \qquad V(\tau, \xi) = \sup_{\eta\in\mathbb{R}^n} \big\{\xi \cdot \eta - \widetilde{V}(\tau, \eta)\big\}. \tag{20}$$

These results were shown by Rockafellar and Wolenski [27]. The Hamiltonian $\widetilde{H}$ associated with a dual Lagrangian $\widetilde{L}$ is exactly $\widetilde{H}(y, x) = -H(x, y)$, and thus it has the same smoothness properties as H. Note also that the Lagrangian dual to $\widetilde{L}$ is the original L.

Example 4.2 (strong convexity and Lipschitz differentiability). A convex function f is differentiable and ∇f is Lipschitz continuous with constant σ if and only if f^* is strongly convex with constant $1/\sigma$. This and (20) automatically proves one of the statements (a), (b) of Theorem 3.6 once the other is in place and similarly for Lemma 3.5. For example, we show (b) of 3.6 with the help of (a). Suppose g is strongly convex with constant δ_0, and ∇H is Lipschitz with constant K. Then ∇g^* is Lipschitz with constant $\gamma_0 = 1/\delta_0$, while the dual Hamiltonian $\widetilde{H}$ also has a Lipschitz gradient, with constant K. Application of (a) shows that the dual value function $\widetilde{V}(\tau, \cdot)$ is differentiable, with $\nabla_\eta \widetilde{V}(\tau, \cdot)$ Lipschitz with constant $\gamma(\tau)$ as described at the end of Theorem 3.6. Now (20) implies that $V(\tau, \cdot)$ is strongly convex, with constant $\delta(\tau) = 1/\gamma(\tau)$. This yields the expression for $\delta(\tau)$ as described in the other formula at the end of Theorem 3.6. The lower bound on $\delta(\tau)$ can be obtained in a similar fashion from the upper bound on $\gamma(\tau)$.

We now focus on sequences of Bolza problems. Given a sequence of Lagrangians $\{L_i\}$, for each i we let $\widetilde{L}_i$ be the Lagrangian dual to L_i as described in (19), and H_i be the Hamiltonian corresponding to L_i as suggested by (2). The value function V_i is defined by (17), while $\widetilde{V}_i$ is defined similarly in terms of g_i^* and $\widetilde{L}_i$.

Assumption 4.3 (uniform growth assumption). Each of the functions g_i and L_i, $i = 1, 2, \ldots$, is proper, l.s.c., and convex. There exist functions $\underline{L}$ and $\overline{L}$, each satisfying Assumption 3.1, such that, for every $i = 1, 2, \ldots$,

$$\underline{L} \le L_i \le \overline{L}.$$

As L satisfies Assumption 3.1 if and only if $\widetilde{L}$ does, the second condition above is equivalent to the existence of $\underline{M}$ and $\overline{M}$, each satisfying Assumption 3.1, such that $\underline{M} \le \widetilde{L}_i \le \overline{M}$; take $\underline{M}$ to be the Lagrangian dual to $\overline{L}$, $\overline{M}$ dual to $\underline{L}$.

LEMMA 4.4 (convergence equivalence). *If Assumption* 4.3 *holds, the following statements are equivalent:*

(a) *Lagrangians L_i epi-converge to L,*

(b) *dual Lagrangians $\widetilde{L}_i$ epi-converge to $\widetilde{L}$,*

(c) *Hamiltonians H_i converge pointwise to H.*

The proof is postponed until section 5. Also there we discuss the convergence of Lagrangians (11) and Hamiltonians (6) corresponding to extended piecewise linear-quadratic functions under perturbations of all defining data; see Theorem 5.6.

Assumption 4.5 (epi-convergence of cost functions). Sequences $\{g_i\}$, $\{L_i\}$ epi-converge, respectively, to g and L.

Equivalently, we could assume that sequences $\{g_i^*\}$ and $\{\widetilde{L}_i\}$ epi-converge, respectively, to g^* and $\widetilde{L}$. We are now ready to state the main result of this section.

THEOREM 4.6 (value function epi-convergence). *Let Assumptions* 4.3 *and* 4.5 *hold. For any $\tau \le T$ and a sequence $\tau_i \to \tau$ (in particular for $\tau_i = \tau$) we have*

$$\text{e-lim}\, V_i(\tau_i, \cdot) = V(\tau, \cdot). \tag{21}$$

Equivalently, $\text{e-lim}\, \widetilde{V}_i(\tau_i, \cdot) = \widetilde{V}(\tau, \cdot)$. *This implies* $\text{e-lim}\, V_i = V$ *and* $\text{e-lim}\, \widetilde{V}_i = \widetilde{V}$.

We prove the theorem by taking advantage of the representation

$$V(\tau, \xi) = \inf_{\xi' \in \mathbb{R}^n} \{E(\tau, \xi, \xi') + g(\xi')\}, \tag{22}$$

where the *fundamental kernel* $E : (-\infty, T] \times \mathbb{R}^n \times \mathbb{R}^n$ is given by

$$E(\tau, \xi, \xi') = \inf \left\{ \int_\tau^T L(x(t), \dot{x}(t))\, dt \mid x(\tau) = \xi,\ x(T) = \xi' \right\},$$

with the infimum taken over all arcs with prescribed endpoints. A symmetric representation of $\widetilde{V}(\tau, \eta)$ is available, with $\widetilde{E}(\tau, \eta, \eta')$ defined in terms of $\widetilde{L}$. The following conjugacy relationship is a direct consequence of (20); to see this, consider $E(\cdot, \cdot, \xi')$ as the value function associated with the terminal cost $g(x) = \delta_{\xi'}(x)$ and use the fact that $g^*(y) = \xi' \cdot y_0$:

$$\widetilde{E}(\tau, \eta, \eta') = \sup_{\xi, \xi'} \{\eta \cdot \xi - \eta' \cdot \xi' - E(\tau, \xi, \xi')\},$$
$$E(\tau, \xi, \xi') = \sup_{\eta, \eta'} \left\{\xi \cdot \eta - \xi' \cdot \eta' - \widetilde{E}(\tau, \eta, \eta')\right\}.$$

We will need some facts about continuity and convergence of integral functionals. It is known that for a fixed $\tau > 0$, the functional $\Phi(\tau, \cdot)$ defined on the space of absolutely continuous arcs on $[0, \tau]$ by $\Phi(\tau, z(\cdot)) = \int_0^\tau L(z(t), \dot{z}(t))dt$ is weakly sequentially lower semicontinuous. This can be shown as a consequence of the conjugacy between L and H, and by interchanging the integration and maximization,

$$\int_0^\tau \sup_w \{w \cdot \dot{z}(t) - H(z(t), w)\}\, dt = \sup_{w(\cdot)} \int_0^\tau \{w(t) \cdot \dot{z}(t) - H(z(t), w(t))\}\, dt, \tag{23}$$

where the latter supremum is taken over all arcs w in $L^\infty[0, T]$ (see, for example, [26, Theorem 14.60]). Now consider a sequence of functionals $\Phi_i(\tau, z(\cdot)) = \int_0^\tau L_i(z(t), \dot{z}(t))dt$ a sequence of arcs x_i on $[0, \tau]$ weakly convergent to an arc x (meaning that $\dot{x}_i$ converge weakly to $\dot{x}$ in L^1 and $x_i(0)$ converge to $x(0)$). Then

$$\liminf_i \Phi_i(\tau, x_i(\cdot)) \geq \Phi(\tau, x(\cdot)). \tag{24}$$

We only need to consider the case where $\liminf \Phi_i(\tau, x_i(\cdot)) < +\infty$. As in (23), we have, for any w in $L^\infty[0, T]$, $\Phi_i(\tau, x_i(\cdot)) \geq \int_0^\tau \{w(t) \cdot \dot{x}_i(t) - H_i(x_i(t), w(t))\}\, dt$. Then, as $\dot{x}_i(\cdot)$ converge weakly in L^1 to $\dot{x}(\cdot)$, $x_i(\cdot)$ converge pointwise to $x(\cdot)$, and H_i converge to H pointwise and also uniformly on compact sets (Lemma 5.4), we get $\liminf \Phi_i(\tau, x_i(\cdot)) \geq \int_0^\tau \{w(t) \cdot \dot{x}(t) - H(x(t), w(t))\}\, dt$, and this holds for any w in $L^\infty[0, T]$. By (23), we conclude (24). In the proof of Lemma 4.7 we extend these arguments to varying time intervals.

LEMMA 4.7 (fundamental kernel epi-convergence). *Let E_i and $\widetilde{E}_i$ be the fundamental kernels associated, respectively, with L_i and $\widetilde{L}_i$. Under assumptions of Theorem* 4.6*, for any $\tau < T$ and a sequence $\tau_i \to \tau$ (in particular for $\tau_i = \tau$) we have*

$$\text{e-lim}\, E_i(\tau_i, \cdot, \cdot) = E(\tau, \cdot, \cdot). \tag{25}$$

Equivalently, $\text{e-lim}\, \widetilde{E}_i(\tau_i, \cdot, \cdot) = \widetilde{E}(\tau, \cdot, \cdot)$. *Consequently,* $\text{e-lim}\, E_i = E$ *and* $\text{e-lim}\, \widetilde{E}_i = \widetilde{E}$.

Proof. Fix $\tau < T$ and $\tau_i \to \tau$. First, we show that $\text{e-lim inf}_i\, E_i(\tau_i, \cdot, \cdot) \geq E(\tau, \cdot, \cdot)$, that is, for any point $(\xi, \xi') \in \mathbb{R}^{2n}$ and a sequence $(\xi_i, \xi_i') \to (\tau, \xi, \xi')$, we have

$$\liminf_{i \to \infty} E_i(\tau_i, \xi_i, \xi_i') \geq E(\tau, \xi, \xi'). \tag{26}$$

We only need to consider the case where $\liminf_{i\to\infty} E_i(\tau_i, \xi_i, \xi_i') = m < +\infty$, and if necessary we pass to a subsequence so that $E_i(\tau_i, \xi_i, \xi_i') \to m$. There exist arcs x_i on $[\tau_i, T]$ such that $E_i(\tau_i, \xi_i, \xi_i') = \Phi_i(\tau_i, x_i(\cdot)) = \int_0^{\tau_i} L_i(x_i(t), \dot{x}_i(t))dt$. Setting $a_i = (T - \tau_i)/(T - \tau)$ and defining $x_i^0(\tau + s) = x_i(\tau_i + a_i s)$, $L_i^0(x, v) = a_i L_i(x, v/a_i)$ leads to

$$\Phi_i(\tau_i, x_i(\cdot)) = \int_{\tau_i}^T L_i(x_i(t), \dot{x}_i(t))dt = \int_\tau^T L_i^0(x_i^0(t), \dot{x}_i^0(t))dt = \Phi_i^0(\tau, x_i^0(\cdot)),$$

with L_i^0 epiconverging to L [26, Exercise 7.47]. Corresponding Hamiltonians are $H_i^0(x, y) = a_i H(x, y)$, while the dual Lagrangians are $\tilde{L}_i^0(x, v) = a_i \widetilde{L}_i(x, v/a_i)$. As $\{L_i\}$ satisfies Assumption 4.3, so does $\{L_i^0\}$; this is a direct calculation. Consequent uniform growth assumptions imply in particular that some subsequence of rescaled arcs x_i^0 on $[0, \tau]$ weakly converges to an arc x on $[0, \tau]$ with $x(\tau) = \xi$,

$x(T) = \xi'$ (this follows from Theorem 1 in [21]). Moreover, as in (24), we have $\lim_i \Phi_i(\tau_i, x_i(\cdot)) = \lim_i \Phi_i^0(\tau, x_i^0(\cdot)) \geq \Phi(\tau, x(\cdot))$. But the arc x is feasible for the problem defining $E(\tau, \xi, \xi')$, and (26) follows.

The same argument applied to dual problems gives e-$\liminf \widetilde{E}_i(\tau_i, \cdot, \cdot) \geq \widetilde{E}(\tau, \cdot, \cdot)$. Lemma 4.1 (a) will conclude (25) (and the equivalent dual statement) if we show that neither $E_i(\tau_i, \cdot, \cdot)$ nor $\widetilde{E}_i(\tau_i, \cdot, \cdot)$ escapes to the horizon. Uniform growth in Assumption 4.3 and the rescaling arguments above imply that $\{E_i(\tau_i, \cdot, \cdot)\}$ is uniformly bounded below by $\widehat{E}(\tau, \cdot, \cdot)$, a fundamental function corresponding to some Lagrangian satisfying Assumption 3.1. As the latter function is proper and convex, it is bounded below by an affine function. A similar bound is in place for $\widetilde{E}_i(\tau_i, \cdot, \cdot)$, and thus the desired conclusions hold.

Lastly, the very definition of epi-convergence explains that (25) implies e-$\lim E_i = E$. □

Proof (Theorem 4.6). As in Lemma 4.7, we begin by showing that for any $(\tau, \xi) \in (0, +\infty) \times \mathbb{R}^n$ and a sequence $(\tau_i, \xi_i) \to (\tau, \xi)$, we have

$$\liminf_i V_i(\tau_i, \xi_i) \geq V(\tau, \xi). \tag{27}$$

It suffices to consider, passing to a subsequence if necessary, the case of $\lim_i V_i(\tau_i, \xi_i) < +\infty$. Recall (22). Functions g_i epi-converge to g by assumption, while Lemma 4.7 and the definition of epi-convergence yield e-$\liminf_i E_i(\tau_i, \xi_i, \cdot) \geq E(\tau, \xi, \cdot)$. Now by Theorem 7.46 of [26], we obtain

$$\text{e-}\liminf_i \{E_i(\tau_i, \xi_i, \cdot) + g_i(\cdot)\} \geq E(\tau, \xi, \cdot) + g(\cdot). \tag{28}$$

As mentioned in the proof of Lemma 4.7, $\{E_i(\tau_i, \cdot, \cdot)\}$ is uniformly bounded below by $\widehat{E}(\tau, \cdot, \cdot)$, a fundamental kernel corresponding to some Lagrangian satisfying Assumption 3.1. Proposition 4.2 in [27] implies that

$$E_i(\tau_i, \xi_i, \xi') \geq \widehat{E}(\tau, \xi_i, \xi') \geq \theta\left(\max\{0, |\xi'| - \alpha|\xi_i|\}\right) - \beta|\xi_i| \tag{29}$$

for a proper, nondecreasing, and coercive $\theta : [0, +\infty) \mapsto \mathbb{R}$ and constants α, β. As ξ_i converge, there exist a, b such that $\widehat{E}(\tau_i, \xi, \xi') \geq \theta(\max\{0, |\xi'| - a\}) - b$. A similar bound is in place for $E(\tau, \cdot, \cdot)$, and consequently, $E_i(\tau_i, \xi_i, \cdot)$ and $E(\tau, \xi, \cdot)$ are bounded below by a coercive function. Convexity and epi-convergence of g_i to g implies, by 7.34 in [26], that g_i and g are bounded below (uniformly in i) by $-\rho(|\cdot| + 1)$, for some constant ρ. As $\inf_{\xi'}\{E_i(\tau_i, \xi_i, \xi') + g_i(\xi')\}$ converge to a finite value, there exists a compact set S such that

$$\inf_{\xi'} \{E_i(\tau_i, \xi', \xi_i) + g_i(\xi')\} = \inf_{\xi' \in S} \{f_i(\xi') + E_i(\tau_i, \xi', \xi_i)\},$$

and a similar condition holds for $E(\tau, \xi, \xi') + g(\xi')$. Consequently, infimum in (22) can be taken over S, similarly for $V_i(\tau_i, \cdot)$. Now (28) and Proposition 7.29 in [26] yield (27).

Growth conditions $g_i(\xi') \geq -\rho(|\xi'| + 1)$, (29), and the fact that since θ in (29) is coercive, there exists $\gamma > 0$ such that $\theta \geq \rho|\cdot| - \gamma$ imply that

$$\begin{aligned} V_i(\tau_i, \xi) &\geq \inf_{\xi'} \{-\rho(|\xi'| + 1) + \rho \max\{0, |\xi'| - \alpha|\xi|\} - \gamma - \beta|\xi|\} \\ &\geq -(\alpha\rho + \beta)|\xi| - (\rho + \gamma). \end{aligned}$$

A similar bound holds for $\widetilde{V}_i(\tau_i,\cdot)$. This and (27) show the desired epi-convergence of $V_i(\tau_i,\cdot)$ as well as $\widetilde{V}_i(\tau_i,\cdot)$, by Lemma 4.1 (b). Epi-convergence of V_i and $\widetilde{V}_i$ follows directly from the definition of epi-convergence. □

We now describe how any problems fitting the general Assumption 3.1 can be approximated by problems with value functions possessing regularity as discussed in section 3. We will rely on Moreau–Yosida envelopes of convex and saddle functions. For any proper, l.s.c., and convex f and $\lambda > 0$, $e_\lambda f(x) = \inf_q \left\{ f(q) + \frac{1}{2\lambda}\|x-q\|^2 \right\}$ is finite and differentiable; see [26, Theorem 2.26]. A generalization of this smoothing technique to saddle functions was introduced by Attouch and Wets [2]. Applied to a concave-convex Hamiltonian H (and simplified to single parameter λ vs. the original two), it yields a differentiable concave-convex function

$$e_\lambda H(x,y) = \sup_p \inf_q \left\{ H(p,q) - \frac{1}{2\lambda}\|x-p\|^2 + \frac{1}{2\lambda}\|y-q\|^2 \right\}. \tag{30}$$

(We use the same notation for Moreau–Yosida regularization of convex and saddle functions; it should be clear which one is considered.) The key fact is that $\nabla e_\lambda f$ and $\nabla e_\lambda H$ are globally Lipschitz with constant $1/\lambda$. This is the case since $\nabla e_\lambda f$ is the Yosida regularization of the monotone subdifferential ∂f (see Exercise 12.23 in [26]), while $(x,y) \mapsto (-\nabla_x e_\lambda H(x,y), \nabla_y e_\lambda H(x,y))$ is the Yosida regularization of the monotone mapping $(x,y) \mapsto (-\tilde{\partial}_x H(x,y), \partial_y H(x,y))$.

COROLLARY 4.8 (regularization of value functions). *Let L be any Lagrangian satisfying Assumption* 3.1*; $\widetilde{L}$ be the associated dual Lagrangian; g be any proper, l.s.c., and convex function; and V, $\widetilde{V}$ be the associated value functions. There exists a sequence of finite convex functions g_i and a sequence of Lagrangians L_i satisfying Assumption* 4.3 *such that the following hold.*

(a) *Conclusions of Theorem* 4.6 *hold for sequences $\{V_i\}$, $\{\widetilde{V}_i\}$ of value functions corresponding, respectively, to L_i, g_i and their dual costs g_i^*, $\widetilde{L}_i$.*

(b) *For each i, V_i and $\widetilde{V}_i$ are continuously differentiable, and there exist continuous and positive functions $\gamma_i : (-\infty, T] \mapsto \mathbb{R}$, $\delta : (-\infty, T] \mapsto \mathbb{R}$ such that*

(i) *$\nabla_\xi V_i(\tau,\cdot)$ and $\nabla_\xi \widetilde{V}_i(\tau,\cdot)$ are Lipschitz with constant $\gamma(\tau)$,*

(ii) *$V(\tau,\cdot)$ and $\widetilde{V}(\tau,\cdot)$ are strongly convex with constant $\delta(\tau)$.*

This can be achieved by considering (with H associated to L)

$$g_i(x) = e_{1/i}\, g(x) + \|x\|^2/i, \qquad H_i(x,y) = e_{1/i}\, H(x,y),$$

and letting L_i and $\widetilde{L}_i$ be the Lagrangians associated with H_i.

Proof. Condition ($A3$) in Assumption 3.1 (by the proof of Theorem 2.3 in [27]) and the definition of H_i imply, respectively, that

$$H(x,y) \le \theta^*(\|y\|) + (\alpha\|y\| + \beta)\,\|x\|, \quad H_i(x,y) \le \sup_p \left\{ H(p,y) - \frac{1}{2\lambda}\|p-x\|^2 \right\}.$$

Combining the two inequalities yields

$$\begin{aligned} H_i(x,y) &\le \theta^*(\|y\|) + \sup_p \left\{ (\alpha\|y\| + \beta)\,\|p\| - \frac{1}{2\lambda}\|p-x\|^2 \right\} \\ &= \theta^*(\|y\|) + \frac{\lambda}{2}\,(\alpha\|y\| + \beta)^2 + (\alpha\|y\| + \beta)\,\|x\|. \end{aligned}$$

This in turn implies that $(A3)$ holds for L_i, with θ replaced by

$$\theta'(r) = \inf_{s\in[0,r]} \left\{ \theta^*(s) + \frac{\lambda}{2}(\alpha s + \beta)^2 \right\}.$$

Coercivity of both θ^* and the quadratic implies that of θ', which is obviously nondecreasing. A symmetric argument shows that $\widetilde{L}_i$ satisfies $(A3)$ uniformly, and consequently, Assumption 4.3 is satisfied. Moreau–Yosida approximations of H hypo/epi-converge to H, and as all these functions are finite, the convergence is pointwise (Lemma 5.4). Functions g_i epi-converge to g by Theorem 1.25 and Exercise 7.47 in [26]. This shows (a).

To see (b), note that g has a Lipschitz gradient (with constant i) as well as strongly convex (with constant $1/i$). Now, invoke Theorem 3.6 and the symmetry between strong convexity and Lipschitz continuity of the gradient of the dual as outlined in Example 4.2. □

Example 4.9 (regularization of control problems). Recall that the Hamiltonian (6) corresponding to an extended piecewise linear-quadratic problem (3) had the special structure $H(x,y) = y \cdot Ax + J^*(B^*y, Cx)$. The regularization, as described in Corollary 4.8, can be applied to such H, but a more explicit smoothing technique is available. One may regularize J^* directly, using the convex-concave counterpart of (30)—the infimum is to be taken over the first variable, supremum over the second. Such regularization, with parameter $1/i$ can be equivalently obtained by defining functions J_i^* in (7) with matrices P and Q replaced, respectively, by positive definite $P + I/i$, $Q + I/i$. (Here I denotes an identity matrix of appropriate size.)

5. Convex analysis tools. We say that a function $K : \mathbb{R}^k \times \mathbb{R}^l \mapsto [-\infty, +\infty]$ is convex-concave if, for any fixed $z \in \mathbb{R}^l$, the function $K(\cdot, z)$ is convex, while for any fixed $w \in \mathbb{R}^k$, $K(w, \cdot)$ is concave. We call a convex-concave function K proper if the effective domain of K, defined as

$$\operatorname{dom} K = \left\{ w \in \mathbb{R}^k \mid K(w,z) < +\infty \ \forall z \in \mathbb{R}^l \right\} \times \left\{ z \in \mathbb{R}^l \mid -\infty < K(w,z) \ \forall w \in \mathbb{R}^k \right\},$$

is nonempty.

Convex function duality gives a one-to-one correspondence between a proper lsc convex function and its conjugate (also proper and l.s.c.). Saddle function duality describes a one-to-one correspondence between *equivalence classes* of proper *closed* saddle functions. Closedness is a notion corresponding, in a sense, to lower semicontinuity of convex functions. For the somewhat technical definition, and the reasons for considering equivalence classes, see Rockafellar [19]. Here, we limit ourselves to the facts crucial to the developments in what follows.

Any equivalence class $[K]$ of closed saddle functions contains the lowest and the highest element, denoted $\underline{K}$ and $\overline{K}$, and consists of all closed saddle functions K such that $\underline{K} \le K \le \overline{K}$. If a saddle function K is finite, then it is closed, $K = \underline{K} = \overline{K}$, and the class $[K]$ of all closed functions equivalent to K is just $\{K\}$. A saddle function k, defined on $W \times Z$, for some nonempty closed convex sets $W \subset \mathbb{R}^K$, $Z \subset \mathbb{R}^L$, gives rise to an equivalence class $[K]$ of saddle functions on $\mathbb{R}^K \times \mathbb{R}^l$, whose lowest and highest elements, $\underline{K}$, $\overline{K}$, are given by

$$\underline{K}(w,z) = \begin{cases} k(w,z) & \text{for } w \in W, z \in Z, \\ -\infty & \text{for } z \notin Z, \\ +\infty & \text{for } w \notin W, z \in Z; \end{cases} \qquad \overline{K}(w,z) = \begin{cases} k(w,z) & \text{for } w \in W, z \in Z, \\ +\infty & \text{for } w \notin W, \\ -\infty & \text{for } w \in W, z \notin Z. \end{cases}$$

Equivalent saddle functions have the same effective domains, on the relative interior of which they are equal to each other (and finite).

For a given saddle function K, the lower conjugate $\underline{K^*}$ and the upper conjugate $\overline{K^*}$ are defined by

$$\underline{K^*}(a,b) = \sup_{u\in\mathbb{R}^k}\inf_{v\in\mathbb{R}^l}\{a\cdot u+b\cdot v-K(u,v)\},\quad \overline{K^*}(a,b) = \inf_{v\in\mathbb{R}^l}\sup_{u\in\mathbb{R}^k}\{a\cdot u+b\cdot v-K(u,v)\}. \tag{31}$$

The lower and upper conjugate functions are equivalent to each other and are, respectively, the lowest and the highest elements of $[K^*]$, the class of saddle functions conjugate to K. In fact, $\underline{K^*}$, $\overline{K^*}$ do not depend on the choice of $K \in [K]$, so $[K^*]$ should be thought of as conjugate to $[K]$. The lower and upper conjugates of any $K^* \in [K^*]$ are, in turn, the lowest and highest elements of $[K]$.

Example 5.1 (Hamiltonian in terms of a conjugate function). Recall that the Hamiltonian (6) was expressed in terms of a function J^*, which can be viewed as a conjugate of J (a unique conjugate, if we request that J^* be finite), where

$$J(u,v) = p\cdot u + \frac{1}{2}u\cdot Pu + q\cdot v - \frac{1}{2}v\cdot Qv - v\cdot Du \quad \text{for } (u,v)\in U\times V, \tag{32}$$

and has appropriately assigned $\pm\infty$ values outside $U\times V$.

Subdifferentials of K^* are exactly the saddle points in the expressions in (31); see Rockafellar [19, Theorem 37.2]. In particular, as finite saddle functions have nonempty subdifferentials, Theorem 2.2 can be viewed as saying that J has a saddle point on $U\times V$ for any (p,q). In other words, the function J_0 below has a saddle point under any affine perturbation. Similarly, Theorem 2.4 states the Lipschitz continuity of saddle points of J^* under perturbations. From a numerical viewpoint, finding the gradients of the Hamiltonian (6) amounts to solving a quadratic minimax problem.

As the linear terms $p\cdot u$ and $q\cdot v$ in (32) do not influence the finiteness and differentiability of $J^*(\cdot,\cdot)$, in proofs of Theorems 2.2 and 2.4 we work with

$$J_0(u,v) = \frac{1}{2}u\cdot Pu - \frac{1}{2}v\cdot Qv - v\cdot Du.$$

(From (7), we get that $J^*(a,b) = J_0^*(a-p,b+q)$.) We will need the following technical lemma.

LEMMA 5.2. *Assume that sets W and Z in $\mathbb{R}^n$ are polyhedral. Then $W+Z=\mathbb{R}^n$ is equivalent to $W^\infty+Z^\infty=\mathbb{R}^n$. For a linear mapping L we have $(LW)^\infty = LW^\infty$.*

Proof. For a polyhedral set W we can conclude that $W\subset W^\infty+\epsilon_w\mathbb{B}$ for some $\epsilon>0$, this follows for example from Corollary 3.53 in [26]. Thus if $W+Z=\mathbb{R}^n$, then $W^\infty+Z^\infty+(\epsilon_w+\epsilon_z)\mathbb{B}=\mathbb{R}^n$. But since $W^\infty+Z^\infty$ is a cone, we must have $W^\infty+Z^\infty=\mathbb{R}^n$. Now assume the latter. We have $W^\infty\subset W-w$ for any $w\in W$. Similarly for Z. Then $W^\infty+Z^\infty\subset W+Z-(w+z)$, which shows that $W+Z=\mathbb{R}^n$. The fact about linear mappings follows directly from the representation of a polyhedral set in Corollary 3.53 in [26]. ∎

For a proper lsc and convex function f, finiteness of f^* is equivalent to coercivity of f. Generalization of this fact to saddle functions, shown by Goebel [13, Proposition 2.7], states that for a proper closed convex-concave function $K:\mathbb{R}^k\times\mathbb{R}^l\mapsto[-\infty,\infty]$, the following conditions are equivalent:

(a) The class $[K^*]$ of convex-concave functions conjugate to K consists of a unique finite-valued function.

(b) The convex function $\alpha(u) = \sup_v K(u,v)$ and the concave $\beta(v) = \inf_u K(u,v)$ are both proper and coercive (respectively, in the convex and concave sense).

A concave function g is coercive (in the concave sense) if $-g$ is coercive as a convex function. Condition (a) can be translated to the following: for every $(a,b) \in \mathbb{R}^k \times \mathbb{R}^l$, $\underline{K^*}(a,b) = \overline{K^*}(a,b)$, and the common value is finite).

Proof (Theorem 2.2). By the result quoted above, $J_0^*(\cdot,\cdot)$ is finite if and only if the convex function

$$\phi(u) = \sup_{v \in V} \left\{ \frac{1}{2} u \cdot Pu - \frac{1}{2} v \cdot Qv - v \cdot Du + \delta_U(u) \right\}$$

and the concave function

$$\psi(v) = \inf_{u \in U} \left\{ \frac{1}{2} u \cdot Pu - \frac{1}{2} v \cdot Qv - v \cdot Du - \delta_V(v) \right\}$$

are proper and coercive. By symmetry, it will suffice to analyze $\phi(\cdot)$. We have

$$\phi(u) = \frac{1}{2} u \cdot Pu + \delta_U(u) + \sup_{v \in V} \left\{ v \cdot (-Du) - \frac{1}{2} v \cdot Qv \right\}.$$

Let $\phi_1(u) = \frac{1}{2} u \cdot Pu + \delta_U(u)$ and $\phi_2(u) = \sup_{v \in V} \left\{ v \cdot (u) - \frac{1}{2} v \cdot Qv \right\}$. Properness of $\phi(\cdot)$ is equivalent to the existence of some $u \in U$ with $\phi_2(-Du)$ finite. As $\operatorname{dom} \phi_2 = (V^\infty \cap \ker Q)^*$ [26, Example 11.18], we get that $\phi(\cdot)$ is proper if and only if $-DU \cap (V^\infty \cap \ker Q)^* \neq \emptyset$. Assuming that this holds, we obtain, through Corollary 11.33 in [26], that the conjugate of the function $u \mapsto \phi_2(-Du)$ at a point w is given by

$$\inf_{v \in V} \left\{ \frac{1}{2} v \cdot Qv \mid w = -D^* v \right\}$$

and the domain of this function is $-D^*V$. The domain of $\phi_1^*(\cdot)$ is $(U^\infty \cap \ker P)^*$. Then the domain of $\phi^*(\cdot)$ is $(U^\infty \cap \ker P)^* + (-D^*V)$. Now the properness and coercivity of $\phi(\cdot)$ is equivalent to $\operatorname{dom} \phi^*(\cdot) = \mathbb{R}^k$. We get that $\phi(\cdot)$ is proper and coercive if and only if

$$-DU \cap (V^\infty \cap \ker Q)^* \neq \emptyset, \qquad -D^*V + (U^\infty \cap \ker P)^* = \mathbb{R}^k.$$

Analogous statements for $\psi(\cdot)$ follow after analyzing the convex function $-\psi(\cdot)$ in the above way. We obtain

$$D^*V \cap (U^\infty \cap \ker P)^* \neq \emptyset, \qquad DU + (V^\infty \cap \ker Q)^* = \mathbb{R}^l.$$

Now note that $-D^*V + (U^\infty \cap \ker P)^* = \mathbb{R}^k$ implies $D^*V \cap (U^\infty \cap \ker P)^* \neq \emptyset$. Indeed, since $0 \in \mathbb{R}^k$, there exists a $v \in V$ such that $0 \in -D^*v + (U^\infty \cap \ker P)^*$. But this means that $D^*v \in (U^\infty \cap \ker P)^*$, so $D^*V \cap (U^\infty \cap \ker P)^* \neq \emptyset$. The latter condition is then superfluous, and a similar statement can be made about $-DU \cap (V^\infty \cap \ker Q)^* \neq \emptyset$.

Using the properties of polyhedral sets in Lemma 5.2, we can translate the condition $DU + (V^\infty \cap \ker Q)^* = \mathbb{R}^l$ to

$$DU^\infty + (V^\infty \cap \ker Q)^* = \mathbb{R}^l.$$

By polarizing both sides of this equation according to the rules in Corollary 11.25 in [26], we get one of the conditions in (8). The other one is obtained symmetrically

from $-D^*V + (U^\infty \cap \ker P)^* = \mathbb{R}^k$. The expression for $(DU^\infty)^*$ and $(-D^*V^\infty)^*$ also come from Corollary 11.25. ∎

Saddle points in the definition (7) of J^* are exactly the subgradients of that function. This allows us to use a result of Dontchev and Rockafellar [11] on the stability of saddle points; we quote it below in a form specialized for our current setting. By $(a,b) \in \partial^s K(w,z)$ we mean that $a \in \partial_w K(w,z)$, $b \in \tilde{\partial}_z K(w,z)$.

LEMMA 5.3 (see [11, Theorem 3.2]). *Assume that $(\bar{u},\bar{v}) \in \partial^s J_0^*(a,b)$. Then a necessary and sufficient condition for $\partial^s J_0^*$ to be single-valued and Lipschitz continuous on a neighborhood of (a,b) is the following:*

$$\begin{cases} u \in U_0 - U_0, \quad Pu = 0, \quad Du \in [V_0 \cap -V_0]^\perp & \Rightarrow \quad u = 0, \\ v \in V_0 - V_0, \quad Qv = 0, \quad D^*v \in [U_0 \cap -U_0]^\perp & \Rightarrow \quad v = 0, \end{cases} \tag{33}$$

where $U_0 = T_U(\bar{u}) \cap (a - P\bar{u} + R^*\bar{v})^\perp$ and $V_0 = T_V(\bar{v}) \cap (b + Q\bar{v} + R\bar{u})^\perp$.

The subspace $U_0 - U_0$ is the smallest subspace containing U_0, whereas $U_0 \cap -U_0$ is the largest subspace contained in the cone U_0. Similarly for V_0.

Proof (Theorem 2.4). For a convex set S, the lineality space S_l of S is the set of all those vectors y, such that for all $x \in S$, the line from x in the direction of y is contained in S. If S is a polyhedral set, $S_l = S^\infty \cap -S^\infty$. Using this notation,

$$[D^*(V^\infty \cap -V^\infty)]^\perp = \left\{u \mid Du \in V_l^\perp\right\},$$

and similarly for the other similar expression in condition (9). Thus, this condition can be restated as

$$\begin{cases} Pu = 0, Du \in V_l^\perp \Rightarrow u = 0, \\ Qv = 0, D^*v \in U_l^\perp \Rightarrow v = 0. \end{cases}$$

We first show that for a closed convex set S and any $w \in N_S(s)$, $S_l \subset w^\perp$. The condition for $w \in N_S(s)$ is that for all $x' \in S$, $(x' - x) \cdot w \le 0$, in particular, for every $l \in S_l$, $l \cdot w \le 0$. But S_l is a subspace, so it must be that $l \cdot w = 0$. This shows that $S_l \subset w^\perp$. Also note that $S_l \subset T_S(s)$.

Pick any (a,b) with $J_0^*(a,b)$ finite. As J_0^* is piecewise linear-quadratic, $\partial^s J_0^*(a,b)$ is nonempty. Pick any $(\bar{u},\bar{v}) \in \partial^s J_0^*(a,b)$. This is equivalent to $(a,b) \in \partial^s J_0(\bar{u},\bar{v})$, meaning $a - P\bar{u} + D^*\bar{v} \in N_U(\bar{u})$ and $b + Q\bar{v} + D\bar{u} \in -N_V(\bar{v})$, and consequently $U_l \subset (a - P\bar{u} + D^*\bar{v})^\perp$ and $V_l \in (b + Q\bar{v} + D\bar{u})^\perp$. This implies that $U_l \subset U_0$ and $V_l \subset V_0$, so then $U_l \subset U_0 \cap -U_0$, $V_l \subset V_0 \cap -V_0$ and also $U_l^\perp \supset (U_0 \cap -U_0)^\perp$, $V_l^\perp \supset (V_0 \cap -V_0)^\perp$.

In view of the above inclusions, condition (9) implies that (33) holds everywhere. That is, in the neighborhood of every point where J^* is finite, this function is also differentiable—therefore, in particular, finite. But the domain of J_0^* is a polyhedral, so also closed, set. Then J_0^* is finite and differentiable everywhere. ∎

A corresponding notion of convergence for convex-concave functions is that of epi/hypo-convergence. We will only use it for sequences of convex-concave functions which are modulated (in the sense of Rockafellar [24]), that is, for sequences which satisfy the following: for some $\rho \ge 0$ and some i_0, we have, for all $i > i_0$,

$$\inf_{|w| \le \rho} \overline{K_i}(w,z) \le \rho(1 + |z|) \quad \forall z, \qquad \sup_{|z| \le \rho} \underline{K_i}(w,z) \ge -\rho(1 + |w|) \quad \forall w. \tag{34}$$

Under Assumption 4.3, the sequence of functions $(y,x) \to H_i(x,y)$ is modulated. This can be seen by looking at the equivalent to Assumption 3.1 growth conditions

on the Hamiltonian, as described in Rockafellar and Wolenski [27], Theorem 2.3; see also our proof of Corollary 4.8. A sequence of (equivalence classes of) convex-concave functions K_i is said to *epi/hypo-converge* to K if

$$\lim_{\epsilon\searrow 0}\left[\limsup_{z_i\to z,i\to\infty}\left(\inf_{|w_i-w|\le\epsilon}\overline{K_i}(w_i,z_i)\right)\right]\le\overline{K}(w,z), \tag{35}$$

$$\lim_{\epsilon\searrow 0}\left[\liminf_{w_i\to w,i\to\infty}\left(\sup_{|z_i-z|\le\epsilon}\underline{K_i}(w_i,z_i)\right)\right]\ge\underline{K}(w,z). \tag{36}$$

LEMMA 5.4 (convergence of finite saddle functions). *Let K_i, $i=1,2,\dots$ and K be finite-valued convex-concave functions on $\mathbb{R}^k\times\mathbb{R}^l$. The following are equivalent:*

(a) *K_i converge epi/hypo-graphically to k,*

(b) *K_i converge pointwise to k,*

(c) *K_i converge uniformly to k on every compact subset of $\mathbb{R}^k\times\mathbb{R}^l$.*

Proof. Assume (a). Subdifferentials of K_i converge graphically to that of K, this follows from an extension of Attouch's theorem for convex functions; see [24, Theorem 4.3]. As subdifferentials of K are convex-valued, Exercise 5.34 in [26] implies the existence of $N>0$, $\epsilon_0>0$ such that, $\|\partial_w K_i(w',z')\|<N$ for $(w',z')\in(w,z)+\epsilon_0\mathbb{B}$. For $\epsilon<\epsilon_0$ we have $\inf_{|w_i-w|\le\epsilon}K_i(w_i,z_i)\ge K_i(w,z_i)-\epsilon N$. Using this in (35) we get

$$\begin{aligned}K(w,z)&\ge\lim_{\epsilon\searrow 0}\left[\limsup_{z_i\to z,i\to\infty}\left(K_i(w,z_i)-\epsilon N\right)\right]\\&\ge\lim_{\epsilon\searrow 0}\left[\limsup_{i\to\infty}\left(K_i(w,z)-\epsilon N\right)\right]=\limsup_{i\to\infty}K_i(w,z).\end{aligned}$$

Symmetric argument shows that $K(w,z)\le\liminf_{i\to\infty}K_i(w,z)$, and thus K_i converge to K pointwise. Implication (b)⇒(c) was shown in [19, Theorem 35.1], while (c)⇒ (a) is simple—it follows from the uniform continuity of K and the definition of epi/hypo-convergence. □

Proof (Lemma 4.4). The equivalence of (a) and (b) follows from the definitions of $\widetilde{L}_i$, $\widetilde{L}$ and the fact that convex conjugacy preserves epi-convergence; see, for example, Theorem 11.34 in [26]. An extension of this fact to partial conjugacy, first shown by Attouch, Aze, and Wets [1] and specialized to modulated sequences in [24, Theorem 4.1], implies that (a) is equivalent to the "hypo/epi-convergence" of H_i to H. As the Hamiltonians are finite, hypo/epi-convergence is equivalent to their pointwise convergence. □

We conclude by discussing the convergence of extended piecewise linear-quadratic problems. Let $\mathcal{C}_i(\tau,\xi)$ be defined as in (3) by matrices A_i, B_i, C_i, D_i, P_i, Q_i, vectors p_i, q_i and sets U_i, V_i. To study the convergence of $\{\mathcal{C}_i(\tau,\xi)\}$ to $\mathcal{C}(\tau,\xi)$ one could analyze the sequence of Lagrangians $\{L_i\}$ defined as in (11), with the help of the calculus of epi-convergence, as described for example in [26, Chapter 7]. We propose an alternate way, suggested by Lemma 4.4 and Example 5.1—we focus on Hamiltonians and rely on the lemma below.

LEMMA 5.5 (convergence of constrained saddle functions and their conjugates). *Suppose that*

(a) *$k_i:\mathbb{R}^k\times\mathbb{R}^l\mapsto\mathbb{R}$, $i=1,2,\dots$, are convex-concave functions converging pointwise to a finite-valued convex-concave function k;*

(b) *$W_i\in\mathbb{R}^k$, $Z_i\in\mathbb{R}^l$, $i=1,2,\dots$, are nonempty closed convex sets converging, respectively, to nonempty closed convex sets W, Z.*

Let $[K_i]$ be the equivalence class of convex-concave functions determined by k_i and $W_i \times Z_i$, similarly define $[K]$ by k and $W \times Z$, and assume that $\{[K_i]\}$ is modulated. Then the sequence $\{[K_i]\}$ epi/hypo-converges to K. Consequently, the sequence $\{[M_i]\}$ given by

$$\underline{M_i}(a,b) = \sup_{w\in W_i}\inf_{z\in Z_i}\{a\cdot w+b\cdot z-k_i(w,z)\},\ \overline{M_i}(a,b) = \inf_{z\in Z_i}\sup_{w\in W_i}\{a\cdot w+b\cdot z-k_i(w,z)\}$$

epi/hypo-converges to $[M]$ described by

$$\underline{M}(a,b) = \sup_{w\in W}\inf_{z\in Z}\{a\cdot w+b\cdot z-k(w,z)\},\ \overline{M}(a,b) = \inf_{z\in Z}\sup_{w\in W}\{a\cdot w+b\cdot z-k(w,z)\}.$$

If all four of the functions above are finite-valued, the equivalence classes $[M_i]$ and $[M]$ consist of just one function each, and the convergence is pointwise.

Proof. We show that (35) holds for $\{K_i\}$ and K; the argument for (36) is symmetrical. When $w \notin W$, there is nothing to prove, as $K(w,z) = +\infty$. Suppose that $w \in W$ and fix $\epsilon > 0$. There exists a sequence $\bar{w}_i \to w$ with $\bar{w}_i \in W_n$, and we have

$$\limsup_{z_i\to z, i\to\infty}\left(\inf_{|w_i-w|\le\epsilon}\overline{K_i}(w_i,z_i)\right) \le \limsup_{z_i\to z,i\to\infty}\overline{K_i}(\bar{w}_i,z_i).$$

If $z \notin Z$, any sequence $z_i \to z$ must eventually satisfy $z_i \notin Z_i$, and thus $\overline{K_i}(\bar{w}_i, z_i) = -\infty$. Thus

$$\limsup_{z_i\to z, i\to\infty}\left(\inf_{|w_i-w|\le\epsilon}\overline{K_i}(w_i,z_i)\right) = -\infty = \overline{K}(w,z).$$

Now note that

$$\begin{aligned}\limsup_{z_i\to z, i\to\infty}\left(\inf_{|w_i-w|\le\epsilon}\overline{K_i}(w_i,z_i)\right) &\le \limsup_{z_i\to z,i\to\infty}\overline{K_i}(\bar{w}_i,z_i)\\ &\le \limsup_{z_n\to z,n\to\infty} k_n(\bar{w}_n,z_n) = k(w,z),\end{aligned}$$

where the equality follows from the fact that k_i converge to k uniformly on any compact neighborhood of (w,z) (Lemma 5.4). If $z \in Z$, $k(w,z) = \overline{K}(w,z)$, and this shows the epi/hypo-convergence of $\{K_i\}$ to K.

Epi/hypo-convergence is preserved under saddle function conjugacy [24, Theorem 4.2]. As $\{M_i\}$ are saddle conjugates of $\{K_i\}$ (in (31) the infimum and supremum need to be taken only over the sets where the function values are finite), epi/hypo-convergence of $\{K_i\}$ to K implies that of $\{M_i\}$ to M. The last statement follows from Lemma 5.4. □

A related result was shown by Wright [29]. It concluded the convergence of $\{K_i\}$, if each K_i had the form $k_i'(w) - k_i''(z) - w \cdot Dz$ (separable saddle function plus a constant biaffine term); convergence of M_i was not addressed there. Also in [29], epi/hypo-convergence was employed to study discrete approximations of $C(\tau,\xi)$.

THEOREM 5.6 (convergence of piecewise linear-quadratic Hamiltonians). *Assume that matrices A_i, B_i, C_i, D_i, P_i, Q_i, vectors p_i, q_i and sets U_i, V_i defining the problem $\mathcal{C}_i(\tau,\xi)$ converge, respectively, to A, B, C, D, P, Q, p, q, U, V defining $\mathcal{C}(\tau,\xi)$. Suppose also that the data in $\mathcal{C}_i(\tau,\xi)$, $i = 1,2,\ldots$, and $C(\tau,\xi)$ satisfies the conditions of Theorem 2.2. Then Hamiltonians H_i converge pointwise to H.*

Proof. The sequence of functions J_i corresponding to $\mathcal{C}_i(\tau,\xi)$ as in (32) is modulated (too see this, note that there exist $u_i \in U_i$ converging to some $u \in U$, and for

some $\rho > 0$, $\inf_{|u|\le\rho} \overline{J}_i(u,v)$ is bounded above by $p_i \cdot u_i + \frac{1}{2} u_i \cdot P_i u_i + q_i \cdot v - v \cdot D_i u_i$; this shows the first inequality in (34)). The quadratic expressions defining J_i in (32) converge pointwise (on the whole space) to that of J. Lemma 5.5 guarantees that $\{J_i\}$ as well as $\{J_i^*\}$ epi/hypo-converge to, respectively, J and J^*. As the functions J_i^* and J^* are finite, their convergence is uniform on compact sets by Lemma 5.4. But then, it also implies the pointwise convergence of Hamiltonians H_i. □

REFERENCES

[1] H. Attouch, D. Azé, and R. J.-B. Wets, *Convergence of convex-concave saddle functions: Applications to convex programming and mechanics*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 5 (1988), pp. 537–572.

[2] H. Attouch and R. J.-B. Wets, *A convergence theory for saddle functions*, Trans. Amer. Math. Soc., 280 (1983), pp. 1–41.

[3] A. Bemporad, M. Morari, V. Dua, and E. Pistikopoulos, *The explicit linear quadratic regulator for constrained systems*, Automatica J. IFAC, 38 (2002), pp. 3–20.

[4] A. Briani, *Convergence of Hamilton-Jacobi equations for sequences of optimal control problems*, Commun. Appl. Anal., 4 (2000), pp. 227–244.

[5] G. Buttazzo and G. Dal Maso, Γ*-convergence and optimal control problems*, J. Optim. Theory Appl., 38 (1982), pp. 385–407.

[6] C. Byrnes, *On the Riccati partial differential equation for nonlinear Bolza and Lagrange problems*, J. Math. Systems Estim. Control, 8 (1998), pp. 1–54.

[7] C. Byrnes and H. Frankowska, *Unicité des solutions optimales et absence de chocs pour les équations d'Hamilton-Jacobi-Bellman et de Riccati*, C. R. Acad. Sci. Paris Ser. I Math, 315 (1992), pp. 427–431.

[8] N. Caroff and H. Frankowska, *Optimality and characteristics of Hamilton-Jacobi-Bellman equations*, in Optimization, Optimal Control, and Partial Differential Equations Internat. Ser. Numer. Math., 107, Birkhäuser, Basel, 1992, pp. 169–180.

[9] N. Caroff and H. Frankowska, *Conjugate points and shocks in nonlinear optimal control*, Trans. Amer. Math. Soc., 348 (1996), pp. 3133–3153.

[10] F. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[11] A. Dontchev and R. Rockafellar, *Primal-dual solution perturbations in convex optimization*, Set-Valued Anal., 9 (2001), pp. 49–65.

[12] R. Goebel, *Convexity, Convergence and Feedback in Optimal Control*, Ph.D. thesis, University of Washington, Seattle, WA, 2000.

[13] R. Goebel, *Convexity in zero-sum differential games*, SIAM J. Control Optim., 40 (2002), pp. 1491–1504.

[14] R. Goebel, *Regularity of the optimal feedback and the value function in convex problems of optimal control*, Set-Valued Anal., 12 (2004), pp. 127–145.

[15] R. Goebel and R. Rockafellar, *Generalized conjugacy in Hamilton-Jacobi theory for fully convex Lagrangians*, J. Convex Anal., 9 (2002), pp. 463–473.

[16] W. Heemels, S. V. Eijndhoven, and A. Stoorvogel, *Linear quadratic regulator with positive controls*, Internat. J. Control, 70 (1998), pp. 551–578.

[17] J. Joly and F. Thelin, *Convergence of convex integrals in l^p spaces*, J. Math. Anal. Appl., 54 (1976), pp. 230–244.

[18] R. Rockafellar, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.

[19] R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[20] R. Rockafellar, *Generalized Hamiltonian equations for convex problems of Lagrange*, Pacific J. Math., 33 (1970), pp. 411–427.

[21] R. Rockafellar, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.

[22] R. Rockafellar, *Linear-quadratic programming and optimal control*, SIAM J. Control Optim., 25 (1987), pp. 781–814.

[23] R. Rockafellar, *Hamiltonian trajectories and duality in the optimal control of linear systems with convex costs*, SIAM J. Control Optim., 27 (1989), pp. 1007–1025.

[24] R. Rockafellar, *Generalized second derivatives of convex functions and saddle functions*, Trans. Amer. Math. Soc., 322 (1990), pp. 51–77.

[25] R. Rockafellar, *Large-scale extended linear-quadratic programming and multistage optimization*, in Advances in Numerical Partial Differential Equations and Optimization (Merida, 1989), SIAM, Philadelphia, 1991, pp. 247–261.

[26] R. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[27] R. Rockafellar and P. Wolenski, *Convexity in Hamilton–Jacobi theory,* I: *Dynamics and duality*, SIAM J. Control Optim., 39 (2000), pp. 1323–1350.
[28] R. Rockafellar and C. Zhu, *Primal-dual projected gradient algorithms for extended linear-quadratic programming*, SIAM J. Optim., 3 (1993), pp. 751–783.
[29] S. Wright, *Consistency of primal-dual approximations for convex optimal control problems*, SIAM J. Control Optim., 33 (1995), pp. 1489–1509.
[30] C. Zhu, *On a certain parameter of the discretized extended linear-quadratic problem of optimal control*, SIAM J. Control Optim., 34 (1996), pp. 62–73.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1812–1843

A NONDEGENERATE MAXIMUM PRINCIPLE FOR THE IMPULSE CONTROL PROBLEM WITH STATE CONSTRAINTS*

A. ARUTYUNOV†, D. KARAMZIN‡, AND F. PEREIRA§

Abstract. In this article, a free-time impulsive control problem with state constraints and equality and inequality constraints on the trajectory endpoints is considered. A weakened maximum principle is obtained for problems with data measurable in the time variable, being the time transversality conditions deduced with the help of some extra convexity assumption on the state constraints. In the case of smooth problems a nondegenerate maximum principle is derived by using a penalty function method.

Key words. optimal control, impulsive control, maximum principle, state constraints, nonfixed time, time transversality conditions, nondegeneracy

AMS subject classifications. 49K15, 49N25

DOI. 10.1137/S0363012903430068

1. Statement of the problem.

We shall address the following impulse control optimization problem:

$$J(p,u,\mu) = e_0(p) \to \min, \tag{1.1}$$

$$dx = f(x,u,t)dt + g(x,t)d\mu, \quad t \in [t_0,t_1], \tag{1.2}$$

$$e_1(p) \le 0, \quad e_2(p) = 0, \tag{1.3}$$

$$\varphi(x,t) \le 0, \quad t \in [t_0,t_1], \tag{1.4}$$

$$u(t) \in U(t) \quad [t_0,t_1]\text{-a.e.}, \quad \mu \ge 0, \tag{1.5}$$

$$p = (x_0,x_1,t_0,t_1), \quad x_0 = x(t_0),\ x_1 = x(t_1).$$

Here e_1, e_2, φ are given vector-functions with values in $\Re^{k_j}$, $j = 1,2,3$, respectively, $t \in \Re^1$ is the time variable, μ is a nonnegative scalar valued Borel measure on time interval $[t_0,t_1]$, referred to by impulse control, and x is the state variable with values in $\Re^n$. The notation a.e. stands for almost all $t \in [t_0,t_1]$ with respect to Lebesgue measure. The vector u with values in $\Re^m$ is called control. An admissible control is an essentially bounded measurable function $u(t)$ such that $u(t) \in U(t)$ a.e. The vector $p \in \Re^{2n+2}$ is called an endpoint.

*Received by the editors June 16, 2003; accepted for publication (in revised form) July 19, 2004; published electronically March 22, 2005.
http://www.siam.org/journals/sicon/43-5/43006.html

†Differential Equations and Functional Analysis Department, Peoples Friendship University of Russia, Mikluka-Maklai, 6, Moscow 117198, Russia (arutun@orc.ru). This author's research was supported by Russian Foundation for Basic Research grant N02-01-00334.

‡Department of Calculus Mathematics and Cybernetics, Moscow State University, Vorob'yovi gory, Moscow 119899, Russia (dmitry_karamzin@mail.ru). This author's research was supported by Russian Foundation for Basic Research grant N02-01-00334.

§Institute for Systems and Robotics, Faculdade de Engenharia, Universidade do Porto, R. Dr. Roberto Frias, 4200-465 Porto, Portugal (flp@fe.up.pt). This author's research was supported by INVOTAN and by the Portuguese Science and Technology Foundation under project CorDyAL.

The functions e_0, e_1, e_2, φ, g, defining the minimizing functional, endpoint, and state constraints, and the impulse driven dynamics are continuously differentiable in all their arguments. The vector function f is linear in u, i.e.,

$$f(t,x,u) = f_0(t,x) + F(t,x)u, \tag{1.6}$$

continuously differentiable in x for almost all t, and, together with its partial derivatives in x, is measurable in t for all fixed (x,u). Both f and its partial derivatives in x are bounded on any bounded set and continuous in (x,u) uniformly in t. The set-valued mapping U is measurable in t and bounded, i.e., $\exists\ c > 0$, s.t. $|U(t)| \leq c$ $\forall t$. The set $U(t)$ is convex and closed for almost all t.

The triple (p,u,μ) is said to be a *control process* if $\exists\ x(t) : x(t_0) = x_0$, $x(t_1) = x_1$ and x,u,μ satisfy (1.2). A control process is called *admissible* if it satisfies all constraints of the considered problem. The admissible process (p^*,u^*,μ^*) is said to be *optimal* if for any admissible process (p,u,μ), the inequality $e_0(p^*) \leq e_0(p)$ is true.

The goal for this article is to derive necessary conditions for optimality in the form of a nondegenerate maximum principle for the free-time impulsive control problem (1.1)–(1.5) with control constraints. The maximum principle is said to degenerate if condition (4.3) in our main result, Theorem 4.1, does not hold.

If condition (4.3) does not hold, then it can be easily seen that the maximum condition may not be informative in the sense that it holds for any admissible controls (u,μ). Indeed, that is so when $\psi(t) = 0$ a.e., $\sigma_r(s) = 0$ a.e. (see Theorem 4.1). For example, we may point to a specific class of problems for which the classical maximum principle always degenerates: the autonomous time-optimal problem with state constraints and fixed endpoints which belong to the boundary of the state constraints.

This article pertains to the significant effort made over the years to extend the conventional optimal control theory, addressing systems with trajectories which are continuous, more precisely, absolutely continuous, to systems whose trajectories might present discontinuities.

Despite the already significant body of theory addressing optimal impulsive control problems developed during the last two decades (consider, for example, [2, 4, 5, 7, 13, 10, 14, 15, 16, 17, 18, 20, 21, 22, 23], none of the authors that derived necessary conditions of optimality for impulsive control problems with state constraints, namely, [4, 15, 17], addressed the issue of nondegeneracy of the optimality conditions.

The issue of nondegeneracy is an important one in optimal control and, for conventional optimal control problems with state constraints, it was addressed by, among others, [1, 3, 6]. Furthermore, a weakened maximum principle (Theorem 7.1) for free-time impulsive control problems with data merely measurable in the time variable was derived, being the time transversality conditions obtained with the help of some extra convexity assumption on the state constraints. This result is also new.

This article is organized as follows. After addressing in the next section the set of hypotheses under which our results are proved, we present and discuss some preliminary concepts. In section 4, we state our main result and present some remarks as well as an illustrative example. Then, some auxiliary results are stated in section 5. Their proof is presented in the appendix. In section 6, we state and prove two results of increasing order of complexity. While in the first result we consider the free-time impulsive optimal control problem without constraints, in the second result only add endpoint state constraints. The weakened version of our main result is stated and proved for the full problem in section 7. Finally, in section 8, we derive our main result.

2. Hypotheses. Let us formulate basic definitions and assumptions.

DEFINITION 2.1. *Endpoint constraints are said to be* regular *if, for any endpoint vector* $p = (x_0, x_1, t_0, t_1)$ *satisfying* (1.3),

(1) *the vectors* $\frac{\partial e_2^j}{\partial p}(p)$, $j = 1, \dots, k_2$, *are linearly independent;*

(2) *there exists vector* $\bar{p} \in \Re^{2n+2}$ *such that*

$$\frac{\partial e_2}{\partial p}(p)\bar{p} = 0, \qquad \left\langle \frac{\partial e_1^j}{\partial p}(p), \bar{p} \right\rangle > 0 \quad \forall j \ \textit{s.t.}\ e_1^j(p) = 0.$$

DEFINITION 2.2. *State constraints are said to be* regular *if, for any* (x, t) *satisfying* (1.4), *there exists vector* $q = q(x, t) \in \Re^n$ *such that*

$$\langle \varphi_x^j(x, t), q \rangle > 0 \quad \forall j \ \textit{s.t.}\ \varphi^j(x, t) = 0.$$

DEFINITION 2.3. *Let vector* $p^* = (x_0^*, x_1^*, t_0^*, t_1^*)$ *satisfy endpoint constraints* (1.3) *and* $\varphi(x_k^*, t_k^*) \le 0$, $k = 0, 1$. *We say that state constraints are* compatible *with endpoint constraints at* p^* *if there exists* $\varepsilon > 0$ *such that*

$$\{p \in \Re^{2n+2} : |p - p^*| \le \varepsilon,\ e_1(p) \le 0,\ e_2(p) = 0\} \subseteq \{p : \varphi(x_k, t_k) \le 0,\ k = 0, 1\}.$$

Assumption S. The function f is continuously differentiable in all arguments, and the set-valued map $U(\cdot)$ is constant. To be more precise, $U(t) \equiv U$, being U convex and compact.

DEFINITION 2.4. *Let Assumption* S *be in force. The admissible trajectory* $x(t)$, $t \in [t_0, t_1]$, *is called* controllable *at the endpoints (with regard to state constraints) if there exist* $u_k \in U$ *and* $m_k \in [0, +\infty)$, $k = 0, 1$, *such that*

$$(-1)^k[\langle f(x_k, u_k, t_k) + g(x_k, t_k)m_k, \varphi_x^j(x_k, t_k)\rangle + \varphi_t^j(x_k, t_k)] < 0$$

for all j *such that* $\varphi^j(x_k, t_k) = 0$. *Here* $x_k = x(t_k)$, $k = 0, 1$.

When $g \equiv 0$ or μ is in class of absolutely continuous measures having density in L_∞, then Definition 2.4 becomes a definition of controllability for the associated conventional control optimization problem [1]. A simple argument gives the following sufficient condition for controllability: a trajectory $x(\cdot)$ is controllable at the endpoints if at least one of the two following conditions holds:

- There exists $u_k \in U$ such that $\forall j$ satisfying $\varphi^j(x_k, t_k) = 0$,

$$(-1)^k[\langle f(x_k, u_k, t_k), \varphi_x^j(x_k, t_k)\rangle + \varphi_t^j(x_k, t_k)] < 0.$$

- $(-1)^k W^j(x_k, t_k) < 0$ $\forall j$ satisfying $\varphi^j(x_k, t_k) = 0$.

From now on, let $W^j(x, t) = \langle \varphi_x^j(x, t), g(x, t)\rangle$, $j = 1, \dots, k_3$.

We need Definitions 2.1–2.4 and Assumption S to prove nondegeneracy (Theorem 4.1).

The following *convexity* assumption is used along with Definition 2.3 to obtain time transversality conditions for problems with data measurable in t and state constraints (Theorem 7.1).

Assumption C. Let $x^*(\cdot)$ be the reference optimal trajectory on $[t_0^*, t_1^*]$ and $E_0^* = (x_0^{*+}, t_0^*)$, $E_1^* = (x_1^{*-}, t_1^*)$. Let the function φ be twice continuously differentiable.

There exists $\delta > 0$ such that

$$\dot{W}^j(E) \geq 0 \quad \forall E \in \Re^n \times \Re \text{ s.t. } |E - E_k^*| \leq \delta \quad \forall j,k \text{ s.t. } \varphi^j(E_k^*) = W^j(E_k^*) = 0, \tag{2.1}$$

where $\dot{W}^j = \langle g_x g, \varphi_x^j \rangle + \langle g, \varphi_{xx}^j g \rangle$ is the evolution derivative of W^j at any point of the arc joining the jump endpoints.

Here are a few comments about condition (2.1). It always holds when $W^j(E_k^*) \neq 0$. If $\varphi^j(E_k^*) = W^j(E_k^*) = 0$, then (2.1) means that function φ^j is convex along jump evolutions in some neighborhood of the endpoint. In this case, examples where (2.1) holds are

- linear systems—g does not depend on x, and φ is linear in x;
- conventional system—$g \equiv 0$ (then in both cases $\dot{W}^j \equiv 0$); and
- the function g does not depend on x, and the matrix φ_{xx}^j is nonnegative defined.

Finally, (2.1) is true when $\dot{W}^j(E_k^*) > 0$.

3. Preliminaries. Let us introduce the adopted notation and present the main concepts to be used in this article.

Denote by $T = [t_0, t_1]$ the time interval and by $C(T)$ the Banach space of continuous functions $f : T \to \Re^1$ with the usual norm $\|f\|_C = \max_{t \in T} |f(t)|$. Let $V(T)$ be the linear space of functions of bounded variation on T, right continuous on interval (t_0, t_1), and $V^n(T)$ be the space of n-vector valued functions $x(t) = (x^1(t), \ldots, x^n(t))$ such that $x^j \in V(T)$, $j = 1, \ldots, n$. Denote by $C^*(T)$ the topologically dual space to $C(T)$ (its elements are Borel measures on T) and by $C_+^*(T)$ the class of nonnegative Borel measures on T.

Given $\mu \in C^*(T)$, let the distribution function $F(t;\mu)$ of the Borel measure μ be defined by

$$F(t;\mu) = \int_{[0,t]} d\mu = \mu([0,t]), \quad t \in (0,1], \qquad F(0;\mu) = 0.$$

The variation of μ is an element in $C_+^*(T)$ defined as the Borel measure $\operatorname{Var}\mu = \mu^+ + \mu^-$, where $\mu = \mu^+ - \mu^-$ is the Jordan decomposition of μ. The total variation of μ is given by the number $|\mu| = \operatorname{Var}\mu(T)$. The variation of a vector-measure (vector-function) is defined as the sum of the variation of its components.

We denote by μ_d and μ_c, respectively, the discrete and the continuous part of the measure μ. Let us put $\operatorname{Ds}(\mu) = \{r \in T : \operatorname{Var}\mu(\{r\}) > 0\}$, $\operatorname{Cont}(\mu) = [T \setminus \operatorname{Ds}(\mu)] \cup \{t_0\} \cup \{t_1\}$. By $\mu_1 \leq \mu_2$ it is meant that $\mu_2 - \mu_1 \in C_+^*(T)$. The weak star convergence of a sequence of measures $\{\mu_i\}$ to μ is represented by $\mu_i \xrightarrow{w} \mu$.

Given $g \in V(T)$, let $g(s^+) = \lim_{t \to s,\, t > s} g(t)$, $g(s^-) = \lim_{t \to s,\, t < s} g(t)$ be, respectively, the right and the left limits of g at point s. Function g defines the Borel measure $\nu[g]$ by the formula

$$\nu[g]([0,t]) = g(t^+) - g(0), \quad t \in [0,1].$$

Thus, according to our notation, $\nu[F(t;\mu)] = \mu$. The measure corresponding to the length is denoted by $\mathcal{L}$, $\mathcal{L} = \nu[t]$. If $x \in V^n(T)$, then $\nu[x]$ is a vector measure with components $\nu[x^j]$, $j = 1, \ldots, n$. $\operatorname{Var}|_a^b[g]$ denotes the variation of the measure $\nu[g]$ on $[a,b]$.

We present the concept of the solution to (1.2), following the one given first in [8, 9] and later completed and further discussed in [2, 13, 11, 12, 20, 23, 5, 16, 19]. Let

us consider the function $\xi(x,r,\Delta):\Re^n\times\Re^1\times\Re^1\to\Re^n$, with $\xi(x,r,\Delta)=\alpha_r(1)$, being $\alpha_r(s)=\alpha_r(s;x,\Delta)$ the solution to the following conventional differential equation:

$$\dot\alpha_r=g(\alpha_r,r)\Delta,\qquad s\in[0,1],\qquad \alpha_r(0)=x.$$

DEFINITION 3.1. *The function $x(t)$ is called the solution to* (1.2) *if $x(t_0)=x_0$ and, $\forall t>t_0$,*

$$x(t)=x_0+\int_{t_0}^{t}f(x,u,s)ds+\int_{[t_0,t]}g(x,s)d\mu_c+\sum_{r\in\mathrm{Ds}(\mu),\ r\le t}[\xi(x(r^-),r,\mu(\{r\}))-x(r^-)].$$

Here, $\mu(\{r\})$ is the atom of μ at point r.

Let $\phi(x,t)$ be a scalar continuous function on $\Re^n\times\Re^1$ and consider the following concept of graph completion supremum:

$$\operatorname*{gc\,sup}_{t\in T}\phi(x,t)=\max_{t\in T}\max_{s\in[0,1]}\phi(\alpha_t(s;x(t^-),\mu(\{t\})),t).$$

Now, we are in position to precise the definition of trajectory admissability introduced in section 1.

DEFINITION 3.2. *Inequality* (1.4), *i.e., $\varphi(x,t)\le 0$, is understood in the sense that for $j=1,\dots,k_3$, $\operatorname{gc\,sup}\varphi^j(x,t)\le 0$. This means that the trajectory $x(t)$ satisfies the state constraints* (1.4) *if and only if*

(1) $\varphi(x(t),t)\le 0\ \forall t\in[t_0,t_1]$, *and*
(2) $\varphi(\alpha_r(s),r)\le 0\ \forall s\in[0,1]$ *and* $\forall r\in\mathrm{Ds}(\mu)$.

4. The main result. In this section, we state nondegenerate necessary conditions for optimal control problems with state constraints and equality and inequality endpoint constraints derived in its full generality.

Let $\lambda=(\lambda_0,\lambda_1,\lambda_2)$ and consider the following scalar functions:

$$H(x,u,\psi,t)=\langle f(x,u,t),\psi\rangle,$$
$$Q(x,\psi,t)=\langle g(x,t),\psi\rangle,$$
$$l(p,\lambda)=\sum_{j=0}^{2}\langle e_j(p),\lambda_j\rangle.$$

THEOREM 4.1. *Let (p^*,u^*,μ^*) be a solution to problem* (1.1)–(1.5). *Suppose that Assumption* S *is in force, the state constraints are compatible with endpoint constraints, the state and endpoint constraints are regular, and the optimal trajectory is controllable.*

Then, there exist

- *number $\lambda_0\ge 0$, vectors $\lambda_1\in\Re^{k_1}$, $\lambda_1\ge 0$, $\lambda_2\in\Re^{k_2}$,*
- *vector function $\psi\in V^n(T^*)$,*
- *scalar function $\phi\in V(T^*)$,*
- *vector measure $\eta=(\eta^1,\dots,\eta^{k_3})$, $\eta^j\in C_+^*(T^*)$, s.t. $\mathrm{Ds}(\mu^*)\cap\mathrm{Ds}(\eta^j)=\emptyset\ \forall j$,*

and

- *for every atom $r\in\mathrm{Ds}(\mu^*)$, there exist its own*
  - *vector function $\sigma_r\in V^n([0,1])$,*
  - *scalar function $\theta_r\in V([0,1])$, and*
  - *vector measure $\eta_r=(\eta_r^1,\dots,\eta_r^{k_3})$, $\eta_r^j\in C_+^*([0,1])$, $j=1,\dots,k_3$,*

such that, for all $t \in (t_0^*, t_1^*]$ *and all* $s \in [0,1]$,

$$\psi(t) = \psi_0 - \int_{t_0^*}^{t} H_x(s)ds - \int_{[t_0^*,t]} Q_x(s)d\mu_c^* + \int_{[t_0^*,t]} \varphi_x^\top(x^*,s)d\eta + \Sigma(\psi,t),$$

$$\Sigma(\psi,t) = \sum_{r\in \mathrm{Ds}(\mu^*),\ r\le t} [\sigma_r(1) - \psi(r^-)],$$

$$\phi(t) = \phi_0 + \int_{t_0^*}^{t} H_t(s)ds + \int_{[t_0^*,t]} Q_t(s)d\mu_c^* - \int_{[t_0^*,t]} \varphi_t^\top(x^*,s)d\eta + \Theta(\phi,t), \tag{4.1}$$

$$\Theta(\phi,t) = \sum_{r\in \mathrm{Ds}(\mu^*),\ r\le t} [\theta_r(1) - \phi(r^-)],$$

$$\begin{cases} d\alpha_r^*(s) = g(\alpha_r^*(s),r)\Delta_r^* ds, \\ d\sigma_r(s) = -g_x^\top(\alpha_r^*(s),r)\sigma_r(s)\Delta_r^* ds + \varphi_x^\top(\alpha_r^*(s),r)d\eta_r, \\ d\theta_r(s) = \langle g_t(\alpha_r^*(s),r),\sigma_r(s)\rangle \Delta_r^* ds - \varphi_t^\top(\alpha_r^*(s),r)d\eta_r, \\ \alpha_r^*(0) = x^*(r^-), \qquad \sigma_r(0) = \psi(r^-), \qquad \theta_r(0) = \phi(r^-), \qquad \Delta_r^* = \mu^*(\{r\}), \end{cases}$$

$$\psi_0 = \frac{\partial l}{\partial x_0}(p^*,\lambda), \qquad \psi_1 = -\frac{\partial l}{\partial x_1}(p^*,\lambda),$$

$$\phi_0 = -\frac{\partial l}{\partial t_0}(p^*,\lambda), \qquad \phi_1 = \frac{\partial l}{\partial t_1}(p^*,\lambda),$$

$$\langle g(\alpha_r^*(s),r),\sigma_r(s)\rangle = 0 \quad \forall r \in \mathrm{Ds}(\mu^*),$$

$$\mathrm{supp}(\eta_r^j) \subseteq \{s\in[0,1] : \varphi^j(\alpha_r^*(s),r) = W^j(\alpha_r^*(s),r) = 0\} \quad \forall j,$$

$$\langle \lambda_1, e_1(p^*)\rangle = 0,$$

$$\varphi^j(x^*(t),t) = 0 \ \ \eta^j\text{-a.e.} \quad \forall j,$$

$$\max_{u\in U} H(u,t) = H(t) \ \ a.e., \qquad \max_{u\in U} H(u,t) = \phi(t) \quad \forall t \in (t_0^*, t_1^*), \tag{4.2}$$

$$Q(t) \le 0 \quad \forall t, \qquad Q(t) = 0 \quad \mu^*\text{-a.e.},$$

$$\lambda_0 + \mathcal{L}(\{t : |\psi(t)| > 0\}) + \sum_{r\in \mathrm{Ds}(\mu^*)} \mathcal{L}(\{s : |\sigma_r(s)| > 0\})\Delta_r^* = 1. \tag{4.3}$$

In this result and from now on, we adopt the following short notation: $T^* = [t_0^*, t_1^*]$, $\psi_k = \psi(t_k^*)$, $\phi_k = \phi(t_k^*)$, $\Delta_k^* = \mu^*(\{t_k^*\})$, $k = 0,1$, $H(t,u) = H(x^*(t),u,\psi(t),t)$, $H(t) = H(t,u^*(t))$, $Q(t) = Q(x^*(t),\psi(t),t)$, $H_x(t) = H_x(x^*(t),u^*(t),\psi(t),t)$, ..., etc. In other words, if H, Q, or their partial derivatives miss some of arguments x, ψ, u, then it is understood that the values $x^*(t)$, $\psi(t)$, and $u^*(t)$ are considered in their place.

The proof is presented in sections 6, 7, and 8. It is based on a penalty function method [1], results from [16, 20], and on the reduction to the so-called v-problem (Proposition 5.7).

Remark 1. From Theorem 4.1, we deduce $\mathrm{Var}\,\nu[Q] \le \mathrm{const}(\mathrm{Var}\,\eta + \mathcal{L})$. This implies that the function $Q(t)$ is continuous on $\mathrm{Ds}(\mu^*)$.

Remark 2. The time transversality conditions in Theorem 4.1 can be rewritten as follows:

$$\begin{aligned} \max_{u\in U} H(\beta_k^*,u,\gamma_k,t_k^*) + (-1)^{k+1}\Bigg[\Delta_k^* \int_0^1 \langle g_t(\alpha_k^*(s),t_k^*),\sigma_k(s)\rangle ds \\ - \int_{[0,1]} \varphi_t^\top(\alpha_k^*(s),t_k^*)d\eta_k - \varphi_t^\top(x_k^*,t_k^*)\eta(\{t_k^*\}) \\ - \frac{\partial l}{\partial t_k}(p^*,\lambda)\Bigg] = 0, \quad k = 0,1. \end{aligned} \tag{4.4}$$

From now on, $\beta_0^* = x^*(t_0^{*+})$, $\beta_1^* = x^*(t_1^{*-})$, $\gamma_0^* = \psi(t_0^{*+})$, $\gamma_1^* = \psi(t_1^{*-})$, $\Delta_k^* = \mu^*(\{t_k^*\})$, $k = 0, 1$, being α_k^*, σ_k, and η_k, the jump evolution elements at point t_k^*, $k = 0, 1$.

Below, we give a simple example showing that the maximum principle proved in Theorem 4.1 may degenerate when the controllability condition (Definition 2.4) does not hold.

Example 4.2. Consider optimal problem

$$dx = 2td\mu, \quad t \in [0,1]; \qquad x_0 = 0, \qquad x_1 = 1;$$
$$x \geq t^2, \qquad \int_{[0,1]} 1 d\mu \to \min.$$

Now, we show that minimum is reached for the Lebesgue measure $\mu^* = \mathcal{L}$. The corresponding optimal trajectory is the parabola $x^*(t) = t^2$. In fact, this statement is a consequence of the inequality $F(t;\mu) \geq t \ \forall t \in [0,1]$, which holds for any admissible measure μ. Let us prove it. We have

$$x(t) \geq t^2 \ \Rightarrow \ \int_{[0,t]} 2td\mu \geq t^2.$$

By integrating by parts, we have $2F(t;\mu)t \geq t^2 + 2\int_{[0,t]} F(s;\mu)ds$ and, thus,

$$F(t;\mu) \geq \frac{t}{2} + \frac{1}{t}\int_{[0,t]} F(s;\mu)ds.$$

It follows that $F(t;\mu) \geq t/2$. By substituting in the right part of the obtained inequality, we arrive at the more precise estimate $F(t;\mu) \geq 3t/4$. By repeating the procedure, we get, for step n, $F(t;\mu) \geq (2^n - 1)t/2^n$, and, by passing to the limit as $n \to \infty$, we establish $F(t;\mu) \geq t$. Thus $\mu^* = \mathcal{L}$.

However, the maximum principle proved in Theorem 4.1 degenerates for this problem. Indeed, the impulsive maximum condition yields $2t\psi(t) - \lambda_0 = 0$ for a.e. $t \in [0,1]$. By passing to the limit when $t \to 0$ and by bearing in mind that $\psi(t)$ is bounded, we conclude that $\lambda_0 = 0$. Hence, $\psi(t) = 0 \ \forall t \in (0,1)$. It follows that (4.3) does not hold. This happens because the optimal trajectory is not controllable.

5. Lemmas and propositions. In this section, we compile a set of auxiliary lemmas and propositions that we will use in the proof of the main result. Their proofs appear in the appendix. While Lemmas 5.1 to 5.5 are of independent interest in themselves, Propositions 5.6 and 5.7 support the progressive reduction strategy adopted in the proof of Theorems 4.1, 6.1, 6.2, and 7.1.

LEMMA 5.1. *Given a sequence of measures $\mu_i \in C_+^*(T)$, $\mu_i \xrightarrow{w} \mu$, and a sequence of functions $f_i \in C(T)$, such that*

(1) *there exists a sequence of absolutely continuous measures $\eta_i \in C_+^*(T)$, $\eta_i \xrightarrow{w} \eta$ such that $\mathrm{Ds}(\mu) \cap \mathrm{Ds}(\eta) = \emptyset$ and $\mathrm{Var}\,\nu[f_i] \leq c\eta_i \ \forall i$;*

(2) *$f_i(t) \to f(t) \ \forall t \in \mathrm{Cont}(\eta)$, where $f \in V(T)$.*

Then

$$\int_{[t_0,t_1]} f_i(t) d\mu_i \to \int_{[t_0,t_1]} f(t) d\mu.$$

Remark 3. The following example shows that we cannot omit requirement (1) in Lemma 5.1.

Example 5.2. Let $T = [0,1]$, and $d\mu_i = \delta_{t_i}(t)$ is the sequence of Dirac's measures, concentrated at points $t_i = \frac{1}{i}$. Consider the following sequence of (bump) functions $\{f_i\}$:

$$f_i(t) = \begin{cases} it, & t \in [0, i^{-1}], \\ 2 - it, & t \in [i^{-1}, 2i^{-1}], \\ 0, & t \in [2i^{-1}, 1]. \end{cases}$$

It is clear that $\mu_i \xrightarrow{w} \mu$, where $d\mu = \delta_0(t)$, and $f_i(t) \to 0\ \forall t \in [0,1]$. Nevertheless $\int_{[0,1]} f_i(t)d\mu_i = 1\ \forall i$, while $\int_{[0,1]} f(t)d\mu = 0$. The fact is that for the sequences $\{\mu_i\}$ and $\{f_i\}$ constructed above, there is no sequence $\{\eta_i\}$ satisfying the requirements of Lemma 5.1. Indeed, $\operatorname{Var}\nu[f_i] \xrightarrow{w} \eta$ with $d\eta = 2\delta_0(t)$, but $\operatorname{Ds}(\mu) \cap \operatorname{Ds}(\eta) = \{0\} \neq \emptyset$.

LEMMA 5.3. *Let all the hypotheses of Lemma* 5.1 *hold. Let* $x_i(t_0) = x(t_0) = 0,\ i \in \mathbf{N}$, *and* $\forall t \in (t_0, t_1]$,

$$x_i(t) = \int_{[t_0,t]} f_i(s)d\mu_i, \qquad x(t) = \int_{[t_0,t]} f(s)d\mu.$$

For any given convergent sequence $\{t_i\}$, $t_i \to s \in T$, *where* $s \notin \operatorname{Ds}(\mu)$, *we have that* $x_i(t_i) \to x(s)$.

LEMMA 5.4. *Let* $\mu_i \in C_+^*(T)$, $\mu_i \xrightarrow{w} \mu$, $u_i \xrightarrow{w} u$ *weakly in* $L_2(T)$, $u_i(t) \in U(t)$ *a.e.,* $x_{0,i} \to x_0$, *and given a sequence of vector-functions* $x_i \in V^n(T)$ *such that*

$$dx_i = f(x_i, u_i, t)dt + g(x_i, t)d\mu_i, \quad t \in T,\ x_i(t_0) = x_{0,i}.$$

If $\operatorname{gc\,sup}_{t\in T} |x_i(t)| \leq \text{const}\ \forall i$, *then* $x_i(t) \to x(t)\ \forall t \in \operatorname{Cont}(\mu)$, *where* $x(t)$ *is the solution to* (1.2). *Furthermore,* $\operatorname{gc\,sup}_{t\in T} |x_i(t)| \to \operatorname{gc\,sup}_{t\in T} |x(t)|$.

LEMMA 5.5. *Let the sequences* $\{f_i\}$ *and* $\{\eta_i\}$ *be such that* $f_i \in C(T)$, $\eta_i \in C_+^*(T)$, *and* $f_i \to f \in C(T)$ *uniformly and* $\eta_i \xrightarrow{w} \eta \in C_+^*(T)$, *respectively. Assume also that* $\operatorname{supp}(\eta) \subset W_0 = \{t : f(t) = 0\}$. *Consider the sequence of functions* $\{x_i\}$ *defined by*

$$x_i(t) = \int_{[t_0,t]} f_i(s)d\eta_i, \quad t > t_0, \qquad x_i(t_0) = 0.$$

If the function f *is Lipschitz continuous, then* $\operatorname{Var}|_{\tau}[x_i] \to 0$.

Reduction R_1. In parallel with initial problem (1.1)–(1.4), denoted (P), we shall consider problem (P_1):

$$(P_1): \quad \begin{cases} e_0(p) \to \min, \\ dx = f(x,u,t)dt + g(x,t)d\mu, \\ d\alpha_k = g(\alpha_k, \theta_k)d\upsilon_k, \quad \alpha_k(t_{1-k}) = x_k, \\ d\theta_k = 0, \quad \theta_k = t_k,\ k = 0,1,\ t \in [t_0, t_1], \\ e_1(p) \leq 0, \quad e_2(p) = 0, \\ \varphi(x,t) \leq 0, \quad \varphi(\alpha_k, \theta_k) \leq 0, \quad k = 0,1, \\ u(t) \in U(t) \text{ a.e.}, \quad \mu, \upsilon_k \geq 0,\ k = 0,1, \\ p = (\xi_0, \xi_1, t_0, t_1), \quad x_k = x(t_k), \quad \xi_k = \alpha_k(t_k), \quad k = 0,1. \end{cases}$$

PROPOSITION 5.6. *Problems* (P) *and* (P_1) *are equivalent. This means that for every admissible process* (p, u, μ) *of* (P), *there exists an admissible process* $(\tilde{p}, \tilde{u}, \tilde{\mu}, \upsilon_0, \upsilon_1)$ *of* (P_1) *such that* $e_0(p) = e_0(\tilde{p})$, *and, conversely, for each admissible process of* (P_1), *there exists one of* (P) *yielding the same cost.*

Reduction R_1 permits us to focus on simpler problems in which the control measure has no atoms at the initial and the final points of the time interval.

Remark 4. All the results obtained in this paper for problem (P) can easily be written to problem (P_1) because the systems corresponding to measures μ, υ_0, and υ_1 are dynamically independent.

Reduction R_2. This is a reduction to the so-called v-problem [1]. Let Assumption S be in force. Consider problem (P_2) (v-problem).

$$(P_2): \quad \begin{cases} e_0(p) \to \min, \\ dx = (v+1)f(x,u,\chi)dt + g(x,\chi)d\mu, \quad t \in [t_0,t_1], \\ d\chi = (v+1)dt, \\ e_1(p) \le 0, \quad e_2(p) = 0, \quad p = (x_0,x_1,\chi_0,\chi_1), \quad \varphi(x,\chi) \le 0, \\ u(t) \in U, \quad |v(t)| \le 1/2 \text{ a.e.}, \quad \mu \ge 0. \end{cases}$$

PROPOSITION 5.7. *Problems* (P) *and* (P_2) *are equivalent.*

6. Primary problems. We start by considering the optimal control problem without state and endpoint constraints, i.e., (1.1)–(1.2) and (1.5), for which a maximum principle is given in Theorem 6.1. Then, the complexity of the problem is increased by adding endpoint constraints (1.3), and the necessary conditions of optimality presented in Theorem 6.2 are derived. State constraints will not be considered in this section.

THEOREM 6.1. *Let* (p^*,u^*,μ^*) *be an optimal process for the problem specified by* (1.1), (1.2) *and* (1.5). *Then, there exists a function* $\psi \in V^n(T^*)$ *such that*

$$\begin{cases} dx^* = H_\psi(t)dt + Q_\psi(t)d\mu^*, \\ d\psi = -H_x(t)dt - Q_x(t)d\mu^*, \end{cases} \qquad t \in T^*, \tag{6.1}$$

$$\psi_0 = \frac{\partial e_0}{\partial x_0}(p^*), \qquad \psi_1 = -\frac{\partial e_0}{\partial x_1}(p^*), \tag{6.2}$$

$$\max_{u\in U(t)} H(u,t) = H(t) \ \textit{a.e.}, \tag{6.3}$$

$$Q(t) \le 0 \quad \forall t, \qquad Q(t) = 0 \quad \forall t \in \operatorname{supp}(\mu^*), \tag{6.4}$$

$$\begin{cases} \operatorname{ess\,lim\,sup}\limits_{t\to t_k^*} \max\limits_{u\in U(t)} H(\beta_k^*,u,\gamma_k,t) \\ \quad +(-1)^{k+1}\left[\Delta_k^* \displaystyle\int_0^1 \langle g_t(\alpha_k^*,t_k^*),\sigma_k\rangle ds - \frac{\partial e_0}{\partial t_k}(p^*)\right] \ge 0, \\ \operatorname{ess\,lim\,inf}\limits_{t\to t_k^*} \max\limits_{u\in U(t)} H(\beta_k^*,u,\gamma_k,t) \\ \quad +(-1)^{k+1}\left[\Delta_k^* \displaystyle\int_0^1 \langle g_t(\alpha_k^*,t_k^*),\sigma_k\rangle ds - \frac{\partial e_0}{\partial t_k}(p^*)\right] \le 0, \end{cases} \qquad k = 0,1, \tag{6.5}$$

$$\operatorname*{ess\,lim\,inf}_{t\to t_k^*} \max_{u\in U(t)} H(x_k^*,u,\psi_k,t) + (-1)^k \frac{\partial e_0}{\partial t_k}(p^*) \le 0. \tag{6.6}$$

Here, $\beta_0^* = x^*(t_0^{*+})$, $\beta_1^* = x^*(t_1^{*-})$, $\gamma_0 = \psi(t_0^{*+})$, $\gamma_1 = \psi(t_1^{*-})$, $\Delta_k^* = \mu^*(\{t_k^*\})$, $k = 0,1$, *are the control atoms at the time endpoints (see last remark in section* 4*), and the pair* (α_k^*,σ_k), $k = 0,1$, *in* (6.5) *corresponds to the solution* (6.1) *in atom* t_k^*. *According to our solution concept, this means that the pair* (α_k^*,σ_k) *satisfies*

$$\begin{cases} \dot\alpha_k^* = g(\alpha_k^*,t_k^*)\Delta_k^*, \quad \alpha_k^*(0) = x^*(t_k^{*-}), \\ \dot\sigma_k = -g_x^\top(\alpha_k^*,t_k^*)\sigma_k\Delta_k^*, \quad \sigma_k(0) = \psi(t_k^{*-}), \end{cases} \qquad s \in [0,1], \quad k = 0,1. \tag{6.7}$$

Proof. The proof of (6.1)–(6.4) is given in [16, 20]. So we only have to show that the time-transversality conditions (6.5) hold. We shall do it using the scheme from [1].

We start by assuming that measure μ^* has no atoms at points t_k^*, i.e., $\Delta_k^* = 0$, $k = 0, 1$. Let us prove (6.5) when $k = 1$.

Fix $\varepsilon > 0$ such that $t_1^* - \varepsilon \in \mathrm{Cont}(\mu^*)$ and let $T_\varepsilon = [t_1^* - \varepsilon, t_1^*]$. Define the measure $\mu_\varepsilon := \mu^* + \mu^*(T_\varepsilon)\delta(t_1^* - \varepsilon)$ on $[t_0^*, t_1^* - \varepsilon]$. Let $p_\varepsilon = (x_0^*, x_\varepsilon(t_1^* - \varepsilon), t_0^*, t_1^* - \varepsilon)$, where x_ε is the trajectory associated to $(x_0^*, u^*, \mu_\varepsilon)$. For ε sufficiently small and from the optimality of the process (p^*, u^*, μ^*), i.e.,

$$e_0(p_\varepsilon) - e_0(p^*) \geq 0 \quad \forall \varepsilon > 0,$$

we have that

$$\left\langle \frac{\partial e_0}{\partial x_1}(p^*), \Delta x_\varepsilon \right\rangle - \varepsilon \frac{\partial e_0}{\partial t_1}(p^*) + o(|\Delta x_\varepsilon|) + o(\varepsilon) \geq 0, \tag{6.8}$$

where $\Delta x_\varepsilon = x_\varepsilon(t_1^* - \varepsilon) - x_1^*$.

Let $\xi_\varepsilon^* = \xi(x^*(t_1^* - \varepsilon), t_1^* - \varepsilon, \mu^*(T_\varepsilon))$. Then, by definition $\Delta x_\varepsilon = \Delta x_\varepsilon^1 + \Delta x_\varepsilon^2$, where

$$\Delta x_\varepsilon^1 = -\int_{T_\varepsilon} f(x^*, u^*, t)dt, \qquad \Delta x_\varepsilon^2 = \xi_\varepsilon^* - \left(x^*(t_1^* - \varepsilon) + \int_{T_\varepsilon} g(x^*, t)d\mu^* \right).$$

The estimate $|\Delta x_\varepsilon^1| \leq \mathrm{const}\, \varepsilon$ is obvious. Let us show that $|\Delta x_\varepsilon^2| \leq \mathrm{const}\, \varepsilon \mu^*(T_\varepsilon)$. Since $|\Delta x_\varepsilon^2| = 0$ whenever $\mu^*(T_\varepsilon) = 0$, we assume that $\mu^*(T_\varepsilon) > 0$ $\forall \varepsilon > 0$.

Let us construct a sequence of absolutely continuous measures $\{\mu_i\}$ having density $\dot{\mu}_i = m_i > 0$ a.e., defined on the segment T_ε such that $\mu_i \xrightarrow{w} \mu^*$ weakly* on T_ε, $\mu_i(T_\varepsilon) = \mu^*(T_\varepsilon)$. Let x_i be a continuation of the solution x^* on segment T_ε, corresponding to the measure μ_i (such a continuation exists when i is sufficient large). By Lemma 5.4, $x_i(t) \to x^*(t)$ $\forall t \in \mathrm{Cont}(\mu^*)$. Then, $\Delta x_{\varepsilon,i} = x_\varepsilon(t_1^* - \varepsilon) - x_i(t_1^*) \to \Delta x_\varepsilon$ as $i \to \infty$. Furthermore, $\Delta x_{\varepsilon,i}^1 \to \Delta x_\varepsilon^1$ and $\Delta x_{\varepsilon,i}^2 \to \Delta x_\varepsilon^2$, being

$$\Delta x_{\varepsilon,i}^1 = -\int_{T_\varepsilon} f(x_i, u^*, t)dt, \qquad \Delta x_{\varepsilon,i}^2 = \xi_\varepsilon^* - \left(x^*(t_1^* - \varepsilon) + \int_{T_\varepsilon} g(x_i, t)m_i dt \right).$$

To estimate $|\Delta x_{\varepsilon,i}^2|$, let us consider the equation

$$\dot{x}_{2,i} = g(x_{1,i} + x_{2,i}, t)m_i, \quad t \in T_\varepsilon, \qquad x_{2,i}(t_1^* - \varepsilon) = x^*(t_1^* - \varepsilon). \tag{6.9}$$

Here, $x_{1,i}(t) = \int_{t_1^* - \varepsilon}^t f(x_i, u^*, \tau)d\tau$ and $x_i = x_{1,i} + x_{2,i}$.

Define the functions $\pi_i : T_\varepsilon \to [0, 1]$ as follows:

$$\pi_i(t) = \frac{F(t; \mu_i) - F(t_1^* - \varepsilon; \mu_i)}{\mu^*(T_\varepsilon)}, \quad t \in T_\varepsilon.$$

Obviously, π_i is absolutely continuous and $\frac{d\pi_i}{dt} > 0$ a.e. Therefore, there exists the inverse function $\pi_i^{-1} : [0, 1] \to T_\varepsilon$ which is also strictly monotone and absolutely continuous. By the change of variable $s = \pi_i(t)$ in (6.9), and by putting $\alpha_i(s) = x_{2,i}(\pi_i^{-1}(s))$, we arrive at the equation

$$\dot{\alpha}_i = g(x_{1,i}(\pi_i^{-1}(s)) + \alpha_i, \pi_i^{-1}(s))\mu^*(T_\varepsilon), \quad s \in [0, 1], \qquad \alpha_i(0) = x^*(t_1^* - \varepsilon).$$

By definition $\xi_\varepsilon^* = \alpha(1)$, where $\alpha(s)$ satisfies

$$\dot\alpha = g(\alpha, t_1^* - \varepsilon)\mu^*(T_\varepsilon), \quad s \in [0,1], \qquad \alpha(0) = x^*(t_1^* - \varepsilon).$$

From here, $|\alpha_i(s) - \alpha(s)| \le \mu^*(T_\varepsilon) \int_0^s c(|\alpha_i(\tau) - \alpha(\tau)| + \int_{T_\varepsilon} |f(x_i, u^*, \theta)| d\theta + \varepsilon) d\tau \ \forall s$.

From Gronwall's inequality, $|\Delta x^2_{\varepsilon,i}| = |\alpha_i(1) - \alpha(1)| \le \text{const}\, \varepsilon \mu^*(T_\varepsilon)$ and, therefore,

$$|\Delta x^2_\varepsilon| \le \text{const}\, \varepsilon \mu^*(T_\varepsilon) = o(\varepsilon).$$

The last equality holds from the fact that $\mu^*(T_\varepsilon) \to 0$ as $\varepsilon \to 0$.

In view of the obtained estimates and the already proved transversality conditions (6.2), inequality (6.8) becomes

$$\left\langle \psi_1, \int_{T_\varepsilon} f(x^*, u^*, t) dt \right\rangle - \varepsilon \frac{\partial e_0}{\partial t_1}(p^*) + o(\varepsilon) \ge 0.$$

From here, we may write

$$\int_{t_1^* - \varepsilon}^{t_1^*} \max_{u \in U(t)} H(u, \psi_1, t) dt - \varepsilon \frac{\partial e_0}{\partial t_1}(p^*) + o(\varepsilon) \ge 0,$$

and, by dividing the last inequality by $\varepsilon > 0$ and taking the limit as ε goes to zero, we obtain

$$\operatorname{ess\,lim\,sup}_{t \to t_1^*} \max_{u \in U(t)} H(x_1^*, u, \psi_1, t) - \frac{\partial e_0}{\partial t_1}(p^*) \ge 0.$$

This corresponds to the first inequality in (6.5).

Now, we consider the other inequality for the final endpoint. Let us put $\mu^*([t_1^*, +\infty)) = 0$. The control function u^* is continued to the right beyond t_1^* as follows:

$$u^*(t) \in \operatorname*{Arg\,max}_{u \in U(t)} H(x_1^*, u, \psi_1, t), \quad t > t_1^*.$$

The trajectory x^* is continued as a solution to (1.2), corresponding to the constructed u^*, on the segment $[t_1^*, t_1^* + \delta]$ for a sufficient small $\delta > 0$. For $\varepsilon \in (0, \delta]$, let $p_\varepsilon = (x_0^*, x^*(t_1^* + \varepsilon), t_0^*, t_1^* + \varepsilon)$. By using the same arguments as above, we arrive at the second inequality in (6.5) for $k = 1$. Similar arguments allow us to show these inequalities for $k = 0$.

Now, let us show that conditions (6.5) hold when $\Delta_k^* > 0$, $k = 0, 1$. For this purpose let us consider the problem (P_1) as a reduction (via R_1) of (P).[1]

If the process (p^*, u^*, μ^*) is a solution to (P), then the process $(\beta_0^*, \beta_1^*, t_0^*, t_1^*, u^*, \Delta_0^*, \Delta_1^*, \tilde\mu)$ is a solution to (P_1). Here, $\tilde\mu = \mu^* - \sum_{k=0}^1 \Delta_k^* \delta_{t_k^*}(t)$, and β_0^*, and β_1^* are as defined above. The measure $\tilde\mu$ has no atoms in t_k^*, $k = 0, 1$, and the maximum principle for this problem has just been proved.

[1]Here, (P) and (P_1) are considered as in section 5 but without constraints. To be more precise, if (P) is a problem without constraints, then (P_1) is defined as follows (equivalent definition):

$$(P_1): \begin{cases} e_0(p) \to \min, \\ dx = f(x,u,t)dt + g(x,t)d\mu, \ d\Delta_k = 0, \ \Delta_k \ge 0, \ k = 0,1, \\ p = (\xi_0, \xi_1, t_0, t_1), \ \xi_k = \xi(x_k, t_k, (-1)^{k+1}\Delta_k). \end{cases}$$

By applying this maximum principle to (P_1) and decoding its conditions in terms of the data of problem (1.1)–(1.2), we conclude the existence of a function $\psi^R \in V^n(T^*)$ such that

$$
\begin{aligned}
&d\psi^R = -H_x(\tilde{x},\psi^R,t)dt - Q_x(\tilde{x},\psi^R,t)d\tilde{\mu}, \quad t \in T^*,\\
&\begin{cases} \psi_0^R = \xi_x^\top(\beta_0^*,t_0^*,-\Delta_0^*)\dfrac{\partial e_0}{\partial x_0}(p^*),\\ \psi_1^R = -\xi_x^\top(\beta_1^*,t_1^*,\Delta_1^*)\dfrac{\partial e_0}{\partial x_1}(p^*), \end{cases} \qquad (6.10)\\
&\left\langle \xi_\Delta(\beta_k^*,t_k^*,(-1)^{k+1}\Delta_k^*), \frac{\partial e_0}{\partial x_k}(p^*)\right\rangle = 0, \quad k=0,1, \qquad (6.11)\\
&\max_{u\in U(t)} H(\tilde{x}(t),u,\psi^R(t),t) = H(\tilde{x}(t),u^*(t),\psi^R(t),t), \quad \text{a.e.},\\
&Q(\tilde{x}(t),\psi^R(t),t) \le 0, \quad t \in T^*,\\
&Q(\tilde{x}(t),\psi^R(t),t) = 0 \quad \forall t \in \operatorname{supp}(\tilde{\mu}),\\
&\begin{cases} \operatorname{ess\,lim\,sup}\limits_{t\to t_k^*} \max\limits_{u\in U(t)} H(\beta_k^*,u,\psi_k^R,t)\\ \quad +(-1)^k\left[\dfrac{\partial e_0}{\partial t_k}(p^*) + \left\langle \xi_t(\beta_k^*,t_k^*,(-1)^{k+1}\Delta_k^*), \dfrac{\partial e_0}{\partial x_k}(p^*)\right\rangle\right] \ge 0,\\ \operatorname{ess\,lim\,inf}\limits_{t\to t_k^*} \max\limits_{u\in U(t)} H(\beta_k^*,u,\psi_k^R,t)\\ \quad +(-1)^k\left[\dfrac{\partial e_0}{\partial t_k}(p^*) + \left\langle \xi_t(\beta_k^*,t_k^*,(-1)^{k+1}\Delta_k^*), \dfrac{\partial e_0}{\partial x_k}(p^*)\right\rangle\right] \le 0, \end{cases} \qquad k=0,1. \end{aligned}
$$

Here, $\tilde{x}$ is the optimal trajectory for problem (P_1), and the matrix function ξ_x and the vector functions ξ_Δ, ξ_t are the derivatives of ξ with respect to, respectively, x, Δ, and t.

Clearly, $\tilde{x}(t) = x^*(t)$ $\forall t \in (t_0^*, t_1^*)$, and $\tilde{x}(t_k^*) = \beta_k^*$, $k = 0, 1$. Let us take $\psi(t) = \psi^R(t)$ $\forall t \in (t_0^*, t_1^*)$, $\psi(t_k^*) = (-1)^k \frac{\partial e_0}{\partial x_k}(p^*)$, $k = 0, 1$, and show that ψ satisfies (6.1)–(6.5). Clearly, we need to verify them only at the two endpoints t_k^*, $k = 0, 1$.

Consider the case $k = 1$ (the case $k = 0$ is similar). Let (α_1^*, σ_1) be a solution to (6.7). It is straightforward to verify that function $(\xi_x, \xi_\Delta, \xi_t)$ is given by

$$(\xi_x, \xi_\Delta, \xi_t)(\beta_1^*, t_1^*, \Delta_1^*) = (\bar{\xi}_x, \bar{\xi}_\Delta, \bar{\xi}_t)(1),$$

where $(\bar{\xi}_x, \bar{\xi}_\Delta, \bar{\xi}_t)(0) = (E, 0, 0)$ (E is the n-dimensional identity matrix) and, for $s \in [0, 1]$,

$$
\begin{cases} \dot{\bar{\xi}}_x(s) = g_x(\alpha_1^*(s), t_1^*)\bar{\xi}_x(s)\Delta_1^*,\\ \dot{\bar{\xi}}_\Delta(s) = g_x(\alpha_1^*(s), t_1^*)\bar{\xi}_\Delta(s)\Delta_1^* + g(\alpha_1^*(s), t_1^*),\\ \dot{\bar{\xi}}_t(s) = g_x(\alpha_1^*(s), t_1^*)\bar{\xi}_t(s)\Delta_1^* + g_t(\alpha_1^*(s), t_1^*)\Delta_1^*. \end{cases}
$$

From (6.10) we get $\sigma_1(1) = \psi_1$. Thus (6.1)–(6.3) are proved.

To prove (6.4), we will show that $\langle \sigma_1, g(\alpha_1^*, t_1^*)\rangle \equiv 0$. Indeed, by direct computation and by using the fact that $\frac{d}{ds}\langle \sigma_1, g(\alpha_1, t_1^*)\rangle = 0$ $\forall s \in [0, 1]$, we have that $\frac{d}{ds}\langle \sigma_1, \bar{\xi}_\Delta\rangle = \langle \sigma_1, g(\alpha_1^*, t_1^*)\rangle = \text{const}$ $\forall s \in [0, 1]$. On the other hand, from (6.11) and the definition of ξ_Δ, we conclude that $\langle \sigma_1(0), \bar{\xi}_\Delta(0)\rangle = \langle \sigma_1(1), \bar{\xi}_\Delta(1)\rangle = 0$.

Hence, $\langle \sigma_1, \bar{\xi}_\Delta\rangle \equiv 0$ and, as consequence, $\langle \sigma_1, g(\alpha_1^*, t_1^*)\rangle \equiv 0$ on $[0, 1]$, i.e., condition (6.4) is proved.

To obtain inequalities in (6.5), note that these follow from the fact that

$$\frac{d}{ds}\langle \sigma_1, \bar{\xi}_t \rangle = \langle \sigma_1, g_t(\alpha_1^*, t_1^*)\rangle \Delta_1^*.$$

The condition (6.6) can be easily deduced as its analogue in (6.5).

The proof is complete. □

THEOREM 6.2. *Let (p^*, u^*, μ^*) be an optimal process for the problem specified by (1.1)–(1.3) and (1.5). Then, there exist a number $\lambda_0 \geq 0$, vectors $\lambda_j \in \Re^{k_j}$, $j = 1, 2$, with $|\lambda| = 1$, and a function $\psi \in V^n(T^*)$ satisfying*

$$\begin{aligned}
&d\psi = -H_x(t)dt - Q_x(t)d\mu^*, \quad t \in T^*,\\
&\psi_0 = \frac{\partial l}{\partial x_0}(p^*, \lambda), \quad \psi_1 = -\frac{\partial l}{\partial x_1}(p^*, \lambda),\\
&\lambda_1 \geq 0, \quad \langle \lambda_1, e_1(p^*)\rangle = 0;\\
&\max_{u \in U(t)} H(u, t) = H(t) \ \textit{a.e.},\\
&Q(t) \leq 0 \quad \forall t, \qquad Q(t) = 0 \quad \forall t \in \operatorname{supp}(\mu^*),
\end{aligned}$$

$$\begin{cases}
\operatorname{ess\,lim\,sup}\limits_{t \to t_k^*} \max\limits_{u \in U(t)} H(\beta_k^*, u, \gamma_k, t)\\
\quad +(-1)^{k+1}\left[\Delta_k^* \int_0^1 \langle g_t(\alpha_k^*, t_k^*), \sigma_k\rangle ds - \frac{\partial l}{\partial t_k}(p^*, \lambda)\right] \geq 0,\\
\operatorname{ess\,lim\,inf}\limits_{t \to t_k^*} \max\limits_{u \in U(t)} H(\beta_k^*, u, \gamma_k, t)\\
\quad +(-1)^{k+1}\left[\Delta_k^* \int_0^1 \langle g_t(\alpha_k^*, t_k^*), \sigma_k\rangle ds - \frac{\partial l}{\partial t_k}(p^*, \lambda)\right] \leq 0,
\end{cases} \qquad k = 0, 1. \tag{6.12}$$

Furthermore, inequalities (6.12) *may be replaced by*

$$\operatorname{ess\,lim\,inf}_{t \to t_k^*} \max_{u \in U(t)} H(x_k^*, u, \psi_k, t) + (-1)^k \frac{\partial l}{\partial t_k}(p^*) \leq 0, \ \ k = 0, 1. \tag{6.13}$$

However, we cannot guarantee that both conditions (6.12) *and* (6.13) *hold at the same time, i.e., we have two different versions of the maximum principle.*

Proof. The proof is based on the penalty function method [1]. Assume, first, that $\Delta_k^* = 0$, $k = 0, 1$. Take any natural i and define the function $e_{0,i}$ as follows:

$$e_{0,i}(p) = e_0(p) + |p - p^*|^2 + i(|e_1^+(p)|^2 + |e_2(p)|^2).$$

Here, for a given vector $a = (a^1, \ldots, a^k)$, a^+ stands for the vector with components $(a^j)^+ = \max\{0, a^j\}$, $j = 1, \ldots, k$.

Let $F(t) = F(t; \mu^*)$ be the distribution function of the optimal measure. Take $c = \operatorname{gc\,sup}_{t \in T^*} |x^*(t)| + |\mu^*| + |p^*| + 1$ and a sufficiently small $\varepsilon > 0$. Consider, for each $i \in \mathbf{N}$, the following penalty problem:

$$\begin{cases}
J_i(p, u, \mu) = e_{0,i}(p) + |y_1 - F(t_1^*)|^2\\
\qquad\qquad +\varepsilon \int_{t_0}^{t_1} [|u - u^*(t)|^2 + |y - F(t)|^2]dt \to \min,\\
dx = f(x, u, t)dt + g(x, t)d\mu, \quad t \in [t_0, t_1],\\
dy = d\mu, \qquad y_0 = 0, \qquad t \in [t_0, t_1],\\
\max\{|p|, \operatorname{gc\,sup}|x|, y_1\} \leq c,\\
u(t) \in U(t) \text{ a.e.}, \quad \mu \geq 0.
\end{cases} \tag{6.14}$$

This problem is called the *i-problem.*

For every i-problem, there exists a solution (p_i, u_i, μ_i).[2] By using compactness and by extracting a subsequence, we obtain a process (p, u, μ) such that $p_i \to p$, $u_i \xrightarrow{w} u$ weakly in $L_2^m(T_c)$, $\mu_i \xrightarrow{w} \mu$ weakly* in $C^*(T_c)$.[3] It is a straightforward task to establish that $p = p^*$, $u = u^*$, and $\mu = \mu^*$. Moreover, from the penalty method we conclude that indeed $u_i \to u^*$ in $L_2^m(T_c)$. Then, by extracting a subsequence if necessary, we conclude that $u_i(t) \to u^*(t)$ a.e.

All inequality constraints of problem (6.14) become strict for i sufficiently large. For this reason we can apply Theorem 6.1 to the i-problem. By writing down the necessary conditions of Theorem 6.1 and bearing in mind that $i \to \infty$, we conclude Theorem 6.2 when $\Delta_k^* = 0$, $k = 0, 1$. If $\Delta_k^* > 0$, then, by using reduction R_1 (the same way as in Theorem 6.1), we obtain the time-transversality conditions in the conventional case. To prove (6.13) we do not need to use reduction R_1.

The proof is complete. □

7. The problem with state constraints. In this section, we prove a *weakened* maximum principle for problems with state and endpoint constraints.

THEOREM 7.1. *Consider the optimal control problem* (1.1)–(1.5) *with compatible state and endpoint constraints and let triple* (p^*, u^*, μ^*) *be its solution.*

Then, there exist a number $\lambda_0 \geq 0$, *vectors* $\lambda_1 \in \Re^{k_1}$, $\lambda_1 \geq 0$, $\lambda_2 \in \Re^{k_2}$, *a vector function* $\psi \in V^n(T^*)$, *a vector measure* $\eta = (\eta^1, \dots, \eta^{k_3})$, $\eta^j \in C_+^*(T^*)$ *such that* $\mathrm{Ds}(\mu^*) \cap \mathrm{Ds}(\eta^j) = \emptyset$ $\forall j$, *and, for every atom* $r \in \mathrm{Ds}(\mu^*)$, *there exist its own vector function* $\sigma_r \in V^n([0,1])$ *and its own vector measure* $\eta_r = (\eta_r^1, \dots, \eta_r^{k_3})$, $\eta_r^j \in C_+^*([0,1])$, $j = 1, \dots, k_3$, *such that*

$$\psi(t) = \psi_0 - \int_{t_0^*}^{t} H_x(s)ds - \int_{[t_0^*,t]} Q_x(s)d\mu_c^* + \int_{[t_0^*,t]} \varphi_x^\top(x^*, s)d\eta + \Sigma(\psi, t), \quad t \in (t_0^*, t_1^*], \tag{7.1}$$

$$\Sigma(\psi, t) = \sum_{r \in \mathrm{Ds}(\mu^*),\ r \leq t} [\sigma_r(1) - \psi(r^-)],$$

$$\begin{cases} \dot{\alpha}_r^*(s) &= g(\alpha_r^*(s), r)\Delta_r^*, \quad s \in [0,1], \\ d\sigma_r(s) &= -g_x^\top(\alpha_r^*(s), r)\sigma_r(s)\Delta_r^* ds + \varphi_x^\top(\alpha_r^*(s), r)d\eta_r, \quad s \in [0,1], \\ \alpha_r^*(0) &= x^*(r^-), \\ \sigma_r(0) &= \psi(r^-), \\ \Delta_r^* &= \mu^*(\{r\}), \end{cases}$$

$$\psi_0 = \frac{\partial l}{\partial x_0}(p^*, \lambda), \qquad \psi_1 = -\frac{\partial l}{\partial x_1}(p^*, \lambda), \tag{7.2}$$

$$\langle g(\alpha_r^*(s), r), \sigma_r(s)\rangle = 0 \quad \forall s \in [0,1],\ \forall r \in \mathrm{Ds}(\mu^*) \tag{7.3}$$

$$\mathrm{supp}(\eta_r^j) \subseteq \{s \in [0,1] : \varphi^j(\alpha_r^*(s), r) = W^j(\alpha_r^*(s), r) = 0\} \quad \forall j, \tag{7.4}$$

$$\langle \lambda_1, e_1(p^*)\rangle = 0,$$

$$\varphi^j(x^*(t), t) = 0 \ \ \eta^j\text{-a.e.} \quad \forall j \tag{7.5}$$

$$\max_{u \in U(t)} H(u, t) = H(t) \quad \text{a.e.}, \tag{7.6}$$

[2] This is due to the compactness and Lemma 5.4. In fact, since $x(t)$ are bounded uniformly (due to $\mathrm{gc}\sup|x| \leq$ const and by using $|\mu| \leq$ const) and $\mathrm{Var}\,|_a^b[x_i(t)] \leq \sup_{s\in[a,b]}|f_i(s)| \times |\mu_i|([a,b])$ $\forall a, b \in T$ (from the uniform bound of (u_i, μ_i)), we can use Helly's theorem and extract appropriate subsequences.

[3] Here, $T_c = [-c, c]$.

$$Q(t) \le 0 \quad \forall t, \qquad Q(t) = 0 \quad \mu^*\text{-a.e.}, \tag{7.7}$$

$$\operatorname{ess}\liminf_{t\to t_k^*} \max_{u\in U(t)} H(x_k^*, u, \psi_k, t) + (-1)^k \frac{\partial l}{\partial t_k}(p^*, \lambda) \le 0, \quad k = 0, 1, \tag{7.8}$$

$$|\lambda| + |\eta| + \sum_{r\in \mathrm{Ds}(\mu^*)} |\eta_r| = 1. \tag{7.9}$$

Furthermore, when Assumption C *is in force, then in place of* (7.8) *we can obtain* time transversality conditions, *which are written as follows for the case where* $\Delta_k^* = 0$:[4]

$$\begin{cases} \operatorname{ess}\limsup_{t\to t_k^*} \max_{u\in U(t)} H(x_k^*, u, \psi_k, t) + (-1)^k \dfrac{\partial l}{\partial t_k}(p^*, \lambda) \ge 0, \\ \operatorname{ess}\liminf_{t\to t_k^*} \max_{u\in U(t)} H(x_k^*, u, \psi_k, t) + (-1)^k \dfrac{\partial l}{\partial t_k}(p^*, \lambda) \le 0, \end{cases} \quad k = 0, 1. \tag{7.10}$$

Besides, as in Theorem 6.2, *we cannot guarantee that* (7.8) *and* (7.10) *are satisfied at the same time, so again we have two readings of the theorem.*

Proof. The proof is based on a penalty function method [1].

Let us start by proving this result for the most difficult case, i.e., to consider the second variant of the theorem with time transversality conditions (7.10).

Bearing in mind Assumption C and Definition 2.3, let us take a sufficiently small $\delta > 0$. Since μ^* is continuous at t_k^*, $k = 0, 1$, there are points $t_{0,\delta} > t_0^*$, $t_{1,\delta} < t_1^*$, and a small number $\varepsilon > 0$ such that $(\mu^* + \mathcal{L})([t_0^*, t_{0,\delta} + \varepsilon] \cup [t_{1,\delta} - \varepsilon, t_1^*]) < \delta$. Define the function φ_δ^+ as follows:

$$\varphi_\delta^+(x,t) = \begin{cases} \varphi^+(x,t), & t_{0,\delta} < t < t_{1,\delta}, \\ 0 \quad \text{otherwise.} \end{cases}$$

The components of φ_δ^+ are nonnegative and lower semicontinuous.

Let $F(t) = F(t; \mu^*)$ and $c = \operatorname{gc\,sup} |x^*| + |p^*| + |\mu^*| + 1$. Take any $i, A \in \mathbf{N}$ and construct the following auxiliary penalty problem:

$$\begin{cases} J_{i,A}(p,u,\mu) = e_0(p) + |p - p^*|^2 + |y_1 - F(t_1^*)|^2 \\ \qquad\qquad + A\displaystyle\int_{t_0}^{t_1} |\varphi^+(x,t)|^2 dt + A \int_{[t_0,t_1]} |\varphi_\delta^+(x,t)|^2 d\mu \\ \qquad\qquad + i^{-1} \displaystyle\int_{t_0}^{t_1} [|u - u^*(t)|^2 + |y - F(t)|^2] dt \to \min, \\ \qquad\qquad dx = f(x,u,t)dt + g(x,t)d\mu, \quad dy = d\mu, \ t \in [t_0, t_1], \\ \qquad\qquad y_0 = 0, \qquad e_1(p) \le 0, \qquad e_2(p) = 0, \\ \qquad\qquad |p - p^*| \le \delta, \quad \mu([t_0, t_{0,\delta} + \varepsilon) \cup (t_{1,\delta} - \varepsilon, t_1]) \le \delta, \\ \qquad\qquad \max\{\operatorname{gc\,sup} |x|, y_1\} \le c, \qquad u(t) \in U(t) \text{ a.e.}, \quad \mu \ge 0. \end{cases} \tag{7.11}$$

The functional $J_{i,A}$ is weakly lower semicontinuous in (u, μ) and, thus, for every $i, A \in \mathbf{N}$, the problem (7.11) has a solution. Denote it by $(p_{i,A}, u_{i,A}, \mu_{i,A})$ and let $(x_{i,A}, y_{i,A})$ be the corresponding optimal trajectory. By using compactness, extract a subsequence as i is fixed and obtain a process (p, u, μ) satisfying $p_{i,A} \to p$, $u_{i,A} \xrightarrow{w} u$ weakly in $L_2^m(T_c)$, $\mu_{i,A} \xrightarrow{w} \mu$ weakly* in $C^*(T_c)$[5] as $A \to \infty$. Furthermore, $x_{i,A}(t) \to x(t)$ $\forall t \in \mathrm{Cont}(\mu)$. We proceed by showing that $p = p^*$, $u = u^*$, $\mu = \mu^*$ (for every fixed i).

[4] Case $\Delta_k^* > 0$ may be researched with the help of reduction R_1.

[5] Again, $T_c = [-c, c]$.

Let us first show that (p, u, μ) is an admissible process. In fact, from (7.11), we have

$$\int_{T_{i,A}} |\varphi^+(x_{i,A}, t)|^2 dt + \int_{T_{i,A}} |\varphi_\delta^+(x_{i,A}, t)|^2 d\mu_{i,A} \le \frac{\text{const}}{A},$$

and, by passing to the limit (with the help of Lemma 5.4) as $A \to \infty$, we obtain

$$\int_{t_0}^{t_1} |\varphi^+(x, t)|^2 dt + \int_{[t_0,t_1]} |\varphi_\delta^+(x, t)|^2 d\mu = 0.$$

From here, and by using Assumption C and the compatibility condition, we deduce that $\operatorname{gc}\sup \varphi^j(x) \le 0$, $j = 1, \dots, k_3$. Thus, (p, u, μ) is an admissible process. Hence, $e_0(p) \ge e_0(p^*)$. Furthermore, since

$$J_{i,A}(p_{i,A}, u_{i,A}, \mu_{i,A}) \le J_{i,A}(p^*, u^*, \mu^*) = e_0(p^*),$$

we have that

$$e_0(p_{i,A}) + |p_{i,A} - p^*|^2 + |y_{1,i,A} - F(t_1^*)|^2 + A \int_{T_{i,A}} |\varphi^+(x_{i,A}, t)|^2 dt$$
$$+ A \int_{T_{i,A}} |\varphi_\delta^+(x_{i,A}, t)|^2 d\mu_{i,A} + \frac{1}{i} \int_{T_{i,A}} (|u_{i,A} - u^*(t)|^2 + |y_{i,A} - F(t)|^2) dt \le e_0(p^*).$$

By passing to the limit in this inequality as $A \to \infty$, and by using the weak lower semicontinuity of the left part in u, we obtain

$$e_0(p) + |p - p^*|^2 + |y_1 - F(t_1^*)|^2 + \frac{1}{i} \int_{t_0}^{t_1} (|u - u^*(t)|^2 + |y - F(t)|^2) dt$$
$$+ \limsup_{A \to \infty} \left[A \int_{T_{i,A}} |\varphi^+(x_{i,A}, t)|^2 dt + A \int_{T_{i,A}} |\varphi_\delta^+(x_{i,A}, t)|^2 d\mu_{i,A} \right] \le e_0(p^*).$$

Since $e_0(p) \ge e_0(p^*)$, we conclude immediately that $p = p^*$, $u = u^*$, $\mu = \mu^*$, and

$$\lim_{A \to \infty} A \int_{T_{i,A}} |\varphi^+(x_{i,A}, t)|^2 dt + A \int_{T_{i,A}} |\varphi_\delta^+(x_{i,A}, t)|^2 d\mu_{i,A} = 0 \quad \forall i.$$

So, for every i, we can take a number A_i satisfying

$$|p_{i,A_i} - p^*|^2 + \|u_{i,A_i} - u^*\|_{L_2}^2 + \|y_{i,A_i} - F(t)\|_{L_2}^2$$
$$+ A_i \int_{T_{i,A_i}} |\varphi^+(x_{i,A_i}, t)|^2 dt + A_i \int_{T_{i,A_i}} |\varphi_\delta^+(x_{i,A_i}, t)|^2 d\mu_{i,A_i} \le \frac{1}{i}$$

and define a new diagonal sequence denoted by (p_i, u_i, μ_i), i.e.,

$$p_i = p_{i,A_i}, \qquad u_i = u_{i,A_i}, \qquad \mu_i = \mu_{i,A_i}. \tag{7.12}$$

By extracting a subsequence, we obtain $p_i \to p^*$, $u_i \to u^*$ a.e., $\mu_i \xrightarrow{w} \mu^*$ as $i \to \infty$. For this sequence, we have $x_i(t) = x_{i,A_i}(t) \to x^*(t)$ $\forall t \in \operatorname{Cont}(\mu^*)$. By replacing A in problem (7.11) by this specially chosen A_i, a new problem, called *i-problem*, is obtained.

All the additional inequality constraints of i-problem become strict inequalities for i sufficiently large. Therefore, Theorem 6.2 can be applied to the i-problem. It states that for every such i, there are functions $\psi_i \in V^n(T_i)$, $\zeta_i \in C(T_i)$, a number $\lambda_{0,i} \geq 0$, and vectors $\lambda_{1,i}$ and $\lambda_{2,i}$ such that

$$d\psi_i = -H^i_x(t)dt - Q^i_x(t)d\mu_i + \varphi^\top_x(x_i,t)d\eta_i, \quad t \in T_i, \tag{7.13}$$

$$\psi_{0,i} = \frac{\partial l}{\partial x_0}(p_i,\lambda_i) + 2\lambda_{0,i}(x_{0,i} - x^*_0), \quad \psi_{1,i} = -\frac{\partial l}{\partial x_1}(p_i,\lambda_i) - 2\lambda_{0,i}(x_{1,i} - x^*_1), \tag{7.14}$$

$$d\zeta_i = \frac{2\lambda_{0,i}}{i}[y_i - F(t)]dt, \qquad \zeta_i(t_{1,i}) = -2\lambda_{0,i}[y_{1,i} - F(t^*_1)],$$

$$\lambda_{1,i} \geq 0, \quad \langle e_1(p_i), \lambda_{1,i}\rangle = 0,$$

$$\max_{u\in U(t)}[H_i(u,t) - \lambda_{0,i}i^{-1}|u - u^*(t)|^2] = H_i(t) - \lambda_{0,i}i^{-1}|u_i(t) - u^*(t)|^2 \text{ a.e.}, \tag{7.15}$$

$$Q_i(t) + \zeta_i(t) - \lambda_{0,i}A_i|\varphi^+_\delta(x_i(t),t)|^2 \leq 0 \quad \forall t, \tag{7.16}$$

$$Q_i(t) + \zeta_i(t) - \lambda_{0,i}A_i|\varphi^+_\delta(x_i(t),t)|^2 = 0 \quad \forall t \in \operatorname{supp}(\mu_i), \tag{7.17}$$

$$\left\{\begin{array}{l} \operatorname{ess\,lim\,sup}\limits_{t\to t_{k,i}} \max\limits_{u\in U(t)}[H(\beta_{k,i},u,\gamma_{k,i},t) - \lambda_{0,i}i^{-1}|u - u^*(t)|^2] \\ \quad -\lambda_{0,i}A_i|\varphi^+(\beta_{k,i},t_{k,i})|^2 + (-1)^k\left[\dfrac{\partial l}{\partial t_k}(p_i,\lambda_i) + 2\lambda_{0,i}(t_{k,i} - t^*_k)\right] \geq 1(i), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 0,1, \\ \operatorname{ess\,lim\,inf}\limits_{t\to t_{k,i}} \max\limits_{u\in U(t)}[H(\beta_{k,i},u,\gamma_{k,i},t) - \lambda_{0,i}i^{-1}|u - u^*(t)|^2] \\ \quad -\lambda_{0,i}A_i|\varphi^+(\beta_{k,i},t_{k,i})|^2 + (-1)^k\left[\dfrac{\partial l}{\partial t_k}(p_i,\lambda_i) + 2\lambda_{0,i}(t_{k,i} - t^*_k)\right] \leq 1(i), \end{array}\right. \tag{7.18}$$

$$|\lambda_i| + \operatorname{gc\,sup}_{t\in T_i}|\psi_i(t)| + |\eta_i| = 1. \tag{7.19}$$

Here, $T_i = [t_{0,i}, t_{1,i}]$, $\lambda_i = (\lambda_{0,i}, \lambda_{1,i}, \lambda_{2,i})$, $\beta_{0,i} = x_i(t^+_{0,i})$, $\beta_{1,i} = x_i(t^-_{1,i})$, $\gamma_{0,i} = \psi_i(t^+_{0,i})$, $\gamma_{1,i} = \psi_i(t^-_{1,i})$, and $\eta_i = (\eta^1_i, \dots, \eta^{k_3}_i)$ is a vector measure whose components are defined, $j = 1, \dots, k_3$, by the distribution functions

$$F(t;\eta^j_i) = 2\lambda_{0,i}A_i\int_{[t_{0,i},t]}[\varphi^j(x_i,t)]^+ds + 2\lambda_{0,i}A_i\int_{[t_{0,i},t]}[\varphi^j_\delta(x_i,t)]^+d\mu_i, \quad t > t_{0,i}.$$

The expression $1(i)$ denotes a sequence of numbers converging to zero as $i \to \infty$, and the index i in H and Q (and its partial derivatives) indicates that they are evaluated at $x_i(t)$, $u_i(t)$, and $\psi_i(t)$, whenever the corresponding arguments x, u, and ψ are missing.

Note that because the function φ^+_δ is discontinuous, the i-problem is a nonstandard one. However, Theorem 6.2 applies in the same way for this problem. Furthermore, conditions (7.13)–(7.19) hold independently of the values $\mu_i(\{t_{k,\delta}\})$, $k = 0, 1$. This is so due to the nonnegativity and lower semicontinuity of the function $|\varphi^+_\delta|^2$ and also to the following fact: if $|\varphi^+|^2$ is zero at some point (x,t), then its derivative is zero at this point.

From (7.19), we get that $\zeta_i \to 0$ uniformly. Again from (7.19), and by using compactness, we conclude that, after subsequence extraction, $\lambda_i \to \lambda$, and $\eta^j_i \overset{w}{\to} \tilde\eta^j$, $j = 1, \dots, k_3$, as $i \to \infty$. Let, for every Borel set $B \subseteq \Re^1$, $\eta(B) = \tilde\eta(B) - \tilde\eta(B \cap \operatorname{Ds}(\mu^*))$, where $\tilde\eta = (\tilde\eta^1, \dots, \tilde\eta^{k_3})$. Thus, $\eta = (\eta^1, \dots, \eta^{k_3})$ is a vector measure satisfying $\operatorname{Ds}(\mu^*) \cap \operatorname{Ds}(\eta^j) = \emptyset$ $\forall j$.

Also from (7.19), it follows that the variation of ψ_i is bounded uniformly in i. By using the second Helly theorem and by extracting a subsequence, we obtain a function ψ_H such that $\psi_i(t) \to \psi_H(t)\ \forall t \in T^*$ as $i \to \infty$. Let us find $\psi \in V^n(T^*)$ such that $\psi(t) = \psi_H(t)\ \forall t \in \mathrm{Cont}(\mu^*) \cap \mathrm{Cont}(\eta)$.[6] Such function $\psi(t)$ exists because of the estimate $\mathrm{Var}\,|_a^b[\psi_H] \le c(\mu^* + \mathrm{Var}\,\tilde{\eta} + \mathcal{L})([a,b])\ \forall a \le b$, which follows from the inequality $\mathrm{Var}\,\nu[\psi_i] \le c(\mu_i + \mathrm{Var}\,\eta_i + \mathcal{L})$ as $i \to \infty$.

Let us prove that $\psi(t)$ satisfies the conditions of our theorem. For this purpose, let us construct a countable family of absolutely continuous measures $\{\hat{\mu}_i\}$, $i \in \mathbf{N}$, which approximately satisfies all the conditions (except, possibly, the ones concerning time transversality (7.18)) of the maximum principle for the i-problem, as follows.

Let $\{t_k\}$, $k \in \mathbf{N}$, be a countable set of points everywhere dense in T^* such that $t_k \in X = \mathrm{Cont}(\mu^*) \cap \mathrm{Cont}(\eta) \cap [\bigcap_{i=1}^{\infty} \mathrm{Cont}(\mu_i)]$,[7] and let $\{\phi_k\}$ be a countable set of functions everywhere dense in $C(T_c)$. For every i sufficiently large, there exists a sequence of absolutely continuous measures $\mu_{i,\tau}$ having densities $m_{i,\tau} > 0$ a.e., such that $\mu_{i,\tau} \xrightarrow{w} \mu_i$ on T_i, $\mu_{i,\tau} \xrightarrow{w} \mu_i$ on $[t_{0,i}, t_{0,\delta}]$, $\mu_{i,\tau} \xrightarrow{w} \mu_i$ on $[t_{1,\delta}, t_{1,i}]$ as $\tau \to \infty$ for all τ.

Let the pair $(x_{i,\tau}, \psi_{i,\tau})$ satisfy the following differential system:[8]

$$\begin{cases} dx_{i,\tau} = f(x_{i,\tau}, u_i, t)dt + g(x_{i,\tau}, t)d\mu_{i,\tau}, \\ d\psi_{i,\tau} = -H_x(x_{i,\tau}, u_i, \psi_{i,\tau}, t)dt - Q_x(x_{i,\tau}, \psi_{i,\tau}, t)d\mu_{i,\tau} + \varphi_x^{\top}(x_{i,\tau}, t)d\eta_{i,\tau}, \\ x_{i,\tau}(t_{0,i}) = x_i(t_{0,i}),\ \ \psi_{i,\tau}(t_{0,i}) = \psi_i(t_{0,i}), \end{cases}$$

where the measure $\eta_{i,\tau}^j$ is given by its distribution function

$$\begin{aligned} F(t; \eta_{i,\tau}^j) = {} & 2\lambda_{0,i}A_i \int_{[t_{0,i},t]} [\varphi^j(x_{i,\tau}, t)]^+ ds \\ & + 2\lambda_{0,i}A_i \int_{[t_{0,i},t]} [\varphi_\delta^j(x_{i,\tau}, t)]^+ d\mu_{i,\tau}, \quad j = 1, \dots, k_3. \end{aligned}$$

By using Lemma 5.4, choose a number τ_i such that

$$\begin{cases} \displaystyle\sum_{j=1}^{k_3} \sum_{k=1}^{i} \left[\left| \int_{T_i} \phi_k(t) d(\mu_{i,\tau_i} - \mu_i) \right| + \left| \int_{T_i} \phi_k(t) d(\eta_{i,\tau_i}^j - \eta_i^j) \right| \right] \le \frac{1}{i}, \\ \displaystyle\sum_{k=1}^{i} |\psi_{i,\tau_i}(t_k) - \psi_i(t_k)| + \left| \operatorname{gc}\sup_{t \in T_i} |\psi_i(t)| - \max_{t \in T_i} |\psi_{i,\tau_i}(t)| \right| \le \frac{1}{i}, \\ \displaystyle\left| \int_{T_i} \tilde{Q}_i(x_{i,\tau_i}, \psi_{i,\tau_i}, t) d\mu_{i,\tau_i} - \int_{T_i} \tilde{Q}_i(x_i, \psi_i, t) d\mu_i \right| \\ \qquad \displaystyle + \sum_{j=1}^{k_3} \left| \int_{T_i} \varphi^j(x_{i,\tau_i}, t) d\eta_{i,\tau_i}^j - \int_{T_i} \varphi^j(x_i, t) d\eta_i^j \right| \le \frac{1}{i}, \\ |\psi_{i,\tau_i}(t_{1,i}) - \psi_i(t_{1,i})| + \max_{t \in T_i} |\tilde{Q}_i(x_{i,\tau_i}, \psi_{i,\tau_i}, t) - \tilde{Q}_i(x_i, \psi_i, t)| \le \frac{1}{i}, \end{cases} \tag{7.20}$$

where $\tilde{Q}_i(x, \psi, t) = Q(x, \psi, t) - \lambda_{0,i}A_i|\varphi_\delta^+(x, t)|^2$.

[6] From now on, $\mathrm{Cont}(\eta) = \bigcap_{j=1}^{k_3} \mathrm{Cont}(\eta^j)$.

[7] Such an everywhere dense set exists, because, being a countable union of countable sets, $T^* \setminus X$ is countable.

[8] For every i, the solution of this system exists as τ becomes sufficiently large.

Put $\hat{\psi}_i = \psi_{i,\tau_i}$, $\hat{x}_i = x_{i,\tau_i}$, $\hat{\mu}_i = \mu_{i,\tau_i}$, $\hat{\eta}_i = \eta_{i,\tau_i}$. By extracting a subsequence, we have that $\hat{\psi}_i(t) \to \psi_A(t)$ $\forall t \in T^*$.

Now, we show that $\psi_A(t) = \psi(t)$ $\forall t \in \mathrm{Cont}(\mu^*) \cap \mathrm{Cont}(\eta)$. Indeed, since $\mathrm{Var}\,\nu[\psi_i] \le c(\mu_i + \mathrm{Var}\,\eta_i + \mathcal{L})$, $\psi(t)$ is continuous on the set $\mathrm{Cont}(\mu^*) \cap \mathrm{Cont}(\eta) \setminus \{t_0^*, t_1^*\}$. The same can be said about ψ_A. Then, it follows from (7.20) that $\psi_A(t) = \psi(t)$ $\forall t \in \mathrm{Cont}(\mu^*) \cap \mathrm{Cont}(\eta)$. For such a sequence, $\hat{x}_i(t) \to x^*(t)$, $\forall t \in \mathrm{Cont}(\mu^*)$, $\hat{\mu}_i \xrightarrow{w} \mu^*$, $\hat{\eta}_i^j \xrightarrow{w} \tilde{\eta}^j$, $j = 1, \dots, k_3$, as $i \to \infty$.

Let us prove (7.1). We start by showing that there are sequences of absolutely continuous measures $\{\bar{\mu}_i\}$, $\{\bar{\eta}_i^j\}$, $j = 1, \dots, k_3$, and also a sequence of natural numbers $k_i \ge i$ such that

(1) $\bar{\mu}_i \xrightarrow{w} \mu_d^*$, and, $\forall j$, $\bar{\eta}_i^j \xrightarrow{w} \tilde{\eta}^j - \eta^j$ as $i \to \infty$,

(2) $\hat{\mu}_{k_i} \ge \bar{\mu}_i$, $\hat{\eta}_{k_i}^j \ge \bar{\eta}_i^j$ $\forall i, j$.

If $\mathrm{Ds}(\mu^*) = \emptyset$, then we put $\bar{\mu}_i = \bar{\eta}_i^j = 0$, $k_i = i$, $\forall i, j$.

Let $\mathrm{Ds}(\mu^*) \ne \emptyset$. Consider the chain of sets D_i such that $D_0 = \emptyset$, $D_{i-1} \subseteq D_i \subseteq \mathrm{Ds}(\mu^*)$ and $\sum_{r \in \mathrm{Ds}(\mu^*) \setminus D_i} \mu^*(\{r\}) \le \frac{1}{i}$, $i \in \mathbf{N}$. Define sets $S_{r,i} = [r - \rho_i, r + \rho_i]$, $r \in D_i$, as a system of pairwise disjoint closed neighborhoods of points r such that

(i) $\rho_i > 0$, $\rho_i \to 0$ as $i \to \infty$,

(ii) $|\mu^*(S_i) - \sum_{r \in D_i} \mu^*(\{r\})| + \sum_{j=1}^{k_3} |\tilde{\eta}^j(S_i) - \sum_{r \in D_i} \tilde{\eta}^j(\{r\})| \le \frac{1}{i}$, where $S_i = \bigcup_{r \in D_i} S_{r,i}$,

(iii) $r \pm \rho_i \in \mathrm{Cont}(\mu^*) \cap \mathrm{Cont}(\eta)$.

The existence of such sets S_i follows from the regularity of Borel measures $\mu^*, \tilde{\eta}^j$, $j = 1, \dots, k_3$. Due to the weak star convergence of the considered sequences of measures, it is possible to take a natural number $k_i \ge i$ such that

$$\sum_{r \in D_i} \left(\left| \hat{\mu}_{k_i}(S_{r,i}) - \mu^*(S_{r,i}) \right| + \sum_{j=1}^{k_3} \left| \hat{\eta}_{k_i}^j(S_{r,i}) - \tilde{\eta}^j(S_{r,i}) \right| \right) \le \frac{1}{i}.$$

Define $\bar{\mu}_i$ and $\bar{\eta}_i = \{\bar{\eta}_i^1, \dots, \bar{\eta}_i^{k_3}\}$, respectively, by $\bar{\mu}_i(B) := \hat{\mu}_{k_i}(B \cap S_i)$, and $\bar{\eta}_i(B) := \hat{\eta}_{k_i}(B \cap S_i)$, for any Borel set $B \subseteq \Re^1$. Clearly, $\bar{\mu}_i \xrightarrow{w} \mu_d^*$, $\bar{\eta}_i^j \xrightarrow{w} \tilde{\eta}^j - \eta^j$ as $i \to \infty$ (and, hence, $\hat{\mu}_{k_i} - \bar{\mu}_i \xrightarrow{w} \mu_c^*$, $\hat{\eta}_{k_i}^j - \bar{\eta}_i^j \xrightarrow{w} \eta^j$) for $j = 1, \dots, k_3$. It is also clear that $\hat{\mu}_{k_i} - \bar{\mu}_i \in C_+^*(T_c)$, $\hat{\eta}_{k_i}^j - \bar{\eta}_i^j \in C_+^*(T_c)$ $\forall i, j$.

From now on, consider subsequences extracted from $\{\hat{\mu}_i\}$, $\{\hat{\eta}_i\}$, $\{\hat{x}_i\}$, $\{\hat{\psi}_i\}$, $\{u_i\}$, and $\{T_i\}$ (denote them again by the old index i) according to the constructed sequence $\{k_i\}$ and let us rewrite (1.2), (7.13) as follows:

$$\begin{cases} \hat{x}_i(t) = x_{0,i} + \displaystyle\int_{t_{0,i}}^{t} f(\hat{x}_i, u_i, s)ds + \int_{[t_{0,i},t]} g(\hat{x}_i, s)d(\hat{\mu}_i - \bar{\mu}_i) + \int_{[t_{0,i},t]} g(\hat{x}_i, s)d\bar{\mu}_i, \\ \hat{\psi}_i(t) = \psi_{0,i} - \displaystyle\int_{t_{0,i}}^{t} \hat{H}_x^i(s)ds - \int_{[t_{0,i},t]} \hat{Q}_x^i(s)d(\hat{\mu}_i - \bar{\mu}_i) - \int_{[t_{0,i},t]} \hat{Q}_x^i(s)d\bar{\mu}_i \\ \qquad + \displaystyle\int_{[t_{0,i},t]} \varphi_x^\top(\hat{x}_i, s)d(\hat{\eta}_i - \bar{\eta}_i) + \int_{[t_{0,i},t]} \varphi_x^\top(\hat{x}_i, s)d\bar{\eta}_i, \quad t \in T_i. \end{cases}$$

From now on, the index i and the sign hat in H and Q (and also its partial derivatives) stand for their values when $\hat{x}_i(t)$, $u_i(t)$, and $\hat{\psi}_i(t)$ replace the omitted arguments x,

u, and ψ respectively. Put, $\forall t \in T_i$,

$$\hat{x}_i^c(t) = x_{0,i} + \int_{t_{0,i}}^{t} f(\hat{x}_i, u_i, s)ds + \int_{[t_{0,i},t]} g(\hat{x}_i, s)d(\hat{\mu}_i - \bar{\mu}_i),$$

$$\hat{x}_i^d(t) = \int_{[t_{0,i},t]} g(\hat{x}_i, s)d\bar{\mu}_i,$$

$$\hat{\psi}_i^{\eta}(t) = \psi_{0,i} - \int_{t_{0,i}}^{t} \hat{H}_x^i(s)ds - \int_{[t_{0,i},t]} \hat{Q}_x^i(s)d(\hat{\mu}_i - \bar{\mu}_i) + \int_{[t_{0,i},t]} \varphi_x^{\top}(\hat{x}_i, s)d(\hat{\eta}_i - \bar{\eta}_i),$$

$$\hat{\psi}_i^{d}(t) = -\int_{[t_{0,i},t]} \hat{Q}_x^i(s)d\bar{\mu}_i + \int_{[t_{0,i},t]} \varphi_x^{\top}(\hat{x}_i, s)d\bar{\eta}_i.$$

Thus, $\hat{x}_i(t) = \hat{x}_i^c(t) + \hat{x}_i^d(t)$, $\hat{\psi}_i(t) = \hat{\psi}_i^{\eta}(t) + \hat{\psi}_i^d(t)$ $\forall t \in T_i$. Bearing in mind that $\hat{x}_i(t) \to x^*(t)$ $\forall t \in \mathrm{Cont}(\mu^*)$, by Lebesgue theorem and Lemma 5.1, we deduce that

$$\hat{x}_i^c(t) \to x_c^*(t) = x_0^* + \int_{t_0^*}^{t} f(x^*, u^*, s)ds + \int_{[t_0^*,t]} g(x^*, s)d\mu_c^* \quad \forall t \in T^*,\ i \to \infty.$$

In a similar way, from Lemma 5.1, we conclude that $\forall t \in \mathrm{Cont}(\eta)$,

$$\hat{\psi}_i^{\eta}(t) \to \psi_{\eta}(t) = \psi_0 - \int_{t_0^*}^{t} H_x(s)ds - \int_{[t_0^*,t]} Q_x(s)d\mu_c^* + \int_{[t_0^*,t]} \varphi_x^{\top}(x^*, s)d\eta$$

as $i \to \infty$. Let $\psi_d = \psi - \psi_{\eta}$. Then, $\hat{\psi}_i^d(t) \to \psi_d(t)$ $\forall t \in \mathrm{Cont}(\mu^*) \cap \mathrm{Cont}(\eta)$.

Let us show $\psi_d = \Sigma(\psi, t)$. In fact, fix $t \in \mathrm{Cont}(\mu^*) \cap \mathrm{Cont}(\eta)$ and $\varepsilon > 0$. Take a number $N = N(\varepsilon)$ such that $\sum_{r \in \mathrm{Ds}(\mu^*) \setminus D_N} [\mu^*(\{r\}) + \mathrm{Var}\, \tilde{\eta}(\{r\})] \le \varepsilon$. In this case, then

$$\limsup_{i \to \infty} \left| \sum_{r \in D(N,t)} \left[\int_{S_{r,i}} -\hat{Q}_x^i(s)d\bar{\mu}_i + \int_{S_{r,i}} \varphi_x^{\top}(\hat{x}_i, s)d\bar{\eta}_i \right] - \hat{\psi}_i^d(t) \right| \le \mathrm{const}\, \varepsilon, \tag{7.21}$$

where, from now on, $D(N,t) = \{r \in D_N : r \le t\}$.

Let $r_i^- = r - \rho_i$ and, for $r \in D(N,t)$, consider, on segment $S_{r,i}$, the system

$$\begin{cases} \hat{x}_i^d(s) = \hat{x}_i^d(r_i^-) + \displaystyle\int_{[r_i^-,s]} g(\hat{x}_i(\tau), \tau)d\bar{\mu}_i, \\ \hat{\psi}_i^d(s) = \hat{\psi}_i^d(r_i^-) - \displaystyle\int_{[r_i^-,s]} g_x^{\top}(\hat{x}_i(\tau), \tau)\hat{\psi}_i(\tau)d\bar{\mu}_i + \int_{[r_i^-,s]} \varphi_x^{\top}(\hat{x}_i(\tau), \tau)d\bar{\eta}_i. \end{cases} \tag{7.22}$$

For each i sufficiently large, we define the function[9]

$$\pi_{r,i}(\tau) = \frac{F(\tau; \bar{\mu}_i) - F(r_i^-; \bar{\mu}_i)}{\bar{\mu}_i(S_{r,i})}, \quad r \in D(N,t),\ \tau \in S_{r,i}.$$

The function $\pi_{r,i} : S_{r,i} \to [0,1]$ is absolutely continuous and strictly increasing, $\frac{d\pi_{r,i}}{d\tau} = \frac{\bar{m}_i(\tau)}{\bar{\mu}_i(S_{r,i})} > 0$ ($\bar{m}_i(\tau)$ is the density of measure $\bar{\mu}_i$). Hence, there exists an inverse mapping, $\theta_{r,i} : [0,1] \to S_{r,i}$, $\theta_{r,i} = (\pi_{r,i})^{-1}$, which is also absolutely continuous and

[9]Let us take numbers i such that $t \notin S_{r,i}$ $\forall i \ge i_0$, $r \in D(N,t)$, where i_0 is sufficiently large.

strictly increasing. By performing the change of variable $\omega = \pi_{r,i}(\tau)$ in (7.22), we obtain the system defined for all $s \in [0,1]$, and $r \in D(N,t)$,

$$\alpha_{r,i}(s) = \hat{x}_i^d(r_i^-) + \hat{x}_i^c(\theta_{r,i}(s)) + \bar{\mu}_i(S_{r,i}) \int_0^s g(\alpha_{r,i}(\omega), \theta_{r,i}(\omega))d\omega,$$
$$\sigma_{r,i}(s) = \hat{\psi}_i^d(r_i^-) + \hat{\psi}_i^\eta(\theta_{r,i}(s)) - \bar{\mu}_i(S_{r,i}) \int_0^s g_x^\top(\alpha_{r,i}(\omega), \theta_{r,i}(\omega))\sigma_{r,i}(\omega)d\omega$$
$$+ \int_{[0,s]} \varphi_x^\top(\alpha_{r,i}(\omega), \theta_{r,i}(\omega))d\eta_{r,i},$$

where

$$\alpha_{r,i}(\omega) = \hat{x}_i(\theta_{r,i}(\omega)), \quad \sigma_{r,i}(\omega) = \hat{\psi}_i(\theta_{r,i}(\omega)),$$
$$F(\omega; \eta_{r,i}^j) = F(\theta_{r,i}(\omega); \bar{\eta}_i^j), \quad j = 1, \dots, k_3.$$

Clearly, $\theta_{r,i}(\omega) \to r$ uniformly on $[0,1]$, and, by Lemma 5.3, $\hat{x}_i^c(\theta_{r,i}(\omega)) \to x_c^*(r)$, and $\hat{\psi}_i^\eta(\theta_{r,i}(\omega)) \to \psi_\eta(r)$ uniformly on $[0,1]$ as $i \to \infty$. From Gronwall's inequality, it follows that $\alpha_{r,i}$ is a Cauchy sequence. Its limit is α_r^*.[10] Compactness yields the existence of measures η_r^j and of a function $\sigma_r \in V^n([0,1])$ such that, after subsequence extraction, $\hat{\eta}_{r,i}^j \xrightarrow{w} \eta_r^j$, $j = 1, \dots, k_3$, $\sigma_{r,i}(s) \to \sigma_r(s)$, $\forall s \in \bigcap_{j=1}^{k_3} \mathrm{Cont}(\eta_r^j)$. By passing to the limit in the last system as $i \to \infty$, we deduce that[11] $\forall s \in [0,1]$ and $r \in D(N,t)$,

$$\alpha_r^*(s) = x^*(r^-) + \mu^*(\{r\}) \int_0^s g(\alpha_r^*, r)d\omega,$$
$$\sigma_r(s) = \psi(r^-) - \mu^*(\{r\}) \int_0^s g_x^\top(\alpha_r^*, r)\sigma_r d\omega + \int_{[0,s]} \varphi_x^\top(\alpha_r^*, r)d\eta_r.$$

From here and from (7.21), we have

$$\left| \psi_d(t) - \sum_{r \in D(N,t)} [\sigma_r(1) - \psi(r^-)] \right| \le \mathrm{const}\, \varepsilon.$$

But $\varepsilon > 0$ is arbitrary. Then, by definition of Σ, this means $\psi_d(t) = \Sigma(\psi, t)$. Thus, (7.1) and (7.2) are proved.

A simple contradiction argument implies (7.9). In fact if $|\lambda_i| \to 0$, $|\hat{\eta}_i| \to 0$, then $\max_{t \in T_i} |\hat{\psi}_i(t)| \to 0$ as $i \to \infty$, and (7.19) does not hold.

Let us prove (7.3), and (7.4). Fix $r \in \mathrm{Ds}(\mu^*)$, $j = 1, \dots, k_3$. From the penalty method and (7.20), we have that

$$A_i \int_{T_i} |\varphi^+(\hat{x}_i, t)|^2 dt + A_i \int_{T_i} |\varphi_\delta^+(\hat{x}_i, t)|^2 d\hat{\mu}_i \to 0. \tag{7.23}$$

Furthermore,

$$2A_i \int_{T_i} |\varphi^+(\hat{x}_i, t)|^2 dt + 2A_i \int_{T_i} |\varphi_\delta^+(\hat{x}_i, t)|^2 d\hat{\mu}_i$$
$$\ge 2\lambda_{0,i} A_i \left(\int_{T_i} \varphi^j(\hat{x}_i, t)[\varphi^j(\hat{x}_i, t)]^+ dt + \int_{T_i} \varphi_\delta^j(\hat{x}_i, t)[\varphi_\delta^j(\hat{x}_i, t)]^+ d\hat{\mu}_i \right)$$
$$= \int_{T_i} \varphi^j(\hat{x}_i, t) d\hat{\eta}_i^j \ge 0.$$

[10] This is the corollary of the uniqueness of the solution [23].

[11] Here, we used the fact that by construction, $\hat{x}_i^d(r_i^-) \to x^*(r^-) - x_c^*(r)$ and $\hat{\psi}_i^d(r_i^-) \to \psi_d(r^-)$.

Then, from (7.23), $\int_{S_{r,i}} \varphi^j(\hat{x}_i,t)d\hat{\eta}_i^j \to 0$ as $i \to \infty$. By changing variable $s = \pi_{r,i}(t)$ and by passing to the limit we get $\int_0^1 \varphi^j(\alpha_r^*,r)d\eta_r^j = 0$. Since $\varphi^j(\alpha_r^*(s),r) \le 0$, we conclude that $\varphi^j(\alpha_r^*(s),r) = 0\ \forall s \in \mathrm{supp}(\eta_r^j)$.

Next, we show that $Q(t) = Q(x^*(t),\psi(t),t) \le 0$, $\forall t \in T^*$. Indeed, $\zeta_i \to 0$ uniformly on T^*. Let $\hat{\kappa}_i(t) = \lambda_{0,i}A_i|\varphi_\delta^+(\hat{x}_i(t),t)|^2$. From (7.23), $\hat{\kappa}_i(t) \to 0$ strongly in $L_1(T^*)$, and, by extracting subsequences, $\hat{\kappa}_i(t) \to 0$ a.e. From (7.16), (7.20), and the right continuity of $Q(t)$ on interval $[t_{0,\delta},t_{1,\delta}]$, we establish that $Q(t) \le 0\ \forall t \in T^*$.

From (7.17), (7.20), and $\int_{T_i} \hat{Q}_i(t)d\hat{\mu}_i \to 0$, we have that $\int_0^1 \langle g(\alpha_r^*,r),\sigma_r\rangle ds = 0$. From (7.16), (7.20), (7.23), and $Q(t) \le 0\ \forall t$, it is not hard to deduce by arguments analogous to the ones that show $\varphi^j(\alpha_r^*(s),r) = 0\ \forall s \in \mathrm{supp}(\eta_r^j)$ that $q_r(s) = \langle g(\alpha_r^*(s),r),\sigma_r(s)\rangle \le 0\ \forall s \in [0,1]$. Thus, $q_r(s) = 0\ \forall s \in (0,1)$. Let us show that $q_r(0) = q_r(1) = 0$. Indeed, if $\eta_r(\{0\}) = 0$, then $q_r(0) = 0$. Let $\eta_r^j(\{0\}) > 0$ for some $j = 1,\dots,k_3$. Then, $\varphi^j(\alpha_r^*(0),r) = 0$ and $W^j(\alpha_r^*(0),r) \le 0$. Otherwise the state constraint would not be satisfied. Thus, $q_r(0) \ge 0$. But, at the same time, the complementary inequality is true. Hence, $q_r(0) = W^j(\alpha_r^*(0),r) = 0$. Similarly, $q_r(1) = 0$. We proved statements (7.3), and (7.4).

Let us prove (7.5). Since $\mathrm{Var}\,\nu[\varphi^j(\hat{x}_i,t)] \le c(\hat{\mu}_i + \mathcal{L})$, we have, by Lemma 5.1, that

$$\int_{T_i} \varphi^j(\hat{x}_i,t)d[\hat{\eta}_i^j - \bar{\eta}_i^j] \to \int_{T^*} \varphi^j(x^*,t)d\eta^j, \quad i \to \infty,\ j = 1,\dots,k_3.$$

On the other hand, it follows from (7.23) that $\int_{T_i} \varphi^j(\hat{x}_i,t)d[\hat{\eta}_i^j - \bar{\eta}_i^j] \to 0$. Now, (7.5) follows from nonpositivity of $\varphi^j(x^*,t)$, $j = 1,\dots,k_3$.

By passing to the limit in (7.15), we establish (7.6).

Let us show how to prove (7.7) with the help of Lemma 5.1. It has already been asserted that $\mathrm{Ds}(\nu[Q]) \cap \mathrm{Ds}(\mu^*) = \emptyset$. However, it is not enough to use Lemma 5.1 to obtain the desired conclusion. Next, we show that $\forall r \in \mathrm{Ds}(\mu^*)$, $\mathrm{Var}\,|_0^1[q_{r,i}] \to 0$ as $i \to \infty$, where $q_{r,i}(s) = \langle \sigma_{r,i}(s), g(\alpha_{r,i}(s),\theta_{r,i}(s))\rangle$, $s \in [0,1]$. In fact,

$$q_{r,i}(s) = q_{r,i}(0) + \sum_{j=1}^{k_3} \int_0^s W^j(\alpha_{r,i},\theta_{r,i})d\eta_{r,i}^j + \int_0^s \langle \sigma_{r,i}, g_t(\alpha_{r,i},\theta_{r,i})\rangle d\theta_{r,i}, \quad s \in [0,1].$$

From this we conclude that

$$\mathrm{Var}\,|_0^1[q_{r,i}] \le \sum_{j=1}^{k_3} \mathrm{Var}\,|_0^1 \left[\int_0^s W^j(\alpha_{r,i},\theta_{r,i})d\eta_{r,i}^j\right] + 1(i).$$

Since $W^j(\alpha_{r,i},\theta_{r,i}) \to W^j(\alpha_r^*,r)$ uniformly on $[0,1]$ as $i \to \infty$, and since $\eta_{r,i}^j \xrightarrow{w} \eta_r^j$, we have that $\mathrm{supp}(\eta_r^j) \subseteq \{s : W^j(\alpha_r^*(s),r) = 0\}\ \forall j$. From Lemma 5.5, $\mathrm{Var}\,|_0^1[q_{r,i}] \to 0$. Hence,

$$\mathrm{Var}\,\nu[\hat{Q}_i] \le c([\hat{\mu}_i - \bar{\mu}_i] + \mathrm{Var}[\hat{\eta}_i - \bar{\eta}_i] + \mathcal{L}) + 1(i).$$

By using Lemma 5.1, (7.17), (7.20), and (7.23), we deduce that

$$1(i) = \int_{T_i} \hat{Q}_i(t)d\hat{\mu}_i + \int_{T_i} \hat{\kappa}_i(t)d\hat{\mu}_i \to \int_{T^*} Q(t)d\mu^*$$

and, therefore, $\int_{T^*} Q(t)d\mu^* = 0$. Bearing in mind that $Q(t) \le 0\ \forall t$, we conclude (7.7).

Let us prove (7.10). We consider $k = 0$.

We return to the initial sequence of solutions (7.12) and consider conditions (7.18). Note that $\Delta_{0,i} = \mu_i(\{t_{0,i}\}) \to 0$ since μ^* is continuous at t_0^*. Hence, $x_{0,i}^+ = \beta_{0,i} \to x_0^*$ and $\psi_{0,i}^+ = \gamma_{0,i} \to \psi_0$ as $i \to \infty$. We put $\kappa_i = \lambda_{0,i} A_i[\varphi^+(x_{0,i}^+, t_{0,i})]^2$ and show that $\kappa_i \to 0$ as $i \to \infty$. Two cases can be considered:

(1) The sequence $\Delta_{0,i}$ has a zero subsequence.

(2) $\Delta_{0,i} > 0$ for all i greater some i_0.

By extracting a subsequence we have, in the first case, that $x_{0,i}^+ = x_{0,i}$ and, from the compatibility between state and endpoint constraints, $\kappa_i = 0$.

Let us consider the second case. When i is sufficiently large, the function $Q_i(t)$ is absolutely continuous in some neighborhood of point $t_{0,i}$ and its derivative in this set is given by

$$\dot{Q}_i(t) = \sum_{j=1}^{k_3} 2\lambda_{0,i} A_i [\varphi^j(x_i(t), t)]^+ W^j(x_i(t), t) + \omega_i(t), \tag{7.24}$$

where $\omega_i(t)$ is an essentially bounded measurable function, $\|\omega_i\|_{L_\infty} \le \text{const}\ \forall i$.

Since $\Delta_{0,i} > 0$, we have, from (7.17), $Q_i(t_{0,i}) + \zeta_{0,i} = 0$. By using this and (7.16), we conclude by integrating (7.24) with the Newton–Leibnitz formula that

$$\sum_{j=1}^{k_3} \int_{t_{0,i}}^{t_{0,i}+\Delta} 2\lambda_{0,i} A_i [\varphi^j(x_i, s)]^+ W^j(x_i, s) ds \le \text{const}\, \Delta.$$

By dividing both sides by $\Delta > 0$ and passing to the limit as $\Delta \to 0$, we obtain

$$\sum_{j=1}^{k_3} \lambda_{0,i} A_i [\varphi^j(x_{0,i}^+, t_{0,i})]^+ W^j(x_{0,i}^+, t_{0,i}) \le \text{const}. \tag{7.25}$$

From Assumption C and the compatibility of state and endpoint constraints, it follows that if $W^j(x_{0,i}^+, t_{0,i}) \le 0$, then $[\varphi^j(x_{0,i}^+, t_{0,i})]^+ = 0$. In this way, from (7.25), we conclude that

$$\lambda_{0,i} A_i [\varphi^j(x_{0,i}^+, t_{0,i})]^+ W^j(x_{0,i}^+, t_{0,i}) \le \text{const}, \quad j = 1, \dots, k_3. \tag{7.26}$$

Let $\kappa_i^j = \lambda_{0,i} A_i [\varphi^{j+}(x_{0,i}^+, t_{0,i})]^2$. Notice that $\kappa_i = \sum_{j=1}^{k_3} \kappa_i^j$. If $W^j(x_{0,i}^+, t_{0,i}) \le 0$, then, as it was seen above, $\kappa_i^j = 0$. Suppose that $\kappa_i^j > 0$. Then, $W^j(x_{0,i}^+, t_{0,i}) > 0$. Furthermore,

$$\begin{aligned}
\kappa_i^j &= \lambda_{0,i} A_i [\varphi^+(x_{0,i}^+, t_{0,i})]^2 \\
&= \lambda_{0,i} A_i \left[\varphi^+(x_{0,i}, t_{0,i}) + \Delta_{0,i} \int_0^1 h[\varphi^j(\alpha_{0,i}, t_{0,i})] W^j(\alpha_{0,i}, t_{0,i}) ds \right]^2,
\end{aligned}$$

where $h(t)$ is the Heavyside function, i.e., $h(t) = 0$ when $t \le 0$ and $h(t) = 1$ otherwise. From the last equality, we have, by the constraints compatibility assumption and Assumption C, that for sufficiently large i,

$$\kappa_i^j \le 2\lambda_{0,i} A_i \Delta_{0,i}^2 [W^j(x_{0,i}^+, t_{0,i})]^2.$$

Now, from (7.26), we have

$$\kappa_i^j \le \text{const}\, \frac{\lambda_{0,i} A_i \Delta_{0,i}^2}{[\lambda_{0,i} A_i \varphi^{j+}(x_{0,i}^+, t_{0,i})]^2} = \text{const}\, \frac{\Delta_{0,i}^2}{\kappa_i^j},$$

and, from this, we conclude that

$$(\kappa_i^j)^2 \le \text{const}\, \Delta_{0,i}^2 \to 0.$$

Now, by passing to the limit in (7.18), we obtain (7.10).

Now note that Assumption C and the requirement $\Delta_k^* = 0$, $k = 0, 1$, are used only to obtain the time transversality conditions and, therefore, can be omitted in the remaining parts of the proof. Indeed, let us modify the penalty term in (7.11) $\int_T [\varphi_\delta^+(x,t)]^2 d\mu$ onto $\int_T [\varphi^+(x,t)]^2 d\mu$. It is clear that such a transformation of the function φ_δ^+ into φ^+ does not invalidate the arguments proving (7.1)–(7.9). The only issue here is to ensure that the compatibility of state and endpoint constraints stays in force and cannot be omitted. Indeed, we use it to prove (a) the maximum condition for the impulsive component (more precisely, to get inequalities $Q(t_k^*) \le 0$, $k = 0, 1$) and (b) inequalities (7.8).

The proof is complete. □

8. Nondegenerate maximum principle. Here we prove our main result, Theorem 4.1. This theorem is formulated under the smooth Assumption S. Therefore, we shall need the following.

Remark 5. Under Assumption S, inequalities (7.8) in Theorem 7.1 become

$$\max_{u \in U} H(x_k^*, u, \psi_k, t_k^*) + (-1)^k \frac{\partial l}{\partial t_k}(p^*, \lambda) \le 0, \quad k = 0, 1. \tag{8.1}$$

Proof of Theorem 4.1. Let us verify the technique suggested in [1]. For this, we shall use the first statement of Theorem 7.1, i.e., conditions (7.1)–(7.9).

Equations (4.2) can be obtained by applying Theorem 7.1 to v-problem.[12] This proof is absolutely analogous to the corresponding one in [1].

We shall consider the case $\Delta_k^* = 0$[13] and use a contradiction argument to show (4.3). Assume that (4.3) does not hold. Then, because of the right continuity of ψ on $B^* = (t_0^*, t_1^*)$, we have that $\lambda_0 = 0$, $\psi(t) = 0$ $\forall t \in B^*$ and, therefore, $\sigma_r(s) = 0$, $\forall s \in [0,1]$, $\forall r \in \mathrm{Ds}(\mu^*)$. By using the regularity of state constraints, we conclude that $\eta(B^*) = 0$, $\eta_r = 0$ $\forall r \in \mathrm{Ds}(\mu^*)$. Furthermore, from the maximum principle,

$$\varphi_x^\top(x_k^*, t_k^*)\eta(\{t_k^*\}) = -\frac{\partial l}{\partial x_k}(p^*, \lambda), \quad k = 0, 1.$$

In the same way, from (4.2) we obtain

$$\langle \varphi_t(x_k^*, t_k^*), \eta(\{t_k^*\})\rangle = -\frac{\partial l}{\partial t_k}(p^*, \lambda), \quad k = 0, 1.$$

By substituting these expressions into (8.1), we deduce that for $k = 0, 1$,

$$\max_{u \in U} H(x_k^*, u, (-1)^{k+1}\varphi_x^\top(x_k^*, t_k^*)\eta(\{t_k^*\}), t_k^*) + (-1)^{k+1}\langle \varphi_t(x_k^*, t_k^*), \eta(\{t_k^*\})\rangle \le 0.$$

Analogously, by substituting in inequalities $Q(t_k^*) \le 0$, $k = 0, 1$, we deduce, for $k = 0, 1$,

$$\langle g(x_k^*, t_k^*), (-1)^{k+1}\varphi_x^\top(x_k^*, t_k^*)\eta(\{t_k^*\})\rangle \le 0.$$

[12] This is obtained by reduction R_2. The equivalence is guaranteed by Proposition 5.7.

[13] The case $\Delta_k^* > 0$ is dealt with in a similar way.

From the controllability of $x^*(t)$, we conclude that $\eta(\{t_k^*\}) = 0$, $k = 0, 1$, and, therefore, $\frac{\partial l}{\partial p}(p^*, \lambda) = 0$. From the regularity of endpoint constraints, we get $\lambda = 0$. Thus, we obtain a contradiction with (7.9).

The proof is complete. □

Remark 6. Here, we want to emphasize the following important class of problems. Suppose that all state constraints are convex along the jump evolutions (i.e., $\dot{W}^j \geq 0$ $\forall j$). For such problems, we have the following.

(1) The controllability concept can be simplified as follows. An optimal trajectory $x^*(t)$ is controllable if there exist points $s_k \in [0, 1]$, $k = 0, 1$, such that Definition 2.4 holds at $(\tilde{x}_k, t_k^*)$, where $\tilde{x}_k = \alpha_k^*(s_k)$. This can be proved because of the reduction R_1 (where the extra state constraints can be omitted for the considered class of problems).

(2) Assumption C always holds under mentioned convexity supposition, hence for such problems we may obtain the time transversality conditions for problems merely measurable in t problem (Theorem 7.1). Note that, here, time transversality conditions can always be derived with the help of reduction R_2 (see Remark 2 in section 4). However, we need smooth Assumption S to use this reduction.

A simple example of the problem with pointed convexity properties is given by linear systems: g does not depend on x, φ is linear in x. Furthermore for linear problems, (4.3) becomes

$$\lambda_0 + \mathcal{L}(\{t : |\psi(t)| > 0\}) = 1.$$

In conclusion, we say that the following question pertaining to Assumption C remains open: *How do we get time transversality conditions for the case when data of the problem is measurable in t without any convexity assumptions on state constraints?*

9. Appendix. *Proof of Lemma* 5.1. We start by noting that if $s \in T$ and $t_i \to s$ as $i \to \infty$, then

$$F(s^-; \mu) \leq \liminf_{t_i \to s} F(t_i; \mu_i) \leq \limsup_{t_i \to s} F(t_i; \mu_i) \leq F(s^+; \mu).$$

This statement can be easily obtained with the help of Lemma 7.1 in [1] and with a simple contradiction arguments. By using the last statement, it is not hard to establish that

$$\limsup_{i \to \infty} \sup_{t \in K} |F(t; \mu_i) - F(t; \mu)| \leq \varepsilon, \tag{9.1}$$

where K is a closed subset of T such that $\mu(\{t\}) \leq \varepsilon$ $\forall t \in K$ $\varepsilon \geq 0$.

Furthermore, the triangle inequality gives

$$\left| \int_T f_i(t) d\mu_i - \int_T f(t) d\mu \right| \leq \left| \int_T f_i(t) d(\mu_i - \mu) \right| + \left| \int_T [f_i(t) - f(t)] d\mu \right|. \tag{9.2}$$

By the Lebesgue theorem, the second term in the right-hand side converges to zero. In fact, $f_i(t) \to f(t)$, μ-a.e. Furthermore, the f_i are uniformly bounded.

Thus, we have to show that $\int_T f_i(t) d(\mu_i - \mu) \to 0$. Let $F_i(t) = F(t; \mu_i)$ and $F(t) = F(t; \mu)$. By integrating by parts, we deduce

$$\int_T f_i(t) d(\mu_i - \mu) = f_i(t_1)[F_i(t_1) - F(t_1)] - \int_T [F_i(t) - F(t)] d\nu[f_i].$$

From the weak convergence $F_i(t_1) \to F(t_1)$. Let us show that $\int_T [F_i(t) - F(t)] d\nu[f_i] \to 0$.

There are two cases: either $\mathrm{Ds}(\mu) = \emptyset$ or $\mathrm{Ds}(\mu) \neq \emptyset$. If $\mathrm{Ds}(\mu) = \emptyset$, then this means that measure μ is continuous, and, by using (9.1) (with $\varepsilon = 0$, $K = T$), we conclude the proof. Let $\mathrm{Ds}(\mu) \neq \emptyset$ and fix $\varepsilon > 0$ sufficiently small. There exists a finite, not empty set of points $t_j \in \mathrm{Ds}(\mu)$, $j = 1, \ldots, N$, $N = N(\varepsilon) \geq 1$, such that $\mu(\{t_j\}) \geq \varepsilon$. Since $\mathrm{Ds}(\mu) \cap \mathrm{Ds}(\eta) = \emptyset$, then $\forall j \leq N$, $\exists \delta_j > 0$ such that $\eta(C_j) \leq \varepsilon N^{-1}$, where $C_j = [t_j - \delta_j, t_j + \delta_j]$, $t_j \pm \delta_j \in \mathrm{Cont}(\eta)$ and C_j are pairwise disjoint. Let O_j be a neighborhood of t_j such that $O_j \subset C_j$. Set $O = \bigcup_{j=1}^N O_j$ is open. Hence, $K = T \setminus O$ is closed. For such $O \subset C$, $\eta(O) \leq \eta(C) \leq \varepsilon$, where $C = \bigcup_{j=1}^N C_j$. From this, we conclude that

$$\left| \int_T [F_i(t) - F(t)] d\nu[f_i] \right| \leq \int_K |F_i(t) - F(t)| d \operatorname{Var} \nu[f_i] + \int_C |F_i(t) - F(t)| d \operatorname{Var} \nu[f_i].$$

By using (9.1) and $\operatorname{Var} \nu[f_i] \leq c\eta_i$ $\forall i$, we deduce that

$$\begin{aligned} \limsup_{i \to \infty} \left| \int_T [F_i(t) - F(t)] d\nu[f_i] \right| &\leq \limsup_{i \to \infty} \left(\sup_{t \in K} \left| F_i(t) - F(t) \right| \operatorname{Var} \nu[f_i](K) \right) \\ &\quad + \limsup_{i \to \infty} \left(\sup_{t \in C} \left| F_i(t) - F(t) \right| \operatorname{Var} \nu[f_i](C) \right) \\ &\leq \varepsilon c |\eta| + 2\varepsilon c |\mu| \leq c_1 \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, the right-hand part of (9.2) converges to zero. The proof is complete. □

Proof of Lemma 5.3. Fix $\varepsilon > 0$ such that $s \pm \varepsilon \in \mathrm{Cont}(\mu)$. Furthermore,

$$x_i(s + \varepsilon) - x_i(t_i) = \int_{[t_i, s+\varepsilon]} f_i(t) d\mu_i = A_{i,\varepsilon}.$$

Clearly, $|A_{i,\varepsilon}| \leq c\mu_i([t_i, s + \varepsilon])$ $\forall i$. By passing to the limit,

$$-c\mu([s, s+\varepsilon]) \leq \liminf_{i \to \infty} [x_i(s+\varepsilon) - x_i(t_i)] \leq \limsup_{i \to \infty} [x_i(s+\varepsilon) - x_i(t_i)] \leq c\mu([s, s+\varepsilon]).$$

By Lemma 5.1, $x_i(s + \varepsilon) \to x(s + \varepsilon)$ as $i \to \infty$, and, therefore,

$$x(s+\varepsilon) - c\mu([s, s+\varepsilon]) \leq \liminf_{i \to \infty} x_i(t_i) \leq \limsup_{i \to \infty} x_i(t_i) \leq x(s+\varepsilon) + c\mu([s, s+\varepsilon]).$$

Bearing in mind that $c\mu([s, s+\varepsilon]) \to 0$ as $\varepsilon \to 0$ (indeed, $s \notin \mathrm{Ds}(\mu)$), we conclude the proof. □

To prove Lemma 5.4, we need the following propositions.

PROPOSITION 9.1. *Let $x \in V^n(T)$, $\operatorname{Var} \nu[x] \leq c\mu$. Then, if function $h(x,t)$ is Lipschitz continuous in (x,t), there exists a constant denoted* const *such that* $\operatorname{Var} \nu[h(x, \cdot)] \leq \mathrm{const}(\mu + \mathcal{L})$.

Proof. By definition of the variation

$$\operatorname{Var}|_a^b [h(x(t), t)] = \sup_{T_N} \sum_{k=1}^N |h(x(s_k), s_k) - h(x(s_{k-1}), s_{k-1})|,$$

where supremum is considered in every possible finite partitions of segment $[a,b]$. By using the Lipschitz continuity of h and condition $\operatorname{Var}\nu[x] \le c\mu$, we deduce that

$$\sum_{k=1}^{N} |h(x^1(s_k),\dots,x^n(s_k),s_k) - h(x^1(s_{k-1}),\dots,x^n(s_{k-1}),s_{k-1})|$$
$$\le \operatorname{const}\left(\sum_{j=1}^{n} \operatorname{Var}|_a^b[x^j] + a - b\right) \le \operatorname{const}[\mu([a,b]) + \mathcal{L}([a,b])].$$

The proof is complete. □

PROPOSITION 9.2. *If the solution to* (1.2) *exists, then it is unique.*

Proof. Suppose that there are functions $x, y \in V^n(T)$ such that $x \ne y$, and x, and y do satisfy (1.2). From Definition 3.1, we have

$$|x(t) - y(t)| \le \int_{t_0}^{t} |f(x,u,s) - f(y,u,s)|ds + \int_{[t_0,t]} |g(x,s) - g(y,s)|d\mu_c$$
$$+ |\Phi(x,t) - \Phi(y,t)| \quad \forall t \in T.$$

Here and from now on, $\Phi(x,t)$ denotes the last item in formula specifying $x(t)$ in Definition 3.1. By using Gronwall's inequality, we arrive at the estimate

$$\sup_{s\in[t_0,t]} |x(s) - y(s)| \le \int_{[t_0,t]} c|x(s) - y(s)|(d\mu_c + ds)$$
$$+ \sum_{s\in D(t)} c\mu(\{s\})e^{c\mu(\{s\})}|x(s^-) - y(s^-)|$$
$$\le c \sup_{s\in[t_0,t]} |x(s) - y(s)| \times (\mu_c([t_0,t]) + t - t_0) + c_1 \sum_{s\in D(t)} \mu(\{s\})A_s,$$

where $A_s = |x(s^-) - y(s^-)|$, $s \in D(t)$, $D(t) = \operatorname{Ds}(\mu) \cap [t_0,t]$, and $c_1 = ce^{c|\mu|}$. Let $\phi(t) = \sup_{s\in[t_0,t]} |x(s) - y(s)|$, and $\gamma(t) = c(\mu_c([t_0,t]) + t - t_0)$. Let us enumerate the atoms of the measure μ in $[t_0,t]$ by the decreasing order of its values and denote by $\{a_j\}$ such a sequence, $a_j = \mu(\{s_j\})$, $s_j \in D(t)$, $j = 1,2,\dots$. In this constructive process, it may happen that the measure μ has no atoms or only a finite number of atoms s_j, $j = 1,\dots,r$, $r \ge 1$ on $[t_0,t]$. In the first case, we put $a_j = 0\ \forall j$, and in the second case, we let $a_j = 0\ \forall j > r$. By construction, $a_{j+1} \le a_j$, $a_j \to 0+$, as $j \to \infty$, and $\sum_{j=1}^{\infty} a_j = \mu_d([t_0,t])$. Now, we can write the inequality obtained above as

$$\phi(t) \le \phi(t)\gamma(t) + c_1 \sum_{j=1}^{\infty} a_j A_{s_j}. \tag{9.3}$$

Now, let us estimate an upper bound of the series $\sum_{j=1}^{\infty} a_j A_{s_j}$. Fix $\varepsilon > 0$. Since, $\forall j$, $A_{s_j} \le \phi(t)$, there exists a number $N = N(\varepsilon)$ such that $\sum_{j=N+1}^{\infty} a_j A_{s_j} \le \varepsilon\phi(t)$. To evaluate the partial sum $S_N = \sum_{j=1}^{N} a_j A_{s_j}$, let us consider $s_j < s_{j+1}$, $j = 1,\dots,N-1$. (If this is not the case, then we may renumber the finite set of a_j and s_j, $j = 1,\dots,N-1$. It is not hard to see that this relabeling does not affect the proof.) Let $Q(t) = \phi(t)\gamma(t) + c_1\varepsilon\phi(t)$. From (9.3), it follows that $A_{s_1} \le Q(t)$, $A_{s_2} \le Q(t) + a_1 A_{s_1} \le Q(t)(1 + a_1)$, $A_{s_3} \le Q(t) + a_1 A_{s_1} + a_2 A_{s_2} \le Q(t)(1 + a_1)(1 + a_2), \dots$, and so on,

until we get

$$A_{s_N} \le Q(t) \prod_{j=1}^{N-1} (1 + a_j).$$

Thus $S_N \le Q(t) \sum_{j=1}^{N} a_j \prod_{i=1}^{j-1}(1+a_i)$. Since the series $\sum_{j=1}^{\infty} \ln(1+a_j) \le \sum_{j=1}^{\infty} a_j \le |\mu|$ converges, the infinite product $\prod_{j=1}^{\infty}(1 + a_j)$ also converges. Then, $S_N \le Q(t)$ $\sum_{j=1}^{N} a_j e^{|\mu|} \le Q(t)|\mu|e^{|\mu|}$. Now, from (9.3), we obtain

$$\phi(t) \le \phi(t)\gamma(t) + c_1[\varepsilon\phi(t) + Q(t)|\mu|e^{|\mu|}],$$

and hence

$$\phi(t) \le \phi(t)[\gamma(t) + \varepsilon c_1][1 + c_1|\mu|e^{|\mu|}].$$

By letting $\varepsilon \to 0$, we finally arrive at

$$\phi(t) \le c[1 + c|\mu|e^{c(1+|\mu|)}](\mu_c([t_0, t]) + t - t_0)\phi(t) \quad \forall t \in T. \tag{9.4}$$

The uniqueness follows from the continuity of the Borel measure μ_c. Indeed, there exists τ_1 such that $c[1+c|\mu|e^{c(1+|\mu|)}](\mu_c([t_0, \tau_1]) + \tau_1 - t_0) < 1$. Hence $\phi(\tau_1) = 0$, and, on the segment $[t_0, \tau_1]$, the solutions $x(t)$ and $y(t)$ are congruent. All the conditions of the proposition are satisfied on the segment $[\tau_1, t_1]$. By applying the described procedure to $[\tau_1, t_1]$, we find τ_2 such that $x(t) = y(t)$ $\forall t \in [\tau_1, \tau_2]$, and so on. It is obvious that we may cover the whole segment T in a finite number of steps. Therefore, $x(t) \equiv y(t)$ and the proof is complete. □

PROPOSITION 9.3. *Consider a sequence of absolutely continuous measures* $\{\mu_i\}$, $\mu_i \in C_+^*(T)$, $i = 1, 2, \ldots$, *converging weakly star to* μ, *a sequence of controls* $\{u_i\}$, $u_i \xrightarrow{w} u$ *weakly in* $L_2(T)$, $u_i(t) \in U(t)$ *a.e., a sequence of vectors* $\{x_{0,i}\}$, $x_{0,i} \to x_0 \in \mathbf{R}^n$, *and the corresponding sequence of vector functions* $\{x_i\}$, $x_i \in V^n(T)$, *satisfying*

$$x_i(t) = x_{0,i} + \int_{t_0}^{t} f(x_i, u_i, s)ds + \int_{[t_0,t]} g(x_i, s)d\mu_i, \quad t \in T. \tag{9.5}$$

Then, if $|x_i(t)| \le \text{const}$, $x_i(t) \to x(t)$ $\forall t \in \text{Cont}(\mu)$, *where* $x(t)$ *is the solution to* (1.2). *Furthermore,* $\max_{t\in T} |x_i(t)| \to \text{gc}\sup_{t\in T} |x(t)|$.

Proof. Since x_i are solutions to (9.5) and are uniformly bounded, it follows that $\text{Var}\, \nu[x_i] \le \text{const}(\mu_i + \mathcal{L})$. From Proposition 9.1, $\text{Var}\, \nu[g(x_i, t)] \le \text{const}(\mu_i + \mathcal{L})$ and from Helly's second theorem, we may extract a subsequence so that $x_i(t) \to \tilde{x}(t)$ $\forall t \in T$ for some $\tilde{x} \in V^n(T)$.

Let the function $x \in V^n(T)$ be such that $x(t) = \tilde{x}(t)$ $\forall t \in \text{Cont}(\mu)$. The function $x(t)$ exists since $\text{Var}\,|_a^b[\tilde{x}] \le c(\mu+\mathcal{L})([a, b])$ $\forall a \le b$. Now, let us show that $x(t)$ satisfies (1.2).

Without any loss of generality, we may consider that $\frac{dF(t;\mu_i)}{dt} > 0$ a.e. on T, $i = 1, 2, \ldots$.[14]

[14]Otherwise, we can consider the sequence $\tilde{\mu}_i = \mu_i + i^{-1}\mathcal{L}$, and, from (9.5), we obtain

$$x_i(t) = x_{0,i} + \int_{t_0}^{t} f(x_i, u_i, s)ds + \int_{t_0}^{t} g(x_i, s)d\tilde{\mu}_i - p_i(t),$$

where $p_i(t) := \frac{1}{i}\int_{t_0}^{t} g(x_i, s)ds$ clearly converges to 0 uniformly on T, thus preserving all the required reasoning.

We start by showing that there exist sequences of absolutely continuous measures $\{\bar\mu_k\}$ and of numbers $k_i \geq i$ such that $\bar\mu_k \xrightarrow{w} \mu_d$ and $\mu_{i_k} - \bar\mu_k \xrightarrow{w} \mu_c$ on T.

If $\mathrm{Ds}(\mu) = \emptyset$, we put $\bar\mu_i = 0$, $k_i = i$, $\forall i$.

Let $\mathrm{Ds}(\mu) \neq \emptyset$. Consider the chain of sets D_i such that $D_{i-1} \subseteq D_i \subseteq \mathrm{Ds}(\mu)$ and $\sum_{r\in \mathrm{Ds}(\mu)\setminus D_i} \mu(\{r\}) \leq i^{-1}$, being $D_0 = \emptyset$. Define the sets $S_{r,i} = [r-\rho_i, r+\rho_i]$, $r \in D_i$ as a system of closed pairwise disjoint neighborhoods of points r, such that

(1) $\rho_i > 0$, $\rho_i \to 0$ as $i \to \infty$,
(2) $|\mu(S_i) - \sum_{r\in D_i} \mu(\{r\})| \leq \frac{1}{i}$, where $S_i = \bigcup_{r\in D_i} S_{r,i}$, and
(3) $r \pm \rho_i \in \mathrm{Cont}(\mu)$.

The existence of such sets S_i follows from the regularity of μ. Take $k_i \geq i$ such that $\sum_{r\in D_i} |\mu_{k_i}(S_{r,i}) - \mu(S_{r,i})| \leq \frac{1}{i}$. This is possible because $\mu_i \xrightarrow{w} \mu$. Let $\bar\mu_k(B) = \mu_{k_i}(B \cap S_i)$ for any Borel set $B \subset \mathbf{R}^1$. It is easy to verify that $\bar\mu_i \xrightarrow{w} \mu_d$. Hence, $\mu_{k_i} - \bar\mu_i \xrightarrow{w} \mu_c$. Note also that $\mu_{k_i} - \bar\mu_i \in C_+^*(T)$.

From now on we shall relabel the subsequence extracted (with the help of the sequence k_i) as the initial one (by index i) and rewrite (9.5) as

$$x_i(t) = x_{0,i} + \int_{t_0}^t f(x_i,u_i,s)ds + \int_{[t_0,t]} g(x_i,s)(d\mu_i - d\bar\mu_i) + \int_{[t_0,t]} g(x_i,s)d\bar\mu_i,\ t\in T, \quad \forall i.$$

Let $x_i^c(t) = x_{0,i} + \int_{t_0}^t f(x_i,u_i,s)ds + \int_{[t_0,t]} g(x_i,s)(d\mu_i - d\bar\mu_i)$, and $x_i^d(t) = \int_{[t_0,t]} g(x_i,s)d\bar\mu_i$, $t \in T$.

Thus, $x_i(t) = x_i^c(t) + x_i^d(t)$ $\forall t \in T$. By the Lebesgue theorem, the weak-L_2 convergence of controls u_i, formula (1.6), the weak star convergence of μ_i, and Lemma 5.1, $x_i^c(t) \to x_c(t) = x_0 + \int_{t_0}^t f(x,u,s)ds + \int_{[t_0,t]} g(x,s)d\mu_c$ $\forall t \in T$, $i \to \infty$. Then, $x_i^d(t) \to x(t) - x_c(t) = x_d(t)$ $\forall t \in \mathrm{Cont}(\mu)$, being $x_d \in V^n(T)$.

Next, we show that $x_d = \Phi(x,t)$ with Φ defined as in Proposition 9.2 above. Let $t \in \mathrm{Cont}(\mu)$. Fix $\varepsilon > 0$, and take a number $N = N(\varepsilon)$ such that $\sum_{r\in \mathrm{Ds}(\mu)\setminus D_N} \mu(\{r\}) \leq \varepsilon$.

Then,

$$\limsup_{i\to\infty} \left| \sum_{r\in D(N,t)} \int_{S_{r,i}} g(x_i,s)m_i(s)ds - x_i^d(t) \right| \leq \mathrm{const}\,\varepsilon. \tag{9.6}$$

Here and from now on, $D(N,t) = \{r \in D_N : r \leq t\}$, and $m_i(s)$ is the density of μ_i.

For $r \in D(N,t)$ consider, on $S_{r,i}$, the system

$$x_i^d(s) = x_i^d(r-\rho_i) + \int_{r-\rho_i}^s g(x_i,\sigma)m_i(\sigma)d\sigma, \quad s \in S_{r,i}. \tag{9.7}$$

For all sufficiently large i, define the functions

$$\pi_{r,i}(\sigma) = \frac{F(\sigma;\mu_i) - F(r-\rho_i;\mu_i)}{\mu_i(S_{r,i})}, \quad r \in D(N,t).$$

The function $\pi_{r,i} : S_{r,i} \to [0,1]$, is absolutely continuous and strictly increasing. Moreover, $\frac{d\pi_{r,i}}{d\sigma} = \frac{m_i(\sigma)}{\mu_i(S_{r,i})} > 0$ a.e. Hence, there exists an inverse map $\theta_{r,i} : [0,1] \to S_{r,i}$, $\theta_{r,i} = (\pi_{r,i})^{-1}$. Let us perform the change of variables $\tau = \pi_{r,i}(\sigma)$ in (9.7). We have, for $s \in [0,1]$, and $r \in D(N,t)$,

$$\alpha_{r,i}(s) = x_i^d(r-\rho_i) + x_i^c(\theta_{r,i}(s)) + \mu_i(S_{r,i}) \int_0^s g(\alpha_{r,i}(\tau),\theta_{r,i}(\tau))d\tau,$$

where $\alpha_{r,i}(\tau) = x_i(\theta_{r,i}(\tau))$. It is clear that $\theta_{r,i}(\tau) \to r$ uniformly on $[0,1]$, and, from Lemma 5.3, $x_i^c(\theta_{r,i}(\tau)) \to x_c(r)$ uniformly on $[0,1]$, as $i \to \infty$. It is straightforward to show that $\alpha_{r,i}$ is in $C([0,1])$.

Denote its limit as in i by α_r. By passing to the limit in the last formula as $i \to \infty$, we deduce

$$\alpha_r(s) = x(r^-) + \mu(\{r\}) \int_0^s g(\alpha_r, r)d\tau, \ s \in [0,1], \ r \in D(N,t).$$

(Here we use the fact that by construction, $x_i^d(r-\rho_i) \to x_d(r^-)$.) From this and from (9.6), we have

$$\left| x_d(t) - \sum_{r \in D(N,t)} [\alpha_r(1) - x(r^-)] \right| \le \text{const}\, \varepsilon.$$

By taking the limit as $\varepsilon \to 0$, we obtain, from the definition of Φ, that $x_d(t) = \Phi(x(t), t)$. The statement of the proposition is proved for some subsequence. However, Proposition 9.2 allows its extension to the whole sequence $\{x_i\}$.

It remains to verify that $\max_{t\in T} |x_i(t)| \to \operatorname{gc\,sup}_{t\in T} |x(t)|$. Clearly, this follows from Lemma 5.3, the reasoning presented above, and a simple contradiction argument. The proof is complete. □

Proof of Lemma 5.4. Since the x_i are solutions to (1.2) and are uniformly bounded, it follows that $\operatorname{Var} \nu[x_i] \le \text{const}(\mu_i + \mathcal{L})$. From Proposition 9.1, we have that $\operatorname{Var} \nu[g(x_i,t)] \le \text{const}(\mu_i + \mathcal{L})$, and by applying Helly's second theorem, we may extract a subsequence so that $x_i(t) \to \tilde{x}(t)$ $\forall t \in T$. Let $x \in V^n(T)$ be a function such that $x(t) = \tilde{x}(t)$ $\forall t \in \text{Cont}(\mu)$. Such a function $x(t)$ exists since $\operatorname{Var}|_a^b[\tilde{x}] \le c(\mu + \mathcal{L})([a,b])$, $\forall a \le b$. Let us show that $x(t)$ satisfies (1.2).

For each number i, let us construct a sequence of absolutely continuous measures $\mu_{i,j} \in C_+^*(T)$ such that $\mu_{i,j} \xrightarrow{w} \mu_i$ on T, as $j \to \infty$. Let $\{\phi_k : k \in \mathbf{N}\}$ be a countable everywhere dense subset of $C(T)$ set of functions, and let $\{t_k : k \in \mathbf{N}\}$ be countable everywhere dense subset of points of T set such that $t_k \in \text{Cont}(\mu) \cap [\bigcap_{i=1}^\infty \text{Cont}(\mu_i)]$.

By using Proposition 9.3, we find, for each i, a number $j = j(i)$ so that, for $k = 1, \dots, i$,

$$\max\left\{ \left| \int_T \phi_k d\mu_{i,j} - \int_T \phi_k d\mu_i \right|, |x_i(t_k) - x_{i,j}(t_k)| \right\} \le \frac{1}{i},$$

$$\left| \max_{t\in T} |x_{i,j}(t)| - \operatorname{gc\,sup}_{t\in T} |x_i(t)| \right| \le \frac{1}{i},$$

where $x_{i,j}$ is the solution to the equation

$$x_{i,j}(t) = x_{0,i} + \int_{t_0}^t f(x_{i,j}, u_i, s)ds + \int_{[t_0,t]} g(x_{i,j}, s)d\mu_{i,j}, \quad t \in T. \tag{9.8}$$

Let us show that this solution exists for j sufficiently large. Assume $c = \operatorname{gc\,sup}_{t\in T} |x_i(t)| + 1$. Consider functions

$$\tilde{f}(x,u,t) = \begin{cases} f(x,u,t), & |x| \le c, \\ f(p_c(x),u,t), & |x| > c, \end{cases} \qquad \tilde{g}(x,t) = \begin{cases} g(x,t), & |x| \le c, \\ g(p_c(x),t), & |x| > c, \end{cases}$$

where $p_c(x)$ is the projection of vector x on the sphere $|x| = c$. Note that $\tilde{f}$ and $\tilde{g}$ are Lipschitz continuous in x uniformly in (x,u,t). This is why, for each j, there exists

a solution $x_{i,j}$ to (9.8), with functions $\tilde{f}$ and $\tilde{g}$ replacing f and g, respectively. This is so due to a theorem of existence of solution for convenient differential equations. From (9.4) it is not hard to deduce that the $x_{i,j}$ are uniformly bounded in j. By using Propositions 9.2 and 9.3 for large j we conclude that $\max_{t\in T}|x_{i,j}(t)| \le c$, and hence the solution to (9.8) exists.

Let $\mu_i^a = \mu_{i,j}$, and $x_i^a = x_{i,j}$. By construction $\mu_i^a \xrightarrow{w} \mu$. By Proposition 9.3, there exists a solution $\bar{x}$ of (1.2) on T and $x_i^a(t) \to \bar{x}(t)$ $\forall t \in \mathrm{Cont}(\mu)$. But, on the other hand, $\forall k$, $x_i^a(t_k) \to x(t_k)$, as $i \to \infty$. Then, from inequality $\mathrm{Var}\,\nu[x_i^a] \le \mathrm{const}(\mu_i^a + \mathcal{L})$, it follows that $x_i^a(t) \to x(t)$ $\forall t \in \mathrm{Cont}(\mu)$. Hence, $x(t)$ is a solution to (1.2).

The lemma is proved for some subsequence. However, by using Proposition 9.2, we extend it to the whole sequence $\{x_i\}$. The proof is complete. □

Proof of Lemma 5.5. Fix any $\varepsilon > 0$ and consider the set $S(W_0,\varepsilon)$, the ε-neighborhood of W_0. Since f is Lipschitz continuous, $|f(t)| \le \mathrm{const}\,\varepsilon$ $\forall t \in S(W_0,\varepsilon)$. From the weak convergence $\eta_i(T \setminus S(W_0,\varepsilon)) \to 0$ as $i \to \infty$. From uniform convergence, we conclude that $\|f_i - f\|_C \le \varepsilon$ for i sufficiently large. Then, for any $a < b$,

$$\left|\int_{[a,b)} f_i(s)d\eta_i\right| \le c[\varepsilon\eta_i([a,b)) + \eta_i([a,b) \cap [T \setminus S(W_0,\varepsilon)])].$$

By definition of variation of a function $\limsup_{i\to\infty} \mathrm{Var}\,|_\tau[x_i] \le \mathrm{const}\,\varepsilon$. But $\varepsilon > 0$ is arbitrary, and hence, $\mathrm{Var}\,|_\tau[x_i] \to 0$. The proof is complete. □

Proof of Proposition 5.6. Let (p,u,μ) be an admissible process in problem (P). Associated with it, let us construct an admissible process $(\tilde{p},\tilde{u},\tilde{\mu},\upsilon_0,\upsilon_1)$ for problem (P_1). Let $\tilde{u} = u$, $\tilde{\mu} = \mu - \sum_{k=0,1}\mu(\{t_k\})\delta_{t_k}$,[15] $\tilde{x}_0 = x(t_0^+)$, $\tilde{x}_1 = x(t_1^-)$, $\upsilon_k = \mu(\{t_k\})(t_1-t_0)^{-1}\mathcal{L}$, $\tilde{t}_k = t_k$, $k = 0,1$. Clearly, $(\tilde{p},\tilde{u},\tilde{\mu},\upsilon_0,\upsilon_1)$ is an admissible process for (P_1), and $e_0(p) = e_0(\tilde{p})$.

Conversely, let $(\tilde{p},\tilde{u},\tilde{\mu},\upsilon_0,\upsilon_1)$ be an admissible process in (P_1) and take $u = \tilde{u}$, $\mu = \tilde{\mu} + \sum_{k=0,1}|\upsilon_k|\delta(t_k)$, $x_k = \xi_k = \xi(\tilde{x}_k,t_k,(-1)^{k+1}|\upsilon_k|)$, $t_k = \tilde{t}_k$, $k = 0,1$. Then, the process (p,u,μ) is admissible for (P), and $e_0(p) = e_0(\tilde{p})$. The proof is complete. □

Proof of Proposition 5.7. Let $(\tilde{x}_0,\tilde{x}_1,\chi_0,\chi_1,\tilde{t}_0,\tilde{t}_1,\tilde{u},v,\tilde{\mu})$ be an admissible process for (P_2). Since $\frac{3}{2} \ge \dot{\chi} \ge \frac{1}{2} > 0$ a.e., χ strictly increases on $[\tilde{t}_0,\tilde{t}_1]$. Hence, there exists inverse mapping $\pi(t) = \chi^{-1}(t)$, $t \in [\chi(\tilde{t}_0),\chi(\tilde{t}_1)]$, which is also strictly increasing and absolutely continuous. Let $t_0 = \chi(\tilde{t}_0)$, $t_1 = \chi(\tilde{t}_1)$, $x_0 = \tilde{x}_0$, $u(t) = \tilde{u}(\pi(t))$, and $F(t;\mu) = F(\pi(t);\tilde{\mu})$. We show that $x(t) = \tilde{x}(\pi(t))$. In fact, by changing variable $\tau = \pi(s)$, deduce

$$\begin{aligned}\tilde{x}(\pi(t)) &= \tilde{x}_0 + \int_{\tilde{t}_0}^{\pi(t)}(v+1)f(\tilde{x},\tilde{u},\chi)d\tau + \int_{[\tilde{t}_0,\pi(t)]} g(\tilde{x},\chi)d\tilde{\mu}\\ &= x_0 + \int_{t_0}^{t} f(x,u,s)ds + \int_{[t_0,t]} g(x,s)d\mu\\ &= x(t), \quad t \in [t_0,t_1].\end{aligned}$$

From here, it follows that (p,u,μ) is an admissible process for (P), and $e_0(p) = e_0(\tilde{p})$.

Conversely, let (p,u,μ) be an admissible process for (P) and consider $v = 0$. Then, $(p,t_0,t_1,u,0,\mu)$ is admissible for (P_2). The proof is complete. □

[15] δ_r indicates Dirac's measure in r.

REFERENCES

[1] A.V. Arutyunov, *Optimality Conditions: Abnormal and Degenerate Problems*, Math. Appl. 526, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[2] A. Bressan and F. Rampazzo, *Impulsive control systems with commutative vector fields*, J. Optim. Theory Appl., 71 (1991), pp. 67–83.

[3] A.Ya. Dubovitskii and V.A. Dubovitskii, *Necessary conditions of strong minimum in optimal control problem with degenerate endpoint and state constraints*, Uspekhi Mat. Nauk, 40 (1985), pp. 175–176 (in Russian).

[4] V.A. Dykhta, *Necessary optimality conditions for impulsive processes under constraints on the image of the control measure*, Izv. Vuz. Mat., 12 (1996), pp. 1–9.

[5] V.A. Dyhta and O.N. Samsonyuk, *Impulse Control Optimization and Applications*, Physmatlit, Moscow, 2000 (in Russian).

[6] M.M. Ferreira and R.B. Vinter, *When is the maximum principle for state constrained problems nondegenerate?*, J. Math. Anal. Appl., 187 (1994), pp. 438–467.

[7] G. Kolokolnikova, *A variational maximum principle for discontinuous trajectories of unbounded asymptotically linear control systems*, J. Differential Equations, 33 (1997), pp. 1633–1640.

[8] B. Miller, *Sampled-data control processes described by ordinary differential equations* I, Automat. Remote Control, 39 (1978), pp. 57–67.

[9] B. Miller, *Sampled-data processes described by ordinary differential equations* II, Automat. Remote Control, 39 (1978), pp. 338–344.

[10] B.M. Miller, *The optimality conditions in a problem of control of a system that can be described with a differential equation with measure*, Autom. Telemekh., 6 (1982), pp. 752–761.

[11] B.M. Miller, *Optimality conditions in generalized control problems* I, Automat. Remote Control, 53 (1992), pp. 362–370.

[12] B.M. Miller, *Optimality conditions in generalized control problems* II, Automat. Remote Control, 53 (1992), pp. 505–513.

[13] B.M. Miller, *Method of discontinuous time change in problems of control for impulse and discrete continuous system*, Automat. Remote Control, 54 (1993), pp. 1727–1750.

[14] B.M. Miller and E.Y. Rubinovitch, *Impulsive Control in Continuous and Discrete-Continuous Systems*, Kluwer Academic Publishers, Amsterdam, 2002.

[15] Y. Orlov, *Theory of Optimal Systems with Generalized Controls*, Nauka, Moscow, 1988 (in Russian).

[16] F.M.F.L. Pereira and G.N. Silva, *Necessary conditions of optimality for vector-valued impulsive control problems*, Systems Control Lett., 40 (2000), pp. 205–215.

[17] F.M.F.L. Pereira, *A maximum principle for impulsive control problems with state constraints*, Comput. Appl. Math., 19 (2000), pp. 1–19.

[18] A.V. Sarychev, *Optimization of generalized controls in a nonlinear time optimal problem*, Differential Equations, 27 (1991), pp. 539–550.

[19] G.N. Silva and R.B. Vinter, *Measure differential inclusions*, J. Math. Anal. Appl., 202 (1996), pp. 727–746.

[20] G.N. Silva and R.B. Vinter, *Necessary conditions for optimal impulsive control problems*, SIAM J. Control Optim., 35 (1997), pp. 1829–1846.

[21] R.B. Vinter and F.M.F.L. Pereira, *A maximum principle for optimal processes with discontinuous trajectories*, SIAM J. Control Optim., 26 (1988), pp. 205–229.

[22] S. Zavalischin and A. Sesekin, *Impulsive Processes: Models and Applications*, Nauka, Moscow, 1991.

[23] S.T. Zavalischin and A.N. Sesekin, *Dynamic Impulse Systems: Theory and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1844–1866

QUOTIENTS OF FULLY NONLINEAR CONTROL SYSTEMS*

PAULO TABUADA† AND GEORGE J. PAPPAS‡

Abstract. In this paper, we introduce and study quotients of fully nonlinear control systems. Our definition is inspired by categorical definitions of quotients as well as recent work on abstractions of affine control systems. We show that quotients exist under mild regularity assumptions and characterize the structure of the quotient state/input space. This allows one to understand how states and inputs of the quotient system are related to states and inputs of the original system. We also introduce a notion of projectability which turns out to be equivalent to controlled invariance. This allows one to regard previous work on symmetries, partial symmetries, and controlled invariance as leading to special types of quotients. We also show the existence of quotients that are not induced by symmetries or controlled invariance. Such decompositions have a potential use in a theory of hierarchical control based on quotients.

Key words. quotient control systems, control systems category, controlled invariance, symmetries

AMS subject classifications. 93A10, 93A30, 93B11, 93C10

DOI. 10.1137/S0363012901399027

1. Introduction.

The analysis and synthesis problems for nonlinear control systems are often very difficult due to the size and complicated nature of the equations describing the processes to be controlled. It is therefore desirable to have a methodology that decomposes control systems into smaller subsystems while preserving the properties relevant for analysis or synthesis. From a theoretical point of view, the problem of decomposing control systems is also extremely interesting since it reveals system structure that must be understood and exploited.

In this paper we will focus on the study of quotient control systems since they can be seen as lower dimensional models that may still carry enough information about the original system. We will build on several accumulated results of different authors that in one way or another have made contributions to this problem. One of the first approaches was given in [17] were the analysis of the Lie algebra of a control system lead to a decomposition into smaller systems. At the same time in [35], quotients of control systems induced by observability equivalence relations where introduced in the more general context of realization theory. In [31], Lie algebraic conditions are formulated for the parallel and cascade decomposition of nonlinear control systems, while the feedback version of the same problem was addressed in [24]. A different approach was based on reduction of mechanical systems by symmetries. In [39], symmetries were introduced for mechanical control systems and further developed in [9] for general control systems. The existence of such symmetries was then used to decompose control systems as the interconnection of lower dimensionality subsystems. The notion of symmetry was further generalized in [26], where it was shown that the existence of symmetries implies that a certain distribution associated with the

*Received by the editors December 1, 2001; accepted for publication (in revised form) May 19, 2004; published electronically March 22, 2005. This research was supported by Fundação para a Ciência e Tecnologia under grant PRAXIS XXI/BD/18149/98.
http://www.siam.org/journals/sicon/43-5/39902.html

†Department of Electrical Engineering, 268 Fitzpatrick Hall, University of Notre Dame, Notre Dame, IN 46556 (ptabuada@nd.edu).

‡Department of Electrical and Systems Engineering, 200 South 33rd Street, University of Pennsylvania, Philadelphia, PA 19104 (pappasg@seas.upenn.edu).

symmetries was controlled invariant. This related the notion of symmetry with the notion of controlled invariance for nonlinear systems. Controlled invariance [23, 12] was also used to decompose systems into smaller components. A different approach was taken in [22] where it was shown how to study controllability of systems evolving on principle fiber bundles through their projection on the base space. More recently, a modular approach to the modeling of mechanical systems has been proposed in [40], by studying how the interconnection of Hamiltonian control systems can still be regarded as a Hamiltonian control system. A different research direction was taken in [29], where instead of using structural properties of control systems, a constructive procedure was proposed to compute smaller control systems called abstractions.

In several of the above approaches, some notion of quotienting is involved. When symmetries exist, one of the blocks of the decompositions introduced in [9] is simply the original control system factored by the action of a Lie group representing the symmetry. If a control system admits a controlled invariant distribution, it is shown in [23, 12] that it has a simpler local representation. This simpler representation can be obtained by factoring the original control system by the equivalence relation defined by considering the leaves of the foliation induced by the controlled invariant distribution, equivalence classes. The notion of abstraction introduced in [29] can also be seen as a quotient since the abstraction is a control system on a smaller dimensional state space defined by an equivalence relation on the state space of the original control system. These facts motivate fundamental questions such as existence and characterization of quotient systems. Existence questions have already been addressed in [35] but in a different setting. Only specific equivalence relations were considered (those induced by indistinguishability), and the input space remained unaltered by the factorization process. Furthermore, the quotients discussed in [35] are of a particular nature being characterized by the notion of projectability introduced in section 6.

A thorough understanding of quotient systems also has important consequences for hierarchical control, since the construction of quotients proposed in [29] implicitly indicates that certain states of the original system may become inputs on the quotient control system. It is perhaps surprising that this methodology interchanges the role of state and input. However, this fact is the crucial factor that allows the development of a hierarchical control theory based on quotients. Since states of the original system may become inputs of the quotient system, a control design performed on a quotient system can serve as a design specification for the original system. We can therefore regard a control design as a *specification* for the evolution of certain state variables on the more detailed model. A complete and thorough understanding of how the states and inputs propagate from control systems to their quotients will enable such a hierarchical design scheme. Preliminary work exploiting such a hierarchical approach has been reported in [37].

In this paper, we take a new approach to the study of quotients by introducing the category of control systems as the natural setting for such problems in systems theory. The use of category theory for the study of problems in system theory also has a long history which can be traced back to the works of Arbib and Manes (see [2] for an introduction). More recently, several authors have also adopted a categorical approach as in [19], where the category of affine control systems is investigated. We also mention [33], where a categorical approach has been used to provide a general theory of systems.

We define the category of control systems whose objects are fully (nonaffine) nonlinear control systems and morphisms map trajectories between objects. The

morphisms in this category extend the notion of ϕ-related systems from [28]. In this categorical setting we formulate the notion of quotient control systems and show, in one of the main results, that

under some regularity (constant rank) assumptions quotient control systems always exist.

This result implies that, given a nonlinear projection map from the state space to some reduced state space, we can always construct a new control system on the reduced state space with the property that the nonlinear projection map carries trajectories of the original system into trajectories of the reduced system. This should be contrasted with several other approaches which rely on the existence of symmetries or controlled invariance to assert the existence of quotients. We also introduce the notion of projectable control sections, which will be a fundamental ingredient to characterize the structure of quotients. This notion is in fact equivalent to controlled invariance, and this allows one to regard quotients based on symmetries or controlled invariance as a special type of quotients. General quotients, however, are not necessarily induced by symmetries or controlled invariance and have the property that some of their inputs are related to states of the original model. This fact, implicit in [29], is explicitly characterized in this paper by understanding how the state and input space of the quotient is related to the state and input space of the original control system. In particular, this paper's main contribution states that

in the absence of symmetries, states that are factored out in the quotient construction can be regarded as inputs of the quotient control system.

This result clearly distinguishes general quotients from previously studied quotients based on symmetries or partial symmetries in which inputs of the quotient system are the inputs (or a quotient) of the original system inputs. Since existence of symmetries can be regarded as rare phenomena,[1] as shown in [32] for single-input systems, construction of quotients enables a widely applicable hierarchical approach to control design based on reconstruction of trajectories for the original system from quotient trajectories [37].

The outline of the paper is as follows. We start by introducing the relevant notions from differential geometry and control theory in section 2. We then review the notion of ϕ-related control systems in section 3 which was originally introduced in [28] and will motivate the definition of the category of control systems presented in section 4. In section 5, we introduce the notion of quotient control systems and prove an existence and uniqueness result regarding quotients which roughly asserts that given a regular equivalence relation on the state space of a control system a quotient system exists (under some regularity conditions) and is unique up to isomorphism. The characterization of quotients will be the goal of the remaining sections of the paper. We first introduce the notion of projectable control section in section 6 and prove the main result of the paper characterizing the structure of the quotient state/input space in section 7. We end with conclusions and some open questions for further research in section 8.

2. Control systems. In this section we introduce all the relevant notions from differential geometry and control systems necessary for the remainder of the paper. The reader may wish to consult numerous books on these subjects, such as [1] for differential geometry and [14, 27] for control theory.

[1] We thank one of the anonymous reviewers for bringing this fact to our attention.

2.1. Differential geometry. We will consider that all the manifolds will be C^∞ and that all the maps will be smooth. Let M be a manifold and T_xM its tangent space at $x \in M$. The tangent bundle of M is denoted by $TM = \cup_{x\in M} T_xM$, and π_M is the canonical projection map $\pi_M : TM \longrightarrow M$ taking a tangent vector $X(x) \in T_xM \subset TM$ to the base point $x \in M$. Now, letting M and N be manifolds and $\phi : M \longrightarrow N$ a map, we denote by $T_x\phi : T_xM \longrightarrow T_{\phi(x)}N$ the induced tangent map which maps tangent vectors X at T_xM to tangent vectors $T_x\phi \cdot X$ at $T_{\phi(x)}N$. If ϕ is such that $T_x\phi$ is surjective at $x \in M$, then we say that ϕ is a submersion at x. When ϕ is a submersion at every $x \in M$ we simply say that it is a submersion. Similarly, we say that ϕ has constant rank if the rank of the pointwise linear map $T_x\phi$ is constant for every $x \in M$. When ϕ has an inverse which is also smooth we call ϕ a diffeomorphism. We say that a manifold M is diffeomorphic to a manifold N, denoted by $M \cong N$, when there is a diffeomorphism between M and N. When this is the case we can use $\phi^{-1} : N \longrightarrow M$ to define a vector field on M from a vector field $Y \in TN$, denoted by $\phi^*Y = (\phi^{-1})_*Y$ and defined by $T_{\phi(x)}\phi^{-1} \cdot Y(\phi(x))$.

A fibered manifold is a manifold B equipped with a surjective submersion $\pi_B : B \longrightarrow M$. Manifolds B and M are called the total space and the base space, respectively. The surjection π_B defines a submanifold $\pi_B^{-1}(x) = \{b \in B : \pi_B(b) = x\} \subseteq B$ for every $x \in M$ called the fiber at $x \in M$. We will usually denote a fibered manifold simply by $\pi_B : B \longrightarrow M$. Since a surjective submersion is locally the canonical projection from $\mathbb{R}^i$ to $\mathbb{R}^j$, $i = \dim(B)$ and $j = \dim(M)$, we can always find local coordinates (x, y), where x are coordinates for the base space and y are coordinates for the fibers over the base space. We shall call these coordinates adapted coordinates.

A map $\varphi : B_1 \longrightarrow B_2$ between two fibered manifolds is fiber preserving iff there exists a map $\phi : M_1 \longrightarrow M_2$ between the base spaces such that the following diagram commutes:

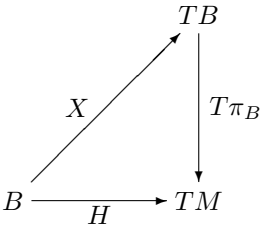
$$\begin{array}{ccc} B_1 & \xrightarrow{\varphi} & B_2 \\ {\scriptstyle \pi_{B_1}}\downarrow & & \downarrow{\scriptstyle \pi_{B_2}} \\ M_1 & \xrightarrow[\phi]{} & M_2 \end{array} , \tag{2.1}$$

that is to say, iff $\pi_{B_2} \circ \varphi = \phi \circ \pi_{B_1}$. In such a case we also refer to φ as a fiber preserving lift of ϕ. Given fibered manifolds B_1 and B_2, we will say that B_1 is a fibered submanifold of B_2 if the inclusion map $i : B_1 \hookrightarrow B_2$ is fiber preserving.

Given a map $h : M \longrightarrow N$ defined on the base space of a fibered manifold, its extension to the total space B is given by $\pi_B^* h = h \circ \pi_B$. We now consider the extension of a map $H : B \longrightarrow TM$ to a vector field in B. We will define local and global extensions of H. Globally, we define H^e as the set of all vector fields[2] $X : B$

[2]Global existence of such vector fields X follows from the existence of a horizontal space $\mathcal{H} \subseteq TB$, $\mathcal{H} \cong TM$ that allows the decomposition of TB as $TB = \mathcal{H} \oplus \ker(T\pi_B)$. A global extension of a map $H : B \longrightarrow TM$ to a vector field $X : B \longrightarrow TB$ is now uniquely defined as the vector field $X = H : B \longrightarrow TM \cong \mathcal{H} \subseteq TB$. Such horizontal space can be obtained, for example, as the orthogonal complement to $\ker(T\pi_B)$ given by a Riemannian metric on B.

$\longrightarrow TB$ such that the following diagram commutes:

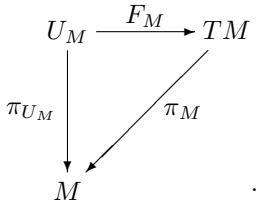
(2.2)


that is, $T\pi_B(X) = H$. When working locally, one can be more specific and select a distinguished element of H^e, denoted by H^l, which satisfies in adapted local coordinates (x, y), $H^l = H\frac{\partial}{\partial x} + 0\frac{\partial}{\partial y}$. A vector field $Y : M \longrightarrow TM$ on the base space M of a fibered manifold can also be extended to a vector field on the total space. It suffices to compose Y with the projection $\pi_B : B \longrightarrow M$ and recover the previous situation since $Y \circ \pi_B$ is a map from B to TM.

2.2. Control systems. Since the early days of control theory it was clear that in order to give a global definition of control systems the notion of input could not be decoupled from the notion of state [4, 41]. Although the coupling between states and inputs is usually modeled through the use of fiber bundles, we shall consider more general spaces.

In any case a control system can be globally defined as follows.

DEFINITION 2.1 (control system). *A control system $\Sigma_M = (U_M, F_M)$ consists of a fibered manifold $\pi_{U_M} : U_M \longrightarrow M$ called the control bundle and a map $F_M : U_M \longrightarrow TM$ making the following diagram commutative:*

(2.3)


That is, $\pi_M \circ F_M = \pi_{U_M}$, where $\pi_M : TM \longrightarrow M$ is the tangent bundle projection.

The input space U_M is modeled as a fibered manifold since in general the available control inputs may depend on the current state of the system. In adapted coordinates (x, v), Definition 2.1 reduces to the familiar expression $\dot{x} = f(x, v)$ with $v \in \pi^{-1}_{U_M}(x)$. The lack of local triviality assumptions on π_{U_M} is motivated by the need to model the construction of abstractions of control affine systems, as described in [29], in a fully nonlinear context. As the following example illustrates, even in simple situations the inputs of a control system resulting from an abstraction or quotient process can depend on the states in a way that cannot be modeled by a fiber bundle.

Consider control system $F_M : U_M \longrightarrow TM$ with $U_M = M \times U$, $M = \mathbb{R}^3$, $U =]0, 1[$ defined by:

$$F_M(x, y, z, u) = \left(x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} + z\frac{\partial}{\partial z} \right) u.$$

On the state space we define the following map $\phi : \mathbb{R}^3 \longrightarrow \mathbb{R}$ based on Reeb's foliation:

$$\phi(x, y, z) = (1 - r^2)e^z, \quad r = x^2 + y^2. \tag{2.4}$$

Computing the derivative of ϕ,

$$\mathrm{d}\phi = e^z(-4rx\,\mathrm{d}x - 4ry\,\mathrm{d}y + (1-r^2)\,\mathrm{d}z),$$

we see that ϕ is a submersion since $1-r^2=0$ for $r^2=1$, which implies that $x\neq 0$ or $y\neq 0$, and this in turn implies that $\mathrm{d}\phi\neq 0$. This shows that we can see $\phi:\mathbb{R}^3\longrightarrow\mathbb{R}$ as a fibered manifold. If we now compute the projection of F_M on $\mathbb{R}$ by ϕ, we obtain

$$\mathrm{d}\phi\cdot F_M = e^z(-4rx^2-4ry^2+(1-r^2)z)u.$$

The set of vectors defined by the previous expression can be seen as a control system on $\mathbb{R}$ up to control parameterization, as it defines the possible directions of motion achievable by control. This is the principle underlying the notion of abstraction described in [29]. Such a collection of vector fields admits the natural parameterization $\pi^{-1}_{U_M}(\phi^{-1}(w))$ for every $w\in\mathbb{R}$. However, such a set of inputs cannot be given the structure of a fiber bundle. To see this, it suffices to note that the fibers $\phi^{-1}(w)$ are not homeomorphic for $w>0$ and $w=0$. For $w>0$ we can solve $\phi(x,y,z)=w$ to obtain $z=\log\frac{w}{1-r^2}$ which defines $\phi^{-1}(w)$ as

$$\left\{(x,y,z)\in\mathbb{R}^3 \;:\; z=\log\frac{w}{1-r^2} \quad\wedge\quad 0\leq r^2<1\right\}$$

and which is homeomorphic to the open unit disk in $\mathbb{R}^2$. If $w=0$, solving $\phi(x,y,z)=0$, we obtain $r=1$ which is diffeomorphic to a cylinder. We thus see that for any open set O in $\mathbb{R}$ containing 0, $\pi_{U_M}(\phi^{-1}(0))$ cannot be diffeomorphic to $O\times L$ for some manifold L describing the typical fibers of $\phi\circ\pi_{U_M}$ as they are not diffeomorphic for different points in O. It is precisely the need to capture and analyze situations like this that forces one to consider models for the state/input space other than fiber bundles. The need to model these and other couplings between states and inputs has led to alternative approaches where the notion of control system and its properties are defined independently of states and inputs as in Willem's behavioral theory [30] and Fliess' differential algebraic approach [7].

We now return to our discussion of control systems by introducing the notion of control section[3] that is closely related to control systems and which will be fundamental in our study of quotients.

Definition 2.2 (control section). *Given a manifold M, a control section on M is a fibered submanifold $\pi_{\mathcal{S}_M}:\mathcal{S}_M\longrightarrow M$ of TM.*

We denote by $\mathcal{S}_M(x)$ the set of vectors $X\in T_xM$ such that $X\in\pi^{-1}_{\mathcal{S}_M}(x)$. We now see that under certain regularity assumptions, a control system (U_M,F_M) defines a control section by the pointwise assignment $\mathcal{S}_M(x)=F_M(\pi^{-1}_{U_M}(x))$. Conversely, a control section also defines a control system as we shall see in detail in section 4. The notion of a control section allows one to refer in a concise way to the set of all tangent vectors that belong to the image of F_M by saying that $X\in T_xM$ belongs to $\mathcal{S}_M(x)$ iff there exists a $u\in U_M$ such that $\pi_M(u)=x$ and $F_M(u)=X$. When $\mathcal{S}_M(x)$ defines an affine distribution on TM, we call control system F_M control affine and fully nonlinear otherwise.

Having defined control systems the concept of trajectories or solutions of a control system is naturally expressed by the following definition.

[3]In some literature this notion is also known as the field of admissible velocities.

DEFINITION 2.3 (trajectories of control systems). *A smooth curve $c : I \longrightarrow M$, $I \subseteq \mathbb{R}^+_0$, is called a trajectory of control system $\Sigma_M = (U_M, F_M)$ if there exists a (not necessarily smooth) curve $c^U : I \longrightarrow U_M$ making the following diagrams commutative:*

$$\begin{array}{ccc} & & U_M \\ & \overset{c^U}{\nearrow} & \downarrow \pi_{U_M} \\ I & \xrightarrow[c]{} & M \end{array} \qquad \begin{array}{ccc} & & U_M \\ & \overset{c^U}{\nearrow} & \downarrow F_M \\ I & \xrightarrow[Tc]{} & TM \end{array} \quad , \tag{2.5}$$

where we have identified I with TI.

The above commutative diagrams are equivalent to the following equalities:

$$\begin{aligned} \pi_{U_M} \circ c^U &= c, \\ Tc &= F_M(c^U), \end{aligned}$$

which mean in adapted coordinates that $x(t)$ is a trajectory of a control system if there exists an input $v(t)$ such that $x(t)$ satisfies $\dot{x}(t) = f(x(t), v(t))$ and $v(t) \in \pi^{-1}_{U_M}(x(t))$ for all $t \in I$.

3. ϕ-related control systems. We start by reviewing the notion of ϕ-related control systems originally introduced in [28] and which motivates the construction of the category of control systems to be later presented.

DEFINITION 3.1 (ϕ-related control systems). *Let Σ_M and Σ_N be two control systems defined on manifolds M and N, respectively. Given a map $\phi : M \longrightarrow N$ we say that Σ_N is ϕ-related to Σ_M iff for every $x \in M$,*

$$T_x\phi(\mathcal{S}_M(x)) \subseteq \mathcal{S}_N \circ \phi(x). \tag{3.1}$$

In [28] it is shown that this notion is equivalent to a more intuitive relation between Σ_M and Σ_N.

PROPOSITION 3.2 (see [28]). *Let Σ_M and Σ_N be two control systems defined on manifolds M and N, respectively, and let $\phi : M \longrightarrow N$ be a map. Control system Σ_N is ϕ-related to Σ_M iff for every trajectory $c(t)$ of Σ_M, $\phi(c(t))$ is a trajectory of Σ_N.*

Propagating trajectories from one system to another is clearly desirable. Since most control system properties are properties of its trajectories, relating trajectories of different control systems also allows one to relate the corresponding properties. If, in fact, system Σ_N is lower dimensional than system Σ_M, then we are clearly reducing the complexity of Σ_M. We can therefore regard Σ_N as an *abstraction* of Σ_M in the sense that some aspects of Σ_M have been collapsed or abstracted away, while others remain in Σ_N. This motivated the notion of abstraction based on trajectory propagation in [28], which defined an abstraction of a control system Σ_M as a ϕ-related control system Σ_N by a surjective submersion ϕ.

The idea of sending trajectories from one system to trajectories of another system has been used many times in control theory to study equivalence of control systems. We mention, for example, linearization by diffeomorphism [16] or feedback linearization [5, 10, 13]. In these examples the maps ϕ relating the control systems were in fact diffeomorphisms so that no aggregation or abstraction was involved. Related to the feedback linearization problem is the partial feedback linearization problem where only partial linearization is thought of. Such a problem can be reduced to the feedback linearization problem by considering feedback linearization of a subsystem of

the original control system [20]. The notion of a subsystem can also be described by defining how subsystem trajectories relate to the original system trajectories. In this case, we require the existence of a map (satisfying certain injectivity assumptions) transforming subsystem trajectories into trajectories of the original system. The use of trajectory propagating maps can already be traced back to the works of Arbib and Manesl (see [2] for an introduction), where by the use of category theoretic ideas it is shown that (discrete time) control systems and finite state automata are just different manifestations of the same phenomena.

4. The category of control systems. Informally speaking, a category is a collection of *objects* and *morphisms* between the objects that relate to the structure of the objects. If one is interested in understanding vector spaces, it is natural to consider vector spaces as objects and linear maps as morphisms, since they preserve the vector space structure. This choice for objects and morphisms defines **Vect**, the category of vector spaces. Choosing manifolds for objects leads to the natural choice of smooth maps for morphisms and defines **Man**, the category of smooth manifolds. In this section we introduce the category of control systems which we regard as the natural framework to study quotients of control systems. Besides providing an elegant language to describe the constructions to be presented, category theory also offers a conceptual methodology for the study of objects, control systems in this case. We refer the reader to [18] and [3] for further details on the elementary notions of category theory used throughout the paper.

The category of control systems, denoted by **Con**, has as objects control systems as described in Definition 2.1. The morphisms in this category extend the concept of ϕ-related control systems described by Definition 3.1. Since the notion of ϕ-related control systems relates control sections and these can be parameterized by controls, the lifted notion should relate control sections as well as its parameterizations by inputs.

DEFINITION 4.1 (morphisms of control systems). *Let Σ_M and Σ_N be two control systems defined on manifolds M and N, respectively. A morphism f from Σ_M to Σ_N is a pair of maps $f = (\phi, \varphi)$, $\phi : M \longrightarrow N$ and $\varphi : U_M \longrightarrow U_N$ making the following diagrams commutative:*

(4.1)
$$\begin{array}{ccc} U_M & \xrightarrow{\varphi} & U_N \\ {\scriptstyle \pi_{U_M}}\downarrow & & \downarrow{\scriptstyle \pi_{U_N}} \\ M & \xrightarrow[\phi]{} & N \end{array} \qquad \begin{array}{ccc} U_M & \xrightarrow{\varphi} & U_N \\ {\scriptstyle F_M}\downarrow & & \downarrow{\scriptstyle F_N} \\ TM & \xrightarrow[T\phi]{} & TN \end{array}$$

It will be important for later use to also define isomorphisms.

DEFINITION 4.2 (isomorphisms of control systems). *Let Σ_M and Σ_N be two control systems defined on manifolds M and N, respectively. System Σ_M is isomorphic to system Σ_N iff there exist morphisms f_1 from Σ_M to Σ_N and f_2 from Σ_N to Σ_M such that $f_1 \circ f_2 = (id_N, id_{U_N})$ and $f_2 \circ f_1 = (id_M, id_{U_M})$.*

In this setting, feedback transformations[4] can be seen as special isomorphisms. Consider an isomorphism (ϕ, φ) with $\varphi : U_M \longrightarrow U_M$ such that $\phi = id_M$. In adapted

[4]Some authors use the expression feedback transformation to denote any isomorphism in **Con**. We consider the more restrictive use where changes of coordinates in the state space are disallowed as they cannot be realized by feedback.

coordinates (x, v), where x represents the base coordinates (the state) and v the coordinates on the fibers (the inputs), the isomorphism has a coordinate expression for φ of the form $\varphi = (x, \beta(x, v))$. The fiber term $\beta(x, v)$ representing the new control inputs is interpreted as a feedback transformation since it depends on the state at the current location as well as the former inputs v. We shall therefore refer to feedback transformations as isomorphisms over the identity since we have $\phi = id_M$. The control theoretic notion of feedback equivalence is captured in this framework by noting that two control systems are feedback equivalent iff there exists an isomorphism (although not necessarily a feedback transformation) between the two systems. A related notion is that of system immersion. Although we cannot capture such a notion in our framework, as we have not equipped control systems with observation maps, a restricted version of system immersion can still be defined within our framework. Recall that, according to [6], system Σ_M is said to be immersed in system Σ_N if there exists an injective map $\phi : M \longrightarrow N$ such that the input-output behavior of Σ_M, when initialized at x, equals the input-output behavior of Σ_N, when initialized at $\phi(x)$. If we assume that $U_M = U \times M$ and $U_N = U \times N$ for some common input manifold U, that M is a submanifold of N, and that i is the canonical injection of M into N, then Σ_M is immersed into Σ_N, when (id_U, i) is a morphism from Σ_M to Σ_N. Note that the existence of morphism (id_U, i) implies that $F_M(x, u) = Ti \cdot F_M(x, u) = F_N(i(x), id_U(u)) = F_N(x, u)$ for local coordinates $(x, u) \in U \times M \subseteq U \times N$, and this implies that Σ_M and Σ_N have the same input-output behavior when initialized at x and $i(x)$, respectively.

A control system can alternatively be defined by a control section $\mathcal{S}_M$ on M in the sense that at each point $x \in M$, $\mathcal{S}_M(x)$ defines all the possible directions along which we can flow or steer our system. However, there can be several control parameterizations for $\mathcal{S}_M$ and it is important to understand in what sense all those parameterizations represent the same control system. In order to obtain such equivalence we make the following assumptions about control systems that will be explicitly mentioned when needed:

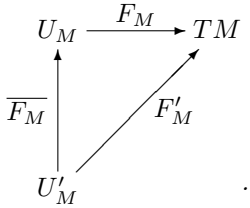
AI: The fibers $\pi_{U_M}^{-1}(x)$ are connected for every $x \in M$.

AII: The map $F_M : U_M \longrightarrow T_M$ is an embedding.

Control systems satisfying assumption **AII** enjoy the following property.

PROPOSITION 4.3. *Let (U_M, F_M) be a control system on manifold M satisfying* **AII** *and let (U'_M, F'_M) be any control system on manifold M such that $\mathcal{S}'_M(x) \subseteq \mathcal{S}_M(x)$ for every $x \in M$. Then, there exists a unique fiber preserving map $\overline{F_M}$ making the following diagram commutative:*

(4.2)



The previous result is an immediate consequence of the fact that $F_M(U_M)$ is an embedded submanifold of TM. This is sufficient for the previous result to hold but not necessary. In fact, the existence of a unique map $\overline{F_M}$ is the property of interest and could be used as a definition. However, it would be difficult to check in concrete examples if a given control system would satisfy such a property. A different approach

would relax the requirement that $F_M(U_M)$ is an embedded submanifold by the weaker assumption of initial submanifold (see [15] for the definition of initial submanifolds and its properties).

Since assumption **AII** implies the universal property [18] stated in Proposition 4.3, any two control systems satisfying **AII** and defining the same control section are isomorphic. It is in this sense that we do not need to distinguish between different parameterizations of the same control section. They are the same control system, up to a change of control coordinates, that is, up to an isomorphism over the identity. This will be important when considering the effect of feedback since, as we have already seen, this change of control coordinates can be regarded as a feedback transformation.

The relation between the notions of ϕ-related control systems (3.1) and **Con** morphisms (4.1) is stated in the next proposition.

PROPOSITION 4.4. *Let Σ_M and Σ_N be two control systems defined on M and N, respectively. If $f = (\phi, \varphi)$ is a* **Con** *morphism from Σ_M to Σ_N, then Σ_N is ϕ-related to Σ_M. Conversely, if Σ_N satisfies* **AII** *and Σ_N is ϕ-related to Σ_M by a smooth map $\phi : M \longrightarrow N$, then there exists a unique fiber preserving lift φ of ϕ such that $f = (\phi, \varphi)$ is a* **Con** *morphism from Σ_M to Σ_N.*

Proof. Definition 4.1 trivially implies Definition 3.1, so let us prove that Definition 3.1 implies Definition 4.1. If Σ_N is ϕ-related to Σ_M, then by Definition 3.1, $T_x\phi(\mathcal{S}_M(x)) \subseteq \mathcal{S}_N \circ \phi(x)$. But $\mathcal{S}_M$ is parameterized by U_M, so that the map $T\phi \circ F_M : U_M \longrightarrow TN$ (see the diagram below) satisfies $T\phi \circ F_M(U_M) \subseteq \mathcal{S}_N$. Therefore, by Proposition 4.3, there is a unique fiber preserving map $\overline{F_N}$ such that

$$\begin{array}{ccc} U_M & \xrightarrow{\overline{F_N}} & U_N \\ {\scriptstyle F_M}\downarrow & & \downarrow{\scriptstyle F_N} \\ TM & \xrightarrow{T\phi} & TN \\ {\scriptstyle \pi_M}\downarrow & & \downarrow{\scriptstyle \pi_N} \\ M & \xrightarrow[\phi]{} & N \end{array}$$

commutes. By taking $\varphi = \overline{F_N}$, $\pi_{U_M} = \pi_M \circ F_M$, and $\pi_{U_N} = \pi_N \circ F_N$ one recovers Definition 4.1 and the equivalence is proved. □

The previous result shows that there is an equivalence between smooth maps ϕ relating control systems and **Con** morphisms provided that we work on a suitable subcategory (where assumption **AII** holds). This means that many properties of nonlinear control systems can be characterized by working with $\mathcal{S}_M$ instead of F_M. We also see that if there is a morphism f from Σ_M to Σ_N, then this morphism carries trajectories of Σ_M to trajectories of Σ_N in virtue of Proposition 3.2.

5. Quotients of control systems. Given a control system Σ_M and an equivalence relation on the manifold M, we can regard the quotient control system as an abstraction since some modeling details propagate from Σ_M to the quotient while other modeling details disappear in the factorization process. This fact motivates the study of quotient control systems as they represent lower complexity (dimension) objects that can be used to verify properties of the original control system. Quotients are also important from a design perspective since a control law for the quotient object can be regarded as a specification for the desired behavior of the original control

system. In this spirit we will address the following questions.

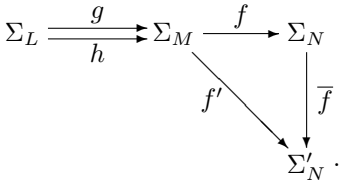
1. *Existence.* Given a control system Σ_M defined on a manifold M and an equivalence relation $\sim_M$ on M, when does there exist a control system on $M/\sim_M$, the quotient manifold, and a fiber preserving lift p_U of the projection $p_M : M \longrightarrow M/\sim_M$ such that (p_M, p_U) is a **Con** morphism?

2. *Uniqueness.* Is the lift p_U of p_M, when it exists, unique?

3. *Structure of the quotient state/input space.* What is the structure of the fibers (input space) of the quotient control system?

To clarify our discussion we formalize the notion of quotient control systems.

DEFINITION 5.1 (quotient control system). *Let Σ_L, Σ_M, Σ_N be control systems defined on manifolds L, M, and N, respectively, and g, h two morphisms from Σ_L to Σ_M. The pair (f, Σ_N), where f is a morphism and Σ_N a control system, is a quotient control system of Σ_M if $f \circ g = f \circ h$ and for any other pair (f', Σ'_N) such that $f' \circ g = f' \circ h$ there exists one and only one morphism $\overline{f}$ from Σ_N to Σ'_N such that the following diagram commutes:*



(5.1)

That is, $f' = \overline{f} \circ f$.

Intuitively, we can read diagram (5.1) as follows. Assume that the set $\sim = \{(u, v) \in U_M \times U_M \;:\; (u, v) = (g(l), h(l)) \text{ for some } l \in U_L\}$ is a regular equivalence relation [1]. Then, the condition $f \circ g = f \circ h$ simply means that f respects the equivalence relation, that is, $u \sim v \;\Rightarrow\; f(u) = f(v)$. Furthermore, it asks that for any other map f' respecting relation $\sim$, there exists a unique map $\overline{f}$ such that $f' = \overline{f} \circ f$. This is a usual characterization of quotient manifolds [1] that we use here as a definition. The same chain of reasoning shows that if we replace control systems by the corresponding state space and the morphisms by the maps between the state spaces, then diagram (5.1) asks for N to also be quotient manifold obtained by factoring M by a regular equivalence relation $\sim_M$ on M defined by g and h. The same idea must therefore hold for control systems. This means that control system Σ_N must also satisfy a unique factorization property in order to be a quotient control system.

From the above discussion it is clear that a necessary condition for the existence of the quotient control system is the existence of the quotient manifold $M/\sim_M$. When $\sim_M$ is a regular equivalence relation the quotient space $M/\sim_M$ will be a manifold [1] and the equivalence relation can be equivalently described by a surjective submersion. We will therefore assume that the regular equivalence relation $\sim_M$ is given by a surjective submersion $\phi : M \longrightarrow N$. Similarly, the fiber preserving lift φ of ϕ will also have to be a surjective submersion. We now consider the following assumption which will be explicitly stated when required.

AIII: The map $T\phi \circ F_M : U_M \longrightarrow TN$ has constant rank and connected fibers.

The first two questions of the previous list are answered in the next theorem which asserts that quotients exist under moderate conditions.

THEOREM 5.2. *Let Σ_M be a control system on a manifold M and $\phi : M \longrightarrow N$ a surjective submersion, and assume that* **AIII** *holds. Then there exist,*

1. *a control system Σ_N on N,*

2. *a* unique *fiber preserving lift* $\varphi : U_M \longrightarrow U_N$ *of* ϕ *such that the pair* $((\phi, \varphi), \Sigma_N)$ *is a quotient control system of* Σ_M.

Proof. By assumption **AIII**, the map $T\phi \circ F_M$ has constant rank and we can define a regular and involutive distribution $\mathcal{D}$ on TU_M by $\mathcal{D} = \ker(TT\phi \circ TF_M)$. Furthermore, as $T\phi \circ F_M$ has connected fibers, also by assumption **AIII**, these are described by the integral manifolds of $\mathcal{D}$. We thus have a regular equivalence relation $\sim \subseteq U_M \times U_M$ obtained by declaring two points equivalent if they lie on the same integral manifold of $\mathcal{D}$. We now consider the manifold $U_M/\sim$ obtained as the quotient of U_M by $\sim$ and denote by $\pi : U_M \longrightarrow U_M/\sim$ the canonical projection. Since $T\phi \circ F_M(u) = T\phi \circ F_M(u')$ iff $\pi(u) = \pi(u')$, it follows from the properties of quotient manifolds [1] that there exists a unique map $\alpha : U_M/\sim \longrightarrow TN$ such that $\alpha \circ \pi = T\phi \circ F_M$. We now define U_N as $U_M/\sim$, π_{U_N} as $\pi_N \circ \alpha$, F_N as α, and φ as π. We note that φ is unique and claim that $((\phi, \varphi), U_M/\sim)$ is a quotient of Σ_M. The pair of maps (ϕ, φ) is a morphism from Σ_M to Σ_N since $T\phi \circ F_M = F_N \circ \varphi$ as required by the second diagram in (4.1), and composing $T\phi \circ F_M = F_N \circ \varphi$ with π_N, we obtain,

$$\begin{aligned}
\pi_N \circ T\phi \circ F_M &= \pi_N \circ F_N \circ \varphi \\
\Leftrightarrow \phi \circ \pi_M \circ F_M &= \pi_N \circ F_N \circ \varphi && \text{since } \pi_N \circ T\phi = \phi \circ \pi_M \\
\Leftrightarrow \phi \circ \pi_{U_M} &= \pi_N \circ F_N \circ \varphi && \text{by commutativity of diagram (2.3)} \\
\Leftrightarrow \phi \circ \pi_{U_M} &= \pi_{U_N} \circ \varphi && \text{by definition of } \pi_{U_N},
\end{aligned}$$

which shows that the first diagram in (4.1) also commutes.

It remains to show that any other morphism $f' = (\phi', \varphi')$ such that ϕ' is compatible with the equivalence relation defined by ϕ factors uniquely through f. Since the equivalence relation defined by ϕ on M induces the equivalence relation $\sim$ on U_M, we see that $\varphi(u) = \varphi(u')$ implies $\varphi'(u) = \varphi'(u')$. It then follows from the universality of φ that φ' factors uniquely through φ; that is, there exists a unique map $\overline{\varphi}$ such that $\varphi' = \overline{\varphi} \circ \varphi$. Similarly, ϕ' factors uniquely trough ϕ via $\overline{\phi}$. It then remains to show that $(\overline{\phi}, \overline{\varphi})$ is a morphism from Σ_N to Σ'_N.

We first show that diagram (4.1) commutes. Let $u_n \in U_N$, as φ is a surjective map; there is a $u_m \in U_M$ such that $\varphi(u_m) = u_n$. We now have, by diagram chasing,

$$\begin{aligned}
F'_N \circ \overline{\varphi} \circ \varphi(u_m) &= F'_N \circ \varphi'(u_m) && \text{since } \varphi' \text{ factors on } \varphi \\
&= T\phi' \circ F_M(u_m) && \text{by commutativity of the 2nd diagram in (4.1)} \\
&= T\overline{\phi} \circ T\phi \circ F_M(u_m) && \text{since } \phi' \text{ factors on } \phi \\
&= T\overline{\phi} \circ F_N \circ \varphi(u_m) && \text{by commutativity of the 2nd diagram in (4.1),}
\end{aligned}$$

and replacing $\varphi(u_m)$ by u_n we see that $\overline{f}$ satisfies the second diagram in (4.1). Commutativity of the first diagram in (4.1) can be obtained similarly by diagram chasing. □

This result provides the first characterization of quotient objects in **Con**. It shows that given a regular equivalence relation on the base (state) space of a control system and a mild regularity condition, there exists a quotient control system on the quotient manifold. Furthermore, it also shows that the regular equivalence relation on M or the map ϕ uniquely determines a fiber preserving lift φ which describes how the state/input pairs of the control system on M relate to the state/input pairs of the quotient control system. Furthermore, we also see that the map F_N is an injective immersion, a fact we will use several times in the remainder of this paper.

Existence of quotients under such weak conditions is perhaps surprising given the fact that in other contexts, quotients exist only in very specific situations. A quotient

group can only be obtained by factoring a group by a normal subgroup and not by a general equivalence relation, a quotient linear space can only be obtained by factoring a linear space by a linear subsubspace and not by a general equivalence relation, etc. This fact highlights the relevance of Theorem 5.2 at the theoretical level but also at the practical level since quotients can be constructively used to hierarchically design trajectories [37].

Having answered the first two questions from the previous list (existence and uniqueness), we concentrate on the characterization of the quotient control system input space. This problem requires a deeper understanding of how ϕ determines φ and will be the goal of the remainder of this paper. Since **Con** was defined over **Man**, that is, morphisms in **Con** are smooth maps and control systems are defined on manifolds, the characterization of φ will require an interplay of tools from differential geometry and category theory.

6. Projectable control sections. We now extend the notion of projectable vector fields from [21] and of projectable families of vector fields from [22] to control sections. The notion of projectable control sections is weaker than projectable vector field or families of vector fields but nonetheless stronger than **Con** morphisms. The motivation for introducing this notion comes from the fact that projectability of control sections will be a fundamental ingredient in characterizing the structure of the quotient system input space . Furthermore, we will also see that projectability, as defined in this categorical setting, will correspond to the well-known notion of controlled invariance.

Given a vector field X on M and a surjective submersion $\phi : M \longrightarrow N$ we say that X is projectable with respect to ϕ when $Y = T\phi \cdot X$, the projection of X, is a well-defined vector field on N that satisfies $T\phi \cdot X = Y \circ \phi$ [21]. The vector field Y is also called ϕ-related to X [1]. This notion was extended to families of vector fields in [22] by requiring that the projection of each vector field in the family is a well-defined vector field on N. However, when working with control sections, which can be regarded as *sets* of vectors at each base point $x \in M$, one should require only that the projection of these *sets* of vectors is the same *set* when the base points on M project on the same base point on N. Intuitively, we are asking for control sections that behave well under the projection defined by ϕ. This is formalized as follows.

DEFINITION 6.1. *Let M be a manifold, $\mathcal{S}_M$ a control section on M, and $\phi : M \longrightarrow N$ a surjective submersion. We say that $\mathcal{S}_M$ is projectable with respect to ϕ iff $\mathcal{S}_M$ induces a control section $\mathcal{S}_N$ on N such that the following diagram commutes:*

$$\begin{array}{ccc} \mathcal{P}(TM) & \xrightarrow{T\phi} & \mathcal{P}(TN) \\ \uparrow{\scriptstyle \mathcal{S}_M} & & \uparrow{\scriptstyle \mathcal{S}_N} \\ M & \xrightarrow[\phi]{} & N \end{array} \quad , \tag{6.1}$$

where $\mathcal{P}(TM)$ denotes the powerset of T_xM for every $x \in M$.

We see that if $\mathcal{S}_M$ is in fact a vector field we recover the notion of projectable vector fields. The notion of projectable control sections is stronger then the notion of **Con** morphism since for any $x_1, x_2 \in M$ such that $\phi(x_1) = \phi(x_2)$ we necessarily have $T\phi(\mathcal{S}_M(x_1)) = \mathcal{S}_N \circ \phi(x_1) = T\phi(\mathcal{S}_M(x_2))$ if $\mathcal{S}_M$ is projectable. On the other hand, if (ϕ, φ) is a **Con** morphism for a fiber preserving lift φ of ϕ, we only have the inclusions $T\phi(\mathcal{S}_M(x_1)) \subseteq \mathcal{S}_N \circ \phi(x_1)$ and $T\phi(\mathcal{S}_M(x_2)) \subseteq \mathcal{S}_N \circ \phi(x_1)$. Therefore, projectability

with respect to ϕ and **AII** implies that ϕ can be extended to a **Con** morphism but given a **Con** morphism $f = (\phi, \varphi)$ from Σ_M to Σ_N it is not true, in general, that $\mathcal{S}_M$ is projectable with respect to ϕ.

To determine the relevant conditions on $\mathcal{S}_M$ that ensure projectability we will need an auxiliary result.

LEMMA 6.2. *Let $f : M \longrightarrow N$ be a map between manifolds and let X_t be the flow of a vector field $X \in TM$ such that $f \circ X_t = f$. Then the following equality holds for every $x \in M$:*

$$T_x f \, T_{X_t(x)} X_{-t} = T_{X_t(x)} f. \tag{6.2}$$

Proof. The equality $f \circ X_t = f$ is equivalent to

$$\begin{aligned} & f \circ X_t(x) = f(x) \\ \Leftrightarrow\ & f(X_t(x)) = f \circ (X_t)^{-1} \circ X_t(x) \\ \Leftrightarrow\ & f(X_t(x)) = f \circ X_{-t}(X_t(x)), \end{aligned} \tag{6.3}$$

and by differentiation of the previous expression we arrive at the desired equality

$$T_{X_t(x)} f = T_x f \, T_{X_t(x)} X_{-t}. \qquad \square \tag{6.4}$$

We can now give sufficient and necessary conditions for projectability of control sections.

PROPOSITION 6.3 (projectable control sections). *Let M be a manifold, $\mathcal{S}_M$ a control section on M, and $\phi : M \longrightarrow N$ a surjective submersion with connected fibers. Given any control system (U_M, F_M) satisfying* **AI** *and defining $\mathcal{S}_M$, and any $\widehat{F_M} \in F_M^e$, $\mathcal{S}_M$ is projectable with respect to ϕ iff*

$$[\widehat{F_M}, \ker(T\pi_{U_M}^* \phi)] \subseteq \ker(T\pi_{U_M}^* \phi) + [\widehat{F_M}, \ker(T\pi_{U_M})]. \tag{6.5}$$

Proof. We show necessity first. Assume that diagram (6.1) commutes. Then we have

$$T_x \phi(\mathcal{S}_M(x)) = T_{x'} \phi(\mathcal{S}_M(x')) \tag{6.6}$$

for all $x, x' \in M$ such that $\phi(x) = \phi(x')$, that is, for any x and x' on the same leaf of the foliation induced by $\ker(T\phi)$. If we denote by K_t the flow of any vector field $K \in \ker(T\pi_{U_M}^* \phi)$, expression (6.6) implies that

$$T_{\pi_{U_M} \circ K_t(u)} \phi(F_M \circ K_t(u)) \in T_x \phi(\mathcal{S}_M(x)) \tag{6.7}$$

for every $t \in \mathbb{R}$ such that K_t is defined and for every $u \in \pi_{U_M}^{-1}(x)$. Since the left-hand side of (6.7) belongs to the right-hand side and $\pi_{U_M}^{-1}(x)$ is connected by **AI**, we can always find a $Y \in \ker(T\pi_{U_M})$ such that its flow Y_t will parameterize the image of the left-hand side. That is

$$T_{\pi_{U_M} \circ K_t(u)} \phi(F_M \circ K_t(u)) = T_{\pi_{U_M} \circ Y_t(u)} \phi(F_M \circ Y_t(u)). \tag{6.8}$$

The previous equality implies that for any $\widehat{F_M} \in F_M^e$ we have

$$T_{K_t(u)} \pi_{U_M}^* \phi(\widehat{F_M} \circ K_t(u)) = T_{Y_t(u)} \pi_{U_M}^* \phi(\widehat{F_M} \circ Y_t(u)); \tag{6.9}$$

however, the equalities $\pi^*_{U_M}\phi \circ K_t = K_t$, $\pi^*_{U_M}\phi \circ Y_t = Y_t$ and Lemma 6.2 allow one to rewrite (6.9) as

$$\begin{aligned} & T_u\pi^*_{U_M}\phi(T_{K_t(u)}K_{-t} \circ \widehat{F_M} \circ K_t(u)) = T_u\pi^*_{U_M}\phi(T_{Y_t(u)}Y_{-t} \circ \widehat{F_M} \circ Y_t(u)) \\ \Leftrightarrow \quad & T_u\pi^*_{U_M}\phi(K_t(u)^*\widehat{F_M}) = T_u\pi^*_{U_M}\phi(Y_t(u)^*\widehat{F_M}). \end{aligned} \tag{6.10}$$

Time differentiation at $t = 0$ now implies

$$\begin{aligned} & T_u\pi^*_{U_M}\phi([K(u), \widehat{F_M}(u)]) = T_u\pi^*_{U_M}\phi([Y(u), \widehat{F_M}(u)]) \\ & \Rightarrow [K, \widehat{F_M}] \in [Y, \widehat{F_M}] + \ker(T\pi^*_{U_M}\phi), \end{aligned} \tag{6.11}$$

which trivially implies inclusion (6.5).

To show sufficiency we use a similar argument. Assume that (6.5) holds; then for any $K \in \ker(T\pi^*_{U_M}\phi)$ there exists a $Y \in \ker(T\pi_{U_M})$ such that

$$\begin{aligned} & T_u\pi^*_{U_M}\phi([\widehat{F_M}(u), K(u)]) = T_u\pi^*_{U_M}\phi([\widehat{F_M}(u), Y(u)]) \\ \Leftrightarrow\ & T_u\pi^*_{U_M}\phi([\widehat{F_M}(u), K(u) - Y(u)]) = 0. \end{aligned} \tag{6.12}$$

Consider now the regular and involutive distribution $\ker(T\pi^*_{U_M}\phi)$. Involutivity and regularity imply that $Z_t^*W \in \ker(T\pi^*_{U_M}\phi)$ for any $W \in \ker(T\pi^*_{U_M}\phi)$ and the flow Z_t of any vector field $Z \in \ker(T\pi^*_{U_M}\phi)$ [34]. Since $K \in \ker(T\pi^*_{U_M}\phi)$ and $Y \in \ker(T\pi^*_{U_M}\phi)$, it follows that $K - Y \in \ker(T\pi^*_{U_M}\phi)$, but from (6.12), $[\widehat{F_M}, K - Y]$ also belongs to $\ker(T\pi^*_{U_M}\phi)$ so that we conclude

$$T_u\pi^*_{U_M}\phi((K-Y)_t(u)^*[\widehat{F_M}, K - Y]) = 0, \tag{6.13}$$

where $(K - Y)_t$ denotes the flow of the vector field $K - Y$. However, the previous expression is equivalent to

$$\begin{aligned} & T_u\pi^*_{U_M}\phi\left(\frac{d}{dt}(K-Y)_t(u)^*\widehat{F_M}\right) = 0 \\ \Leftrightarrow\ & \frac{d}{dt}T_u\pi^*_{U_M}\phi((K-Y)_t(u)^*\widehat{F_M}) = 0, \end{aligned} \tag{6.14}$$

where the last equality follows from the fact that $T\phi$ is a linear map. Since the time derivative is zero, we must have

$$T_u\pi^*_{U_M}\phi((K-Y)_t(u)^*\widehat{F_M}) = T_u\pi^*_{U_M}\phi((K-Y)_0(u)^*\widehat{F_M}) = T_u\pi^*_{U_M}\phi(\widehat{F_M}(u)). \tag{6.15}$$

From the equality $\pi^*_{U_M}\phi = \pi^*_{U_M}\phi \circ (K-Y)_t$ we conclude that $T_u\pi^*_{U_M}\phi T_{(K-Y)_t(u)}(K - Y)_{-t} = T_{(K-Y)_t(u)}\pi^*_{U_M}\phi$ by Lemma 6.2 so that (6.15) can be written as

$$T_{(K-Y)_t(u)}\pi^*_{U_M}\phi(\widehat{F_M} \circ (K-Y)_t(u)) = T_u\pi^*_{U_M}\phi(\widehat{F_M}(u)) \tag{6.16}$$

and projecting on TM we get

$$T_{\pi_{U_M}(K'_t(u))}\phi(F_M \circ (K')_t(u)) = T_x\phi(F_M(u)) \tag{6.17}$$

with $K' = K - Y$. This equality shows that for any $X \in \mathcal{S}_M(x)$, $T_x\phi \cdot X \in T_{x'}\phi(\mathcal{S}_M(x'))$; therefore, $T_x\phi(\mathcal{S}_M(x)) \subseteq T_{x'}\phi(\mathcal{S}_M(x'))$. However, replacing x by x' and K by $-K$ on (6.17), we get $T_{x'}\phi(\mathcal{S}_M(x')) \subseteq T_x\phi(\mathcal{S}_M(x))$ so that we conclude the equality

$$T_x\phi(\mathcal{S}_M(x)) = T_{x'}\phi(\mathcal{S}_M(x')). \tag{6.18}$$

By connectedness of the fibers $\phi^{-1}(y)$ any point x'' satisfying $\phi(x'') = \phi(x)$ can be reached by a concatenation of flows induced by vector fields in $\ker(T\phi)$. Transitivity of equality between sets implies that (6.18) holds for any two points $x, x' \in M$ such that $\phi(x) = \phi(x')$ from which commutativity of diagram (6.1) readily follows. □

By now it is already clear that projectability and local controlled invariance are equivalent concepts. We recall the notion of locally controlled invariant distribution.

DEFINITION 6.4 (locally controlled invariant distributions [27]). *Let $\Sigma_M = (U_M, F_M)$ be a control system on manifold M and let $\mathcal{D}$ be a distribution on M. Distribution $\mathcal{D}$ is locally controlled invariant for F_M if for every $x \in M$ there exist an open set $O \subseteq M$ containing x and a local (feedback) isomorphism over the identity α such that in adapted coordinates (x, v) the new control system $F_M \circ \alpha$ satisfies*

$$[F_M \circ \alpha(x, v), \mathcal{D}(x)] \subseteq \mathcal{D}(x) \tag{6.19}$$

for every (x, v) in the domain of α.

If a control section is projectable, then locally we can always choose $\widehat{F_M} = F_M^l$ and therefore recover the conditions for local controlled invariance from [8].

THEOREM 6.5 (see [8]). *Let Σ_M be a control system on manifold M satisfying* **AI** *and $\phi : M \longrightarrow N$ a surjective submersion with connected fibers. The distribution $\ker(T\phi)$ is locally controlled invariant for F_M iff $\mathcal{S}_M$ is projectable with respect to ϕ.*

Even though controlled invariance and projectability are equivalent concepts, we shall use the notion of projectability to describe control sections that behave well under projection instead of controlled invariance which was introduced to describe certain control enforced invariance properties of control systems [42].

From the study of symmetries of nonlinear control systems [9, 26] it was already known that the existence of symmetries or partial symmetries implies controlled invariance of a certain distribution associated with the symmetries. This shows that control systems that are projectable comprise quotients by symmetry and controlled invariance. Furthermore, quotients induced by indistinguishability, as discussed in [35], are also of this type. However, there are also quotients for which projectability does not hold as we describe in the next section. Furthermore, as the existence of symmetries can be considered a rare phenomena [32], it is especially important to understand the structure of general nonprojectable quotients.

7. The structure of quotient control systems. We have already seen that the notion of **Con** morphisms generalizes the notion of projectable control sections. This shows that it is possible to quotient control systems whose control sections are nonprojectable. In this situation the map φ and the input space of the quotient control system will be significantly different from the projectable case. To understand this difference we start characterizing the fiber preserving lift φ of ϕ. Recall that if $f = (\phi, \varphi)$ is a morphism from Σ_M to Σ_N, we have the following commutative diagram:

$$\begin{array}{ccc} U_M & \xrightarrow{\varphi} & U_N \\ {\scriptstyle F_M}\downarrow & & \downarrow{\scriptstyle F_N} \\ TM & \xrightarrow[T\phi]{} & TN \end{array} \quad . \tag{7.1}$$

Since φ is a surjective submersion, we know that U_N is diffeomorphic to $U_M/\sim$, where $\sim$ is the regular equivalence relation induced by φ. This means that to understand the structure of U_N it is enough to determine the regular and involutive distribution on U_M given by $\ker(T\varphi)$. However, the map φ is completely unknown, so we will resort to the elements that are available, namely, F_M and ϕ, to determine $\ker(T\varphi)$. Differentiating[5] diagram (7.1) we get

$$\begin{array}{ccc} TU_M & \xrightarrow{T\varphi} & TU_N \\ {\scriptstyle TF_M}\downarrow & & \downarrow{\scriptstyle TF_N} \\ TTM & \xrightarrow[TT\phi]{} & TTN \end{array} \tag{7.2}$$

from which we conclude that

$$\ker(TT\phi \circ TF_M) = \ker(TF_N \circ T\varphi) = \ker(T\varphi), \tag{7.3}$$

where the last equality holds since F_N is an immersion, provided that assumption **AIII** holds. We can now attempt to understand what is factored away and what is propagated from U_M to U_N since $\ker(T\varphi)$ is expressible in terms of F_M and ϕ. The first step is to clarify the relation between $\ker(T\varphi)$ and $\ker(T\phi)$. Since φ is a fiber preserving lift of ϕ, the following diagram commutes:

$$\begin{array}{ccc} TU_M & \xrightarrow{T\varphi} & TU_N \\ {\scriptstyle T\pi_{U_M}}\downarrow & & \downarrow{\scriptstyle T\pi_{U_N}} \\ TM & \xrightarrow[T\phi]{} & TN \end{array} \quad , \tag{7.4}$$

which implies that

$$T\pi_{U_M}(\ker(T\varphi)) \subseteq \ker(T\phi). \tag{7.5}$$

However, this only tells us that the reduction on M due to ϕ cannot be "smaller" than the reduction on the base space of U_M due to φ. This leads to the interesting phenomena which occurs when, e.g.,

$$T\pi_{U_M}(\ker(T\varphi)) = \{0\} \subseteq \ker(T\phi). \tag{7.6}$$

[5]The operator sending manifolds to their tangent manifolds and maps to their tangent maps is an endofunctor on **Man**, also called the tangent functor [15].

The above expression implies that the base space of U_M is not reduced by φ. However, U_N is a fibered manifold with base space N and therefore the points reduced by ϕ must necessarily move to the fibers of U_N. This means that points $u, u' \in U_M$ satisfying $\pi_{U_M}(u) \neq \pi_{U_M}(u')$ will be mapped by φ to points satisfying $\pi_{U_N} \circ \varphi(u) = \pi_{U_N} \circ \varphi(u')$ and $\varphi(u) \neq \varphi(u')$. This will not happen if we can ensure the existence of a distribution $\mathcal{D} \subseteq \ker(T\varphi)$ such that $T\pi_{U_M}(\mathcal{D}) = \ker(T\phi)$. The existence of such a distribution turns out to be related with projectability. To show such a fact we need the following characterization of $\ker(T\varphi)$.

LEMMA 7.1. *Let $\Sigma_M = (U_M, F_M)$ be a control system on manifold M, $\phi : M \longrightarrow N$ a surjective submersion and $\varphi : U_M \longrightarrow U_N$ a fiber preserving lift of ϕ which is also a submersion, and assume that* **AIII** *holds. Under these assumptions, a regular distribution $\mathcal{D} \subseteq TU_M$ belongs to* $\ker(T\varphi)$ *iff*

$$[\widehat{F_M}, \mathcal{D}]; \subseteq \ker(T\pi^*_{U_M}), \tag{7.7}$$

where $\widehat{F_M}$ is any vector field in F^e_M.

Proof. Assume the existence of the distribution $\mathcal{D}$; then $\mathcal{D} \subseteq \ker(T\varphi)$ is equivalent to

$$TT\phi \circ TF_M(\mathcal{D}) = \{0\} \tag{7.8}$$

since **AIII** holds. Let $Z \in \mathcal{D}$ and denote by Z_t the flow of Z. Expression (7.8) implies that

$$\frac{d}{dt}\Big|_{t=0} T_{\pi_{U_M} \circ Z_t(u)}\phi(F_M \circ Z_t(u)) = 0 \quad \Rightarrow \quad \frac{d}{dt}\Big|_{t=0} T_{Z_t(u)}\pi^*_{U_M}\phi(\widehat{F_M} \circ Z_t(u)) = 0 \tag{7.9}$$

for any $\widehat{F_M} \in F^e_M$ and for all $t \in \mathbb{R}$ such that Z_t is defined.

Noticing that $Z \in \mathcal{D} \subseteq \ker(T\varphi)$ implies $\varphi = \varphi \circ Z_t$ (since φ is constant on the leaves of the foliation induced by $\ker(T\varphi)$) and $\pi_{U_N} \circ \varphi = \phi \circ \pi_{U_M}$ by commutativity of diagram (4.1), we conclude that $\pi^*_{U_M}\phi$ is also Z_t invariant:

$$\pi^*_{U_M}\phi \circ Z_t = \phi \circ \pi_{U_M} \circ Z_t = (\pi_{U_N} \circ \varphi) \circ Z_t = \pi_{U_N} \circ \varphi = \phi \circ \pi_{U_M} = \pi^*_{U_M}\phi. \tag{7.10}$$

Lemma 6.2 now ensures that

$$T_{Z_t(u)}\pi^*_{U_M}\phi = T_u\pi^*_{U_M}\phi \circ T_{Z_t(u)}Z_{-t}, \tag{7.11}$$

and expression (7.11) allows one to rewrite (7.9) as

$$\begin{aligned}
\frac{d}{dt}\Big|_{t=0} T_{Z_t(u)}\pi^*_{U_M}\phi(\widehat{F_M} \circ Z_t(u)) = 0 &\Leftrightarrow \frac{d}{dt}\Big|_{t=0} T_u\pi^*_{U_M}\phi(T_{Z_t(u)}Z_{-t} \circ \widehat{F_M} \circ Z_t(u)) = 0 \\
&\Leftrightarrow \frac{d}{dt}\Big|_{t=0} T_u\pi^*_{U_M}\phi(Z_t(u)^*\widehat{F_M}) = 0 \\
&\Leftrightarrow T_u\pi^*_{U_M}\phi([Z(u), \widehat{F_M}(u)]) = 0
\end{aligned} \tag{7.12}$$

or, equivalently, $[Z, \widehat{F_M}] \in \ker(T\pi^*_{U_M}\phi)$. Since Z is any vector field in $\mathcal{D}$ it follows that $[\widehat{F_M}, \mathcal{D}] \subseteq \ker(T\pi^*_{U_M}\phi)$ as desired.

The converse is similarly proved. ∎

Using Lemma 7.1, we can now characterize the existence of a distribution $\mathcal{D} \subseteq \ker(T\varphi)$ projecting on $\ker(T\phi)$.

PROPOSITION 7.2. *Let $\Sigma_M = (U_M, F_M)$ be a control system on manifold M satisfying* **AI**, *$\phi : M \longrightarrow N$ a surjective submersion, and $\varphi : U_M \longrightarrow U_N$ a fiber preserving lift of ϕ which is also a submersion. There exists a regular distribution $\mathcal{D}$ on U_M satisfying $\mathcal{D} \subseteq \ker(T\varphi)$ and $T\pi_{U_M}(\mathcal{D}) = \ker(T\phi)$ iff $\mathcal{S}_M$ is projectable with respect to ϕ.*

Proof. We start by showing that projectability implies the existence of $\mathcal{D}$. If $\mathcal{S}_M$ is projectable with respect to ϕ, then for every $x, x' \in M$ such that $\phi(x) = \phi(x')$ we have that $T_x\phi(\mathcal{S}_M(x)) = T_{x'}\phi(\mathcal{S}_M(x'))$. This means that for any $x \in M$, $u \in \pi^{-1}_{U_M}(x)$, and $X \in \ker(T\pi^*_{U_M}\phi)$, there exists a $Y \in \ker(T\pi_{U_M})$ (recall that $\pi^{-1}_{U_M}(x)$ is connected by **AI**) such that

$$T_{\pi_{U_M}\circ X_t(u)}\phi(F_M \circ X_t(u)) = T_x\phi(F_M \circ Y_t(u)) \tag{7.13}$$

for all $t \in \mathbb{R}$ such that the flows X_t and Y_t of X and Y are defined. Considering now $T\phi$ as a map between the manifolds TM and TN, the time derivative of $T_{\alpha(t)}\phi(\beta(t))$ for $(\alpha, \beta) : \mathbb{R} \longrightarrow TM$ provides $T_{(\alpha(t),\beta(t))}T_{\alpha(t)}\phi(T\beta(t))$. The same considerations applied to (7.13) at $t = 0$ give

$$T_{(x,F_M(u))}T_x\phi \circ T_uF_M(X(u)) = T_{(x,F_M(u))}T_x\phi \circ T_uF_M(Y(u)), \tag{7.14}$$

which we rewrite as

$$T_{(x,F_M(u))}T_x\phi \circ T_uF_M(X(u) - Y(u)) = 0 \tag{7.15}$$

by linearity of the involved maps. Since (7.15) is true for any $X \in \ker(T\pi^*_{U_M}\phi)$, we can define the distribution

$$\mathcal{D} = \bigcup_{K \in \ker(T\phi)} \{Z = X - Y \ : \ X \in K^e \ \wedge \ Y \in \ker(T\pi_{U_M}) \text{ is such that (7.15) holds}\}. \tag{7.16}$$

This distribution clearly satisfies

$$TT\phi \circ TF_M(\mathcal{D}) = \{0\} \quad \Leftrightarrow \quad \mathcal{D} \subseteq \ker(T\varphi), \tag{7.17}$$

is regular since $\dim(\mathcal{D}) = \dim(\ker(T\phi))$ by construction, satisfies $T\pi_{U_M}(\mathcal{D}) = \ker(T\phi)$ also by construction, and is therefore the desired distribution.

The converse is proved as follows. Assume the existence of the distribution $\mathcal{D} \subseteq \ker(T\varphi)$. By Lemma 7.1 we have

$$[\widehat{F_M}, \mathcal{D}] \subseteq \ker(T\pi^*_{U_M}).$$

Since $T\pi_{U_M}(\mathcal{D}) = \ker(T\phi)$ it follows that $\mathcal{D} + T\pi_{U_M} = T\pi^*_{U_M}\phi$ and

$$\begin{aligned} & [\widehat{F_M}, \mathcal{D}] \subseteq \ker(T\pi^*_{U_M}) \\ \Rightarrow\ & [\widehat{F_M}, \mathcal{D} + \ker(T\pi_{U_M})] \subseteq \ker(T\pi^*_{U_M}) + [\widehat{F_M}, \ker(T\pi_{U_M})] \\ \Rightarrow\ & [\widehat{F_M}, \ker(T\pi^*_{U_M}\phi)] \subseteq \ker(T\pi^*_{U_M}) + [\widehat{F_M}, \ker(T\pi_{U_M})], \end{aligned}$$

which combined with Proposition 6.3 shows that $\mathcal{S}_M$ is projectable with respect to ϕ as desired. ∎

Collecting the results given by Lemma 7.1 and Proposition 7.2 we can now characterize both φ and U_N. Intuitively, we will use projectability to determine if the fibers of the quotient control system will receive states from M and Lemma 7.1 to characterize the amount of reduction induced by φ on the fibers of π_{U_M}.

THEOREM 7.3 (structure of control systems quotients). *Consider a control system $\Sigma_M = (U_M, F_M)$ over a manifold M satisfying* **AI**, *let $(f, \Sigma_N) = ((\phi, \varphi), (U_N, F_N))$ be a quotient of Σ_M where ϕ has connected fibers, and assume that* **AIII** *holds. Let $\mathcal{E}$ be the involutive distribution defined by $\mathcal{E} = \{X \in \ker(T\pi_{U_M}) : [\widehat{F_M}, X] \in \ker(T\pi^*_{U_M}\phi)\}$, which we assume to be regular, denote by $R_{\mathcal{E}}$ the regular equivalence relation induced by $\mathcal{E}$, and let $\widehat{F_M}$ be any vector field in F^e_M. Under these assumptions the following can be stated:*

1. Reduction from states to states and no reduction on inputs. *Fibered manifold U_N has base space diffeomorphic to N, and fibers $\pi^{-1}_{U_N}(y)$ diffeomorphic to $\pi^{-1}_{U_M}(x)$, $\phi(x) = y$ iff*

(i) *$\mathcal{S}_M$ is projectable with respect to ϕ;*

(ii) *$\mathcal{E} = \{0\}$.*

2. Reduction from states to states and from inputs to inputs. *Fibered manifold U_N has base space diffeomorphic to N, and fibers $\pi^{-1}_{U_N}(y)$ diffeomorphic to $\pi^{-1}_{U_M}(x)/R_{\mathcal{E}}$, $\phi(x) = y$ iff*

(i) *$\mathcal{S}_M$ is projectable with respect to ϕ;*

(ii) *$\mathcal{E} \neq \{0\}$.*

3. Reduction from states to inputs and no reduction on inputs. *Fibered manifold U_N has base space diffeomorphic to N, and fibers $\pi^{-1}_{U_N}(y)$ diffeomorphic to $\pi^{-1}_{U_M}(\phi^{-1}(y))$, $\phi(x) = y$ iff*

(i) *for all $K \in \ker(T\pi^*_{U_M}\phi)$, $\quad [\widehat{F_M}, K] \notin \ker(T\pi^*_{U_M}\phi)$;*

(ii) *$\mathcal{E} = \{0\}$.*

4. Reduction from states to inputs and from inputs to inputs. Fibered manifold U_N has base space diffeomorphic to N, and fibers $\pi^{-1}_{U_N}(y)$ diffeomorphic to $(\pi^{-1}_{U_M}/R_{\mathcal{E}})(\phi^{-1}(y))$, $\phi(x) = y$ iff

(i) for all $K \in ker(T\pi^*_{U_M}\phi)$, $\quad [\widehat{F_M}, K] \notin ker(T\pi^*_{U_M}\phi)$;

(ii) $\mathcal{E} \neq \{0\}$.

Proof. We note that in all four cases the base space of U_N is diffeomorphic to N, since U_N is equipped with a surjective submersion $\pi_{U_N} : U_N \longrightarrow N$. We will, therefore, only discuss the characterization of fibers of π_{U_N}. We follow the enumeration of the theorem.

1 and 2: Since φ is fiber preserving, we denote by $\varphi_x : \pi^{-1}_{U_M}(x) \longrightarrow \pi^{-1}_{U_N}(\phi(x))$ the restriction of φ to the fibers $\pi^{-1}_{U_M}(x)$, $x \in M$. We now claim that projectability implies $\varphi_x(\pi^{-1}_{U_M}(x)) = \varphi_{x'}(\pi^{-1}_{U_M}(x'))$ for every $x, x' \in M$ such that $\phi(x) = \phi(x')$. Recall that by definition of projectability, we have $T_x\phi(\mathcal{S}_M(x)) = T_{x'}\phi(\mathcal{S}_M(x'))$. However, $\mathcal{S}_M(x) = F_M(\pi^{-1}_{U_M}(x))$ so that we conclude $T_x\phi \circ F_M(\pi^{-1}_{U_M}(x)) = T_{x'}\phi \circ F_M(\pi^{-1}_{U_M}(x'))$. From assumption **AIII** follows injectivity of F_N, which combined with commutativity of the second diagram in (4.1) leads to $\varphi_x(\pi^{-1}_M(x)) = \varphi_{x'}(\pi^{-1}_M(x'))$, as desired. This equality also shows that φ_x is surjective since φ is. Furthermore, we conclude that to characterize $\pi^{-1}_{U_N}(y)$ it suffices to characterize the image of φ_x for some $x \in \phi^{-1}(y)$. We now consider $\ker(T\varphi(x)) \cap \ker(T\pi^{-1}_{U_M}(x))$, which by Lemma 7.1 is equal to $\mathcal{E}$ and is regular by assumption. Fields in $\pi^{-1}_{U_M}(x)$ and the equivalence relation $R_{\mathcal{E}}$ can be identified with an equivalence relation on $\pi^{-1}_{U_M}(x)$. We first note that projectability implies via Proposition 7.2 and (7.5) that $T\pi_{U_M}(\ker(T\varphi)) = \ker(T\phi)$. This shows

that

$$\dim(\ker(T\varphi)) = \dim(\ker(T\phi)) + \dim(\mathcal{E}). \tag{7.18}$$

On the other hand,

$$\begin{aligned}
\dim(\pi_{U_N}^{-1}(y)) &= \dim(U_N) - \dim(N) \\
&= \dim(U_M) - \dim(\ker(T\varphi)) - \dim(N) \\
&= \dim(U_M) - \dim(\ker(T\phi)) - \dim(\mathcal{E}) - \dim(N) \qquad \text{by (7.18)} \\
&= \dim(U_M) - \dim(\ker(T\phi)) - \dim(\mathcal{E}) - \dim(M) + \dim(\ker(T\phi)) \\
&= \dim(U_M) - \dim(\mathcal{E}) - \dim(M) \\
&= \dim(\pi_{U_M}^{-1}(x)) - \dim(\mathcal{E}) = \operatorname{rank}(\varphi_x), \qquad (7.19)
\end{aligned}$$

which shows that φ_x is a submersion. We thus see that $\pi_{U_N}^{-1}(y)$ can now be identified with $\pi_{U_M}^{-1}(x)/R_{\mathcal{E}}$ since every vector field $X \in \mathcal{E}$ satisfies $T\pi_{U_M}(X) = 0$ and therefore induces a vector field on $\pi_{U_M}^{-1}(x)$. If $\mathcal{E} = \{0\}$, it follows that $\pi_{U_N}^{-1}(y) \cong \pi_{U_M}^{-1}(x)/R_{\mathcal{E}} \cong \pi_{U_M}^{-1}(x)$ as required by case 1.

Conversely, since the base of U_N is diffeomorphic to the quotient of M by the regular equivalence relation induced by $\ker(T\phi)$ and the fibers of π_{U_N} diffeomorphic to $\pi_{U_M}/R_{\mathcal{E}}$, it follows that $\ker(T\varphi)$ can be locally described by $\mathcal{D} \oplus \mathcal{E}$ for $\mathcal{D} = \ker(T\phi)^l$ and $T\pi_{U_M}(\mathcal{E}) = \{0\}$. From the existence of $\mathcal{D}$ and Proposition 6.3 follows projectability of $\mathcal{S}_M(x)$. Furthermore, if the fibers of π_{U_M} are diffeomorphic to the fibers of π_{U_N}, we have $\mathcal{E} = \{0\}$ (case 1) and otherwise $\mathcal{E} \neq \{0\}$ (case 2).

3 and 4: From assumption (i) and Lemma 7.1 we conclude that there exists no $X \neq 0$ belonging to $K(T\varphi)$ such that $T\pi_{U_M}(X) \in \ker(T\phi)$. Since $T\pi_{U_M}(\ker(T\varphi)) \subseteq \ker(T\phi)$ (see the discussion before (7.5)), it follows that $T\pi_{U_M}(\ker(T\varphi)) = \{0\}$. Consequently, every $X \in \ker(T\varphi)$ is tangent to $\pi_{U_M}^{-1}(x)$ and $\varphi(U_M)$ is diffeomorphic to a fibered manifold with base space M and fibers $\pi_{U_M}^{-1}(x)/R_{\mathcal{E}}$. Let us denote by $\pi : \varphi(U_M) \longrightarrow M$ the projection from total space to base space which clearly satisfies $\pi_{U_M} = \pi \circ \varphi$. We now use the fact $\pi_{U_N} \circ \varphi = \phi \circ \pi_{U_M}$ with $\pi_{U_M} = \pi \circ \varphi$ to get $\pi_{U_N} \circ \varphi = \phi \circ \pi \circ \varphi$ and by surjectivity of φ we finally conclude the equality $\pi_{U_N} = \phi \circ \pi$. It is now clear that $\pi_{U_N}^{-1}(y) \cong \pi^{-1}(\phi^{-1}(y)) \cong (\pi_{U_M}^{-1}/R_{\mathcal{E}})(\phi^{-1}(y))$ as required by case 4. Case 3 is obtained by setting $\mathcal{E} = \{0\}$ and obtaining $\pi_{U_N}^{-1}(y) \cong (\pi_{U_M}^{-1}/R_{\mathcal{E}})(\phi^{-1}(y)) \cong \pi_{U_M}^{-1}(\phi^{-1}(y))$.

The converse is proved as follows. Since the fibers of π_{U_N} are diffeomorphic to $(\pi_{U_M}^{-1}/R_{\mathcal{E}})(\phi^{-1}(y))$, we see that points $u, u' \in U_M$ satisfying $\pi_{U_M}(u) \neq \pi_{U_M}(u')$ and $\phi \circ \pi_{U_M}(u) = \phi \circ \pi_{U_M}(u')$ also satisfy $\varphi(u) \neq \varphi(u')$. This shows that no vector field $X \neq 0$ in $\ker(T\pi_{U_M}^*\phi)$ belongs to $\ker(T\varphi)$ since otherwise different points in a trajectory of X would violate the above remark. The nonexistence of such vectors X implies, via Lemma 7.1, condition (i) and also implies that $\ker(T\varphi) = \mathcal{E}$. It then follows that if $\pi_{U_N}^{-1}(y) \cong \pi_{U_M}^{-1}(\phi^{-1}(y))$, then $\mathcal{E} = 0$ (case 3) and $\mathcal{E} \neq 0$ otherwise (case 4). □

We see that the notion of projectability is fundamentally related to the structure of quotient control systems. If the controlled section $\mathcal{S}_M$ is projectable, then the inputs of the quotient control system are the same or a quotient of the original inputs. Projectability can therefore be seen as a structural property of a control system in the sense that it admits special decompositions [11, 27]. However, for general systems not admitting this special structure, that is, for systems that are not projectable, it is still

possible to construct quotients by moving the neglected state information to the fibers. The states of the original system that are factored out by ϕ are regarded as control inputs in the quotient control system. This shows that from a hierarchical synthesis point of view, control systems that are not projectable are much more appealing since one can design control laws for the abstracted system that when pulled-down to the original one are regarded as specifications for the dynamics on the neglected states [37].

8. Conclusions. In this paper quotients of fully nonlinear control systems were investigated. We showed that under mild conditions quotients exist, and we characterized the structure of the quotient state/input space. This was achieved by introducing the category of control systems which was the natural framework to discuss quotients of control systems. One of the important ingredients of the characterization of quotients was the notion of a projectable control section, which being equivalent to controlled invariance allowed one to understand the difference between general quotients and those induced by symmetries, partial symmetries, or controlled invariance.

There are still innumerous directions to be explored. The correct relations of the results presented in this work with the notion of extended control system [25] are not yet understood. This seems to lead to a possible generalization of the constructive procedures presented in [29] to compute quotients of nonlinear control affine systems to fully nonlinear control systems. Other directions being currently investigated include similar results for mechanical control systems where the Hamiltonian structure is preserved by the factorization process [36] as well as hybrid control systems [38].

Acknowledgments. The authors would like to acknowledge the innumerous enlightening discussions with Gonçalo Tabuada, Esfandiar Hagverdi, and Slobodan Simic on categorical interpretations of dynamical and control systems. Equally appreciated were the constructive comments provided by the anonymous reviewers.

REFERENCES

[1] R. Abraham, J. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, 2nd ed., Applied Mathematical Sciences 75, Springer-Verlag, New York, 1988.

[2] M. Arbib and E. G. Manes, *Machines in a category: An expository introduction*, SIAM Rev., 16 (1974), pp. 163–192.

[3] F. Borceux, *Handbook of Categorical Algebra*, Cambridge University Press, Cambridge, UK, 1994.

[4] R. Brockett, *Control theory and analytical mechanics*, in Geometric Control Theory, C. Martin and R. Hermann, eds., Lie Groups: History, Frontiers and Appl. Vol. VII, Math Sci Press, Brookline, MA, 1977, pp. 1–48.

[5] R. Brockett, *Feedback invariants for nonlinear systems*, in Preprints of the 6th IFAC Congress, Helsinki, Finland, 1978, pp. 1115–1120.

[6] M. Fliess and I. Kupka, *A finiteness criterion for nonlinear input-output differential systems*, SIAM J. Control Optim., 21 (1983), pp. 721–728.

[7] M. Fliess, J. Levine, P. Martin, and P. Rouchon, *Flatness and defect of non-linear systems: Introductory theory and examples*, Internat. J. Control, 61 (1995), pp. 1327–1361.

[8] K. Grasse, *On controlled invariance for fully nonlinear control systems*, Internat. J. Control, 56 (1992), pp. 1121–1137.

[9] J. W. Grizzle and S. I. Marcus, *The structure of nonlinear control systems possessing symmetries*, IEEE Trans. Automat. Control, 30 (1985), pp. 248–258.

[10] L. Hunt, R. Su, and G. Meyer, *Design for multi-input nonlinear systems*, in Differential Geometric Control Theory, R. Brocket, R. Millman, and H. J. Sussmann, eds., Birkhäuser Boston, Boston, 1983, pp. 268–298.

[11] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, Berlin, 1996.

[12] A. Isidori, A. Krener, C. G. Giorgi, and S. Monaco, *Nonlinear decoupling via feedback: A differential geometrical approach*, IEEE Trans. Automat. Control, 26 (1981), pp. 331–345.

[13] B. Jakubczyk and W. Respondek, *On linearization of control systems*, Bull. Acad. Polon. Sci. Ser. Sci. Math., 28 (1980), pp. 517–522.

[14] V. Jurdjevic, *Geometric Control Theory*, Cambridge Studies in Advanced Mathematics 52, Cambridge University Press, Cambridge, UK, 1997.
[15] I. Kolar, P. W. Michor, and J. Slovak, *Natural Operations in Differential Geometry*, Springer-Verlag, Berlin, 1993.
[16] A. Krener, *On the equivalence of control systems and linearization of nonlinear systems*, SIAM J. Control, 11 (1973), pp. 670–676.
[17] A. J. Krener, *A decomposition theory for differentiable systems*, SIAM J. Control Optim., 15 (1977), pp. 813–829.
[18] S. MacLane, *Categories for the Working Mathematician*, Springer-Verlag, New York, 1971.
[19] A. D. Lewis, *The category of affine connection control systems*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 1260–1265.
[20] R. Marino, *On the largest feedback linearizable subsystem*, Systems Control Lett., 6 (1986), pp. 345–351.
[21] G. Marmo, A. Simoni, B. Vitale, and E. J. Saletan, *Dynamical Systems*, John Wiley & Sons, Chichester, UK, 1985.
[22] L. Martin and P. Crouch, *Controllability on principal fiber bundles with compact structure group*, Syst. Control Lett., 5 (1984), pp. 35–40.
[23] H. Nijmeijer, *Controlled invariance for affine control systems*, Internat. J. Control, 34 (1981), pp. 825–833.
[24] H. Nijmeijer, *Feedback decomposition of nonlinear control systems*, IEEE Trans. Automat. Control, 28 (1983), pp. 861–862.
[25] H. Nijmeijer and A. van der Schaft, *Controlled invariance for nonlinear systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 904–914.
[26] H. Nijmeijer and A. van der Schaft, *Partial symmetries for nonlinear systems*, Math. Systems Theory, 18 (1985), pp. 79–96.
[27] H. Nijmeijer and A. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[28] G. J. Pappas, G. Lafferriere, and S. Sastry, *Hierarchically consistent control systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1144–1160.
[29] G. J. Pappas and S. Simic, *Consistent abstractions of affine control systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 745–756.
[30] J. W. Polderman and J. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Springer-Verlag, New York, 1998.
[31] W. Respondek, *On decomposition of nonlinear control systems*, Systems Control Lett., 1 (1981), pp. 301–308.
[32] W. Respondek and I. A. Tall, *Nonlinearizable single-input control systems do not admit stationary symmetries*, Systems Control Lett., 46 (2002), pp. 1–16.
[33] J. Rutten, *Universal coalgebra: A theory of systems*, Theoret. Comput. Sci., 249 (2000), pp. 3–80.
[34] H. J. Sussmann, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–188.
[35] H. J. Sussmann, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1976), pp. 263–284.
[36] P. Tabuada and G. J. Pappas, *Abstractions of Hamiltonian control systems*, Automatica, J. IFAC, 39 (2003), pp. 2025–2033.
[37] P. Tabuada and G. J. Pappas, *Hierarchical trajectory generation for a class of nonlinear systems*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Hawaii, 2003.
[38] P. Tabuada, G. J. Pappas, and P. Lima, *Compositional abstractions of hybrid control systems*, J. Discrete Event Dyn. Syst., 14 (2004), pp. 203–238.
[39] A. van der Schaft, *Symmetries and conservation laws for Hamiltonian systems with inputs and outputs: A generalization of Noether's theorem*, Systems Control Lett., 1 (1982), pp. 108–115.
[40] A. J. van der Schaft and B.Maschke, *Interconnected mechanical systems, part* I*: Geometry of interconnection and implicit Hamiltonian systems*, in Modelling and Control of Mechanical Systems, A. Astolfi, D. Limebeer, C. Melchiorri, A. Tornambe, and R. Vinter, eds., Imperial College Press, London, 1997.
[41] J. Willems, *System theoretic models for the analysis of physical systems*, Ricerche Automat., 10 (1979), pp. 71–106.
[42] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1979.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1867–1887

HYBRID NECESSARY PRINCIPLE*

MAURO GARAVELLO† AND BENEDETTO PICCOLI†

Abstract. We consider a hybrid control system and general optimal control problems for this system. We suppose that the switching strategy imposes restrictions on control sets and we provide necessary conditions for an optimal hybrid trajectory, stating a hybrid necessary principle (HNP). Our result generalizes various necessary principles available in the literature.

Key words. optimal control, necessary conditions, switching strategy

AMS subject classifications. 49J15, 93C65

DOI. 10.1137/S0363012903416219

1. Introduction.

This paper deals with optimal control problems for hybrid systems. Our definition of a hybrid system is the one used in [20]. Roughly speaking, a hybrid system is a collection of control systems called locations, possibly defined on different manifolds, and an automaton that rules the switchings between locations. The term hybrid is commonly used to indicate the presence of both continuous and discrete dynamics, and in our case the continuous part is given by location controlled dynamics and the discrete part by the automaton. An optimal control problem is obtained assigning Lagrangian running costs on each location and final and switching costs.

Recently optimization problems for hybrid systems have attracted a lot of attention; thus both theoretical results and applications were developed (see [6, 10, 11, 12, 14, 17, 23, 29]). For general theory of hybrid systems we refer to [3, 15, 19].

For an optimal (classical) control problem, the main tool for the construction of optimal trajectories, and then optimal synthesis, is the celebrated Pontryagin maximum principle (PMP). The strength of PMP is evident when it permits us to describe completely the structure of optimal trajectories and obtain a finite dimensional reduction of the problem; see, for example, [5, 8, 24]. A hybrid maximum principle (HMP) was developed in [26]; see also [20, 21]. A key role is played by the switching mechanism that permits us to pass from one location to another with possible restrictions on state, time to spend in the next location, and feasible controls for the next location. The first two kinds of restrictions do not affect the general strategy of PMP, and an HMP can be proved in a similar way. However, the restriction on usable controls, after location switchings, dramatically changes the possibility of constructing "needle variations" that are the basic ingredient to prove PMP, and HMP is no more applicable. More precisely, a classical needle variation is no longer prolongable after a location switching time; therefore a new class of "admissible needle variations" must be introduced. As a first example (in section 3), to construct this kind of variation, one can prolong the family of trajectories, originating from a needle variation, via the choice of a suitable family of controls having continuity and weak differentiability (in L^1) properties.

*Received by the editors August 20, 2003; accepted for publication (in revised form) June 14, 2004; published electronically March 22, 2005.
http://www.siam.org/journals/sicon/43-5/41621.html
†IAC, Viale del Policlinico 137, 00161 Roma, Italy (m.garavello@iac.rm.cnr.it, piccoli@iac.rm.cnr.it).

The problem of producing variations more general than needle variations has been extensively considered in the literature for the classical (not hybrid) setting (see [1, 2, 4, 7, 18, 25]), and more recently also for the hybrid settings; see [27].

In section 4 we introduce a general concept of "map of variations." The basic request is again to have weak differentiability properties, but now in the space of bounded Radon measures, seen as the dual of the space of continuous functions. We are then able to prove the hybrid necessary principle (HNP), using results from [22, 27, 28]. The word maximum in this context disappears, since necessary conditions are no longer written in a supremum form.

Section 2 gives basic definitions of *hybrid systems* and states HMP, section 3 deals with *admissible needle variations,* and section 4 deals with the *map of variations.* Finally, an appendix, with two technical lemmas on weak convergence in L^1 and for measures, concludes the paper.

2. Basic definitions and HMP. We start by introducing the definition of a hybrid system.

DEFINITION 2.1. A *hybrid control system* is a 7-tuple $\Sigma = (\mathcal{Q}, M, U, f, \mathcal{U}, J, \mathcal{S})$ such that

- H1 $\mathcal{Q}$ is a finite set;
- H2 $M = \{M_q\}_{q\in\mathcal{Q}}$ is a family of smooth manifolds, indexed by $\mathcal{Q}$;
- H3 $U = \{U_q\}_{q\in\mathcal{Q}}$ is a family of sets;
- H4 $f = \{f_q\}_{q\in\mathcal{Q}}$ is a family of maps $f_q : M_q \times U_q \mapsto TM_q$ (TM_q is the tangent bundle of M_q), such that $f_q(x,u) \in T_xM_q$ for every $(x,u) \in M_q \times U_q$;
- H5 $\mathcal{U} = \{\mathcal{U}_q\}_{q\in\mathcal{Q}}$ is a family of sets $\mathcal{U}_q$ whose members are maps $u : \mathrm{Dom}(u) \to U_q$, defined on some interval $\mathrm{Dom}(u) \subset \mathbb{R}$;
- H6 $J = \{J_q\}_{q\in\mathcal{Q}}$ is a family of subintervals of $\mathbb{R}^+$;
- H7 $\mathcal{S}$ is a subset of $\mathrm{Switch}(\Sigma)$, where

$$\mathrm{Switch}(\Sigma) \stackrel{\mathrm{def}}{=} \{(q,x,q',x',u(\cdot),\tau) : q,q' \in \mathcal{Q}, x \in M_q, x' \in M_{q'}, u(\cdot) \in \mathcal{U}_{q'}, \tau \in J_{q'}\}.$$

The members of $\mathcal{Q}$ are called *locations* and represent the states of the automaton. The families M, U are, respectively, the *family of state spaces* and the *family of control spaces* of Σ. For each q, the manifold M_q, the set U_q, the map f_q, and the set $\mathcal{U}_q$ are, respectively, the *state space*, the *control space*, the *controlled dynamical law*, and the *class of admissible controls* at location q.

The system evolves in a location q according to the corresponding controlled dynamic and then switches as prescribed by $\mathcal{S}$. The intervals J_q indicate the lengths of time intervals on which the system can stay in location q. So, for example, if $J_q = [0, +\infty[$, then the system can evolve in location q on every interval of time.

For $q, q' \in \mathcal{Q}$, we write

$$\mathcal{S}_{q,q'} \stackrel{\mathrm{def}}{=} \{(x,x') \in M_q \times M_{q'} : (q,x,q',x',u(\cdot),\tau) \in \mathcal{S} \text{ for some } u(\cdot) \in \mathcal{U}_{q'} \text{ and } \tau \in J_{q'}\}.$$

The sets $\mathcal{S}_{q,q'}$ are called the *switching sets* of Σ from location q to location q'. Moreover, for $q, q' \in \mathcal{Q}$ and $x \in M_q$, $x' \in M_{q'}$, we write

$$\mathcal{U}_{q,x,q',x'} \stackrel{\mathrm{def}}{=} \{u(\cdot) \in \mathcal{U}_{q'} : (q,x,q',x',u(\cdot),\tau) \in \mathcal{S} \text{ for some } \tau \in J_{q'}\}.$$

The set $\mathcal{U}_{q,x,q',x'}$ is formed by the controls we can use at location q' if there is a switching from the point x of M_q to the point x' of $M_{q'}$.

DEFINITION 2.2. *A hybrid state is a triplet (q, x, τ), where $q \in \mathcal{Q}$ is the location, $x \in M_q$ is the state of the control system, and $\tau \in [0, \sup J_q)$ is the time since last switching. We denote by $\mathcal{HS}$ the set of all hybrid states.*

The evolution of the hybrid system is as follows. Given a hybrid initial state $(q_1, x_0, 0)$, at time t_0, on some time interval $[t_0, t_1[$, with $t_1 - t_0 \in J_{q_1}$, the system evolves solving

$$\begin{cases} q(t) \equiv q_1, \\ \dot{x}(t) = f_{q_1}(x(t), u_1(t)), & x(t_0) = x_0, \\ \dot{\tau}(t) = 1, & \tau(t_0) = 0 \end{cases}$$

for some $u_1(\cdot) \in \mathcal{U}_{q_1}$ such that $\mathrm{Dom}(u_1) \supset [t_0, t_1]$. This means that the system remains in location q_1 until $\tau = t_1 - t_0$ and it evolves on M_{q_1} according to the dynamic $f_{q_1}(x(t), u_1(t))$ for the control $u_1(\cdot) \in \mathcal{U}_{q_1}$. If the solution to the previous system can be prolonged on the whole interval $[t_0, t_1]$, then we can choose another hybrid state $(q_2, x_1, 0)$, a control $u_2(\cdot) \in \mathcal{U}_{q_2}$, and t_2 such that $(q_1, x(t_1), q_2, x_1, u_2(\cdot), t_2 - t_1) \in \mathcal{S}$ and let the system evolve in location q_2 following the corresponding controlled dynamics on the interval $[t_1, t_2]$:

$$\begin{cases} q(t) \equiv q_2, \\ \dot{x}(t) = f_{q_2}(x(t), u_2(t)), & x(t_1) = x_1, \\ \dot{\tau}(t) = 1, & \tau(t_1) = 0. \end{cases}$$

We say that a location switching from q_1 to q_2 occurs at time t_1. Then we can proceed in the same way with a location switching and so on. Notice that the time t_1 (t_2 and so on) can be chosen freely in J_{q_1} (respectively, J_{q_2} and so on); hence it represents a control for the hybrid system.

We assume that if $u \in \mathcal{U}_q$, then every time translation of u is in $\mathcal{U}_q$; more precisely, we assume

(A1) If $u \in \mathcal{U}_q$ for some $q \in \mathcal{Q}$, then for every $\sigma \in \mathbb{R}$ the control $\tilde{u}(t) = u(t + \sigma)$ satisfies $\tilde{u} \in \mathcal{U}_q$.

Hence we can always assume that $t_0 = 0$.

Let us now give a precise definition of trajectories, cost functionals, and optimal control problems.

DEFINITION 2.3. *A trajectory is a map $\boldsymbol{X} : [0, T] \to \mathcal{HS}$, $\boldsymbol{X}(t) = (q(t), x(t), \tau(t))$ such that the following holds. There exist $0 = t_0 < t_1 < \cdots < t_\nu = T$ such that if $i \in \{1, \dots, \nu\}$, then $q(\cdot)$ is constant in $[t_{i-1}, t_i[$ and equal to $q_i \in \mathcal{Q}$, $\tau(t) = t - t_{i-1}$ on $[t_{i-1}, t_i[$, $t_i - t_{i-1} \in J_{q_i}$. Moreover, for every $i \in \{1, \dots, \nu\}$, there exists $u_i \in \mathcal{U}_{q_i}$ such that*

- *$x_i(\cdot) := x|_{]t_{i-1}, t_i[}(\cdot)$ is an absolutely continuous function in $]t_{i-1}, t_i[$, continuously prolongable to $[t_{i-1}, t_i]$;*
- *$\frac{d}{dt} x_i(t) = f_{q_i}(x_i(t), u_i(t))$ for a.e. $t \in]t_{i-1}, t_i[$;*
- *$(x_i(t_i), x_{i+1}(t_i)) \in \mathcal{S}_{q_i, q_{i+1}}$ if $i = 1, \dots, \nu - 1$;*
- *$u_{i+1} \in \mathcal{U}_{q_i, x_i(t_i), q_{i+1}, x_{i+1}(t_i)}$ if $i = 1, \dots, \nu - 1$.*

Remark 1. In this setting, for a Cauchy type problem, it is not appropriate to first choose a sequence of controls and then determine the trajectory associated to it, because a priori the sequence could not be admissible, in the sense that there could exist no trajectory corresponding to it. This is due to the fact that in every location q, it is possible to use as controls only a subset of $\mathcal{U}_q$, depending on the switching strategy.

DEFINITION 2.4. *If Σ is a hybrid system, then a Lagrangian for Σ is a family $L = \{L_q\}_{q\in Q}$, $L_q : M_q \times U_q \to \mathbb{R}$ such that, for every trajectory $\boldsymbol{X}$, for every $i \in \{1,\dots,\nu\}$, and for every control u_i associated to x_i, the function $t \mapsto L_{q_i}(x_i(t), u_i(t))$ is integrable in $]t_{i-1}, t_i[$.*

DEFINITION 2.5. *If Σ is a hybrid system, then a switching cost function is a family $\Phi = \{\Phi_{q,q'}\}_{(q,q')\in\mathcal{Q}\times\mathcal{Q}}$ such that each $\Phi_{q,q'}$ is a real valued function defined on $\mathcal{S}_{q,q'}$.*

DEFINITION 2.6. *If Σ is a hybrid system, then an endpoint cost function is a family $\varphi = \{\varphi_{q,q'}\}_{(q,q')\in\mathcal{Q}\times\mathcal{Q}}$ such that each $\varphi_{q,q'}$ is a real valued function defined on $M_q \times M_{q'}$.*

If $L = \{L_q\}_{q\in\mathcal{Q}}$ is a Lagrangian, $\Phi = \{\Phi_{q,q'}\}_{(q,q')\in\mathcal{Q}\times\mathcal{Q}}$ is a switching cost function, and $\varphi = \{\varphi_{q,q'}\}_{(q,q')\in\mathcal{Q}\times\mathcal{Q}}$ is an endpoint cost function for the hybrid control system Σ, then we can define the corresponding *cost functional* C by letting

$$C(\mathbf{X}) = \sum_{j=1}^{\nu} \int_{t_{j-1}}^{t_j} L_{q_j}(x_j(t), u_j(t))\, dt + \sum_{j=1}^{\nu-1} \Phi_{q_j,q_{j+1}}(x_j(t_j), x_{j+1}(t_j)) + \varphi_{q_1,q_\nu}(x_1(t_0), x_\nu(t_\nu)),$$

where $\mathbf{X}$ is a trajectory for Σ.

DEFINITION 2.7. *Given a hybrid control system Σ, a cost functional C, and two nonempty subsets $\mathcal{N}_{in}, \mathcal{N}_{fin}$ of $\mathcal{HS}$, we call with $\mathcal{P}$ the problem of minimizing $C(\boldsymbol{X})$ over all trajectories $\boldsymbol{X}$ for Σ such that*

(i) $(q_1, x_1(t_0), 0) \in \mathcal{N}_{in}$;

(ii) $(q_\nu, x_\nu(t_\nu), t_\nu - t_{\nu-1}) \in \mathcal{N}_{fin}$.

Remark 2. Note that there could be no trajectory satisfying boundary data. However, we expect that in many applications the set $\mathcal{N}_{fin}$ should be chosen to impose restrictions only on the final location q and point x. So if $(q, x, t) \in \mathcal{N}_{fin}$, then $\mathcal{N}_{fin}$ should also contain all the points (q, x, s) with $s \le \sup J_{q_\nu}$ (with possible equality only if $\sup J_{q_\nu} \in J_{q_\nu}$).

The *maximum principle* gives a necessary condition for a trajectory $\mathbf{X}$ to be a solution of $\mathcal{P}$. The set of variations involves trajectories having the same *history* (see [20]) as the candidate optimal one, that is, having the same switching strategy. As suggested in [20], if there is a finite number of possible switching strategies for the optimization problem $\mathcal{P}$, then the maximum principle can sometimes single out the optimal trajectory.

DEFINITION 2.8. *If Σ is a hybrid system and L is a Lagrangian for Σ, then we say that (ψ, ψ_0) is an adjoint pair along a trajectory $\boldsymbol{X}$ if*

1. *$\psi = (\psi_1, \dots, \psi_\nu)$ is such that, for every $i \in \{1,\dots,\nu\}$, $\psi_i : [t_{i-1}, t_i] \to T^*M_{q_i}$ is an absolutely continuous function, $\psi_i(t) \in T^*_{x_i(t)}M_{q_i}$, and*

$$\dot\psi_i(t) = - < \psi_i(t), \frac{\partial}{\partial x} f_{q_i}(x_i(t), u_i(t)) > + \psi_0 \frac{\partial}{\partial x} L_{q_i}(x_i(t), u_i(t))$$

for a.e. $t \in [t_{i-1}, t_i]$;

2. *$\psi_0 \in \mathbb{R}^+$.*

In order to state the switching condition, we need a concept of a tangent cone. In this paper, as in [27], we use the notion of a Boltyanskiĭ approximating cone.

DEFINITION 2.9. *Let S be a subset of a smooth manifold $\mathcal{X}$ and let $\bar s \in S$. A Boltyanskiĭ approximating cone to S at $\bar s$ is a closed convex cone K in the tangent space $T_{\bar s}\mathcal{X}$ such that there exist a neighborhood W of 0 in $T_{\bar s}\mathcal{X}$ and a continuous map*

$\omega : W \cap K \to S$ with the property that $\omega(0) = \bar{s}$ and $\omega(w) = \bar{s} + w + o(\|w\|)$ as $w \to 0$ via values in $W \cap K$.

DEFINITION 2.10. *If Σ is a hybrid system, L is a Lagrangian, and Φ is a switching cost function, then we say that an adjoint pair (ψ, ψ_0) along a trajectory $\boldsymbol{X}$ satisfies the switching condition if*

$$(-\psi_i(t_i), \psi_{i+1}(t_i)) - \psi_0 \nabla \Phi_{q_i, q_{i+1}}(x_i(t_i), x_{i+1}(t_i)) \in K_i^{\perp}$$

for every $i \in \{1, \dots, \nu - 1\}$, where K_i is a Boltyanskiĭ approximating cone to the set $\mathcal{S}_{q_i, q_{i+1}}$ at the point $(x_i(t_i), x_{i+1}(t_i))$ and $K_i^{\perp}$ is its polar cone.

DEFINITION 2.11. *If (ψ, ψ_0) is an adjoint pair along $\boldsymbol{X}$, and*

$$H_i := \sup\{< \psi_i(t), f_{q_i}(x_i(t), u) > -\psi_0 L_{q_i}(x_i(t), u) : u \in U_{q_i}\},$$

then we say that (ψ, ψ_0) satisfies the Hamiltonian value condition *if, for every $i \in \{1, \dots, \nu - 1\}$,*

- *if $t_i - t_{i-1} \in \mathrm{Int}(J_{q_i})$, then $H_i = H_\nu = 0$;*
- *if $t_i - t_{i-1}$ is the left endpoint of J_{q_i}, but J_{q_i} is nontrivial, then $H_i \le 0$;*
- *if $t_i - t_{i-1}$ is the right endpoint of J_{q_i}, but J_{q_i} is nontrivial, then $H_i \ge 0$.*

As explained in the introduction for "simple" switching constraints an HMP is valid. The condition ensuring this is precisely the following:

Assumption (H). For every fixed $q, q' \in \mathcal{Q}$, $x \in M_q$, $x' \in M_{q'}$, we have $\mathcal{U}_{q,x,q',x'} = \mathcal{U}_{q'}$.

Assumption (H) says that in every location $q \in \mathcal{Q}$ we can always use all the controls which are in $\mathcal{U}_q$. Thus the admissible controls do not depend on the location switchings. In particular, the classical "needle variations" are still admissible variations.

HYBRID MAXIMUM PRINCIPLE. *Consider the problem $\mathcal{P}$ and assume* (H). *Let $\boldsymbol{X}$ be a solution for $\mathcal{P}$. Then, under suitable assumptions, there exists an adjoint pair (ψ, ψ_0) along $\boldsymbol{X}$ that satisfies the switching condition, the Hamiltonian maximization, nontriviality, transversality, and Hamiltonian value conditions for $\mathcal{P}$.*

There are some technical assumptions that are necessary for the HMP to hold true. These are specified in [21], [26], [27].

3. Simple necessary conditions. In this section we present some introductory results about necessary conditions for optimality for hybrid systems that do not satisfy assumption (H). We postpone to the next section the statement and the proof of the HNP, the main result of this paper. So this section is intended as a clarifying introduction to the subject of the next section.

Assumption (H) is not verified by many mechanical control systems. For example, to describe a car with gears, one can use a hybrid system, where each location corresponds to a gear of the car and the control is the acceleration. In this case it is clear that when switching from a low gear to a higher one, not all the controls can be used; see [13].

Here we suppose that every M_q is equal to $\mathbb{R}^{d_q}$ for some $d_q \in \mathbb{N}$, $d_q \ge 1$, and that every U_q is a compact subset of $\mathbb{R}^l$ for some $l \in \mathbb{N}$, $l \ge 1$. So $f_q : \mathbb{R}^{d_q} \times U_q \to \mathbb{R}^{d_q}$ and we assume that

$$f_q \in C^2(\mathbb{R}^{d_q} \times U_q; \mathbb{R}^{d_q}); \tag{3.1}$$

hence, for every compact $K \subseteq \mathbb{R}^{d_q}$ there exists a constant $\Gamma_K > 0$ such that

$$\begin{cases} |f_q(x,u) - f_q(y,u)| \le \Gamma_K |x - y| & \forall x, y \in K, \quad \forall u \in U_q, \\ |f_q(x,u) - f_q(x,v)| \le \Gamma_K |u - v| & \forall x \in K, \quad \forall u, v \in U_q. \end{cases} \tag{3.2}$$

In addition, we consider the case $\mathcal{U}_q = L^{p_q}_{loc}(\mathbb{R}; U_q)$ for some $1 \le p_q \le +\infty$ and $L_q \in C^2(\mathbb{R}^{d_q} \times U_q; \mathbb{R})$. The symbol $L^{p_q}_{loc}(\mathbb{R}; U_q)$ denotes the set of functions from $\mathbb{R}$ to U_q belonging to $L^{p_q}(K; U_q)$ for every compact subset K of $\mathbb{R}$. Therefore we are in the situation of local existence and uniqueness for every Cauchy problem.

Remark 3. In order to avoid too many technicalities, in this paper we prefer to consider simplified hypotheses about the manifolds, the vector fields, and the Lagrangians. However, it is possible to prove all the results of this paper in a similar way using weaker assumptions.

Needle variations are the basic tool to prove the PMP in nonhybrid settings and the HMP in hybrid settings. Needle variations consist of modifying the supposed optimal control in a small interval of times and to understand how the trajectory and the cost vary in this way. Unfortunately, under our hypotheses, since the choice of admissible controls depends on the switching strategy, needle variations do not produce admissible trajectories. Therefore, it is necessary to modify the notion of needle variation.

For the aim of simplicity, we consider only admissible needle variations of the following type: the control is the same as that of the candidate optimal trajectory until a certain time $\bar\tau$; then we produce a constant variation for a small interval of times and finally, in the following locations, we consider controls satisfying the switching conditions and some continuity and differentiability properties. More precisely, we have the following definition.

DEFINITION 3.1. *Let us fix a trajectory* $\boldsymbol{X}$ *and* $i \in \{1, \dots, \nu\}$. *We say that the family of trajectories* $\boldsymbol{X}^\varepsilon = (q, x^\varepsilon, \tau)$, $\boldsymbol{X}^\varepsilon : [0, T] \to \mathcal{HS}$ $(\varepsilon > 0)$ *is an* admissible needle variation *at location* i *if*

1. $\boldsymbol{X}^0 \equiv \boldsymbol{X}$;
2. $\boldsymbol{X}^\varepsilon(t) = \boldsymbol{X}(t)$ *for every* $t \in [0, t_{i-1}]$;
3. *the curves* $\varepsilon \mapsto x^\varepsilon_j(t_{j-1})$ *are differentiable at* $\varepsilon = 0^+$ *for every* $j \in \{1, \dots, \nu\}$;
4. *there exists a time* $\bar\tau \in [t_{i-1} + \varepsilon, t_i]$ *such that*

$$u^\varepsilon_i(t) = \begin{cases} u_i(t), & t \in [t_{i-1}, \bar\tau - \varepsilon[, \\ \omega, & t \in [\bar\tau - \varepsilon, \bar\tau[, \\ u_i(t), & t \in [\bar\tau, t_i] \end{cases} \tag{3.3}$$

for some $\omega \in U_{q_i}$, *where the symbol* u^ε_j $(j \in \{1, \dots, \nu\})$ *denotes the control at location* j *of* x^ε_j;

5. *for every* $j \in \{i+1, \dots, \nu\}$, $u^\varepsilon_j \to u_j$ *strongly in* $L^1([t_{j-1}, t_j])$ *as* $\varepsilon \to 0^+$ *and* $\frac{u^\varepsilon_j - u_j}{\varepsilon} \rightharpoonup \theta_j$ *weakly in* $L^1([t_{j-1}, t_j])$ *as* $\varepsilon \to 0^+$ *for some* $\theta_j \in L^1([t_{j-1}, t_j])$.

Remark 4. Notice that in Definition 3.1 we require that $\mathbf{X}^\varepsilon$, when $\varepsilon > 0$, be a family of trajectories. This means that, for a fixed $\varepsilon > 0$, $\mathbf{X}^\varepsilon$ is a trajectory and hence, by Definition 2.3,

$$(x^\varepsilon_j(t_j), x^\varepsilon_{j+1}(t_j)) \in \mathcal{S}_{q_j, q_{j+1}}$$

for every $j \in \{1, \dots, \nu - 1\}$ *and*

$$u^\varepsilon_{j+1} \in \mathcal{U}_{q_j, x_j(t_j), q_{j+1}, x_{j+1}(t_j)}$$

for every $j \in \{1, \dots, \nu - 1\}$.

Moreover, we require the existence of a location q_i, $i \in \{1, \dots, \nu\}$, in which a variation originates. In particular we demand that, in the fixed location q_i, the variation is a classical needle variation and so the expression of the control u^ε_i is

given in (3.3). In another location q_j, $j \in \{1, \dots, \nu\}$, $j \neq i$, we have the following possibilities.

1. *If $j < i$, then $u_j^\varepsilon = u_j$ and $x_j^\varepsilon = x_j$ since the variation originates in location q_i.*

2. *If $j > i$, then we need some regularity properties of the control with respect to the parameter ε. These properties are described in* 5 *of Definition* 3.1. *Recall that for HMP we may choose $u_j^\varepsilon = u_j$ so that* 5 *is trivially satisfied.*

For a needle variation $\mathbf{X}^\varepsilon$ we define

$$v_j(t) = \frac{d}{d\varepsilon} x_j^\varepsilon(t)\big|_{\varepsilon=0}. \tag{3.4}$$

We have the following lemmas.

LEMMA 3.2. *Let us assume* (3.2). *Let $\boldsymbol{X}^\varepsilon$ be an admissible needle variation. Then x^ε converges to x uniformly as ε goes to* 0.

Proof. It is sufficient to prove that, for every $j \in \{1, \dots, \nu\}$, x_j^ε converges uniformly to x_j in $[t_{j-1}, t_j]$ as $\varepsilon \to 0$.

Obviously, if $1 \leq j < i$, then $x_j^\varepsilon = x_j$ and so the conclusion is true. Therefore we can treat the case $j \geq i$. For $t \in [t_{j-1}, t_j]$, we have, for some $\Gamma > 0$,

$$\begin{aligned}
\left|x_j^\varepsilon(t) - x_j(t)\right| &\leq \left|x_j^\varepsilon(t_{j-1}) - x_j(t_{j-1})\right| \\
&+ \left| \int_{t_{j-1}}^t \left[f_{q_j}(x_j^\varepsilon(s), u_j^\varepsilon(s)) - f_{q_j}(x_j(s), u_j(s)) \right] ds \right| \\
&\leq \left|x_j^\varepsilon(t_{j-1}) - x_j(t_{j-1})\right| + \Gamma \int_{t_{j-1}}^t \left|x_j^\varepsilon(s) - x_j(s)\right| ds + \Gamma \int_{t_{j-1}}^t \left|u_j^\varepsilon(s) - u_j(s)\right| ds.
\end{aligned}$$

Now, using the Gronwall lemma, we obtain

$$\begin{aligned}
\left|x_j^\varepsilon(t) - x_j(t)\right| &\leq \left|x_j^\varepsilon(t_{j-1}) - x_j(t_{j-1})\right| e^{\Gamma(t_j - t_{j-1})} \\
&+ \Gamma \int_{t_{j-1}}^{t_j} \left|u_j^\varepsilon(s) - u_j(s)\right| ds\, e^{\Gamma(t_j - t_{j-1})}.
\end{aligned}$$

Obviously, by definition of admissible needle variation, the term $\left|x_j^\varepsilon(t_{j-1}) - x_j(t_{j-1})\right|$ tends to 0 as $\varepsilon \to 0$ and also the last term goes to 0 as $\varepsilon \to 0$. So we have obtained the thesis. □

LEMMA 3.3. *Let us assume* (3.1). *Let $\boldsymbol{X}^\varepsilon$ be an admissible needle variation. Then $v_j \equiv 0$ if $j < i$, $v_i(t) = 0$ if $t_{i-1} \leq t < \bar{\tau}$, and*

$$\begin{cases} \dot{v}_i(t) = D_x f_{q_i}(x_i(t), u_i(t)) v_i(t), \\ v_i(\bar{\tau}) = f_{q_i}(x_i(\bar{\tau}), \omega) - f_{q_i}(x_i(\bar{\tau}, u_i(\bar{\tau}))) \end{cases} \tag{3.5}$$

in the i-location if $\bar{\tau} \leq t \leq t_i$, while

$$\begin{cases} \dot{v}_j(t) = D_u f_{q_j}(x_j(t), u_j(t)) \theta_j(t) + D_x f_{q_j}(x_j(t), u_j(t)) v_j(t), \\ v_j(t_{j-1}) = \frac{d}{d\varepsilon} x_j^\varepsilon(t_{j-1})\big|_{\varepsilon=0} \end{cases} \tag{3.6}$$

if $j > i$.

Proof. Clearly, if $j < i$, then $v_j \equiv 0$. The case $j = i$ is well known, so we consider only the case $j > i$. In particular we prove the case $j = i + 1$, the other cases being similar.

For simplicity, let us denote with f, x, u, respectively, $f_{q_{i+1}}$, x_{i+1}, u_{i+1}. So, it is sufficient to prove that, if z in $[t_i, t_{i+1}]$ is the solution to

$$\begin{cases} \dot z(t) = D_u f(x(t), u(t))\theta_{i+1}(t) + D_x f(x(t), u(t))z(t), \\ z(t_i) = \frac{d}{d\varepsilon} x^\varepsilon_{i+1}(t_i)\big|_{\varepsilon=0}, \end{cases}$$

then $z(t) = \frac{d}{d\varepsilon} x^\varepsilon_{i+1}(t)\big|_{\varepsilon=0}$ for almost every $t \in [t_i, t_{i+1}]$. In order to prove this, for $t \in [t_i, t_{i+1}]$, we estimate

$$\begin{aligned} \left| \frac{x^\varepsilon_{i+1}(t) - x(t)}{\varepsilon} - z(t) \right| &\le \left| \frac{x^\varepsilon_{i+1}(t_i) - x(t_i)}{\varepsilon} - z(t_i) \right| \\ &+ \frac{1}{\varepsilon} \left| \int_{t_i}^t \left[f(x^\varepsilon_{i+1}(s), u^\varepsilon_{i+1}(s)) - f(x(s), u(s)) \right] ds \right. \\ &\left. - \varepsilon \int_{t_i}^t D_u f(x(s), u(s))\theta_{i+1}(s) ds - \varepsilon \int_{t_i}^t D_x f(x(s), u(s)) z(s) ds \right|. \end{aligned}$$

Fix $\delta > 0$. Then for ε sufficiently small (depending on δ), the first term of the right-hand side is less than δ. Using Taylor's expansion,

$$\begin{aligned} \left| \frac{x^\varepsilon_{i+1}(t) - x(t)}{\varepsilon} - z(t) \right| &\le \delta + \left| \int_{t_i}^t D_x f(x(s), u(s)) \cdot \left[\frac{x^\varepsilon_{i+1}(s) - x(s)}{\varepsilon} - z(s) \right] ds \right| \\ &+ \left| \int_{t_i}^t D_u f(x(s), u(s)) \cdot \left[\frac{u^\varepsilon_{i+1}(s) - u(s)}{\varepsilon} - \theta_{i+1}(s) \right] ds \right| \\ &+ \frac{c}{\varepsilon} \int_{t_i}^t \left[\left| x^\varepsilon_{i+1}(s) - x(s) \right| + \left| u^\varepsilon_{i+1}(s) - u(s) \right| \right]^2 ds, \end{aligned} \tag{3.7}$$

where c is a positive constant, depending on the second derivatives of f. Now $x^\varepsilon_{i+1} \to x$ uniformly, so for ε sufficiently small it holds that

$$\begin{aligned} \frac{c}{\varepsilon} \int_{t_i}^t \left| x^\varepsilon_{i+1}(s) - x(s) \right|^2 ds &\le c\delta \int_{t_i}^t \frac{\left| x^\varepsilon_{i+1}(s) - x(s) \right|}{\varepsilon} ds \\ &\le c\delta \int_{t_i}^t \left| \frac{x^\varepsilon_{i+1}(s) - x(s)}{\varepsilon} - z(s) \right| ds + c\delta \int_{t_i}^t |z(s)|\, ds \\ &\le c\delta \int_{t_i}^t \left| \frac{x^\varepsilon_{i+1}(s) - x(s)}{\varepsilon} - z(s) \right| ds + c_1 \delta \end{aligned}$$

with c_1 a positive constant. Moreover,

$$\frac{c}{\varepsilon} \int_{t_i}^t \left| u^\varepsilon_{i+1}(s) - u(s) \right|^2 ds \le c_2 \delta$$

with c_2 a positive constant, since $\frac{u^\varepsilon_{i+1} - u}{\varepsilon}$ converges weakly in $L^1([t_i, t_{i+1}])$ and $u^\varepsilon_{i+1} - u$ converges strongly to 0 in $L^1([t_i, t_{i+1}])$. For a detailed proof of this fact see the appendix. Analogously,

$$\frac{2c}{\varepsilon} \int_{t_i}^t \left| u^\varepsilon_{i+1}(s) - u(s) \right| \left| x^\varepsilon_{i+1}(s) - x(s) \right| ds \le c_2 \delta.$$

The third addend of the right-hand side of (3.7) is estimated similarly, since $\frac{u^\varepsilon_{i+1}-u}{\varepsilon} \rightharpoonup \theta_{i+1}$ weakly in $L^1([t_i, t_{i+1}])$. Thus,

$$\left|\frac{x^\varepsilon_{i+1}(t)-x(t)}{\varepsilon} - z(t)\right| \le M_1\delta + (M_2+c\delta)\int_{t_i}^{t}\left|\frac{x^\varepsilon_{i+1}(s)-x(s)}{\varepsilon} - z(t)\right| ds,$$

where M_1, M_2 are positive constants, depending on f and $U_{q_{i+1}}$. Using the Gronwall lemma we conclude that

$$\left|\frac{x^\varepsilon_{i+1}(t)-x(t)}{\varepsilon} - z(t)\right| \le M_1\delta e^{(M_2+c\delta)(t_{i+1}-t_i)} \tag{3.8}$$

and so the lemma is proved, by the arbitrariness of $\delta > 0$. □

Remark 5. From the last result, we note that the evolution equation for v_j in general is an affine equation, since a term depending on θ_j appears. This is due to the definition of admissible needle variation. For hybrid systems with assumption (H), we may consider usual needle variations and so the resulting equation for v_j is linear, without the term containing θ_j.

Remark 6. It is useful to recall that (3.6) is valid only if $j > i$, i.e., only if the variation is originated in a previous location. Therefore, to prove (3.6) we do not consider expression (3.3), but we use properties 3 and 5 of Definition 3.1.

Now, we want to evaluate how the Lagrangian cost varies for an admissible needle variation. If we define

$$G_\varepsilon(t) := \sum_{h=1}^{j-1}\int_{t_{h-1}}^{t_h} L_{q_h}(x^\varepsilon_h(s), u^\varepsilon_h(s))ds + \int_{t_{j-1}}^{t} L_{q_j}(x^\varepsilon_j(s), u^\varepsilon_j(s))ds$$

when $t_{j-1} \le t < t_j$ and set $w(t) := \frac{d}{d\varepsilon}G_\varepsilon(t)\big|_{\varepsilon=0^+}$, then we can get the following result.

LEMMA 3.4. *Let $\bar\tau \in]t_{i-1}, t_i[$ be the time at which an admissible needle variation originates. If $t \in]\bar\tau, t_i[$, then w satisfies the following differential equation:*

$$\begin{cases} \dot w(t) = \dfrac{\partial}{\partial x}L_{q_i}(x_i(t), u_i(t))v_i(t), \\ w(\bar\tau) = L_{q_i}(x_i(\bar\tau), \omega) - L_{q_i}(x_i(\bar\tau), u_i(\bar\tau)). \end{cases}$$

Moreover, if $i < j \le \nu$, then we have

$$\begin{cases} \dot w(t) = \dfrac{\partial}{\partial x}L_{q_j}(x_j(t), u_j(t))v_j(t) + \dfrac{\partial}{\partial u}L_{q_j}(x_j(t), u_j(t))\theta_j(t), \ t_{j-1} < t < t_j, \\ w(t_{j-1}) = \lim_{t\to t^-_{j-1}} w(t). \end{cases}$$

Proof. First we evaluate $w(\bar\tau)$. By definition

$$w(\bar\tau) = \frac{d}{d\varepsilon}G_\varepsilon(\bar\tau)\big|_{\varepsilon=0^+} = \lim_{\varepsilon=0^+}\frac{G_\varepsilon(\bar\tau) - G_0(\bar\tau)}{\varepsilon},$$

and, by the fact that x^ε and u^ε coincide, respectively, with x and u before $\bar\tau - \varepsilon$, we conclude that

$$\begin{aligned} w(\bar\tau) &= \lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\int_{\bar\tau-\varepsilon}^{\bar\tau}\left[L_{q_i}(x^\varepsilon_i(s), \omega) - L_{q_i}(x_i(s), u_i(s))\right] ds \\ &= L_{q_i}(x_i(\bar\tau), \omega) - L_{q_i}(x_i(\bar\tau), u_i(\bar\tau)). \end{aligned}$$

Now suppose that $\bar\tau < t < t_i$. In this case we have

$$\begin{aligned} w(t) &= \lim_{\varepsilon\to 0^+} \left\{ \frac{1}{\varepsilon} \int_{\bar\tau-\varepsilon}^{\bar\tau} [L_{q_i}(x_i^\varepsilon(s),\omega) - L_{q_i}(x_i^\varepsilon(s),u_i(s))]\,ds \right. \\ &\qquad \left. + \frac{1}{\varepsilon}\int_{\bar\tau}^{t} [L_{q_i}(x_i^\varepsilon(s),u_i(s)) - L_{q_i}(x_i(s),u_i(s))]\,ds \right\} \\ &= L_{q_i}(x_i(\bar\tau),\omega) - L_{q_i}(x_i(\bar\tau),u_i(\bar\tau)) \\ &\quad + \lim_{\varepsilon\to 0^+} \left\{ \int_{\bar\tau}^{t} \frac{\partial}{\partial x} L_{q_i}(x_i(s),u_i(s)) \frac{x_i^\varepsilon(s)-x_i(s)}{\varepsilon} ds \right. \\ &\qquad \left. + \int_{\bar\tau}^{t} \frac{\partial^2}{\partial x^2} L_{q_i}(\tilde x_i(s),\tilde u(s)) \frac{(x_i^\varepsilon(s)-x_i(s))^2}{2\varepsilon} ds \right\}. \end{aligned}$$

Using estimate (3.8) we conclude by the Lebesgue theorem that

$$w(t) = L_{q_i}(x_i(\bar\tau),\omega) - L_{q_i}(x_i(\bar\tau),u_i(\bar\tau)) + \int_{\bar\tau}^{t} \frac{\partial}{\partial x} L_{q_i}(x_i(s),u_i(s)) v_i(s) ds.$$

So the first part of the lemma is proved.

In order to prove the last statement, we claim that, if $t_{j-1} < t < t_j$, then $w(t)$ is equal to

$$\begin{aligned} w(t) &= [L_{q_i}(x_i(\bar\tau),\omega) - L_{q_i}(x_i(\bar\tau),u_i(\bar\tau))] \\ &+ \int_{\bar\tau}^{t_i} \frac{\partial}{\partial x} L_{q_i}(x_i(s),u_i(s)) v_i(s) ds + \sum_{l=i+1}^{j-1} \int_{t_{l-1}}^{t_l} \frac{\partial}{\partial x} L_{q_l}(x_l(s),u_l(s)) v_l(s) ds \\ &+ \sum_{l=i+1}^{j-1} \int_{t_{l-1}}^{t_l} \frac{\partial}{\partial u} L_{q_l}(x_l(s),u_l(s)) \theta_l(s) ds \\ &+ \int_{t_{j-1}}^{t} \frac{\partial}{\partial x} L_{q_j}(x_j(s),u_j(s)) v_j(s) ds + \int_{t_{j-1}}^{t} \frac{\partial}{\partial u} L_{q_j}(x_j(s),u_j(s)) \theta_j(s) ds. \end{aligned}$$

In fact, if $l \in \{i+1,\dots,j\}$, then we have for $t \in [t_{l-1}, t_l]$ that

$$\begin{aligned} &\lim_{\varepsilon\to 0^+} \int_{t_{l-1}}^{t} \frac{L_{q_l}(x_l^\varepsilon(s),u_l^\varepsilon(s)) - L_{q_l}(x_l(s),u_l(s))}{\varepsilon} ds \\ &\quad = \lim_{\varepsilon\to 0^+} \left\{ \int_{t_{l-1}}^{t} \frac{L_{q_l}(x_l^\varepsilon(s),u_l^\varepsilon(s)) - L_{q_l}(x_l(s),u_l^\varepsilon(s))}{\varepsilon} ds \right. \\ &\qquad \left. + \int_{t_{l-1}}^{t} \frac{L_{q_l}(x_l(s),u_l^\varepsilon(s)) - L_{q_l}(x_l(s),u_l(s))}{\varepsilon} ds \right\} \\ &\quad = \int_{t_{l-1}}^{t} \frac{\partial}{\partial x} L_{q_l}(x_l(s),u_l(s)) v_l(s) ds \\ &\qquad + \lim_{\varepsilon\to 0^+} \int_{t_{l-1}}^{t} \frac{L_{q_l}(x_l(s),u_l^\varepsilon(s)) - L_{q_l}(x_l(s),u_l(s))}{\varepsilon} ds. \end{aligned}$$

Now the last term is equal to

$$\begin{aligned}
&\lim_{\varepsilon\to 0^+}\int_{t_{l-1}}^{t}\frac{L_{q_l}(x_l(s),u_l^\varepsilon(s))-L_{q_l}(x_l(s),u_l(s))}{\varepsilon}ds\\
&\quad=\lim_{\varepsilon\to 0^+}\left\{\int_{t_{l-1}}^{t}\frac{\partial}{\partial u}L_{q_l}(x_l(s),u_l(s))\frac{u_l^\varepsilon(s)-u_l(s)}{\varepsilon}ds\right.\\
&\qquad\qquad\left.+\int_{t_{l-1}}^{t}\frac{\partial^2}{\partial u^2}L_{q_l}(\tilde{x}_l(s),\tilde{u}_l(s))\frac{u_l^\varepsilon(s)-u_l(s)}{\varepsilon}(u_l^\varepsilon(s)-u_l(s))ds\right\}\\
&\quad=\int_{t_{l-1}}^{t}\frac{\partial}{\partial u}L_{q_l}(x_l(s),u_l(s))\theta_l(s)ds\\
&\qquad+\lim_{\varepsilon\to 0^+}\int_{t_{l-1}}^{t}\frac{\partial^2}{\partial u^2}L_{q_l}(\tilde{x}_l(s),\tilde{u}_l(s))\frac{u_l^\varepsilon(s)-u_l(s)}{\varepsilon}(u_l^\varepsilon(s)-u_l(s))ds
\end{aligned}$$

by definition of admissible needle variation. The last integral converges to 0 since $\frac{u_l^\varepsilon(s)-u_l(s)}{\varepsilon}$ converges weakly in $L^1([t_{l-1},t_l])$ and $u_l^\varepsilon(s)-u_l(s)$ converges to 0 strongly in $L^1([t_{l-1},t_l])$ and so the product converges to 0 weakly in $L^1([t_{l-1},t_l])$; see the appendix. □

Putting together all the previous results we have the following proposition.

PROPOSITION 3.5. *Let $\boldsymbol{X}$ be a trajectory and let $\boldsymbol{X}^\varepsilon$ be an admissible needle variation. Then, for every adjoint pair (ψ,ψ_0) along $\boldsymbol{X}$ and for every $j\in\{1,\dots,\nu\}$ the function*

$$\psi_j(t)\cdot v_j(t)-\psi_0 w(t)+q_j(t) \tag{3.9}$$

is constant in $[t_{j-1},t_j]$, where v_j is defined by (3.4) *and q_j is any function defined by*

$$\dot{q}_j(t)=-\psi_j(t)\frac{\partial}{\partial u}f_{q_j}(x_j(t),u_j(t))\theta_j(t)+\psi_0\frac{\partial}{\partial u}L_{q_j}(x_j(t),u_j(t))\theta_j(t) \tag{3.10}$$

if $j>i$, while $q_j\equiv 0$ otherwise.

Proof. It is a simple consequence of Lemmas 3.3 and 3.4. If $j<i$, then $v_j\equiv 0$, $q_j\equiv 0$, and $w(t)=0$ for every $t\in[0,t_j]$.

If $j=i$, where q_i is the location at which the admissible needle variation originates, then we have

$$\begin{aligned}
&\frac{d}{dt}\left[\psi_i(t)\cdot v_i(t)-\psi_0 w(t)\right]=\dot{\psi}_i(t)\cdot v_i(t)+\psi_i(t)\cdot\dot{v}_i(t)-\psi_0\dot{w}(t)\\
&\quad=-\psi_i(t)D_x f_{q_i}(x_i(t),u_i(t))v_i(t)+\psi_0\frac{\partial}{\partial x}L_{q_i}(x_i(t),u_i(t))v_i(t)\\
&\quad+\psi_i(t)D_x f_{q_i}(x_i(t),u_i(t))v_i(t)-\psi_0\frac{\partial}{\partial x}L_{q_i}(x_i(t),u_i(t))v_i(t)=0
\end{aligned}$$

and so we have the thesis when $j=i$.

Now if $j > i$, then

$$\begin{aligned}
&\frac{d}{dt}\left[\psi_j(t)\cdot v_j(t) - \psi_0 w(t) + q_j(t)\right] = \dot\psi_j(t)\cdot v_j(t) + \psi_j(t)\cdot \dot v_j(t) - \psi_0 \dot w(t) + \dot q_j(t)\\
&\quad = -\psi_j(t) D_x f_{q_j}(x_j(t),u_j(t))v_j(t) + \psi_0 \frac{\partial}{\partial x} L_{q_j}(x_j(t),u_j(t))v_j(t)\\
&\quad + \psi_j(t) D_u f_{q_j}(x_j(t),u_j(t))\theta_j(t) + \psi_j(t) D_x f_{q_j}(x_j(t),u_j(t))v_j(t)\\
&\quad - \psi_0 \frac{\partial}{\partial x} L_{q_j}(x_j(t),u_j(t))v_j(t) - \psi_0 \frac{\partial}{\partial u} L_{q_j}(x_j(t),u_j(t))\theta_j(t)\\
&\quad - \psi_j(t)\cdot D_u f_{q_j}(x_j(t),u_j(t))\theta_j(t) + \psi_0 \frac{\partial}{\partial u} L_{q_j}(x_j(t),u_j(t))\theta_j(t) = 0.
\end{aligned}$$

This completes the proof. □

Now, we study how to deduce some necessary conditions from the previous analysis. For clarity, we consider optimal control problems where the cost is formed only by the Lagrangian part; that is, the switching cost and the endpoint cost vanish. We suppose that $\mathbf{X}$ is an optimal trajectory and we consider an admissible needle variation $\mathbf{X}^\varepsilon$. Clearly, by optimality, $C(\mathbf{X}) \le C(\mathbf{X}^\varepsilon)$. This implies that $w(T) \ge 0$. Let us consider an adjoint pair (ψ, ψ_0) along $\mathbf{X}$ with the property that, for every $j \in \{1,\dots,\nu\}$,

$$\psi_j(t_j)\cdot v_j(t_j) \le 0. \tag{3.11}$$

Thus

$$\psi_\nu(t_\nu)\cdot v_\nu(t_\nu) - \psi_0 w(t_\nu) \le 0. \tag{3.12}$$

This implies that, for every $q_\nu(\cdot)$ defined as in Proposition 3.5 with $q_\nu(t_\nu) \le 0$, it holds that

$$\psi_\nu(t)\cdot v_\nu(t) - \psi_0 w(t) + q_\nu(t) \le 0 \tag{3.13}$$

for every $t \in [t_{\nu-1}, t_\nu]$. Therefore in the $\nu - 1$ location we have

$$\psi_{\nu-1}(t_{\nu-1})\cdot v_{\nu-1}(t_{\nu-1}) - \psi_0 w(t_{\nu-1}) + q_\nu(t_{\nu-1}) \le 0 \tag{3.14}$$

and so

$$\psi_{\nu-1}(t)\cdot v_{\nu-1}(t) - \psi_0 w(t) + q_\nu(t_{\nu-1}) + q_{\nu-1}(t) \le 0 \tag{3.15}$$

for every $q_{\nu-1}$ with $q_{\nu-1}(t_{\nu-1}) \le 0$ and for every $t \in [t_{\nu-2}, t_{\nu-1}]$.

Iterating this argument we conclude that

$$\psi_j(t)\cdot v_j(t) - \psi_0 w(t) + \sum_{l=j+1}^{\nu} q_l(t_{l-1}) + q_j(t) \le 0 \tag{3.16}$$

for every $j \in \{1,\dots,\nu\}$, $t \in [t_{j-1}, t_j]$ and for every function q_l with $q_l(t_l) \le 0$.

Equation (3.16) gives a necessary condition for optimality when the hybrid system does not satisfy assumption (H). In the next section we generalize this approach.

4. Hybrid necessary principle. This section is devoted to the statement and proof of a general HNP in the case where the lack of Assumption (H) generalizes some previous results in [27]. Again for sake of simplicity we assume (3.1) and (3.2).

We recall first some basic facts about measure theory and in particular about Radon measures; see [16].

DEFINITION 4.1. *Let $a, b \in \mathbb{R}$, $a < b$. If $\mathcal{B}_{(a,b)}$ denotes the Borel σ-algebra on (a, b), then a signed measure $\mu : \mathcal{B}_{(a,b)} \to \overline{\mathbb{R}}$ is called a Radon measure if*

1. *$|\mu(K)| < +\infty$ for every K, compact set in (a, b);*
2. *$\mu(E) = \inf\{\mu(U) : E \subseteq U, U$ open set in $(a, b)\}$ for every $E \in \mathcal{B}_{(a,b)}$;*
3. *$\mu(E) = \sup\{\mu(K) : K \subseteq E, K$ compact set in $(a, b)\}$ for every $E \in \mathcal{B}_{(a,b)}$.*

Moreover, if $|\mu|$ is finite, then μ is said to be bounded. We denote by $\mathfrak{M}_b(a, b; \mathbb{R})$ the set of bounded Radon measures on (a, b).

If $a, b \in \mathbb{R}$, $a < b$, then $L^1(a, b)$ is contained in $\mathfrak{M}_b(a, b; \mathbb{R})$. The inclusion is intended in the following way: to every function $f \in L^1(a, b)$ we associate a measure μ defined by

$$\mu(A) := \int_A f(t)dt, \tag{4.1}$$

where A is a Borel subset of (a, b). The space $\mathfrak{M}_b(a, b; \mathbb{R})$, equipped with the norm $\|\mu\| := |\mu|(a, b)$, is equal to $(C(a, b))'$ (see [16]). So in $\mathfrak{M}_b(a, b; \mathbb{R})$ we may consider the weak* topology. We have that $\mu_n \rightharpoonup^* \mu$ as $n \to +\infty$ if and only if, for every $g \in C(a, b)$,

$$\int_a^b g(t)d\mu_n(t) \to \int_a^b g(t)d\mu(t) \qquad \text{as } n \to +\infty.$$

In the same way, we consider the space of bounded Radon measures on (a, b) with values in $\mathbb{R}^n$ and we indicate them by $\mathfrak{M}_b(a, b; \mathbb{R}^n)$.

Remark 7. The measure μ defined in (4.1) is clearly a signed measure. In fact if the function f is strictly negative in a Borel set E, whose Lebesgue measure is strictly positive, then $\mu(E) < 0$.

All the properties of Radon measures are clearly satisfied by any measure defined as in (4.1). Moreover, equation (4.1) says that $\mu \ll m$; that is, μ is an absolutely continuous measure with respect to the Lebesgue measure.

Now we recall a result about the differentiability of a trajectory with respect to a parameter, used to prove the main result.

DEFINITION 4.2. *Let P be a normed space, let $\{\bar{x}^p\}_{p\in P}$ be a family in $\mathbb{R}^n$, and let $\boldsymbol{f} = \{\boldsymbol{f}_p\}_{p\in P}$ be a family of time-varying vector fields on $\mathbb{R}^n$. For every $p \in P$, we denote with $x^p : [a, b] \to \mathbb{R}^n$ a solution to*

$$\begin{cases} \dot{x}^p(t) = \boldsymbol{f}_p(t, x^p(t)) & a.e.\ t \in [a, b], \\ x^p(a) = \bar{x}^p. \end{cases}$$

Let us fix an element $p_0 \in P$. We say that $\boldsymbol{f}$ is weakly differentiable at $p_0 \in P$ along $x^{p_0}(\cdot)$ if there exists $\bar{\varepsilon} > 0$ such that

1. *for every p such that $\|p - p_0\| \le \bar{\varepsilon}$ and for every $\xi : [a, b[\to \mathbb{R}^{d_q}$, solution to $\dot{\xi}(t) = \boldsymbol{f}_p(t, \xi(t))$, with $\|\xi(t) - x^{p_0}(t)\| \le \bar{\varepsilon}$ for every $t \in [a, b[$, the limit $\lim_{t\to b^-} \xi(t)$ exists;*

2. *for every p such that $\|p-p_0\| \le \bar\varepsilon$ and for every x such that $\|x-x^{p_0}(t)\| < \bar\varepsilon$ for some $a \le t < b$, there exists a local forward solution to*

$$\begin{cases} \dot\xi = \boldsymbol{f}_p(t,\xi), \\ \xi(t) = x; \end{cases}$$

3. *there exist $A \in L^1([a,b];\mathbb{R}^{n\times n})$ and positive functions $\zeta_\varepsilon \in L^1([a,b];\mathbb{R}^+)$ for every $\varepsilon \in]0,\bar\varepsilon]$, such that*

$$\begin{aligned} &\|\boldsymbol{f}_p(t,x) - \boldsymbol{f}_p(t,x^{p_0}(t)) - A(t)(x-x^{p_0}(t))\| \\ &\qquad \le \zeta_\varepsilon(t)(\|x-x^{p_0}(t)\| + \|p-p_0\|) \end{aligned}$$

whenever $a \le t \le b$, $\|x-x^{p_0}(t)\| \le \varepsilon$, $\|p-p_0\| \le \varepsilon$, and

$$\lim_{\varepsilon\to 0^+} \int_a^b \zeta_\varepsilon(t)dt = 0;$$

4. *the function $\tilde w_p(t) := \boldsymbol{f}_p(t,x^{p_0}(t)) - \boldsymbol{f}_{p_0}(t,x^{p_0}(t))$ is integrable in $[a,b]$ for every $\|p-p_0\| \le \bar\varepsilon$;*

5. *for every $\alpha \in C([a,b];\mathbb{R}^n)$ the map*

$$p \mapsto \int_{[a,b]} \alpha(t)\, d\left(\int_a^t \tilde w_p(s)ds\right)$$

is Fréchet differentiable at p_0.

PROPOSITION 4.3. *We use here the notation of Definition* 4.2. *Fix $p_0 \in P$, let $v \in P$, and assume that*

$$\lim_{\varepsilon\to 0^+} \frac{\bar x^{p_0+\varepsilon v} - \bar x^{p_0}}{\varepsilon}$$

exists in $\mathbb{R}^n$ and moreover that there exists $\alpha_{p_0,v}(\cdot) \in \mathfrak{M}_b(a,b;\mathbb{R}^n)$ such that

$$\frac{\boldsymbol{f}_{p_0+\varepsilon v}(t,x^{p_0}(t)) - \boldsymbol{f}_{p_0}(t,x^{p_0}(t))}{\varepsilon} \rightharpoonup^* \alpha_{p_0,v}(t) \tag{4.2}$$

as $\varepsilon \to 0^+$. If we denote by $y_{p_0,v}(t)$ the derivative of $x^{p_0}(t)$ in the v-direction evaluated at p_0, then we have that

$$\dot y_{p_0,v}(t) = \frac{\partial}{\partial x}\boldsymbol{f}_{p_0}(t,x^{p_0}(t)) \cdot y_{p_0,v}(t) + \alpha_{p_0,v}(t), \tag{4.3}$$

where the last equation is to be intended in the integral sense. Moreover, if $\boldsymbol{f}$ is weakly differentiable at p_0 along $x^{p_0}(\cdot)$, then $x^p(\cdot) \to x^{p_0}(\cdot)$ uniformly on $[a,b]$ when $p \to p_0$ in P.

For a proof, see [22, Appendix A].

Remark 8. The previous proposition gives us a tool to evaluate the evolution of the derivative of a trajectory of an ODE with respect to a parameter. In particular, it is useful in order to understand the behavior of modifications due to "variations" on a supposed optimal trajectory. We apply the previous proposition in the case where the time-varying vector fields are the coupled dynamic-Lagrangian functions evaluated on controls produced by a variation. Notice that the limit in (4.2) is in the weak star topology of Radon measures.

DEFINITION 4.4. *Let X, Y be finite dimensional vector spaces over $\mathbb{R}$ and let Λ be a cone in X. Consider a function $f : x + \Lambda \to Y$ for some $x \in X$. We say that f is differentiable at x in the direction Λ if there exists a linear map $D_\Lambda f(x) : X \to Y$ such that*

$$f(x+\lambda) = f(x) + D_\Lambda f(x)\cdot\lambda + o(\|\lambda\|) \quad \text{as } \lambda \to 0,\ \lambda \in \Lambda. \tag{4.4}$$

Obviously the map $D_\Lambda f(x)$ is uniquely determined on $\operatorname{span}\{\Lambda\}$.

Let $\mathbf{X}$ be an optimal trajectory for the problem $\mathcal{P}$ and let $\bar\varepsilon > 0$. We denote with K a cone in $\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_\nu}$, with $v = (v_1, \dots, v_\nu)$ an element of K and with $(u_1, \dots, u_\nu)$ the controls of the candidate optimal hybrid trajectory $\mathbf{X}$. The aim of next definition is to give a rigorous description of all variations we are able to consider. Analogously to [27], we treat variations depending on two parameters: ε and v. ε is a real positive number, while v belongs to a cone in a finite dimensional manifold. The reader can think of v as the parameter responsible for the variation of the initial points of each trajectory x_j, $j \in \{1, \dots, \nu\}$, and ε as parameterizing the control variation.

DEFINITION 4.5 (map of variations). *A map V defined on $[0, \bar\varepsilon] \times K$, $V(\varepsilon, v) = (x_1^{(\varepsilon,v)}, u_1^{(\varepsilon,v)}, \dots, x_\nu^{(\varepsilon,v)}, u_\nu^{(\varepsilon,v)})$, is called a map of variations if, for every $(\varepsilon, v) \in [0, \bar\varepsilon] \times K$, the following hold:*

1. *for every $i \in \{1, \dots, \nu\}$, $u_i^{(\varepsilon,v)} \in \mathcal{U}_{q_i}$ and $u_i^{(\delta\varepsilon,\delta v)} \to u_i$ in $L^1(t_{i-1}, t_i)$ as $\delta \to 0^+$;*
2. *for every $i \in \{1, \dots, \nu\}$, $x_i^{(\varepsilon,v)} :]t_{i-1}, t_i[\to \mathbb{R}^{d_i}$ is an absolutely continuous function continuously prolongable to $[t_{i-1}, t_i]$ such that $\frac{d}{d\delta} x_i^{(\delta\varepsilon,\delta v)}(t_{i-1})\big|_{\delta=0} = v_i$ and*

$$\frac{d}{dt} x_i^{(\varepsilon,v)}(t) = f_{q_i}(x_i^{(\varepsilon,v)}(t), u_i^{(\varepsilon,v)}(t)) \quad \text{for a.e. } t \in [t_{i-1}, t_i]; \tag{4.5}$$

3. *for every $i \in \{1, \dots, \nu - 1\}$, $u_{i+1}^{(\varepsilon,v)} \in \mathcal{U}_{q_i, x_i^{(\varepsilon,v)}(t_i), q_{i+1}, x_{i+1}^{(\varepsilon,v)}(t_i)}$;*
4. *the map*

$$\begin{aligned} \widetilde{C}_V : [0,\bar\varepsilon]\times K &\to (\mathbb{R}^{d_1}\times\mathbb{R}^{d_2})\times\cdots\times(\mathbb{R}^{d_\nu}\times\mathbb{R}^{d_1})\times\mathbb{R} \\ (\varepsilon, v) &\mapsto ((x_1^{(\varepsilon,v)}(t_1), x_2^{(\varepsilon,v)}(t_1)), \dots, (x_\nu^{(\varepsilon,v)}(t_\nu), x_1^{(\varepsilon,v)}(t_0)), \gamma(\varepsilon, v)) \end{aligned} \tag{4.6}$$

with

$$\gamma(\varepsilon, v) = \sum_{i=1}^{\nu} \int_{t_{i-1}}^{t_i} L_{q_i}(x_i^{(\varepsilon,v)}(t), u_i^{(\varepsilon,v)}(t))dt$$

is differentiable at 0 in the direction $\mathbb{R}^+ \times K$;

5. *for every $i \in \{1, \dots, \nu\}$, there exist $\alpha_{i,f,V}^{(\varepsilon,v)} \in \mathfrak{M}_b(t_{i-1}, t_i; \mathbb{R}^{d_i})$ and $\alpha_{i,L,V}^{(\varepsilon,v)} \in \mathfrak{M}_b(t_{i-1}, t_i; \mathbb{R})$ such that*

$$\frac{f_{q_i}(x_i(t), u_i^{(\delta\varepsilon,\delta v)}(t)) - f_{q_i}(x_i(t), u_i(t))}{\delta} \rightharpoonup^* \alpha_{i,f,V}^{(\varepsilon,v)}(t) \tag{4.7}$$

and

$$\frac{L_{q_i}(x_i(t), u_i^{(\delta\varepsilon,\delta v)}(t)) - L_{q_i}(x_i(t), u_i(t))}{\delta} \rightharpoonup^* \alpha_{i,L,V}^{(\varepsilon,v)}(t) \tag{4.8}$$

as $\delta \downarrow 0$.

We denote by $\mathcal{V}$ the set of all maps of variations.

Remark 9. The main reasoning in the proof of the HNP follows the classical approach. We consider *feasible* cones generated by final points of admissible variations and *profitable* cones formed by points that realize a cost lower than that of the candidate trajectory. To have optimality these two cones must be separated; i.e., there exists a hyperplane that separates the cones. From these considerations we deduce necessary conditions.

In Definition 3.1 (map of variations), we require various assumptions. In particular, the assumptions 1, 2, and 3 guarantee that maps of variations produce admissible trajectories for our hybrid system. Moreover assumption 4 implies the existence of the cone generated by the variations and, finally, assumption 5 is necessary in order to apply Proposition 4.3.

Remark 10. Notice that in order to have the differentiability of the function $\widetilde{C}_V$, the Radon measures $\alpha_{i,f,V}^{(\varepsilon,v)}$ and $\alpha_{i,L,V}^{(\varepsilon,v)}$ must depend continuously on the parameters (ε, v). This is guaranteed if $\alpha_{i,f,V}^{(\varepsilon,v)}$ and $\alpha_{i,L,V}^{(\varepsilon,v)}$ are linear with respect to the parameters (ε, v), and if $(\varepsilon_n, v_n) \in \mathbb{R}^+ \times K$ with $(\varepsilon_n, v_n) \to (\varepsilon, v) \in \mathbb{R}^+ \times K$, then $\alpha_{i,f,V}^{(\varepsilon_n,v_n)} \rightharpoonup^* \alpha_{i,f,V}^{(\varepsilon,v)}$ in $\mathfrak{M}_b(t_{i-1}, t_i; \mathbb{R}^{d_i})$ for every $i = 1, \dots, \nu$ and $\alpha_{i,L,V}^{(\varepsilon_n,v_n)} \rightharpoonup^* \alpha_{i,L,V}^{(\varepsilon,v)}$ in $\mathfrak{M}_b(t_{i-1}, t_i; \mathbb{R})$ for every $i = 1, \dots, \nu$.

Now, if $V \in \mathcal{V}$, then we have that

$$D\widetilde{C}_V(0) : \mathbb{R}^+ \times K \to (\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}) \times \cdots \times (\mathbb{R}^{d_\nu} \times \mathbb{R}^{d_1}) \times \mathbb{R}$$

and if $\varepsilon \in [0, \bar{\varepsilon}]$, then

$$D\widetilde{C}_V(0)(\varepsilon, v) = ((w_1, v_2), \dots, (w_\nu, v_1), \beta(\varepsilon, v)), \tag{4.9}$$

where

$$\begin{aligned}\beta(\varepsilon, v) = &\sum_{i=1}^{\nu} \int_{t_{i-1}}^{t_i} \frac{\partial}{\partial x} L_{q_i}(x_i(s), u_i(s)) M_i(s, t_{i-1}) v_i \, ds \\ &+ \sum_{i=1}^{\nu} \int_{t_{i-1}}^{t_i} \frac{\partial}{\partial x} L_{q_i}(x_i(s), u_i(s)) \int_{t_{i-1}}^{s} M_i(s, r) \, d\alpha_{i,f,V}^{(\varepsilon,v)}(r) \, ds \\ &+ \sum_{i=1}^{\nu} \int_{t_{i-1}}^{t_i} d\alpha_{i,L,V}^{(\varepsilon,v)}(s), \end{aligned} \tag{4.10}$$

$$w_i = M_i(t_i, t_{i-1}) v_i + \int_{t_{i-1}}^{t_i} M_i(t_i, s) d\alpha_{i,f,V}^{(\varepsilon,v)}(s), \tag{4.11}$$

and $M_i(t, s)$ $(i = 1, \dots, \nu)$ is the fundamental matrix solution for the linear system

$$\dot{y}(t) = \frac{\partial}{\partial x} f_{q_i}(x_i(t), u_i(t)) y(t).$$

Indeed, the differential of the components $x_i^{(\varepsilon,v)}(t_{i-1})$ is equal to v_i by hypothesis 2 of the definition of map of variations. For the differential of the components $x_i^{(\varepsilon,v)}(t_i)$ we use Proposition 4.3, while for the differential of $\gamma(\varepsilon, v)$ we have to estimate, for

every $i = 1, \dots, \nu$, the limit as $\delta \to 0^+$ of the expression

$$\begin{aligned}&\int_{t_{i-1}}^{t_i} \frac{L_{q_i}(x_i^{(\delta\varepsilon,\delta v)}(t), u_i^{(\delta\varepsilon,\delta v)}(t)) - L_{q_i}(x_i(t), u_i(t))}{\delta} dt \\ &= \int_{t_{i-1}}^{t_i} \frac{L_{q_i}(x_i^{(\delta\varepsilon,\delta v)}(t), u_i^{(\delta\varepsilon,\delta v)}(t)) - L_{q_i}(x_i(t), u_i^{(\delta\varepsilon,\delta v)}(t))}{\delta} dt \\ &\quad + \int_{t_{i-1}}^{t_i} \frac{L_{q_i}(x_i(t), u_i^{(\delta\varepsilon,\delta v)}(t)) - L_{q_i}(x_i(t), u_i(t))}{\delta} dt.\end{aligned}$$

For the last addend we use hypothesis 5 of the definition of map of variations, while for the other term we have to use Proposition 4.3 and Lemma A.2 of the appendix.

Let us denote by $\mathbf{X}^{(\varepsilon,v)}(\cdot)$ the candidate hybrid trajectory obtained by piecing together $x_i^{(\varepsilon,v)}$, $i = 1, \dots, \nu$. Then $\mathbf{X}^{(\varepsilon,v)}(\cdot)$ is a trajectory if and only if

- $(x_i^{(\varepsilon,v)}(t_i), x_{i+1}^{(\varepsilon,v)}(t_i)) \in \mathcal{S}_{q_i,q_{i+1}}$ for $i = 1, \dots, \nu - 1$;
- $(q_1, x_1^{(\varepsilon,v)}(t_0), 0) \in \mathcal{N}_{in}$;
- $(q_\nu, x_\nu^{(\varepsilon,v)}(t_\nu), t_\nu - t_{\nu-1}) \in \mathcal{N}_{fin}$.

Since $\mathbf{X}$ is optimal we have that

$$C(\mathbf{X}^{(\varepsilon,v)}) \geq C(\mathbf{X}) \tag{4.12}$$

whenever the previous conditions hold.

In what follows we identify, for simplicity, $x_{\nu+1}(t_\nu)$ with $x_1(t_0)$ and $\mathbb{R}^{d_{\nu+1}}$ with $\mathbb{R}^{d_1}$. Moreover,

$$\mathcal{S}_{q_\nu,q_{\nu+1}} := \{(z, z') \in \mathbb{R}^{d_\nu} \times \mathbb{R}^{d_1} : (q_1, z', 0) \in \mathcal{N}_{in}, (q_\nu, z, t_\nu - t_{\nu-1}) \in \mathcal{N}_{fin}\}.$$

Now, fix smooth functions $\sigma_i : \mathbb{R}^{d_i} \times \mathbb{R}^{d_{i+1}} \to \mathbb{R}$ $(i = 1, \dots, \nu)$ such that

$$\sigma_i(x_i(t_i), x_{i+1}(t_i)) = 0$$

and $\sigma_i(z_i, z_i') > 0$ if $(z_i, z_i') \in \mathbb{R}^{d_i} \times \mathbb{R}^{d_{i+1}} \setminus \{(x_i(t_i), x_{i+1}(t_i))\}$. Let P be the set of points $((z_1, z_1'), \dots, (z_\nu, z_\nu'), r)$ of $(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}) \times \cdots \times (\mathbb{R}^{d_\nu} \times \mathbb{R}^{d_1}) \times \mathbb{R}$ such that $(z_i, z_i') \in \mathcal{S}_{q_i,q_{i+1}}$ $(i = 1, \dots, \nu)$ and

$$r \leq C(\mathbf{X}) - \sum_{i=1}^{\nu-1} \Phi_{q_i,q_{i+1}}(z_i, z_i') - \widetilde{\varphi}_{q_1,q_\nu}(z_\nu, z_\nu') - \sum_{i=1}^{\nu} \sigma_i(z_i, z_i'),$$

where $\widetilde{\varphi}_{q_1,q_\nu}(z_\nu, z_\nu') = \varphi_{q_1,q_\nu}(z_\nu', z_\nu)$. Notice that if $\widetilde{C}_V(\varepsilon, v) \in P$, then $\mathbf{X}^{(\varepsilon,v)}$ is a hybrid trajectory and $C(\mathbf{X}^{(\varepsilon,v)}) \leq C(\mathbf{X})$ with strict inequality if $\mathbf{X}^{(\varepsilon,v)} \not\equiv \mathbf{X}$; then from (4.12), we get $\widetilde{C}_V([0, \bar{\varepsilon}] \times K) \cap P = \{p_*\}$, where

$$p_* = ((x_1(t_1), x_2(t_1)), \dots, (x_\nu(t_\nu), x_1(t_0)), C_L(\mathbf{X})).$$

Define

$$K_V = D\widetilde{C}_V(0)([0, \bar{\varepsilon}] \times K). \tag{4.13}$$

Let K_P be the set of all $((z_1, z_1'), \dots, (z_\nu, z_\nu'), r)$ such that (z_i, z_i'), for $i = 1, \dots, \nu$, belongs to a Boltyanskiĭ approximating cone to $\mathcal{S}_{q_i,q_{i+1}}$ at the point $(x_i(t_i), x_{i+1}(t_i))$

and

$$r \le -\sum_{i=1}^{\nu-1} \nabla\Phi_{q_i,q_{i+1}}(x_i(t_i), x_{i+1}(t_i)) \cdot (z_i, z_i') - \nabla\widetilde{\varphi}_{q_1,q_\nu}(x_\nu(t_\nu), x_1(t_0)) \cdot (z_\nu, z_\nu').$$

Then K_P is a Boltyanskiĭ approximating cone to P at p_* and is not a linear subspace. By a general separation theorem (see [28]), for every convex cone $\widehat{K} \subseteq \bigcup_{V\in\mathcal{V}} K_V$, there exists an element $\psi \in (\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}) \times \cdots \times (\mathbb{R}^{d_\nu} \times \mathbb{R}^{d_1}) \times \mathbb{R}$ such that it weakly separates K_P from $\widehat{K}$. In particular we may suppose that

$$(\psi, k) \ge 0 \quad \forall k \in K_P$$

and

$$(\psi, k') \le 0 \quad \forall k' \in \widehat{K},$$

where $(\cdot,\cdot)$ denotes the usual scalar product in $\mathbb{R}^n$. We may write

$$\psi = ((\psi_1^+, \psi_2^-), \dots, (\psi_\nu^+, \psi_1^-), -\psi_0).$$

We note that $(0, \dots, 0, -1) \in K_P$ and so $\psi_0 \ge 0$.

If $V \in \mathcal{V}$ and $(\varepsilon, v) \in [0, \bar{\varepsilon}] \times K$ is such that $D\widetilde{C}_V(0)(\varepsilon, v) \in \widehat{K}$, then

$$\begin{aligned}
&\sum_{i=1}^{\nu} \left(\psi_i^- \cdot v_i + \psi_i^+ \cdot M_i(t_i, t_{i-1}) \cdot v_i + \psi_i^+ \int_{t_{i-1}}^{t_i} M_i(t_i, s) d\alpha_{i,f,V}^{(\varepsilon,v)}(s) \right) \\
&\quad - \psi_0 \sum_{i=1}^{\nu} \int_{t_{i-1}}^{t_i} \frac{\partial}{\partial x} L_{q_i}(x_i(s), u_i(s)) M_i(s, t_{i-1}) v_i ds \\
&\quad - \psi_0 \sum_{i=1}^{\nu} \int_{t_{i-1}}^{t_i} \frac{\partial}{\partial x} L_{q_i}(x_i(s), u_i(s)) \int_{t_{i-1}}^{s} M_i(s, r) d\alpha_{i,f,V}^{(\varepsilon,v)}(r) ds \\
&\quad - \psi_0 \sum_{i=1}^{\nu} \int_{t_{i-1}}^{t_i} d\alpha_{i,L,V}^{(\varepsilon,v)}(s) \le 0.
\end{aligned}$$

Hence, if we define, for every $i = 1, \dots, \nu$, $\psi_i : [t_{i-1}, t_i] \to \mathbb{R}^{d_i}$ to be the Carathéodory solution to

$$\begin{cases} \dot{\psi}_i(t) = -\psi_i(t) \dfrac{\partial}{\partial x} f_{q_i}(x_i(t), u_i(t)) + \psi_0 \dfrac{\partial}{\partial x} L_{q_i}(x_i(t), u_i(t)), \\ \psi_i(t_i) = \psi_i^+, \end{cases}$$

then we obtain that

$$\begin{aligned}
&\sum_{i=1}^{\nu} \left(\psi_i^- \cdot v_i + \psi_i(t_i) \int_{t_{i-1}}^{t_i} M_i(t_i, s) d\alpha_{i,f,V}^{(\varepsilon,v)}(s) \right) \\
&\qquad + \sum_{i=1}^{\nu} \psi_i(t_{i-1}) v_i - \psi_0 \sum_{i=1}^{\nu} \int_{t_{i-1}}^{t_i} d\alpha_{i,L,V}^{(\varepsilon,v)}(s) \\
&\quad - \psi_0 \sum_{i=1}^{\nu} \int_{t_{i-1}}^{t_i} \frac{\partial}{\partial x} L_{q_i}(x_i(t), u_i(t)) \int_{t_{i-1}}^{s} M_i(s, r) d\alpha_{i,f,V}^{(\varepsilon,v)}(r) ds \le 0.
\end{aligned} \tag{4.14}$$

Fix $i \in \{1, \dots, \nu - 1\}$ and a point $(z, z') \in K_i$, where K_i is a Boltyanskiĭ approximating cone to $\mathcal{S}_{q_i,q_{i+1}}$ at the point $(x_i(t_i), x_{i+1}(t_i))$. Let us consider

$$\mathbf{z} = ((0,0), \dots, (z, z'), \dots, (0,0), r)$$

with $r = -\nabla\Phi_{q_i,q_{i+1}}(x_i(t_i), x_{i+1}(t_i)) \cdot (z, z')$. Obviously $\mathbf{z} \in K_P$ and so

$$\psi_i^+ \cdot z + \psi_{i+1}^- \cdot z' - \psi_0 r \geq 0;$$

that is,

$$((\psi_i(t_i), \psi_{i+1}^-) + \psi_0 \nabla\Phi_{q_i,q_{i+1}}(x_i(t_i), x_{i+1}(t_i))) \cdot (z, z') \geq 0$$

and so

$$((-\psi_i(t_i), -\psi_{i+1}^-) - \psi_0 \nabla\Phi_{q_i,q_{i+1}}(x_i(t_i), x_{i+1}(t_i))) \in K_i^\perp, \tag{4.15}$$

where $K_i^\perp$ is the polar of the cone K_i.

Now take a point $(z, z') \in K_\nu$, where K_ν is a Boltyanskiĭ approximating cone to $\mathcal{S}_{q_\nu,q_{\nu+1}}$ at the point $(x_\nu(t_\nu), x_1(t_0))$. Let us consider

$$\mathbf{z} = ((0,0), \dots, (0,0), (z, z'), r)$$

with $r = -\nabla\widetilde{\varphi}_{q_1,q_\nu}(x_\nu(t_\nu), x_1(t_0)) \cdot (z, z')$. Analogously we obtain that

$$((-\psi_\nu(t_\nu), -\psi_1^-) - \psi_0 \nabla\widetilde{\varphi}_{q_\nu,q_1}(x_\nu(t_\nu), x_1(t_0))) \in K_\nu^\perp, \tag{4.16}$$

where $K_\nu^\perp$ is the polar of the cone K_ν.

So we have just proved the following theorem.

THEOREM 4.6 (hybrid necessary principle). *Let $\boldsymbol{X}$ be an optimal trajectory for problem $\mathcal{P}$. For every convex cone $\widehat{K}$ contained in $\cup_{V \in \mathcal{V}} K_V$, where K_V is defined in* (4.13)*, there exist an adjoint pair (ψ, ψ_0) along $\boldsymbol{X}$ and $(\psi_1^-, \dots, \psi_\nu^-) \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_\nu}$ such that* (4.14) *holds for every $V \in \mathcal{V}$, $(\varepsilon, v) \in [0, \bar{\varepsilon}] \times K$ such that $D\widetilde{C}_V(0)(\varepsilon, v) \in \widehat{K}$. Moreover,* (4.15) *and* (4.16) *hold.*

Remark 11. Notice that in the previous theorem we have implicitly supposed that the times of switchings are fixed. Obviously it is possible to consider variations of these times as in [27], using a more complicated covector. We obtain analogous necessary conditions that are more complicated and less readable.

Remark 12. If assumption (H) holds and if we can take $(0, \dots, v_i, \dots, 0) \in K$, then from the previous theorem we can obtain the same result as in [27, Theorem 1.4.1].

Appendix. A lemma on integrable functions.

LEMMA A.1. *Let I be a compact interval of $\mathbb{R}$, and let f_n, f be functions of $L^\infty(I)$ such that $\|f_n\|_\infty \leq c < +\infty$ and $f_n \to f$ strongly in $L^1(I)$. Moreover, let g_n be a sequence such that $g_n \rightharpoonup g$ weakly in $L^1(I)$. Then $f_n g_n \rightharpoonup fg$ weakly in $L^1(I)$.*

Proof. Let $\varphi \in L^\infty(I)$. We have to prove that

$$\int_I \varphi(s)(f_n(s)g_n(s) - f(s)g(s))ds \to 0$$

as $n \to +\infty$. We have that

$$\int_I \varphi(s)(f_n(s)g_n(s) - f(s)g(s))ds = \int_I \varphi(s)(f_n(s) - f(s))g_n(s)ds + \int_I \varphi(s)f(s)(g_n(s) - g(s))ds.$$

The last integral goes to 0 since $g_n \rightharpoonup g$. Fix $\varepsilon > 0$. Then we can write

$$\int_I \varphi(s)(f_n(s) - f(s))g_n(s)ds$$
$$= \int_{|f_n - f| \geq \varepsilon} \varphi(s)(f_n(s) - f(s))g_n(s)ds + \int_{|f_n - f| < \varepsilon} \varphi(s)(f_n(s) - f(s))g_n(s)ds.$$

Moreover,

$$\left| \int_{|f_n - f| < \varepsilon} \varphi(s)(f_n(s) - f(s))g_n(s)ds \right| \leq \varepsilon \|\varphi\|_\infty \int_I |g_n(s)|\, ds \leq M\varepsilon,$$

where M is a positive constant. Besides,

$$\left| \int_{|f_n - f| \geq \varepsilon} \varphi(s)(f_n(s) - f(s))g_n(s)ds \right| \leq \|\varphi\|_\infty M_1 \int_{|f_n - f| \geq \varepsilon} |g_n(s)|\, ds$$

with M_1 a positive constant. Since $f_n \to f$ strongly in $L^1(I)$, f_n converges to f in measure. Moreover g_n is equi-integrable by the Dunford–Pettis theorem (see [9]); hence we can find $\bar{n} \in \mathbb{N}$ such that

$$\left| \int_I \varphi(s)(f_n(s) - f(s))g_n(s)ds \right| \leq (M + M_1\|\varphi\|_\infty)\varepsilon$$

for every $n \geq \bar{n}$ and we conclude by the arbitrariness of ε. □

With an analogous proof, which we omit here, we can generalize the previous lemma to the case of Radon measures in the following way.

LEMMA A.2. *Let I be a compact interval of $\mathbb{R}$, and let f_n, f be functions of $C(I)$ such that $f_n \to f$ uniformly on I as $n \to +\infty$. Moreover, let g_n be a sequence in $\mathfrak{M}_b(I)$ such that $g_n \rightharpoonup^* g$ in $\mathfrak{M}_b(I)$, where $g \in \mathfrak{M}_b(I)$. Then $f_n g_n \rightharpoonup^* fg$ in $\mathfrak{M}_b(I)$.*

REFERENCES

[1] A. A. AGRACHEV, R. V. GAMKRELIDZE, AND A. V. SARYCHEV, *Local invariants of smooth control systems*, Acta Appl. Math., 14 (1989), pp. 191–237.

[2] A. A. AGRACHEV, G. STEFANI, AND P. ZEZZA, *An invariant second variation in optimal control*, Internat. J. Control, 71 (1998), pp. 689–715.

[3] P. J. ANTSAKLIS AND A. NERODE, EDS., *Special issue on hybrid systems*, IEEE Trans. Automat. Control, 43 (1998).

[4] R. M. BIANCHINI, *Good needle-like variations*, in Differential Geometry and Control (Boulder, CO, 1997), Proc. Sympos. Pure Math. 64, AMS, Providence, RI, 1999, pp. 91–101.

[5] U. BOSCAIN AND B. PICCOLI, *Optimal Syntheses for Control Systems on 2-D Manifolds*, Math. Appl. 43, Springer-Verlag, Berlin, 2004.

[6] M. S. BRANICKY, V. S. BORKAR, AND S. K. MITTER, *A unified framework for hybrid control: Model and optimal control theory*, IEEE Trans. Automat. Control, 43 (1998), pp. 31–45.

[7] A. BRESSAN, *A high order test for optimality of bang-bang controls*, SIAM J. Control Optim., 23 (1985), pp. 38–48.

[8] A. BRESSAN AND B. PICCOLI, *A generic classification of time-optimal planar stabilizing feedbacks*, SIAM J. Control Optim., 36 (1998), pp. 12–32.

[9] H. BREZIS, *Analyse fonctionnelle: Théorie et applications*, Masson, Paris, 1983.

[10] C. CASSANDRAS, D. L. PEPYNE, AND Y. WARDI, *Optimal control of a class of hybrid systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 398–415.

[11] F. H. CLARKE AND R. B. VINTER, *Optimal multiprocesses*, SIAM J. Control Optim., 27 (1989), pp. 1072–1091.

[12] F. H. CLARKE AND R. B. VINTER, *Applications of optimal multiprocesses*, SIAM J. Control Optim., 27 (1989), pp. 1048–1071.

[13] C. D'Apice, R. Manzo, M. Garavello, and B. Piccoli, *Hybrid optimal control: Case study of a car with gears*, Internat. J. Control, 76 (2003), pp. 1272–1284.

[14] B. De Schutter, *Optimal control of a class of linear hybrid systems with saturation*, SIAM J. Control Optim., 39 (2000), pp. 835–851.

[15] M. D. Di Benedetto and A. Sangiovanni-Vincentelli, eds., *Hybrid Systems: Computation and Control*, Lecture Notes in Comput. Sci. 2034, Springer-Verlag, Berlin, Heidelberg, 2001.

[16] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, John Wiley, New York, 1984.

[17] S. Hedlund and A. Rantzer, *Optimal control of hybrid systems*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 3972–3977.

[18] A. J. Krener, *The high order maximal principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256–293.

[19] A. S. Morse, C. C. Pantelides, S. Sastry, and J. M. Schumacher, eds., *Special issue on hybrid systems*, Automatica J. IFAC, 35 (1999), pp. 347–536.

[20] B. Piccoli, *Hybrid systems and optimal control*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 13–18.

[21] B. Piccoli, *Necessary conditions for hybrid optimization*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 410–415.

[22] B. Piccoli and H. J. Sussmann, *Regular synthesis and sufficient conditions for optimality*, SIAM J. Control Optim., 39 (2000), pp. 359-410.

[23] P. Riedinger, F. Kratz, C. Iungand, and C. Zanne, *Linear quadratic optimization for hybrid systems*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 2466–2470.

[24] H. Schättler, *Regularity properties of optimal trajectories: Recently developed techniques*, in Nonlinear Controllability and Optimal Control, Monogr. Textbooks Pure Appl. Math. 133, Dekker, New York, 1990, pp. 351–381.

[25] G. Stefani, *Higher order variations: How can they be defined in order to have good properties?*, in Nonsmooth Analysis and Geometric Methods in Deterministic Optimal Control (Minneapolis, MN, 1993), IMA Vol. Math. Appl. 78, Springer, New York, 1996, pp. 227–237.

[26] H. J. Sussmann, *A maximum principle for hybrid optimal control problems*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999.

[27] H. J. Sussmann, *A nonsmooth hybrid maximum principle*, in Stability and Stabilization of Nonlinear Systems (Ghent, 1999), Lecture Notes in Control and Inform. Sci. 246, Springer, London, 1999, pp. 325–354.

[28] H. J. Sussmann, *Multidifferential calculus: Chain rule, open mapping and transversal intersection theorems*, in Optimal Control: Theory, Algorithms and Applications, Appl. Optim. 15, W. W. Hager and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 436–487.

[29] X. Xu and P. J. Antsaklis, *A dynamic programming approach for optimal control of switched systems*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 1822–1827.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1888–1912

ASYMPTOTIC CONTROLLABILITY AND ROBUST ASYMPTOTIC STABILIZABILITY*

CHRISTOPHE PRIEUR†

Abstract. This paper deals with asymptotically controllable systems for which there exists no smooth stabilizing state feedback. To investigate the robustness asymptotic stabilization property, a new class of hybrid feedbacks (with a continuous component and a discrete one) is introduced: the hybrid patchy feedbacks. The notion of solutions is a generalization of π-solutions and Euler solutions. It is proved that the origin of all globally asymptotically controllable systems can be globally asymptotically stabilized via a hybrid feedback with robustness with respect to measurement noise, actuator errors, and external disturbances.

Key words. control systems, feedback stabilization, controllability, measurement noise

AMS subject classifications. 93B52, 93D15

DOI. 10.1137/S0363012901385514

1. Introduction.

Let us consider the system

$$\dot{x} = f(x, u), \tag{1}$$

assuming that the control set $K \subset \mathbb{R}^m$ is a compact subset of $\mathbb{R}^m$ and that the map $f : \mathbb{R}^n \times K \to \mathbb{R}^n$ is locally Lipschitz in x, uniformly with respect to u, and continuous in u. We focus our study on systems that are asymptotically controllable, i.e., that satisfy, for every initial point x_0 in $\mathbb{R}^n$, there exists a measurable $u : [0, +\infty) \to K$ such that the (Carathéodory) solution of

$$\dot{x} = f(x, u(t)), \qquad x(0) = x_0,$$

is defined for all $t \geq 0$ and tends to the origin as t tends to infinity; and that satisfy a stability property (see Definition 2.5).

The general problem under consideration in this paper is the asymptotic stabilization via state feedback. Let us recall that *asymptotic stabilization* means that the following two properties hold:

- stability of the origin of the closed-loop system and
- convergence to the origin of all the solutions.

There exists a necessary condition [6, Theorem 1, (iii)] for the existence of a continuous control law which makes the origin globally asymptotically stable. But there are asymptotically controllable systems which do not satisfy this necessary condition and hence for which there does not exist a continuous stabilizing feedback [23, 6] (consider, e.g., the so-called Brockett's example).

Therefore we must consider discontinuous controllers to stabilize all asymptotically controllable systems. The first result concerning the use of such controllers is [24], but the author assumes that the system is analytic and completely controllable. The following property is proved in [8]:

*Received by the editors February 23, 2001; accepted for publication (in revised form) July 27, 2004; published electronically March 22, 2005.
http://www.siam.org/journals/sicon/43-5/38551.html
†LAAS-CNRS, 7, Avenue du Colonel Roche, 31077 Toulouse, Cedex 4, France (Christophe.Prieur@laas.fr).

($\mathcal{P}$) Any asymptotically controllable systems can be asymptotically stabilized by a discontinuous controller.

The notion of solutions used by the authors is the notion of π-solutions (i.e., solutions with a feedback computed with an arbitrary small sampling schedule) [19]. In [1], the authors prove the property ($\mathcal{P}$) for all Carathéodory solutions by exhibiting a patchy feedback.

The controllers in [8, 1] are robust with respect to actuator and external disturbances (i.e., all systems perturbed by small actuator and external disturbances are asymptotically stable) but are not robust with respect to arbitrary small measurement noise. One way to robustly stabilize the system (1) is to enlarge the class of controllers as in [14], where the authors introduced the notion of a dynamic hybrid controller, which is computed with an external model. This controller compares, at suitable sampling times, the predicted state with the measured state. Due to the measurement noise these can differ substantially; therefore, as remarked in [21], it requires a resetting of the controller which may be difficult to construct. Moreover, with this controller, the origin is a robustly globally asymptotically stable equilibrium for π-solutions only. Here we prove also the existence of a hybrid controller (in the sense that it has a continuous component and a discrete one) which renders the origin a robustly globally asymptotically stable equilibrium for a *larger* class of solutions and, moreover, our feedback does not need a *resetting*.

In [21, 7], the authors proved the existence, for all asymptotically controllable systems, of a controller that is robust with respect to measurement noise and makes the origin of system (1) be a semiglobal *practical* stable equilibrium (i.e., driving all states in a given compact set of initial conditions into a specified neighborhood of the origin). (The case of the state-constraint stabilization is studied separately in [10].) It is proved in [22, section 5.4] that one can get a more general result: one can prove the existence of a sampling feedback making the origin be a robust global asymptotically stable equilibrium for all π-solutions with a sampling rate *sufficiently slow*. We exhibit in this paper a robust global asymptotically stabilizing controller for π-solutions with *any fast enough* sampling schedule, so for a larger class of solutions than those considered in [22].

The main result of this paper is Theorem 2.7: if (1) is asymptotically controllable, then there exists a hybrid feedback which makes the origin be a globally asymptotically stable equilibrium and with robustness with respect to measurement noise, actuator errors, and external disturbances. The class of solutions under consideration in this result includes π-solutions, Euler solutions (i.e., the limit of π-solutions as the sampling schedules tend to zero), and the *generalized* solutions (defined in [11, 12]).

To prove this result, we use some techniques of [1] to deduce from the asymptotic controllability a family of nested patchy vector fields, and we introduce hysteresis between an infinite number of controllers as it is done in [16] for two controllers. This allows us to define a *hybrid patchy feedback*. This gives rise to a hybrid system for which we rewrite the notion of solutions of [5] in the context of π-solutions (see Definition 2.1).

Note that this method was used in [18], where the authors used the special geometry of the chained system in dimension n. (In dimension 3 it is equivalent to the Brockett's example by a change of coordinates.) They exhibit a simple hybrid feedback (with only one discrete variable) making the origin of the chained system be a globally *exponentially* stable equilibrium with a robustness with respect to noise.

The paper is organized as follows. In section 2 we introduce the class of solutions

of a system in closed loop with a hybrid feedback and we state our main result. In section 3 we define the class of hybrid patchy feedbacks and we give properties of π-solutions of systems in closed loop with such a feedback in section 4. Finally we prove our main result in section 5.

2. Definitions and statement of the main result. In this section we make more precise the notions of controller and solutions under consideration.

Let $\mathcal{A}$ be a nonempty totally ordered index set. The controllers under consideration in this paper admit the following description (see [25, 5]):

$$u = u(x, s_d), \qquad s_d = k_d(x, s_d^-), \tag{2}$$

where s_d evolves in the set $\{1,2\}^{\mathcal{A}}$, $u : \mathbb{R}^n \times \{1,2\}^{\mathcal{A}} \to K$ is continuous in x for each fixed s_d, $k_d : \mathbb{R}^n \times \{1,2\}^{\mathcal{A}} \to \{1,2\}^{\mathcal{A}}$ is a function, and s_d^- is defined, at this stage only formally, as

$$s_d^-(t) = \lim_{s<t,\ s\to t} s_d(s). \tag{3}$$

For this to make sense, we equip $\{1,2\}^{\mathcal{A}}$ with the discrete topology, i.e., every set is an open set. We say that the above controller is hybrid because it has a continuous component and a discrete one. Moreover, there is a delay since to evaluate $s_d^-(t)$ at time t, we need to know the past values of $s_d(t)$. Note that time cannot be reversed.

In this paper we are interested in a notion of robustness with respect to small noise. To this end, consider three functions satisfying our *standing regularity assumptions*, i.e.,

- ξ and ζ in $\mathcal{L}^\infty_{loc}(\mathbb{R}^n \times \mathbb{R}_{\geq 0}; \mathbb{R}^n)$ which are continuous in x in $\mathbb{R}^n$ for each t in $\mathbb{R}_{\geq 0}$,
- ψ in $\mathcal{L}^\infty_{loc}(\mathbb{R}^n \times \mathbb{R}_{\geq 0}; \mathbb{R}^m)$ which is continuous in x in $\mathbb{R}^n$ for each t in $\mathbb{R}_{\geq 0}$.

We introduce these functions as a measurement noise ξ, an actuator noise ψ, and an external noise ζ of (1) and study the following perturbed system:

$$\begin{cases} \dot{x}(t) = f(x(t), u(x(t) + \xi(x,t), s_d(t)) + \psi(x,t)) + \zeta(x,t), \\ s_d(t) = k_d(x(t) + \xi(x,t), s_d^-(t)). \end{cases} \tag{4}$$

As noted in [14, Remark 1.4], with the presence of ζ and the continuity of f in u, we can omit any explicit reference to actuator errors. So in the following we suppose that, for all x in $\mathbb{R}^n$ and for all $t \geq 0$, we have

$$\psi(x,t) = 0.$$

We have to clarify what we mean by a solution of the corresponding differential equation. The notion of solution is given in detail in [5] but, here, we want to study the implementation of the controller (2). Therefore we consider π-solutions that have a meaningful physical interpretation: it is an accurate model of the process in computer control. These π-solutions are studied in [8, 21, 15, 22, 13] in the case of an ordinary differential equation. Let π be a sampling schedule of $\mathbb{R}$, i.e., a sequence $(t_n)_{n\in\mathbb{Z}}$ such that, for all n in $\mathbb{Z}$, we have $t_n < t_{n+1}$ and $\lim_{n\to+\infty} t_n = \lim_{n\to-\infty} -t_n = +\infty$. Note that the upper and lower diameters of the sampling schedule are defined by (see [21])

$$\overline{d}(\pi) = \sup_{i\in\mathbb{Z}}(t_{i+1} - t_i), \qquad \underline{d}(\pi) = \inf_{i\in\mathbb{Z}}(t_{i+1} - t_i).$$

We rewrite the notion of solution given in [5] in the context of π-solutions.

DEFINITION 2.1. *Let π be a sampling schedule of $\mathbb{R}$, t_0 in π, $T > t_0$, and $(x_0, s_0) \in \mathbb{R}^n \times \{1,2\}^{\mathcal{A}}$. We say that $(X, S_d) : [t_0, T) \to \mathbb{R}^n \times \{1,2\}^{\mathcal{A}}$ is a* π-solution *of* (4) *on $[t_0, T)$ with initial condition (x_0, s_0) if*

1. *The map X is absolutely continuous on $[t_0, T)$.*
2. *We have, for all t in $[t_0, \min(t_1, T))$,*

$$S_d(t) = S_d(t_0), \tag{5}$$

for all i in $\mathbb{N}_{>0}$ and for all t in $[\min(t_i, T), \min(t_{i+1}, T))$,

$$S_d(t) = k_d(X(t_i) + \xi(X(t_i), t_i), S_d(t_{i-1})). \tag{6}$$

3. *We have, for all i in $\mathbb{N}$ and for almost all t in $[\min(t_i, T), \min(t_{i+1}, T))$,*

$$\dot{X}(t) = f\left(X(t), u(X(t_i) + \xi(X(t_i), t_i), S_d(t_i))\right) + \zeta(X(t), t).$$

4. *We have*

$$X(t_0) = x_0, \qquad S_d(t_0) = k_d(x_0 + \xi(x_0, t_0), s_0). \tag{7}$$

As usual we define Euler solutions as the limits of π-solutions as the sampling schedules tend to zero. More precisely, we have the following definition.

DEFINITION 2.2. *Given t_0 in $\mathbb{R}$, $T > t_0$ and $x_0 \in \mathbb{R}^n$, we say that $X : [t_0, T) \to \mathbb{R}^n$ is an* Euler solution *starting from x_0 of* (4) *on $[t_0, T)$ if, for each compact subinterval J of $[t_0, T)$, there exists a sequence π^n of sampling schedules of $\mathbb{R}$ and a sequence (X^n, S_d^n) of π^n-solutions of* (4) *defined on J such that*

$$\lim_{n \to \infty} \left(\sup_J |X^n - X| + \overline{d}(\pi^n) \right) = 0$$

and such that we have

$$X(t_0) = x_0. \tag{8}$$

Actually we are interested in a notion of solutions which is robust with respect to disturbances. For this reason we introduce a notion of generalized solutions (see [11, 12, 17]).

DEFINITION 2.3. *Let t_0 in $\mathbb{R}$, $T > t_0$ and x_0 in $\mathbb{R}^n$. We say that $X : [t_0, T) \to \mathbb{R}^n$ is a* generalized solution *starting from x_0 of* (4) *if we have* (8) *and if, for each J compact subinterval of $[t_0, T)$, there exist two sequences $(e^n)_{n \in \mathbb{N}}$ and $(d^n)_{n \in \mathbb{N}}$ of measurable functions $[t_0, +\infty) \to \mathbb{R}^n$ and a sequence $(X^n, S_d^n)_{n \in \mathbb{N}}$ of π-solutions of*

$$\begin{cases} \dot{x}(t) = f(x(t), u(x(t) + \xi(x(t), t), s_d(t)) + \zeta(x, t) + d^n(t), \\ s_d(t) = k_d(x(t) + \xi(x, t) + e^n(t), s_d^-(t)) \end{cases} \tag{9}$$

such that we have

$$\lim_{n \to +\infty} \left(\sup_J |X^n - X| + \sup_J |e^n| + \operatorname*{esssup}_J |d^n| \right) = 0. \tag{10}$$

By invoking Zorn's lemma exactly as in the proof of [20, Proposition 1], one can prove that every π-solution can be extended to a maximal solution. More precisely, we define the maximal extension taking account of all sufficiently fast sampling schedules π of $[0, +\infty)$.

DEFINITION 2.4. *Let t_0 in $\mathbb{R}$, $T > t_0$, (x_0, s_0) in $\mathbb{R}^n \times \{1,2\}^{\mathcal{A}}$ and $d_0 > 0$. We say that $(X, S_d) : [t_0, T) \to \mathbb{R}^n \times \{1,2\}^{\mathcal{A}}$ is a* d_0-maximal solution *starting from (x_0, s_0) of (4) on $[t_0, T)$, if the following properties hold:*

- *For all $T' < T$, there exists a sampling schedule π of $[0, +\infty)$ such that*

$$\overline{d}(\pi) \leq d_0 \tag{11}$$

and such that (X, S_d) is a π-solution starting from x_0 of (4) on $[t_0, T')$.

- *For all $T' > T$ and for all sampling schedules π of $[0, +\infty)$ such that (11), there does not exist any π-solution (X', S'_d) starting from (x_0, s_0) and defined on $[t_0, T')$ such that the restriction of (X', S'_d) to $[t_0, T)$ is (X, S_d).*

We say that $X : [t_0, T) \to \mathbb{R}^n$ is a maximal Euler solution *starting from x_0 of (4) on $[t_0, T)$ if the following properties hold:*

- *For all $T' < T$, X is an Euler solution starting from x_0 of (4) on $[t_0, T')$.*
- *For all $T' > T$, there does not exist any Euler solution X' starting from x_0 of (4) on $[t_0, T')$ such that the restriction of X' to $[t_0, T)$ is X.*

We say that $X : [t_0, T) \to \mathbb{R}^n$ is a d_0-maximal generalized solution *starting from x_0 of (4) on $[t_0, T)$ if the following properties hold:*

- *For all $T' < T$, X is a generalized solution obtained as limit of π-solutions whose sampling schedule satisfies (11) starting from x_0 of (4) on $[t_0, T')$.*
- *For all $T' > T$, there does not exist any generalized solution X' obtained as limit of π-solutions whose sampling schedule satisfies (11) starting from x_0 of (4) on $[t_0, T')$ and such that the restriction of X' to $[t_0, T)$ is X.*

Let us recall that a function of class $\mathcal{K}_\infty$ is a function $\delta : [0, +\infty) \to [0, +\infty)$ which is continuous, strictly increasing, satisfying $\delta(0) = 0$ and $\lim_{\varepsilon \to +\infty} \delta(\varepsilon) = +\infty$. In the following we denote the closed ball centered at $x \in \mathbb{R}^n$ with radius $r > 0$ by $B(x, r)$. In our context our definition of robust global asymptotic stability is as follows (see [3]).

DEFINITION 2.5. *The origin is said to be a* robustly globally asymptotically stable equilibrium *of the system (4) if the following properties hold:*

1. Existence of solutions: *For all $C > 0$, there exists $\chi_0 = \chi_0(C) > 0$ such that for all ξ, ζ satisfying our regularity assumptions and such that*

$$\sup_{x \in \mathbb{R}^n,\, t \geq 0} |\xi(x,t)| \leq \chi_0, \qquad \operatorname*{esssup}_{x \in \mathbb{R}^n,\, t \geq 0} |\zeta(x,t)| \leq \chi_0, \tag{12}$$

for all (x_0, s_0) in $B(0, C) \times \{1,2\}^{\mathcal{A}}$, and for all sampling schedules π of $\mathbb{R}$, there exists a π-solution of (4) (resp., an Euler solution, resp., a generalized solution) starting from (x_0, s_0) (resp., starting from x_0) at $t_0 = 0$.

2. Completeness: *Moreover, there exists $d_0 = d_0(C)$ such that all the d_0-maximal solutions (resp., maximal Euler solutions, resp., d_0-maximal generalized solutions) of (4) are defined on $[0, +\infty)$.*

3. Global stability: *There exists δ of class $\mathcal{K}_\infty$ such that, for all $\varepsilon > 0$, there exist $\chi_0 = \chi_0(\varepsilon) > 0$ and $d_0 = d_0(\varepsilon) > 0$ such that, for all ξ, ζ satisfying our regularity assumptions and (12), for all (x_0, s_0) in $B(0, \delta(\varepsilon)) \times \{1,2\}^{\mathcal{A}}$, and for every d_0-maximal solution (X, S_d) of (4) (resp., maximal Euler solution X, resp., d_0-maximal generalized solution) starting from (x_0, s_0) (resp., starting from x_0) at $t_0 = 0$, one has*

$$X(t) \in B(0, \varepsilon) \quad \forall t \geq 0. \tag{13}$$

4. Global attractivity: *For all $\varepsilon > 0$ and for all $C > 0$, there exist $T > 0$, $\chi_0 > 0$, and $d_0 > 0$ such that, for all ξ, ζ satisfying our regularity assumptions and* (12)*, for each (x_0, s_0) in $B(0, C) \times \{1, 2\}^{\mathcal{A}}$, and for d_0-every maximal solution (X, S_d) of* (4) *(resp., maximal Euler solution X, resp., d_0-maximal generalized solution) starting from (x_0, s_0) (resp., starting from x_0) at $t_0 = 0$, one has*

$$X(t) \in B(0, \varepsilon) \quad \forall t \geq T. \tag{14}$$

We recall the definition of global asymptotic controllability of the system (1).

DEFINITION 2.6. *The system* (1) *is said to be* globally asymptotically controllable to the origin *if the following properties hold:*

1. *For each x_0 in $\mathbb{R}^n$, there exists an admissible control u_0 (i.e., a measurable function $[0, +\infty) \to K$) such that the maximal Carathéodory solution X starting from x_0 of*

$$\dot{x} = f(x, u_0) \tag{15}$$

is defined for all $t \geq 0$ and satisfies $X(t) \to 0$ as $t \to +\infty$.

2. *For each $\varepsilon > 0$ there exists $C > 0$ such that for each x_0 in $B(0, C)$, there is an admissible control u_0 as in* 1 *such that*

$$X(t) \in B(0, \varepsilon) \quad \forall t \geq 0.$$

Our main result is as follows.

THEOREM 2.7. *Let* (1) *be a globally asymptotically controllable system to the origin. Then there exists a feedback control, $u : \mathbb{R}^n \times \{1, 2\}^{\mathbb{N}} \to K$, $k_d : \mathbb{R}^n \times \{1, 2\}^{\mathbb{N}} \to \{1, 2\}^{\mathbb{N}}$ such that the origin is a robustly globally asymptotically stable equilibrium for the system* (4).

Remark 2.8.

1. Note that in Theorem 2.7 we have the robust global asymptotic stability for π-solutions for *any fast enough* sampling rate since the only constraint on the sampling schedule is (11).

In [21, 7], only for the π-solutions with a sampling rate *sufficiently slow* are considered since, in these papers, it is assumed moreover that the lower diameters of the sampling schedules have a strictly positive lower bound.

See in particular the assumption in [21, Theorem 1],

$$|\xi(t)| \leq \underline{d}(\pi) \quad \forall t \geq 0. \tag{16}$$

Thus the class of solutions under consideration in Theorem 2.7 is *larger* than those considered in [21, 7].

Let us compare (12) and the inequality (16). Given a sampling schedule whose lower diameter is close to zero, this restriction forces the measurement noise to be close to zero. In our context the measurement noise and the lower diameter are completely independent.

Note that the controller given by [21] is not robust with respect to noise which does not satisfy (16) (consider the example of Artstein's circles). See also the discussion given in [22, section 4].

2. Note that Theorem 2.7 is false if in (12) the supremum *sup* is relaxed by *esssup*. See [17, Theorem 4.2], where it is proved, in an analogous situation, that there exists a noise ξ such that $\operatorname{esssup} |\xi| = 0$, $\sup |\xi| \neq 0$ and such that the origin of the perturbed closed-loop system is not an attractive equilibrium.

To prove Theorem 2.7 we need to introduce a class of hybrid patchy feedbacks (see section 3) whose continuous component is derived from a family of nested patchy vector fields (a slight generalization of patchy vector fields defined in [1]) and whose discrete component allows us to unite vector fields, as done in [17] for two vector fields, with robustness with respect to noise. Then we give basic properties of π-solutions of system (1) with a hybrid patchy feedback in section 4 and we prove Theorem 2.7 in section 5.

3. Definition of the hybrid patchy feedbacks. Let Ω be a nonempty open connected subset of $\mathbb{R}^n$. The closure, the interior, and the boundary of Ω are written as $\mathrm{clos}(\Omega)$, $\mathrm{int}(\Omega)$, and $\partial\Omega$, respectively. We define the set $\mathcal{F} = \{1, \dots, 7\}$. Let $\mathcal{A}$ be a nonempty totally ordered index set. Given a set-valued map $F : \mathbb{R}^n \to 2^{\mathbb{R}^n}$, we can define the solutions X of the differential inclusion

$$\dot{x} \in F(x)$$

as all absolutely continuous functions satisfying $\dot{X}(t) \in F(X(t))$ almost everywhere. We follow the ideas of [1, Definition 2.1], but we extend the definition to allow nested sets (as in [16]).

DEFINITION 3.1. *We say that* $(\Omega, ((\Omega_{\alpha,l})_{l\in\mathcal{F}}, g_\alpha)_{\alpha\in\mathcal{A}})$ *is a* family of nested patchy vector fields *if*

1. *for all* $(\alpha, l) \in \mathcal{A} \times \mathcal{F}$; $\Omega_{\alpha,l}$ *is an open bounded subset of* $\mathbb{R}^n$,
2. *for all* $\alpha \in \mathcal{A}$ *and for all* $m > l \in \mathcal{F}$

$$\Omega_{\alpha,l} \subsetneq \mathrm{clos}(\Omega_{\alpha,l}) \subsetneq \Omega_{\alpha,m}; \tag{17}$$

3. *for all* α *in* $\mathcal{A}$, g_α *is a smooth vector field defined in a neighborhood of* $\mathrm{clos}(\Omega_{\alpha,7})$ *taking values in* $\mathbb{R}^n$;
4. *for all compact subsets* C *of* $\mathbb{R}^n$, *there exist* $r = r(C) > 0$ *and* $T = T(C) > 0$ *such that for all* $(\alpha, l) \in \mathcal{A} \times \mathcal{F}$ *satisfying* $\Omega_{\alpha,l} \subset C$, *all solutions* X *of*

$$\dot{x} \in g_\alpha(x) + B(0, r) \tag{18}$$

starting in $\partial\Omega_{\alpha,l} \setminus \bigcup_{\beta>\alpha} \Omega_{\beta,1}$ *are such that*

$$X(t) \in \mathrm{clos}(\Omega_{\alpha,l}) \quad \forall t \in [0, T].$$

5. *The sets* $(\Omega_{\alpha,1})_{\alpha\in\mathcal{A}}$ *form a locally finite covering of* Ω.

Remark 3.2. Some observations are in order.

- Roughly speaking, property 4 states that a part of $\mathrm{clos}(\Omega_{\alpha,l})$ is positively invariant in $[0, T]$ relative to the system (18). Note that we can characterize this property in terms of proximal normal by [9, Theorem 4.3.8] and we can redefine the notion of the patchy vector fields by using this concept of nonsmooth analysis as done in [4].
- On the one hand, given any compact set C, the positive real number $r(C)$ allows us to get robustness with respect to external disturbances. On the other hand, the gap between the different patches given by (17) allow us to get robustness with respect to measurement noise. See Definition 3.7 for a precise statement of admissible radius of measurement noise and external disturbances.
- Let us explain shortly why, to state our main result, we need to consider a family of seven nested patchy vector fields. Patches 2 and 6 define the dynamics of the discrete component of our hybrid controller (see Definition 3.4). Due to the

measurement noise, the switches (this notion will be precisely introduced in Definition 4.1) of the discrete variable can be located only in a neighborhood of Ω_2 and Ω_6 which are described by patches 1-3 and 5-7 (see Lemma 4.2). Patch 4 is only needed to describe π-solutions after the first switch (see Lemma 4.11). These seven patches are enough to state our main result and we show, for Artstein's circles, that we need to use so many patches (see Example 4.10).

EXAMPLE 3.3. *Let us give an example of such a family of nested patchy vector fields.*

We can construct a family of nested patchy vector fields for Artstein's circles. This system is one of the simplest which is not stabilizable by a continuous feedback and which admits a (nonrobust) discontinuous stabilizing feedback. This system is studied in several papers (see, e.g., [2, 21, 22, 16]*) and is defined by*

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} -x_1^2 + x_2^2 \\ -2x_1x_2 \end{pmatrix}. \tag{19}$$

The integral curves of (19) *are*

- *the origin,*
- *the circles centered on the x_2-axis and tangent to the x_1-axis,*
- *the x_1-axis.*

Let us define the three smooth vector fields g_a, g_b, and $g_c : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$\begin{aligned} g_a(x_1, x_2) &= (-x_1^2 + x_2^2, -2x_1x_2)', \\ g_b(x_1, x_2) &= -g_a(x_1, x_2), \\ g_c(x_1, x_2) &= (0, 0). \end{aligned}$$

Let θ be in $\mathbb{R}$ the polar angle of a point $(x_1, x_2) \neq (0, 0)$. For all l in $\mathcal{F}$, let us define the open bounded sets $\Omega_{a,l}$, $\Omega_{b,l}$, and $\Omega_{c,l} \subset \mathbb{R}^2$ by

$$\begin{aligned} \Omega_{a,l} = &\left\{ x \in \mathbb{R}^2, -\frac{3\pi}{4} - \frac{l\pi}{30} < \theta < \frac{3\pi}{4} l + \frac{l\pi}{30} \right\} \cap \left\{ |x| > 1 - \frac{l}{14} \right\} \\ &\cap \left\{ \left(x_1 < 0 \text{ and } x_1^2 + \left(x_2 - 10 - \frac{l}{14} \right)^2 < \left(10 + \frac{l}{14} \right)^2 \right) \right. \\ &\qquad \left. \text{or } \left(x_1 \geq 0 \text{ and } x_1^2 + x_2^2 < \left(20 + \frac{l}{7} \right)^2 \right) \right\}, \\ \Omega_{b,l} = &\operatorname{sym}_{x_2}(\Omega_{a,l}), \\ \Omega_{c,l} = &\left\{ x \in \mathbb{R}^2, |x| < 1 + \frac{l}{7} \right\}, \end{aligned}$$

where sym_{x_2} is the symmetry with respect to the x_2-axis. Let $\mathcal{A} = \{a, b, c\}$ be lexicographically ordered ($a < b < c$) and $\Omega = \operatorname{int}(B(0, 10))$. It is easy to prove that

$$\left(\Omega, \left((\Omega_{\alpha,l})_{l \in \mathcal{F}}, g_\alpha \right)_{\alpha \in \mathcal{A}} \right) \tag{20}$$

is a family of nested patchy vector fields. This is depicted in Figure 3.1*. To make the figure clearer, only two open sets and some values of the vector field g_a are shown.*

With such a family of nested patchy vector fields, we can define a class of hybrid controllers as those considered in section 2. To do this, we denote for all $s_d \in \{1, 2\}^{\mathcal{A}}$ the αth element of s_d by $s_{d,\alpha}$.

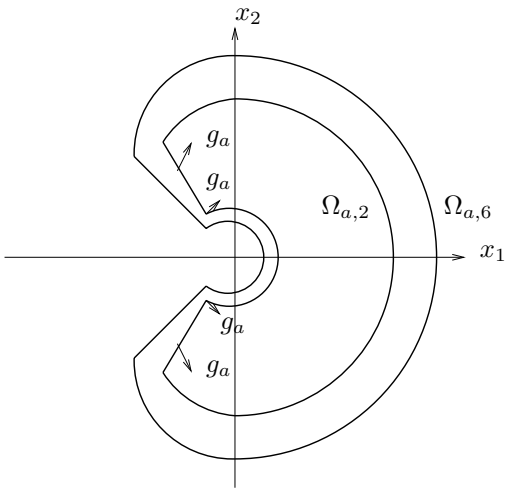


FIG. 3.1. *Artstein's circles as a family of nested patchy vector fields.*

DEFINITION 3.4. *Let* $(\Omega, ((\Omega_{\alpha,l})_{l\in\mathcal{F}}, g_\alpha)_{\alpha\in\mathcal{A}})$ *be a family of nested patchy vector fields. Assume that for each* α *in* $\mathcal{A}$*, we can find a point* k_α *in* K *such that for each* x *in* $\Omega_{\alpha,7}$*, we have*

$$g_\alpha(x) = f(x, k_\alpha). \tag{21}$$

Let k_0 *be an arbitrary point in* K*. Let* (u, k_d) *be the map defined by*

$$\begin{array}{rcll} u : \{1,2\}^{\mathcal{A}} & \to & K, & \\ s_d & \mapsto & k_0 & \text{if } \{\beta \in \mathcal{A},\ s_{d,\beta} = 1\} \text{ is empty or infinite,} \\ & & k_\alpha & \text{if } \alpha = \max\{\beta \in \mathcal{A}, s_{d,\beta} = 1\}, \end{array} \tag{22}$$

and

$$\begin{array}{rcl} k_d : \mathbb{R}^n \times \{1,2\}^{\mathcal{A}} & \to & \{1,2\}^{\mathcal{A}}, \\ (x, s_d) & \mapsto & t_d, \end{array} \tag{23}$$

where t_d *is the sequence defined, for all* α *in* $\mathcal{A}$*, by*

$$\begin{array}{rl} t_{d,\alpha} = 1 & \text{if } x \in \operatorname{clos}(\Omega_{\alpha,2}), \\ t_{d,\alpha} = s_{d,\alpha} & \text{if } x \in \Omega_{\alpha,6} \setminus \operatorname{clos}(\Omega_{\alpha,2}), \\ t_{d,\alpha} = 2 & \text{if } x \notin \Omega_{\alpha,6}. \end{array} \tag{24}$$

We say that (u, k_d) *is a* hybrid patchy feedback *on* Ω*.*

Remark 3.5. This hybrid controller takes advantage of the existence of regions where different controllers k_α exist and, roughly speaking, allows the hybrid variable to choose between the different controllers. This is the main idea of the hysteresis as done in [17] to unite two controllers. Moreover, for any s_d in $\{1,2\}^{\mathcal{A}}$, the function $k_d(., s_d)$ is continuous except on the boundary of the sets defining the hysteresis. This remark is very helpful in particular to establish Lemma 4.2.

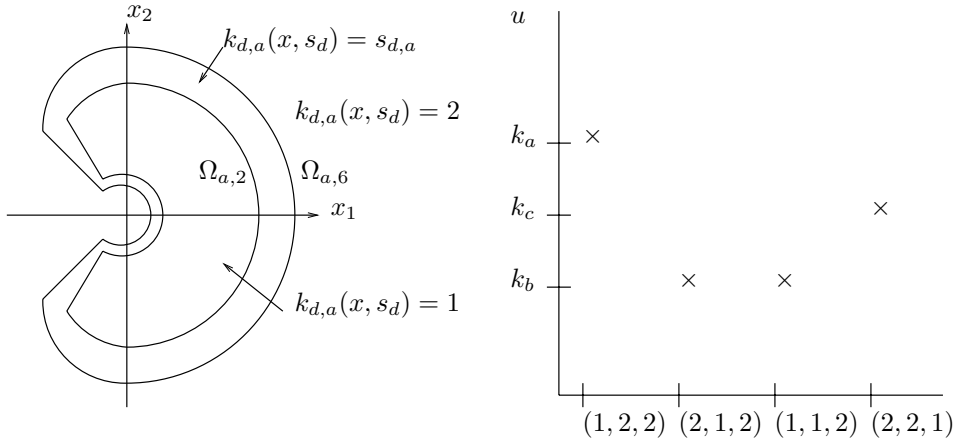


FIG. 3.2. *A hybrid patchy feedback. On the left is the $k_{d,a}$-component and on the right is the u-component.*

EXAMPLE 3.6. *Let us use the family of nested patchy vector fields* (20) *to define a hybrid patchy feedback for Artstein's circles.*

Let us define the controlled Artstein's circles by

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} u(-x_1^2 + x_2^2) \\ -2ux_1x_2 \end{pmatrix} = f((x_1, x_2), u) \tag{25}$$

with u in $\mathbb{R}$. We remark that by denoting $k_a = 1$, $k_b = -1$, and $k_c = 0$, we have (21) *and thus we can define a hybrid patchy feedback, depicted in Figure* 3.2. *The $k_{d,a}$ component is on the left and the u-component (for some values) is on the right.*

Given a family of nested patchy vector fields $(\Omega, ((\Omega_{\alpha,l})_{l\in\mathcal{F}}, g_\alpha)_{\alpha\in\mathcal{A}})$ it is easy to check from Definition 3.1 that for all x in $\mathbb{R}^n$, the set $C_x \subset \mathbb{R}^n$ defined by

$$C_x = \operatorname{clos}\left(\bigcup_{\alpha\in\mathcal{A},\ x\in\Omega_{\alpha,7}} \Omega_{\alpha,7} \right) \tag{26}$$

is a compact set. To investigate the robustness with respect to noise with a family of nested patchy vector fields we generalize [1, Definition 2.3] to a family of nested patchy vector fields and we introduce the next definition.

DEFINITION 3.7. *Let $\chi : \mathbb{R}^n \to \mathbb{R}$ be a continuous map such that for all $x \neq 0$, $\chi(x) > 0$.*

- *We say that χ is an admissible radius for the measurement noise if for all x in $\mathbb{R}^n$ and for all α in $\mathcal{A}$ such that x in $\Omega_{\alpha,7}$, we have*

$$\chi(x) < \frac{1}{2} \min_{l\in\{1,\dots,6\}} d(\mathbb{R}^n \setminus \Omega_{\alpha,l+1}, \Omega_{\alpha,l}). \tag{27}$$

- *We say that χ is an admissible radius for the external disturbances if for all x in $\mathbb{R}^n$, we have $\chi(x) \leq r(C_x)$, where C_x is defined by* (26) *and the corresponding $r > 0$ is guaranteed by* 4 *in Definition* 3.1.

There exists an admissible radius for the measurement noises and for the external disturbances. (Note that with (17), the right-hand side of inequality (27) is strictly positive.)

In Definition 3.4, u does not depend on x. Therefore only the function k_d depend on the measurement noise. Thus the notions of the admissible radius for the measurement noise and for the external disturbances are completely independent.

We need to consider sufficiently fast π-solutions. To define *sufficiently fast* π-solutions, let us introduce the following definition.

DEFINITION 3.8. *Let $p : \mathbb{R}^n \to \mathbb{R}_{>0}$ be a function continuous on $\mathbb{R}^n \setminus \{0\}$. We say that the sampling schedule π of a π-solution (X, S_d) defined on $[t_0, T)$ is* subordinate to p *if for all $i \in \mathbb{N}$ and for all $t \in [\min(t_i, T), \min(t_{i+1}, T))$, we have*

$$t_{i+1} - t_i \leq p(X(t_i) + \xi(X(t_i), t_i)). \tag{28}$$

Now we study the properties of π-solutions.

4. Properties of π-solutions. In this section we study the properties of π-solutions of a system in closed loop with a hybrid patchy feedback. Let Ω be a nonempty open connected subset of $\mathbb{R}^n$ and let

$$(\Omega, ((\Omega_{\alpha,l})_{l\in\mathcal{F}}, g_\alpha)_{\alpha\in\mathcal{A}})$$

be a family of nested patchy vector fields such that (21) holds. Let (u, k_d) be the hybrid patchy feedback on Ω defined by (22)–(24). Let $\chi : \mathbb{R}^n \to \mathbb{R}$ be an admissible radius for the measurement noise and the external disturbances. Consider ξ and ζ satisfying our standing regularity assumptions and such that

$$\forall x \in \mathbb{R}^n, \quad \sup_{t\geq 0} |\xi(x,t)| \leq \chi(x), \quad \operatorname*{esssup}_{t\geq 0} |\zeta(x,t)| \leq \chi(x). \tag{29}$$

The perturbed system under consideration is

$$\begin{cases} \dot{x} = f(x, u(s_d)) + \zeta, \\ s_d = k_d(x + \xi, s_d^-). \end{cases} \tag{30}$$

Let $p : \mathbb{R}^n \to \mathbb{R}$ be a function continuous on $\mathbb{R}^n \setminus \{0\}$ and such that for all (ξ, ζ) with our regularity assumptions and (29), the following inequalities hold:[1]

A1. For all x in $\mathbb{R}^n$, $p(x) > 0$.

A2. For all x in $\mathbb{R}^n$,

$$p(x + \xi(x, 0)) < \frac{1}{4} \min_{l\in\{1,\dots,6\}} \min_{\alpha\in\mathcal{A},\ x+\xi(x,0)\in\Omega_{\alpha,7}} \frac{d(\mathbb{R}^n \setminus \Omega_{\alpha,l+1}, \Omega_{\alpha,l})}{\sup_{y\in\Omega_{\alpha,7},\ u\in K} |f(y,u)|}.$$

A3. For all x in $\mathbb{R}^n$, we have $p(x + \xi(x, 0)) < T(C_x)$, where T is defined by 4 in Definition 3.1 and C_x is defined by (26).

The existence of such a function p results from the fact that for all l in $\{1, \dots, 7\}$, $(\Omega_{\alpha,l})_{\alpha\in\mathcal{A}}$ is locally finite and results from (17).

DEFINITION 4.1. *A map $S_{d,\alpha} : [t_0, T) \to \{1, 2\}$ is said to have a* switch *at time t if $S_{d,\alpha}$ is not continuous at t.*

Given a sampling schedule π of $\mathbb{R}$ and a π-solution (X, S_d) of (30) and t in $[t_0, T)$, we denote the αth element of $S_d(t)$ by $S_{d,\alpha}(t)$. We start by locating the points where there exists α in $\mathcal{A}$ such that $S_{d,\alpha}$ may have a switch. Note that switches can occur only at sampling times, i.e., if there is a switch at time t, then there exists $i \in \mathbb{N}_{>0}$ such that $t = t_i$, $S_{d,\alpha}(t_i) \neq S_{d,\alpha}(t_{i-1})$.

[1] If $\sup |f(y, u)| = 0$, then assumption A2 forces no condition on p.

LEMMA 4.2. *Let (X, S_d) be a π-solution of* (30) *whose sampling schedule is subordinate to p and such that $S_{d,\alpha}$ has a switch at time $t_i \in (t_0, T)$.*

- *If the switch is such that $S_{d,\alpha}(t_{i-1}) = 1$ and $S_{d,\alpha}(t_i) = 2$, then, for all t in $[t_i, \min(t_{i+1}, T))$, $X(t)$ is in $\mathrm{clos}(\Omega_{\alpha,7}) \setminus \Omega_{\alpha,5}$.*
- *If the switch is such that $S_{d,\alpha}(t_{i-1}) = 2$ and $S_{d,\alpha}(t_i) = 1$, then, for all t in $[t_i, \min(t_{i+1}, T))$, $X(t)$ is in $\mathrm{clos}(\Omega_{\alpha,3}) \setminus \Omega_{\alpha,1}$.*

Proof. Let α in $\mathcal{A}$ and $i \in \mathbb{N}_{>0}$ such that $S_{d,\alpha}(t_{i-1}) = 1$ and $S_{d,\alpha}(t_i) = 2$. Then due to (6) and (23)–(24), $X(t_{i-1}) + \xi(X(t_{i-1}), t_{i-1})$ is in $\Omega_{\alpha,6}$. Thus with (27), assumption A2, and (29), it follows directly that, for all t in $[t_{i-1}, \min(t_{i+1}, T))$, $X(t)$ is in $\mathrm{clos}(\Omega_{\alpha,7})$. Similarly we prove that, for all t in $[t_i, \min(t_{i+2}, T))$, $X(t) \notin \Omega_{\alpha,5}$. Thus we obtain that, for all t in $[t_i, t_{i+1})$, $X(t)$ is in $\mathrm{clos}(\Omega_{\alpha,7}) \setminus \Omega_{\alpha,5}$.

The case $S_d(t_{i-1}) = 2$ and $S_d(t_i) = 1$ is established in the same way. ∎

Let us claim a result of existence.

LEMMA 4.3. *For all (x_0, s_0) in $\mathbb{R}^n \times \{1, 2\}^{\mathcal{A}}$ and for all sampling schedules π of $\mathbb{R}$, there exists a π-solution of* (30) *starting from (x_0, s_0).*

Proof. Let (x_0, s_0) be in $\mathbb{R}^n \times \{1, 2\}^{\mathcal{A}}$. Let $s_1 = k_d(x_0 + \xi(x_0, t_0), s_0)$ and α be in $\mathcal{A}$ such that $k_\alpha = u(s_1)$. From our regularity assumptions on f and ζ, the Carathéodory conditions are satisfied for the system

$$\dot{X} = f(X, k_\alpha) + \zeta, \qquad X(t_0) = x_0. \tag{31}$$

Let $t_0 \le T \le t_1$ and X defined on $[t_0, T)$ be a Carathéodory solution of (31). Let S_d be defined, for all t in $[t_0, T)$, by $S_d(t) = s_1$ for all t in $[t_0, T)$. Thus (X, S_d) is a π-solution of (30) starting from (x_0, s_0). ∎

We note that, as usual, maximal solutions of (30) must blow up if their domains of definition are bounded.

LEMMA 4.4. *Let $d_0 > 0$, ξ, and ζ satisfy our regularity assumptions and* (29). *Let $T > t_0$ and (X, S_d) be a d_0-maximal solution of* (30) *defined on $[t_0, T)$. Suppose that $T < +\infty$; then*

$$\limsup_{t \to T} \left(|X(t)| + \frac{1}{d(X(t), \partial\Omega)} \right) = +\infty.$$

Proof. Consider $d_0 > 0$, ξ, ζ satisfying our regularity assumptions and (29), $T > t_0$ and (X, S_d) a d_0-maximal solution defined on $[t_0, T)$. Suppose that the conclusion of Lemma 4.4 does not hold; i.e., there exists a compact subset C of Ω and times t_n in $[t_0, T)$ tending monotonically to T such that $(X(t_n), S_d(t_n))$ is in $C \times \{1, 2\}^{\mathcal{A}}$ for all n. We first establish the following.

Claim 4.5. For some n sufficiently large, for all $t \in [t_n, T)$, $X(t)$ is in the bounded open set $C + \mathrm{int}(B(0, 1))$.

Proof of Claim 4.5. If the conclusion of Claim 4.5 is not true, the continuity of X implies the existence of $s_n \in (t_n, T)$ such that

$$|X(t_n) - X(s_n)| = 1 \quad \text{and} \quad |X(t_n) - X(t)| < 1 \quad \forall t \in [t_n, s_n).$$

It follows that $X(t)$ is in the compact set $C + B(0, 1)$ for all t in $[t_n, s_n]$. Let

$$\rho = \max_{x \in C + B(0,1)} |\chi(x)|, \qquad \sigma = \sup_{\zeta \in B(0,\rho),\ x \in C + B(0,1),\ u \in K} |f(x, u) + \zeta|.$$

Then we have, for all t, s in $[t_n, s_n]$, $|X(t) - X(s)| \le \sigma |t - s|$. Therefore, for n sufficiently large,

$$1 = |X(t_n) - X(s_n)| \le \sigma |s_n - t_n| \le \sigma |T - t_n|.$$

This cannot hold for n large enough and proves Claim 4.5. ∎

Claim 4.5 implies that there exists σ in $\mathbb{R}_{\geq 0}$ such that, for all (s,t) in $[t_n, T)$, we have

$$|X(s) - X(t)| \leq \sigma |s - t|.$$

By invoking the Cauchy criterion, it follows that $X(t)$ has a limit x_0 when t tends to T. Note moreover that by Definition 2.1, there exists $i \in \mathbb{N}$ such that T is in $(t_i, t_{i+1}]$ and thus, for all α in $\mathcal{A}$, $\lim_{t \to T,\, t<T} S_{d,\alpha}(t)$ exists. We denote $s_0 = S_d^-(T)$. Due to Lemma 4.3, there exists a π-solution $(\widetilde{X}, \widetilde{S_d})$ starting from (x_0, s_0) and defined on $[t_0, \widetilde{T})$ with $\widetilde{T} > t_0$. Note that (X', S'_d) defined by

$$\begin{aligned} \forall t \in [t_0, T), \quad X'(t) = X(t), \quad S'_d(t) = S_d(t), \\ \forall t \in (T, T + \widetilde{T}), \quad X'(t) = \widetilde{X}(t - T), \quad S'_d(t) = \widetilde{S_d}(t - T), \end{aligned}$$

is a $\widetilde{\pi}$-solution of (30) defined on $[t_0, T + \widetilde{T})$ for the sampling schedule $\widetilde{\pi} = \pi \cup \{T\}$ whose restriction to $[t_0, T)$ is (X, S_d). Moreover $\widetilde{\pi}$ satisfies (11). So we have obtained a contradiction with the fact that (X, S_d) is a d_0-maximal solution. □

Now we can study the behavior of π-solutions between two switches. For all α in $\mathcal{A}$, let

$$\begin{aligned} \tau_\alpha = \sup \big\{ & T, X \text{ is a Carathéodory solution of } \dot{x} = f(x, k_\alpha) + B(0, \chi(x)) \\ & \text{with } X(t) \in \Omega_{\alpha,7}\ \forall t \in [0, T) \big\}. \end{aligned} \tag{32}$$

Note that there may exist α in $\mathcal{A}$ such that (s.t.) $\tau_\alpha = +\infty$. Let M be the subset of $\Omega \times \{1,2\}^{\mathcal{A}}$ defined by

$$M = \left\{ (x, s_d), \text{s.t.} \left\{ \begin{array}{c} \{\beta \in \mathcal{A},\ s_{d,\beta} = 1\} \text{ is empty or infinite} \\ \text{or} \\ x \in \Omega_{\alpha,5},\ \text{where } \alpha = \max\{\beta,\ s_{d,\beta} = 1\} \end{array} \right\} \right\}. \tag{33}$$

Note that we have the property

$$\forall x_0 \in \Omega, \quad \exists s_0 \in \{1,2\}^{\mathcal{A}}, \quad (x_0, s_0) \in M. \tag{34}$$

Example 4.6. *Let us particularize the set M for Artstein's circles. We have*

$$\begin{aligned} M = \Omega_{a,5} \times \{(1,2,2)\} \cup \Omega_{b,5} \times \{(s_d, 1, 2),\ s_d \in \{1,2\}\} \\ \cup\ \Omega_{c,5} \times \{(s_d, s'_d, 1),\ s_d, s'_d \in \{1,2\}\} \cup \Omega \times \{(2,2,2)\}. \end{aligned}$$

In the following we denote $\overline{m} = \{0, \ldots, m\}$ if $m \in \mathbb{N}$ and $\overline{m} = \mathbb{N}$ if $m = +\infty$.

Lemma 4.7. *Let $0 < T \leq \infty$ and (X, S_d) be a π-solution of (30) whose sampling schedule is subordinate to p, defined on $[0, T)$ and starting in M. Then, there exist $m \in \mathbb{N} \cup \{+\infty\}$, an increasing sequence of time instants $(T_j)_{j \in \overline{m}}$ in $[0, T)$, a sequence $(\alpha_j)_{j \in \overline{m}}$ in $\mathcal{A}$, and a sequence $(k_j)_{j \in \overline{m}}$ in K such that if we let $T_0 = 0$ and $T_{m+1} = T$ (if $m < +\infty$), we have for all $j \in \overline{m}$ the following:*

1. *For all t in (T_j, T_{j+1}), $u(S_d(t)) = k_{\alpha_j}$.*
2. *The map X is a Carathéodory solution of $\dot{x} = f(x, k_{\alpha_j}) + \zeta$ on (T_j, T_{j+1}).*
3. *For all t in $[T_0, T_1)$, $X(t)$ is in $\Omega_{\alpha_0,5}$.*
4. *For all t in $[T_j, T_{j+1})$, $X(t)$ is in $\operatorname{clos}(\Omega_{\alpha_j,3})$, if $j \geq 1$.*
5. *The sequence $(\alpha_j)_{j \in \overline{m}}$ is strictly increasing.*
6. *The inequality $T_{j+1} - T_j < \tau_{\alpha_j}$ holds.*

Proof. Note first that the switches may occur only at a sampling time. Thus we can define $m \in \mathbb{N} \cup \{+\infty\}$ and a sequence of sampling times $(T_j)_{j \in \overline{m}}$ in $[0, T)$ at which switches occur. Between two switches, S_d is constant and thus there exist a sequence α_j in $\mathcal{A}$ and a sequence of admissible controls such that the statements 1 and 2 hold. We denote again by $(T_j)_{j \in \overline{m}}$ the subsequence of $(T_j)_{j \in \overline{m}}$ such that we have, for all $j \in \overline{m}$,

$$\alpha_j \neq \alpha_{j+1}. \tag{35}$$

Let us prove statement 3 and $\alpha_0 < \alpha_1$.

Due to Definition 3.1, there exists a finite number of α in $\mathcal{A}$ such that $X(T_0)$ is in $\Omega_{\alpha,1}$; then due to (23), (27), and (29), there exists α in $\mathcal{A}$ such that $S_{d,\alpha}(T_0) = 1$, and thus by (22), we have $\alpha_0 = \max\{\alpha, S_{d,\alpha}(T_0) = 1\}$.

This implies with (33) that $X(T_0)$ is in $\Omega_{\alpha_0,5}$. Similarly, we can prove that, for all β in $\mathcal{A}$ such that $\alpha < \beta$, we have $X(T_0)$ is not in $\Omega_{\beta,1}$. Thus (29), the fact that χ is an admissible radius for the external noise, (28), and assumption A3 on the function p yield, for all t in $[T_0, T_1)$, $X(t)$ is in $\Omega_{\alpha_0,5}$. Therefore with Lemma 4.2, we deduce that S_{d,α_0} cannot switch at time T_1 and, for all t in $[T_1, T_2)$, $S_{d,\alpha_0}(t) = 1$.

Moreover, due to (22), for all t in (T_1, T_2), we have $S_{d,\alpha_1}(t) = 1$. So, due to (22) and (35), $\alpha_0 < \alpha_1$.

Let us prove the following Claim 4.8, which implies statements 4 and 5 of Lemma 4.7.

Claim 4.8. For all $j > 0$, $j \in \overline{m}$, and for all t in $[T_j, T_{j+1})$, $X(t)$ is in $\mathrm{clos}(\Omega_{\alpha_j,3})$ and $\alpha_j < \alpha_{j+1}$.

Proof of Claim 4.8. Let us prove Claim 4.8 by induction.

The inequality $\alpha_0 < \alpha_1$ implies with (22) that $S_{d,\alpha_1}(T_0) = 2$. Thus with Lemma 4.2, (28), and assumption A3 we have, for all t in $[T_1, T_2)$, $X(t)$ is in $\mathrm{clos}(\Omega_{\alpha_1,3}) \backslash \Omega_{\alpha_1,1}$. Thus with Lemma 4.2, S_{d,α_1} cannot switch at time T_2 and we have, for all t in $[T_2, T_3)$, $S_{d,\alpha_1}(t) = 1$. Moreover due to (22), for all t in (T_2, T_3), we have $S_{d,\alpha_2}(t) = 1$. So, due to (22) and (35), $\alpha_1 < \alpha_2$.

One can inductively deduce statements 4 and 5 for $j \geq 2$ similarly. ∎

To complete the proof of Lemma 4.7, note that statement 6 is a consequence of (32) and statements 2, 3, and 4 ∎

Remark 4.9. Some observations are in order.

- Lemma 4.7 states that for all π-solutions starting in M, the sequence α is strictly increasing and there exists a bound on the time between two switches. This result is analogous to [1, Proposition 3.1]. However, for all π-solutions that do not start in M, the sequence can be nonincreasing. See Example 4.10. Thus we need to add an initial switch to make all solutions enter in M. This is the result stated in Lemma 4.11.
- M is forward invariant for the system (30) for all ξ and ζ satisfying our regularity assumptions and (29).

Example 4.10. *Figure* 4.1 *shows two different* π*-solutions of the hybrid patchy vector field of Artstein's circles by taking account of Lemmas* 4.2 *and* 4.7. *On the left, the* x*-component of* π*-solutions is depicted, and, on the right, we have the evolution of the controllers.*

- *The* π*-solution* (X, S_d)*, a solid line, starts in* M *(with* $x_0 \in \Omega_{a,5}$ *and* $s_0 = (1, 2, 2)$*) at* $T_0 = 0$ *and has one switch at time* T_1 *(with* $X(T_1) + \xi(X(T_1), T_1) \in \Omega_{c,2}$ *and* $S_d(T_1) = (1, 2, 1)$*). With the notations of Lemma* 4.7*, we have* $(\alpha_1, \alpha_2) = (a, c)$ *(see Figure* 3.2*), which is strictly increasing.*

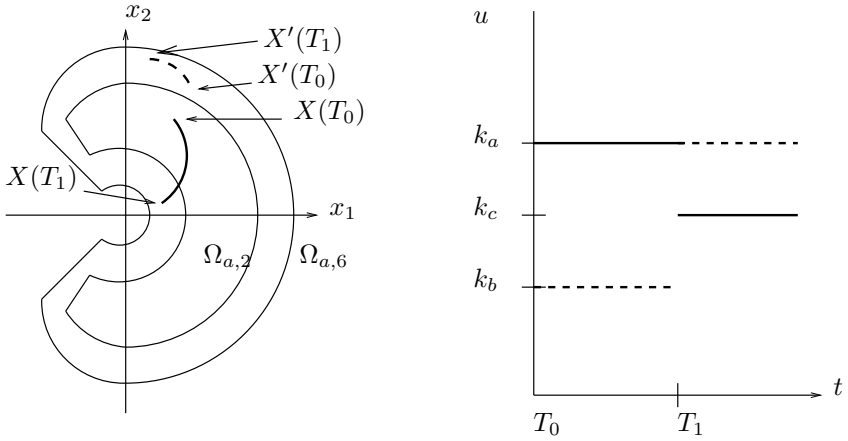


FIG. 4.1. *Two π-solutions with a hybrid patchy feedback. On the left is the x component and on the right is the evolution of the control.*

• *The other π-solution (X', S'_d), a dashed line, starts in the complement of M ($x_0 \in \Omega_{b,7} \setminus \Omega_{b,6}$ and $s_0 = (1,1,2)$) at $T_0 = 0$. There exists a measurement noise ξ satisfying* (29)*, vanishing for $t \neq T_0$, and such that $x_0 + \xi(x_0, T_0) \in \Omega_{b,6}$. Thus $S'_d(T_0) = (1,1,2)$. Therefore, with the notations of Lemma* 4.7*, we have $\alpha_1 = b$. There exists a time $T_1 > T_0$ such that $X'(T_1) \in \Omega_{b,7} \setminus \Omega_{b,6}$. Therefore $S'_d(T_1) = (1,2,2)$ and $\alpha_2 = a$. Thus $\alpha_1 > \alpha_2$ and the sequence is not strictly increasing.*

This example proves that the conclusions of Lemma 4.7 *do not hold for π-solutions which do not start in M (see statement* 5 *in Lemma* 4.7*).*

Due to property (34), we can add a switch to make all π-solutions enter in M. More precisely, let $(\Omega, ((\Omega_{\alpha,l})_{l\in\mathcal{F}}, g_\alpha)_{\alpha\in\mathcal{A}})$ be a family of nested patchy vector fields. Assume that we have (21). Then we can define a map $u : \{1,2\}^{\mathcal{A}} \to K$ by (22) and $\widetilde{k}_d : \mathbb{R}^n \times \{1,2\}^{\mathcal{A}} \to \{1,2\}^{\mathcal{A}}$ by

$$\begin{aligned} \widetilde{k}_d(x, s_d) \;&=\; k_d(x, s_d) \text{ if } \begin{cases} \{\beta \in \mathcal{A},\ s_{d,\beta} = 1\} \text{ is empty or infinite} \\ \text{or} \\ x \in \Omega_{\alpha,4},\ \text{where } \alpha = \max\{\beta, k_{d,\beta}(x, s_d) = 1\} \end{cases} \\ &\text{else} \\ &=\; s_0,\ \text{where } s_0 \text{ is such that } x \in \Omega_{\alpha,2},\ \text{and } \alpha = \max\{\beta, s_{0,\beta} = 1\}. \end{aligned} \tag{36}$$

Consider now the system

$$\begin{cases} \dot{x}(t) = f(x(t), u(s_d(t))) + \zeta(x,t), \\ s_d(t) = \widetilde{k}_d(x(t) + \xi(x,t), s_d^-(t)). \end{cases} \tag{37}$$

We rewrite Lemma 4.7 for all initial conditions.

LEMMA 4.11. *Let $0 < T \leq \infty$ and let (X, S_d) be a π-solution of* (37) *whose sampling schedule is subordinate to p, defined on $[0,T)$ and starting in $\mathbb{R}^n \times \{1,2\}^{\mathcal{A}}$. Then, there exist $m \in \mathbb{N} \cup \{+\infty\}$, an increasing sequence of time-instants $(T_j)_{j\in\overline{m}}$ in $[0,T)$, a sequence $(\alpha_j)_{j\in\overline{m}}$ in $\mathcal{A}$, and a sequence $(k_j)_{j\in\overline{m}}$ in K such that if we let $T_0 = 0$ and $T_{m+1} = T$ (if $m < +\infty$), we have, for all j in $\overline{m}$, the following:*

1. *For all* t *in* (T_j, T_{j+1}), $u(S_d(t)) = k_{\alpha_j}$.
2. *The map* X *is a Carathéodory solution of* $\dot{x} = f(x, k_{\alpha_j}) + \zeta$ *on* (T_j, T_{j+1}).
3. *For all* t *in* $[T_0, T_1)$, $X(t)$ *is in* $\Omega_{\alpha_0,4}$.
4. *For all* t *in* $[T_j, T_{j+1})$, $X(t)$ *is in* $\operatorname{clos}(\Omega_{\alpha_j,3})$ *if* $j \geq 1$.
5. *The sequence* $\alpha_1, \ldots, a_{m+1}$ *is strictly increasing.*
6. *The inequality* $T_{j+1} - T_j < \tau_{\alpha_j}$ *holds.*

Proof. The proof of statements 1 and 2 of Lemma 4.11 is analogous of the proof of statements 1 and 2 of Lemma 4.7.

Due to (36), $X(T_0) + \xi(X(T_0), T_0)$ is in $\Omega_{\alpha_0,4}$. Then due to (27) and (29), we have $X(T_0)$ is in $\Omega_{\alpha_0,5}$. Similarly, we can prove that, for all β in $\mathcal{A}$ such that $\alpha < \beta$, we have $X(T_0)$ is not in $\Omega_{\beta,1}$. Therefore with (29), the fact that χ is an admissible radius for the external noise, (28), and assumption A3 on the function p, we have statement 3.

This implies that $X(T_1)$ is in $\Omega_{\alpha_0,4}$ and therefore $(X(T_1), S_d(T_1))$ is in M and we deduce statements 4 to 6 of Lemma 4.11 from statements 4 to 6 of Lemma 4.7. ∎

Remark 4.12. Note that if there exists a switch (i.e., if $m > 0$ in Lemma 4.11), then, after the first switch, we have $\widetilde{k}_d(X(t) + \xi(X(t), t)) = k_d(X(t) + \xi(X(t), t))$. And thus after the first switch (if it exists), every π-solution of (37) is a π-solution of (30) and in particular we have the conclusion of Lemma 4.4.

5. Use of the asymptotic controllability. Now we use properties of a differential system in closed loop with a hybrid patchy feedback. The purpose of this section is to prove Theorem 2.7. Let us prove a generalization of [1, Proposition 4.1] which yields a feedback that is robust with respect to measurement noise.

PROPOSITION 5.1. *Let* (1) *be globally asymptotically controllable to the origin. Then for every* $0 < r < s$, *there exist* $T, R, \chi, d > 0$, *an open subset of* $\mathbb{R}^n$, $D^{r,s}$, *and a feedback control,* $u = u^{r,s} : \{1,2\}^{\mathbb{N}} \to K$, $k_d = k_d^{r,s} : \mathbb{R}^n \times \{1,2\}^{\mathbb{N}} \to \{1,2\}^{\mathbb{N}}$ *satisfying*

$$B(0,s) \setminus \operatorname{int}(B(0,r)) \subset D^{r,s} \subset B(0,R) \tag{38}$$

such that for any measurable maps $\zeta, \xi : [0, +\infty) \to \mathbb{R}^n$ *satisfying*

$$\sup_{t \geq 0} |\xi(t)| \leq \chi, \qquad \operatorname*{esssup}_{t \geq 0} |\zeta(t)| \leq \chi, \tag{39}$$

and for any initial state x_0 *in* $D^{r,s} \setminus \operatorname{int}(B(0,r))$, *and for any* s_0 *in* $\{1,2\}^{\mathbb{N}}$, *the perturbed system*

$$\begin{cases} \dot{x} = f(x, u(s_d)) + \zeta, \\ s_d = k_d(x + \xi, s_d^-) \end{cases} \tag{40}$$

admits a π*-solution* (X, S_d) *starting from* (x_0, s_0). *Moreover, for all* (x_0, s_0) *in* $\mathbb{R}^n \times \{1,2\}^{\mathbb{N}}$ *and for any d-maximal solution* (X, S_d) *starting from* (x_0, s_0) *and defined on* $[0, T)$, *there exists* $t_{X,S_d} \leq T$, *such that*

$$|X(t_{X,S_d})| < r. \tag{41}$$

Proof. We follow the proof of [1, Proposition 4.1] and we prove Proposition 5.1 in four steps.

Step 1. Fix $0 < r < s$. For each x_0 in $B(0,s)$, there exist a piecewise constant admissible control $u_0 = u_{x_0}$ and some constant $T_0 = T_{x_0}$ such that there exists a solution $X_0 = x(.; x_0, u_0)$ of $\dot{x} = f(x, u_0)$ for which the inequality

$$|X_0(T_0)| < \frac{r}{2} \tag{42}$$

holds. Moreover, by continuity, we can assume that we have

$$m_0 := \inf_{t\in[0,T_0]} |X_0(t)| > \frac{r}{4} \tag{43}$$

and, by possibly redefining u_0, we may assume that X_0 takes different values at any two different points t, t' in $[0, T_0]$. Let $\tau_{0,0} = 0 < \cdots < \tau_{0,N_0} = T_0$ be the points of discontinuity for u_0 on $[0, T_0]$ and $k_{0,j}$ in K be the corresponding values of u_0, i.e., we suppose that, for all j in $\{0, \ldots, N_0 - 1\}$ and for all t in $(\tau_{0,j}, \tau_{0,j+1})$, we have $u_0(t) = k_{0,j}$. Define

$$M_0 = M_{x_0} = \sup_{t\in[0,T_0]} |X_0(t)|. \tag{44}$$

There exist some strictly positive constants $c_0 = c_{x_0}$, $\overline{\rho}_0 = \overline{\rho}_{x_0}$ and $\overline{\chi}_0 = \overline{\chi}_{x_0}$ such that, for any fixed τ in $[0, T_0]$, any strictly positive radius $\rho \leq \overline{\rho}_0$ and $\chi \leq \overline{\chi}_0$, any initial point $\overline{x}$ in $B(X_0(\tau), \rho)$, and any Carathéodory solution $X_{\rho,\chi}(.)$ of

$$\dot{x} \in f(x, u_0(t)) + B(0, \chi) \tag{45}_\chi$$

starting from $\overline{x}$ at time $t = \tau$, we have

$$\sup_{[\tau, T_0+\rho]} |X_{\rho,\chi}(t) - X_0(t)| < c_0(\rho + \chi). \tag{46}$$

Let two strictly positive reals $\rho_0 \leq \overline{\rho}_0$ and $\chi_0 \leq \frac{\overline{\chi}_0}{2}$ be such that, letting

$$\rho_{x_0,1} = \rho_{0,1} = \rho_0, \tag{47}$$

and for all j in $\{2, \ldots, N_0 + 1\}$,

$$\rho_{x_0,j} = \rho_{0,j} = \Sigma_{k=0}^{j-2} 7^k c_0^{k+1} 2\chi_0 + 7^{j-1} c_0^{j-1} \rho_0,$$

we have

$$7\rho_{x_0,N_0+1} < \frac{r}{2} \tag{48}$$

and

$$\max_j \rho_{x_0,j} < \frac{1}{7} \min\left(m_0 - \frac{r}{8}, \overline{\rho}_0\right), \tag{49}$$

where m_0 is defined by (43). Thus it follows directly by induction that if for any fixed $j = 1, \ldots, N_0$ and for any $\overline{x}$ such that

$$\overline{x} \in B(X_0(\tau_{0,j}), 7\rho_{0,j}),$$

we consider any Carathéodory solution $X_{\rho_{0,j},2\chi_0}(.)$ of $(45)_{2\chi_0}$ starting from $\overline{x}$ at time $\tau_{0,j}$, then one has

$$\sup_{t\in[\tau_{0,j}, T_0+\rho_{0,j}]} |X_{\rho_{0,j},\chi_0}(t) - X_0(t)| < \rho_{0,j+1}. \tag{50}$$

Step 2. For j in $\{0, \ldots, N_0 - 1\}$ and for any $\overline{x}$ in $\mathbb{R}^n$, denote $\mathcal{A}_j(\overline{x}, t)$ the attainable set in time t for the Carathéodory solutions of $\dot{x} \in f(x, k_{0,j}) + B(0, 2\chi_0)$ starting from $\overline{x}$. Let us define for all l in $\{1, \ldots, 7\}$ and for all j in $\{1, \ldots, N_0 - 1\}$ the open sets

$$\Gamma_{x_0,j,l} = \bigcup_{\substack{\overline{x}\in \mathrm{int}(B(X_0(\tau_{0,j-1}), l\rho_{0,j})) \\ 0\leq t\leq \tau_{0,j}-\tau_{0,j-1}}} \mathcal{A}_j(\overline{x}, t)$$

and

$$\Gamma_{x_0,N_0,l} = \bigcup_{\substack{\overline{x}\in \mathrm{int}(B(X_0(\tau_{0,N_0-1}),l\rho_{0,j}))\\ 0\le t\le T_0+\rho-\tau_{0,N_0-1}}} \mathcal{A}_{N_0}(\overline{x},t).$$

Note that we have, for all j in $\{1,\dots,N_0\}$ and for all $l<l'$ in $\{1,\dots,7\}$,

$$\Gamma_{x_0,j,l} \subsetneq \mathrm{clos}(\Gamma_{x_0,j,l}) \subsetneq \Gamma_{x_0,j,l'}, \tag{51}$$

and, due to (44) and (50),

$$\Gamma_{x_0,j,l} \subset B(0,7\rho_{x_0,j+1}+M_{x_0}). \tag{52}$$

Moreover (43), (49), and (50) yield

$$B\left(0,\frac{r}{8}\right) \subset \mathbb{R}^n \setminus \Gamma_{x_0,j,l}. \tag{53}$$

Finally, for all N_0-tuple s with entries in $\{1,\dots,7\}$, let us define

$$\Delta_{x_0,s} = \bigcup_{j=1}^{N_0} \Gamma_{x_0,j,s(j)}, \qquad \Delta_{x_0} = \Delta_0 = \Delta_{x_0,\mathbf{1}_{x_0}},$$

where, for any point x_0 in $B(0,s)\setminus \mathrm{int}(B(0,r))$, we denote $\mathbf{1}_{x_0}$ (whose length depends on x_0) the following constant sequence:

$$\begin{array}{rcl} \mathbf{1}_{x_0} = \mathbf{1}_0 : \{1,\dots,N_0\} & \to & \{1,\dots,7\},\\ l & \mapsto & 1. \end{array}$$

Let $g_{0,j} = g_{x_0,j}$ be the vector field on $\mathbb{R}^n$ defined, for all j in $\{1,\dots,N_0-1\}$, by

$$g_{0,j}(x) = f(x,k_{0,j}). \tag{54}$$

Let $\mathcal{F}$ be the finite set $\mathcal{F}=\{1,\dots,7\}$. We can claim that

$$(\Delta_0, ((\Gamma_{0,j,l})_{l\in\mathcal{F}}, g_{0,j})_{j\in\{1,\dots,N_0\}})$$

is a family of nested patchy vector fields. Indeed note first that due to (51) we have (17); second, we have (18) because there exists $T>0$ such that all solutions of $(45)_{2\chi_0}$ starting in $\partial\Gamma_{x_0,j,l}\setminus\bigcup_{j'>j}\Gamma_{x_0,j',1}$ stay in $\mathrm{clos}(\Gamma_{x_0,j,l})$ for all t in $[0,T)$; third, the other properties to fulfill Definition 3.1 are obvious.

Let $p_0:\mathbb{R}^n\to\mathbb{R}$ satisfy assumptions A1 to A3 for this family of nested patchy vector fields and define

$$d_0 = \inf_{x\in\Delta_0} p_0(x).$$

Moreover, due to (54), we define a hybrid patchy feedback (u^0,k_d^0) as considered in Definition 3.4 and thus a feedback control $(u^0,\widetilde{k}_d^0)$ defined by (36). We take χ_0 smaller and suppose that

$$0<\chi_0<\frac{1}{2}\min_{j\in\{1,\dots,N_0\}}\min_{l\in\{1,\dots,6\}} d(\mathbb{R}^n\setminus\Gamma_{0,j,l+1},\Gamma_{0,j,l})$$

and note that χ_0 is an admissible radius for the external disturbances. Consider two measurable maps $\zeta, \xi : [0, +\infty) \to \mathbb{R}^n$ satisfying

$$\sup_{t\geq 0} |\xi(t)| \leq \chi_0, \qquad \operatorname*{esssup}_{t\geq 0} |\zeta(t)| \leq \chi_0,$$

and let (x_0, s_0) be an initial condition in $\Delta_0 \setminus B(0,r) \times \{1,2\}^{N_0}$. Due to Lemma 4.3, there exists (X, S_d) a d_0-maximal solution of (40) in closed loop with $(u^0, \widetilde{k}_d^0)$ starting from (x_0, s_0) and defined on $[0, T)$. Moreover, due to Lemma 4.11, there exist $H \in \mathbb{N} \cup \{+\infty\}$, a sequence of points $t_0 = 0 < \cdots < t_H \leq 2T_0$, and a sequence of indices $j_0, \ldots, j_H$ in $\{1, \ldots, N_0\}$, such that, for all h in $\{1, \ldots, H-1\}$,

$$\forall t \in [t_h, t_{h+1}), \quad X(t) \in \Gamma_{x_0, j_h, 7}, \tag{55}$$

$$t_{h+1} - t_h \leq \tau_{0, j_h}. \tag{56}$$

Note that due to Lemma 4.11, the sequence $j_1, \ldots, j_H$ described below is strictly increasing. Due to (52) and (55), (X, S_d) cannot blow up in $\Gamma_{x_0, j_h, 7}$ for all h in $\{0, \ldots, H-1\}$, and due to Lemma 4.4, (42), (48), (52), and (55), there exists $T^0_{X,S_d} \leq 2T_0$ such that we have the inequalities

$$\forall t \in [0, T^0_{X,S_d}), \quad |X(t)| \leq 7 \max_{j\in\{1,\ldots,N_0\}} \rho_{0,j+1} + M_0, \tag{57}$$

$$|X(T^0_{X,S_d})| < r. \tag{58}$$

Step 3. Since x_0 is in Δ_{x_0}, the family of open tubes $\{\Delta_{x_0},\ r \leq |x_0| \leq s\}$ forms an open covering of the compact set $B(0,s) \setminus \operatorname{int}(B(0,r))$. Let

$$\{\Delta_i,\ i \in \{1, \ldots, N(r,s)\}\}, \quad \Delta_i = \bigcup_{j=1}^{N_i} \Gamma_{i,j,1}, \quad \Gamma_{i,j,1} = \Gamma_{x_i, j, \mathbf{1}_{x_i}(j)},$$

be a finite subcover. Denote

$$k_{i,j} = k_{x_i, j}$$

and the vector field

$$g_{i,j}(x) = f(x, k_{i,j}) \tag{59}$$

defined on $\mathbb{R}^n$. The index set

$$\mathcal{A} = \{(i,j),\ i \in \{1, \ldots, N(r,s)\},\ j \in \{1, \ldots, N_i\}\}$$

can be totally ordered by letting

$$(i,j) < (h,k) \quad \text{if either } i < h \text{ or else } i = h,\ j < k. \tag{60}$$

Let

$$D^{r,s} = \bigcup_{i=1}^{N(r,s)} \Delta_i.$$

We can now define a family of nested patchy vector fields on $\Omega = D^{r,s}$. Let, for all $(\alpha, l) \in \mathcal{A} \times \mathcal{F}$, $\Omega_{\alpha,l}$ be the open set $\Gamma_{i,j,l}$, where $(i,j) = \alpha$. Note that due to (51), for

all $m > l$ in $\mathcal{F}$ and for all α in $\mathcal{A}$, (17) and (18) can be proved as in Step 2. Therefore we can claim that

$$(\Omega, ((\Omega_{\alpha,l})_{l\in\mathcal{F}}, g_\alpha)_{\alpha\in\mathcal{A}})$$

is a family of nested patchy vector fields. Let $p^{r,s} : \mathbb{R}^n \to \mathbb{R}$ satisfy assumptions A1 to A3 for this family of nested patchy vector fields and

$$d^{r,s} = \inf_{x\in\Omega} p^{r,s}(x). \tag{61}$$

Moreover, due to (59), we can define a hybrid patchy feedback $(u^{r,s}, k_d^{r,s})$ and thus a feedback control $(u^{r,s}, \widetilde{k}_d^{r,s})$ as in (36). We let

$$\chi^{r,s} = \min_{1\le i\le N(r,s)} \chi_{x_i}, \tag{62}$$

which is an admissible radius for the external disturbances. We can choose $\chi^{r,s}$ smaller and suppose that

$$0 < \chi^{r,s} < \frac{1}{2} \min_{(i,j)\in\mathcal{A}} \min_{l\in\{1,\dots,6\}} d(\mathbb{R}^n \setminus \Gamma_{i,j,l+1}, \Gamma_{i,j,l}).$$

Then $\chi^{r,s}$ is an admissible radius for the measurement noise on $D^{r,s}$.

Step 4. For all x_0 in $B(0,s)\setminus B(0,r)$, let $T_{x_i} > 0$ be defined just at the beginning of Step 1, and let $\rho_{x_i} >$ be defined by $(45)_{\chi_0}$–(46). Moreover let $\chi^{r,s} > 0$ be defined by (62) and $d^{r,s}$ be defined by (61). Let

$$T = 2 \sum_{i=1}^{N(r,s)} T_{x_i},$$

and consider two measurable maps ξ and $\zeta : [0,+\infty) \to \mathbb{R}^n$ such that

$$\sup_{t\ge 0} |\xi(t)| \le \chi^{r,s}, \qquad \operatorname*{esssup}_{t\ge 0} |\zeta(t)| \le \chi^{r,s}.$$

Let (x_0, s_0) be an initial condition in $D^{r,s} \setminus B(0,r) \times \{1,2\}^{\mathcal{A}}$. Due to Lemma 4.3, there exists (X, S_d) a $d^{r,s}$-maximal solution of (40) in closed loop with $(u^{r,s}, \widetilde{k}_d^{r,s})$ starting from (x_0, s_0). Moreover, due to properties established in Step 3 and Lemma 4.11, there exist $H \in \mathbb{N}\cup\{+\infty\}$, a sequence of points $t_0 = 0 < \cdots < t_H \le T$, and a sequence of indices $\alpha_1, \dots, \alpha_H$ in $\mathcal{A}$, such that, for all h in $\{0, \dots, H-1\}$,

$$\forall t \in [t_h, t_{h+1}), \quad X(t) \in \Gamma_{\alpha_h,7}, \tag{63}$$

$$t_{h+1} - t_h < \tau_{\alpha_h}. \tag{64}$$

Note that due to Lemma 4.11, the sequence $\alpha_1, \dots, \alpha_H$ described above is strictly increasing. Due to (52) and (63), (X, S_d) cannot blow up in $\Gamma_{\alpha_h,7}$ for all h in $\{0, \dots, H-1\}$, and due to Lemma 4.4, (63), there exists $T_{X,S_d} \le T$ such that we have the inequalities

$$\forall t \in [0, T], \quad |X(t)| < R, \tag{65}$$

$$|X(T_{X,S_d})| < r, \tag{66}$$

where R is defined by

$$R = \sup_{1\leq i\leq N(r,s),1\leq j\leq N_i} \{7\rho_{x_i,j} + M_{x_i}\}. \tag{67}$$

This completes the proof of Proposition 5.1. □

PROPOSITION 5.2. *Let* (1) *be globally asymptotically controllable to the origin. Then for any fixed $\varepsilon > 0$, there exists $\sigma > 0$ such that for every $0 < r < s \leq \sigma$, there exist $T, R, \chi, d > 0$, an open subset of $\mathbb{R}^n$, $D^{r,s}$, and a feedback control, $u = u^{r,s} : 2^{\mathbb{N}} \to K$, $k_d = k_d^{r,s} : \mathbb{R}^n \times 2^{\mathbb{N}} \to 2^{\mathbb{N}}$ as in Proposition* 5.1 *with*

$$R < \varepsilon. \tag{68}$$

Proof. The proof is similar to the proof of [1, Proposition 4.2] and consists of properly choosing the piecewise constant admissible control u_{x_0} for each point x_0 in $B(0,s) \setminus \mathrm{int}(B(0,r))$.

To do this, fix $\varepsilon > 0$. Since (1) is globally asymptotically controllable, there exists $\sigma = \sigma(\varepsilon) > 0$ such that, for any fixed $0 < r < s \leq \sigma$, the conclusions of Proposition 5.1 hold together with

$$M_{x_0} < \frac{\varepsilon}{2},$$
$$\max_{j\in\{1,\dots,N_0+1\}} \rho_{x_0,j} < \frac{\varepsilon}{14}$$

for all x_0 satisfying $r < |x_0| < s$. With (65) and (67), this implies (68). □

We are now ready to prove Theorem 2.7.

Proof of Theorem 2.7.

Part 1: Definition of the feedback control. Let $(r_n)_{n\in\mathbb{Z}}$ and $(s_n)_{n\in\mathbb{Z}}$ be two decreasing sequences of strictly positive numbers such that

- for all n in $\mathbb{Z}$, we have $r_{n-1} < s_n$;
- s_n converges to zero as $n \to +\infty$;
- r_{-n} converges to infinity as $n \to +\infty$.

Let $T_n = T(r_n, s_n)$, $R_n = R(r_n, s_n)$, $\chi_n = \chi(r_n, s_n)$ be three sequences of strictly positive numbers and consider a sequence of hybrid patchy feedbacks

$$(D^{r_n,s_n}, u^{r_n,s_n}, k_d^{r_n,s_n}, ((\Gamma_{i,l}^n)_{l\in\{1,\dots,7\}}, k_i^n)_{i\in\{1,\dots,N_n\}})$$

such that

$$R_n < \frac{1}{n} \quad \forall n \in \mathbb{N}_{>0} \tag{69}$$

as in Proposition 5.2. The index set

$$\mathcal{B} = \{(n,i),\ n \in \mathbb{Z},\ i \in \{1,\dots,N_n\}\}$$

can be totally ordered with the same relation of order as (60), i.e., by letting

$$(n,i) < (m,j) \quad \text{if either } n < m \text{ or else } n = m,\ i < j.$$

Then we have the following family of nested patchy vector fields on $\mathbb{R}^n \setminus \{0\}$:

$$(\mathbb{R}^n \setminus \{0\}, ((\Gamma_{i,l}^m)_{l\in\mathcal{F}}, k_i^m),\ (m,i) \in \mathcal{B}).$$

We can define a hybrid patchy feedback (u, k_d) on $\mathbb{R}^n \setminus \{0\}$ as in Definition 3.4. Let $\chi : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}_{>0}$ be a continuous map satisfying

$$\chi(x) \leq \min\left(\chi_n, \frac{|x|}{2}\right) \quad \text{if } x \in D^{r_n,s_n} \setminus \bigcup_{m>n} D^{r_m,s_m}. \tag{70}$$

We define $\chi(0) = 0$. The map χ is continuous at 0 and then χ is an admissible radius for the measurement noise and the external disturbances. Let $(u, \widetilde{k}_d)$ be the feedback control defined by (36) for the hybrid patchy feedback (u, k_d). Let us prove that $(u, \widetilde{k}_d)$ is a global robust stabilizing controller on $\mathbb{R}^n$, i.e., that the origin of system (37) is a robust globally asymptotically stable equilibrium as stated in Theorem 2.7.

Part 2: Theorem 2.7 for π-solutions. Let $p : \mathbb{R}^n \to \mathbb{R}$ be a function continuous on $\mathbb{R}^n \setminus \{0\}$ satisfying the properties A1, A2, and A3.

Existence of π-solutions. Consider ξ, ζ satisfying our regularity assumptions. Let (x_0, s_0) be in $\mathbb{R}^n \times \{1, 2\}^{\mathcal{B}}$. Let $s_1 = k_d(x_0 + \xi(x_0, 0), s_0)$ and α be in $\mathcal{B}$ such that $k_\alpha = u(s_1)$. From our regularity assumptions on f and ζ, the Carathéodory conditions are satisfied for system (31). Let $0 < T \leq t_1$ and $X : [0, T) \to \mathbb{R}^n$ be a Carathéodory solution of (31). Let S_d be defined by $S_d(t) = s_1$ for all t in $[0, T)$. The map (X, S_d) is a π-solution of (37) starting from (x_0, s_0).

Completeness and global stability for π-solutions. Let $\varepsilon > 0$. Let $n \in \mathbb{N}$ be such that $\varepsilon < R_{-n}$. Such an R_{-n} exists because we have $r_{-n} \leq R_{-n}$ and r_{-n} tends to infinity as $n \to +\infty$. Let $\chi_0 > 0$ be defined by

$$\chi_0 = \inf_{x \in B(0,R_{-n}) \setminus B(0,r_{-n})} \chi(x), \tag{71}$$

and let $d_0 > 0$ satisfy the inequalities

$$d_0 < \frac{d(\mathbb{R}^n \setminus B(0, s_{-n}), B(0, r_{-n}))}{\max_{x \in B(0,s_{-n}),\ u \in K} |f(x, u)|}, \tag{72}$$

and for all x in $B(0, R_{-n}) \setminus B(0, r_{-n})$, for all y in $B(0, \chi(x))$,

$$d_0 < p(x + y). \tag{73}$$

Note that due to (70), we have $d_0 > 0$ and $\chi_0 > 0$. Let ξ, ζ satisfy our regularity assumptions and (12). Let (X, S_d) be a d_0-maximal solution of (37) on $[0, T)$ starting from (x_0, s_0) with $|x_0| < s_{-n}$ and (11).

Note that due to (71)–(73), for all $i \in \mathbb{N}$ such that $X(t_i)$ is in $B(0, R_{-n}) \setminus B(0, r_{-n})$, we have (28) and, for all t such that $X(t)$ is in $B(0, R_{-n}) \setminus B(0, r_{-n})$, we have (29).

Therefore, due to Proposition 5.1 and to the definition of the feedback control, if there exists $i \in \mathbb{N}$ such that $X(t_i)$ is in $B(0, s_{-n}) \setminus B(0, r_{-n})$, then there exists $j > i$ such that $X(t_j)$ is in $B(0, r_{-n})$ and for all t in $[i, j]$, $X(t)$ is in $B(0, R_{-n})$. Moreover, due to (72), if there exists $i \in \mathbb{N}$ such that $X(t_i)$ is in $B(0, r_{-n})$, then, for all t in $[t_i, t_{i+1}]$, we have $X(t)$ is in $B(0, s_{-n})$.

Thus we have, for all t in $[0, T)$,

$$|X(t)| \leq R_{-n}. \tag{74}$$

Therefore the conclusion of Lemma 4.4 cannot hold ($\limsup_{t \to T} |X(t)| \neq +\infty$) and thus we have $T = +\infty$ and the maximality property.

Finally, note that $\delta(\varepsilon) = s_{-n}$ tends to $+\infty$ as ε tends to infinity because when ε tends to infinity, n tends to infinity, r_{-n} tends to infinity, and we have $r_{-n-1} < s_{-n}$. Thus we have the stability property.

Global attractivity for π-solutions. Let $\varepsilon > 0$ and $C > 0$. Let $n \in \mathbb{N}$ be such that $\frac{1}{n} < \varepsilon$ and such that $\delta < r_{-n}$. Let $d_0 > 0$ and $\chi_0 > 0$ be defined, respectively, by

$$d_0 = \inf_{x \in B(0,R_{-n}) \setminus B(0,r_n)} p(x), \tag{75}$$

$$\chi_0 = \inf_{x \in B(0,R_{-n}) \setminus B(0,r_n)} \chi(x). \tag{76}$$

Let ξ, ζ satisfying our regularity assumptions and (12). Let (X, S_d) be a π-solution defined on $[0, +\infty)$, starting from (x_0, s_0) whose sampling schedule satisfies $\overline{d}(\pi) < d_0$ and whose initial condition satisfies $|x_0| < C$. Due to Proposition 5.1, there exists $\widetilde{T}$ in $[0, T_{-n}+T_{-n+1}+\cdots+T_n]$ such that $|X(\widetilde{T})| < r_n$. Let $T' = \inf\{t \in [0, \widetilde{T}],\ |X(t)| < s_n\}$. Then due to the stability property and as $R_n < \frac{1}{n}$, we have

$$\forall t \geq \widetilde{T}, \quad |X(t)| \leq \frac{1}{n}.$$

Therefore we have (14) with $T = T_{-n} + \cdots + T_n$.

Part 3: Theorem 2.7 for the generalized solutions.

Existence and completeness for the generalized solutions. This results from the fact that every π-solution of (37) is a generalized solution of (37).

Global stability and global attractivity for the generalized solutions. Let $\varepsilon > 0$. Let $\chi_0 > 0$, $d_0 > 0$ and δ of $\mathcal{K}_\infty$, be such that we have the stability property (13) for all π-solutions of (37) whose sampling schedule satisfies (11) and for all ξ, ζ satisfying our regularity assumptions and (12).

Let X be a generalized solution of (37) starting from $x_0 \in B(0, \frac{\delta(\varepsilon)}{2})$ with ξ, ζ satisfying our regularity assumptions and

$$\sup_{x \in \mathbb{R}^n,\, t \geq 0} |\xi(x,t)| \leq \frac{\chi_0}{2}, \qquad \operatorname*{esssup}_{x \in \mathbb{R}^n,\, t \geq 0} |\zeta(x,t)| \leq \frac{\chi_0}{2} \tag{77}$$

and obtained as limit of π-solutions (X^n, S_d^n) whose sampling schedule satisfies (11). Let us prove (13).

For n sufficiently large, we have

$$\sup_J |e_n(t)| + \operatorname*{esssup}_J |d_n(t)| < \frac{\chi_0}{2} \tag{78}$$

for all J compact subinterval of $[0, T)$. Then for n sufficiently large, (X^n, S_d^n) is a π-solution of (37) whose sampling schedule satisfies (11) with a disturbance satisfying (12). Then we have (13) for this sequence of π-solutions. Therefore we have (13) for the generalized solution X.

The global attractivity property can be proved similarly.

Part 4: Theorem 2.7 for Euler solutions.

Existence and completeness for Euler solutions. Let x_0, s_0 in $\mathbb{R}^n \times \{1, 2\}^{\mathcal{B}}$ and π_n be a sequence of sampling schedules such that $\overline{d}(\pi_n) \to 0$ as n tends to infinity. Let (X^n, S_d^n) be a π_n-solution of (37), starting from (x_0, s_0) and defined on $[0, +\infty)$. Due to Part 2 of the proof of Theorem 2.7, this sequence exists for n sufficiently large and there exists R such that, for all t in $[0, +\infty)$ and for n sufficiently large, we have

$$|X_n(t)| < R.$$

Therefore with Ascoli's theorem, we can define X an Euler solution defined on $[0, +\infty)$ and starting from x_0.

Global stability and global attractivity for Euler solutions. Let $\varepsilon > 0$. Let $\chi_0 > 0$, $d_0 > 0$ and δ of $\mathcal{K}_\infty$ be such that we have the stability property (13) for all d_0-maximal solutions of (37) and for all ξ, ζ satisfying our regularity assumptions and (12).

Let X be an Euler solution of (37) starting from $x_0 \in B(0, \frac{\delta(\varepsilon)}{2})$ with ξ, ζ satisfying our regularity assumptions and (12) and obtained as limit of π-solutions (X^n, S_d^n) satisfying $\overline{d}(\pi_n) \to 0$ as n tends to infinity.

Let us prove (13).

For n sufficiently large, we have $\overline{d}(\pi_n) < d_0$. Then for n sufficiently large, (X^n, S_d^n) is a π-solution of (37) whose sampling schedule satisfies (11) with a disturbance satisfying (12). Then we have (13) for this sequence of π-solutions. Therefore we have (13) for the generalized solution X.

The global attractivity can be proved similarly.

This completes the proof of Theorem 2.7. □

Acknowledgments. The author is deeply grateful to the reviewers for helpful suggestions about the presentation of this paper and to Jean-Michel Coron and Laurent Praly for many stimulating discussions.

REFERENCES

[1] F. Ancona and A. Bressan, *Patchy vectors fields and asymptotic stabilization*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 445–471.

[2] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.

[3] V. Andrian, A. Bacciotti, and G. Beccari, *Global stability and external stability of dynamical systems*, Nonlinear Anal., 28 (1997), pp. 1167–1185.

[4] J. Behrens and F. Wirth, *A globalization procedure for locally stabilizing controllers*, in Nonlinear Control in the Year 2000, Vol. 1, A. Isidori et al., eds., Lecture Notes in Control and Inform. Sci. 258, Springer-Verlag, London, 2000, pp. 171–182.

[5] A. Bensoussan and J.L. Menaldi, *Hybrid control and dynamic programming*, Dynam. Contin. Discrete Impuls. Systems, 3 (1997), pp. 395–442.

[6] R.W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R.W. Brockett, R.S. Millman, and H.J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 181–191.

[7] F.H. Clarke, Yu.S. Ledyaev, L. Rifford, and R.J. Stern, *Feedback stabilization and Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 25–48.

[8] F.H. Clarke, Yu.S. Ledyaev, E.D. Sontag, and A.I. Subbotin, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[9] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern, and P.R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[10] F.H. Clarke, L. Rifford, and R.J. Stern, *Feedback in state constrained optimal control*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 97–133.

[11] O. Hájek, *Discontinuous differential equations, part* I, J. Differential Equations, 32 (1979), pp. 149–170.

[12] H. Hermes, *Discontinuous vector fields and feedback control*, in Differential Equations and Dynamic Systems, J.K. Hale and J.P. La-Salle, eds., Academic Press, New York, London, 1967.

[13] N.N. Krasovkii and A.I. Subbotin, *Game-Theorical Control Problems*, Springer-Verlag, New York, 1988.

[14] Y.S. Ledyaev and E.D. Sontag, *A remark on robust stabilization of general asymptotically controllable systems*, in Proc. Conf. on Information Sciences and Systems (CISS 97), Baltimore, MD, 1997, pp. 246–251.

[15] Y.S. Ledyaev and E.D. Sontag, *A Lyapunov characterization of robust stabilization*, Nonlinear Anal., 37 (1999), pp. 813–840.

[16] C. Prieur, *A robust globally asmptotically stabilizing feedback: The example of the Artstein's circles*, in Nonlinear Control in the Year 2000, Vol. 2, A. Isidori et al., eds., Lecture Notes in Control and Inform. Sci. 258, Springer-Verlag, London, 2000, pp. 279–300.

[17] C. Prieur, *Uniting local and global controllers with robustness to vanishing noise*, Math. Control Signals Systems, 14 (2001), pp. 143–172.

[18] C. Prieur and A. Astolfi, *Robust stabilization of chained systems via hybrid control*, IEEE Trans. Automat. Control, 48 (2003), pp. 1768–1772.
[19] L. Rifford, *Stabilisation des systèmes globalement asymptotiquement commandables*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 211–216.
[20] E.P. Ryan, *Discontinuous feedback and universal adaptive stabilization*, in Control of Uncertain Systems, D. Hinrichsen and B. Mårtensson, eds., Birkhäuser, Boston, 1990, pp. 245–258.
[21] E.D. Sontag, *Clocks and insensitivity to small measurement errors*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 537–557.
[22] E.D. Sontag, *Stability and stabilization: Discontinuities and the effect of disturbances*, in Nonlinear Analysis, Differential Equations, and Control, F.H. Clarke and R.J. Stern, eds., Kluwer, Dordrecht, The Netherlands, 1999, pp. 551–598.
[23] E.D. Sontag and H. Sussmann, *Remarks on continuous feedback*, in Proceedings of the IEEE Conference on Decision and Control, Albuquerque, NM, 1980, pp. 916–921.
[24] H.J. Sussmann, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.
[25] L. Tavernini, *Differential automata and their discrete simulators*, Nonlinear Anal., 11 (1997), pp. 665–683.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 5, pp. 1913–1922

THE VALUE OF ZERO-SUM STOPPING GAMES IN CONTINUOUS TIME*

RIDA LARAKI† AND EILON SOLAN‡

Abstract. We study two-player zero-sum stopping games in continuous time and infinite horizon. We prove that the value in randomized stopping times exists as soon as the payoff processes are right-continuous. In particular, as opposed to existing literature, we do *not* assume any conditions on the relations between the payoff processes.

Key words. Dynkin games, stopping games, optimal stopping, stochastic analysis, continuous time, stochastic duels

AMS subject classifications. Primary, 91A55; Secondary, 91A10

DOI. 10.1137/S0363012903429025

1. Introduction. In many competitive interactions the main strategic issue is timing. To model such situations, Dynkin (1969) introduced stopping games, as a variation of optimal stopping problems. In Dynkin's setup, two players observe the realization of a payoff process in discrete time. Once one of the players decides to stop, player 2 pays player 1 the amount indicated by the payoff process. However, at every given stage only one of the players is allowed to stop; the identity of that player is governed by another process. The strategic choice of each player is the choice of his stopping time. Dynkin (1969) proved that those games admit a value.

Dynkin's seminal paper was extended in various directions. Neveu (1975) allowed the players to stop *simultaneously* and provided a sufficient condition for the existence of the value. Several authors, including Bismut (1977), Alario-Nazaret, Lepeltier, and Marchal (1982), Lepeltier and Maingueneau (1984), and Stettner (1984) studied the problem in *continuous time*.

Yasuda (1985) studied stopping games in discrete time (with either finite horizon or discounted payoff), and allowed the players to choose *randomized* stopping times. Yasuda (1985) proved that the value exists under merely an integrability condition. Rosenberg, Solan, and Vieille (2001) studied the infinite horizon game in discrete time and proved an analogous result. Touzi and Vieille (2002) provided a sufficient condition that ensures the existence of the value in randomized stopping times in continuous time. As their proof utilizes a fixed point argument, it is not constructive.

In the present paper we prove that under merely integrability and continuity conditions, every stopping game in continuous time admits a value in randomized stopping times. In addition, we construct ε-optimal randomized stopping times which are as close as one wishes to (nonrandomized) stopping times; roughly speaking, there

*Received by the editors June 2, 2003; accepted for publication (in revised form) June 17, 2004; published electronically March 22, 2005. The results presented in this paper were proved while the authors attended the workshop on "Stochastic Methods in Decision and Game Theory," organized by Marco Scarsini in June 2002, Erice, Sicily, Italy.

http://www.siam.org/journals/sicon/43-5/42902.html

†CNRS and Laboratoire d'Econométrie de l'Ecole Polytechnique, 1, rue Descartes, 75005 Paris, France (laraki@poly.polytechnique.fr).

‡MEDS Department, Kellogg School of Management, Northwestern University, Evanston, IL 60208, and School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel (e-solan@kellogg.northwestern.edu, eilons@post.tau.ac.il). The research of this author was supported by the Israel Science Foundation (grant 69/01-1).

is a stopping time μ such that for every δ sufficiently small there is an ε-optimal randomized stopping time that stops with probability 1 between times μ and $\mu + \delta$.

Several models that have been extensively studied in different disciplines and that fall into the category of stopping games are wars of attrition (see, e.g., Maynard-Smith (1974), Ghemawat and Nalebuff (1985), and Hendricks, Weiss, and Wilson (1988)), preemption games (see, e.g., Fudenberg and Tirole (1991, section 4.5.3)), duels, and pricing of options. We will illustrate the applicability of our results by discussing the last two models.

We first present the model of duels. In the simplest version, duels are two-player zero-sum games in which each of two gunners is endowed with a single bullet. The two gunners are located at some distance from each other and move closer to one another as time goes on. Since the accuracy of their shots improves as they get closer, it is not clear what the optimal moment is to shoot the opponent. If the accuracy is a stochastic process that depends on, e.g., wind velocity, the gunners face a stopping game.

Although for various classes of duels the existence of the value has been established, and optimal strategies have been computed (see, e.g., Blackwell (1949), Bellman and Girshick (1949), Shapley (1951), Karlin (1959), and the recent survey by Radzik and Raghavan (1994)), the general case is still open.

As we argue below, our results can be applied to any duel, regardless of the number of bullets each player initially has.

We now discuss the relevant literature in pricing of options. In most cases, a holder of an option has the right to exercise the option either on prespecified dates or whenever he chooses, so that the optimization problem reduces to an optimal stopping problem. Callable warrants (see, e.g., Merton (1973)) and convertible bonds (see, e.g., Brennan and Schwartz (1977)) allow for a certain action by the issuer as well. Recently Kifer (2000) introduced game contingent claims, which are general American options in which the issuer can terminate the contract early at some cost. Kifer showed that pricing these options boils down to determining the value of a certain stopping game, and he provided a general characterization for the value. Game contingent claims have been studied also by, e.g., Kallsen and Kühn (2004) and Kühn and Kyprianou (2003). Kyprianou (2004) used Kifer's characterization to explicitly calculate the value of game contingent claims in some cases. McConnell and Schwartz (1986) studied a specific example of callable option notes, which were actually issued in the 1980s.

In the formulation of game contingent claims in Kifer (2000), the right of the holder to exercise the option supersedes the right of the issuer to terminate the contract early, so that if those two events occur simultaneously, the holder gets to exercise the option. However, if the payment when those two events occur simultaneously is different from the payment if the holder were to exercise alone, or the issuer were to terminate the contract alone, Kifer's analysis would no longer be valid. Our result establishes the existence of the value in this case, and may be used, as was done by Kyprianou (2004), to find optimal strategies in given examples.

The paper is arranged as follows. The model and the main result appear in section 2, and the proof of the main result appears in section 3. Further topics, namely, introducing final payoffs and the right-continuity of the value process, are discussed in sections 3.4–3.5. We end by using the right-continuity of the value process to derive a more general existence result for noisy stochastic duels in section 3.6.

2. Model, literature, and main result. A *two-player zero-sum stopping game in continuous time* Γ is given by the following:

- A probability space $(\Omega, \mathcal{A}, P)$: $(\Omega, \mathcal{A})$ is a measurable space, and P is a σ-additive probability measure on $(\Omega, \mathcal{A})$.
- A filtration in continuous time $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$ satisfying "the usual conditions." That is, $\mathcal{F}$ is right-continuous, and $\mathcal{F}_0$ contains all P-null sets: for every $B \in \mathcal{A}$ with $P(B) = 0$ and every $A \subset B$, one has $A \in \mathcal{F}_0$. All stopping times in what follows are of the filtration $\mathcal{F}$.
  Denote $\mathcal{F}_\infty := \vee_{t\geq 0}\mathcal{F}_t$. Assume without loss of generality that $\mathcal{F}_\infty = \mathcal{A}$. Hence $(\Omega, \mathcal{A}, P)$ is a complete probability space.
- Three uniformly bounded $\mathcal{F}$-adapted processes $(a_t, b_t, c_t)_{t\geq 0}$.[1]

DEFINITION 1. *A* randomized stopping time *is a progressively measurable function* $\phi : [0,1] \times \Omega \to [0, +\infty]$ *such that for every* $r \in [0,1]$, $\mu_r(\omega) := \phi(r, \omega)$ *is an optional stopping time.*

For strategically equivalent definitions of randomized stopping times, see Touzi and Vieille (2002). Throughout the paper, the symbols μ and ν stand for stopping times, while ϕ and ψ stand for randomized stopping times.

For every pair (μ, ν) of stopping times we denote

$$\gamma(\mu, \nu) = \mathbf{E}_P\left[a_\mu \mathbf{1}_{\{\mu<\nu\}} + b_\nu \mathbf{1}_{\{\mu>\nu\}} + c_\mu \mathbf{1}_{\{\mu=\nu<+\infty\}}\right].$$

The *expected payoff* that corresponds to a pair of randomized stopping times (ϕ, ψ) is

$$\begin{aligned}(1)\qquad \gamma(\phi, \psi) &= \int_{[0,1]^2} \gamma(\mu_r, \nu_s)\, dr\, ds \\ &= \mathbf{E}_{\lambda\otimes\lambda\otimes P}\left[a_{\mu_r} \mathbf{1}_{\{\mu_r<\nu_s\}} + b_{\nu_s} \mathbf{1}_{\{\mu_r>\nu_s\}} + c_{\mu_r} \mathbf{1}_{\{\mu_r=\nu_s<+\infty\}}\right].\end{aligned}$$

Though the payoff function given by (1) is bilinear, without strong assumptions on the data of the game, the payoff function is not continuous for the same topology which makes the space of randomized stopping times compact.

DEFINITION 2. *If* $\sup_\phi \inf_\psi \gamma(\phi, \psi) = \inf_\psi \sup_\phi \gamma(\phi, \psi)$, *then the common value is the* value in randomized stopping times *and is denoted by* V. *Every randomized stopping time* ϕ *such that* $\inf_\psi \gamma(\phi, \psi)$ *is within* ε *of* V *is* ε-optimal *for player* 1; ε*-optimal randomized stopping times for player* 2 *are defined analogously.*

Observe that for every ϕ one has $\inf_\psi \gamma(\phi, \psi) = \inf_\nu \gamma(\phi, \nu)$, where ν ranges over all stopping times. Indeed, for every ϕ and ψ one has, by Fubini's theorem,

$$\gamma(\phi, \psi) = \mathbf{E}_{\lambda\otimes\lambda\otimes P}[\gamma(\mu_r, \nu_s)] \geq \inf_s \mathbf{E}_{\lambda\otimes\lambda\otimes P}[\gamma(\mu_r, \nu_s)] \geq \inf_\nu \gamma(\phi, \nu) \geq \inf_{\psi'} \gamma(\phi, \psi').$$

This implies that $\sup_\phi \inf_\psi \gamma(\phi, \psi) = \sup_\phi \inf_\nu \gamma(\phi, \nu)$. Similarly, $\inf_\psi \sup_\phi \gamma(\phi, \psi) = \inf_\psi \sup_\mu \gamma(\mu, \psi)$, where μ ranges over all pure stopping times. Recall that one always has $\sup_\phi \inf_\psi \gamma(\phi, \psi) \leq \inf_\psi \sup_\phi \gamma(\phi, \psi)$.

Touzi and Vieille (2002) provided a restrictive condition that ensures the existence of the value. The main result we present is the following.

THEOREM 3. *If the processes* $(a_t)_{t\geq 0}$ *and* $(b_t)_{t\geq 0}$ *are right-continuous, and if* $(c_t)_{t\geq 0}$ *is progressively measurable, then the value in randomized stopping times exists.*

3. Proof of the main result and extensions. From now on we fix a stopping game Γ that satisfies the assumptions of Theorem 3.

[1] Our results hold for a larger class of payoff processes, namely, the class $\mathcal{D}$ that was defined by Dellacherie and Meyer (1975, section II-18). This class contains in particular integrable processes.

3.1. Preliminaries. The following lemma will be used in what follows.

LEMMA 4. *For every stopping time τ and every $\varepsilon > 0$ there is $\delta > 0$ such that $P(\{|a_t - a_\tau| < \varepsilon \ \ \forall t \in [\tau, \tau + \delta]\}) > 1 - \varepsilon$.*

A similar statement holds when one replaces the process $(a_t)_{t\geq 0}$ by the process $(b_t)_{t\geq 0}$.

Proof. For every $n \in \mathbf{N}$, set $A_n = \{\sup\{|a_t - a_\tau|, \tau \leq t \leq \tau + 1/n\} < \varepsilon\}$. Since $(a_t)_{t\geq 0}$ is right-continuous, $P(\cup_{n\in\mathbf{N}} A_n) = 1$, and the result follows. □

One then obtains the following result.

COROLLARY 5. *Let a stopping time τ and $\varepsilon > 0$ be given. There exists $\delta > 0$ sufficiently small such that for every $\mathcal{F}_\tau$-measurable set $A \subseteq \{\tau < +\infty\}$ and every stopping time μ that satisfies $\tau \leq \mu \leq \tau + \delta$,*

$$|\mathbf{E}_P[a_\mu \mathbf{1}_A] - \mathbf{E}_P[a_\tau \mathbf{1}_A]| \leq \varepsilon.$$

3.2. The case $a_t \leq b_t$ for every $t \geq 0$. In this section we prove the following result: when $a_t \leq b_t$ for every $t \geq 0$, the value in randomized stopping times exists and is independent of $(c_t)_{t\geq 0}$.

The idea is as follows. Assume player 1 decides to stop at time t_*. If $c_{t_*} \geq a_{t_*}$, and player 1 stops with probability 1 at time t_*, player 2 has no incentive to stop at t_* as well. However, if $c_{t_*} < a_{t_*}$, player 1 needs to mask the exact time in which he stops, so that player 2 cannot stop at the same time. Since payoffs are right-continuous, player 1 can stop randomly in a small interval after time t_*. This way he makes sure that player 2 does not know the exact moment he will stop, and since $a_t \leq b_t$ for every t, player 2 has no incentive to stop in this time interval. In both cases, whatever the process $(c_t)_{t\geq 0}$ is, if the game has not stopped before time t_* player 1 can guarantee a payoff close to a_{t_*}.

PROPOSITION 6. *If $a_t \leq b_t$ for every $t \geq 0$, then the value in randomized stopping times exists. Moreover, the value is independent of the process $(c_t)_{t\geq 0}$. If $a_t \leq c_t \leq b_t$ for every $t \geq 0$, then there are ε-optimal (nonrandomized) stopping times for both players that are independent of $(c_t)_{t\geq 0}$.*

Proof. Consider an auxiliary stopping game $\Gamma^* = (\Omega, \mathcal{A}, P; \mathcal{F}, (a_t^*, b_t^*, c_t^*)_{t\geq 0})$, where $a_t^* = a_t$ and $b_t^* = c_t^* = b_t$ for every $t \geq 0$.

Lepeltier and Maingueneau (1984) and Stettner (1984) proved that the game Γ^* admits a value, and that there are ε-optimal (nonrandomized) stopping times for both players. We denote the value of Γ^* by v^* and prove that it is the value in randomized stopping times of the original game. Since Γ^* does not depend on the process $(c_t)_{t\geq 0}$, the second assertion of the proposition follows.

Fix $\varepsilon > 0$. Let μ be an ε-optimal (nonrandomized) stopping time for player 1 in Γ^*. In particular, $\inf_\nu \gamma_{\Gamma^*}(\mu, \nu) \geq v^* - \varepsilon$.

We now construct a randomized stopping time ϕ that satisfies $\inf_\nu \gamma_\Gamma(\phi, \nu) \geq v^* - 3\varepsilon$. By Lemma 4 there is $\delta > 0$ such that $P(\{|a_t - a_\mu| < \varepsilon \ \ \forall t \in [\mu, \mu + \delta]\}) > 1 - \varepsilon$. Define a randomized stopping time ϕ by

$$\phi(r, \cdot) = \mu + r\delta \qquad \forall r \in [0, 1].$$

That is, ϕ stops at a random time in the interval $[\mu, \mu + \delta]$. We denote such a randomized stopping time by $\phi = \mu + r\delta$.

Let ν be any stopping time. Since μ is ε-optimal in Γ^*, by the definition of Γ^*, by Fubini's theorem, and since $\lambda \otimes P(\mu + r\delta = \nu) = 0$,

$$\begin{aligned} v^* - \varepsilon &\le \gamma_{\Gamma^*}(\mu, \nu) \\ &= \mathbf{E}_P[a_\mu \mathbf{1}_{\{\mu<\nu\}} + b_\nu \mathbf{1}_{\{\mu\ge\nu\}}] \\ &= \mathbf{E}_{\lambda\otimes P}[a_\mu \mathbf{1}_{\{\mu+r\delta<\nu\}} + a_\mu \mathbf{1}_{\{\mu<\nu<\mu+r\delta\}} + b_\nu \mathbf{1}_{\{\mu\ge\nu\}}]. \end{aligned} \tag{2}$$

Since $\lambda \otimes P(\mu + r\delta = \nu) = 0$ and $(c_t)_{t\ge 0}$ is progressively measurable,

$$\begin{aligned} \gamma_\Gamma(\phi, \nu) &= \mathbf{E}_{\lambda\otimes P}\left[a_{\mu+r\delta} \mathbf{1}_{\{\mu+r\delta<\nu\}} + b_\nu \mathbf{1}_{\{\mu+r\delta>\nu\}} + c_\nu \mathbf{1}_{\{\mu+r\delta=\nu<+\infty\}}\right] \\ &= \mathbf{E}_{\lambda\otimes P}\left[a_{\mu+r\delta} \mathbf{1}_{\{\mu+r\delta<\nu\}} + b_\nu \mathbf{1}_{\{\mu+r\delta>\nu\}}\right] \\ &= \mathbf{E}_{\lambda\otimes P}\left[a_{\mu+r\delta} \mathbf{1}_{\{\mu+r\delta<\nu\}} + b_\nu \mathbf{1}_{\{\mu<\nu<\mu+r\delta\}} + b_\nu \mathbf{1}_{\{\mu\ge\nu\}}\right]. \end{aligned} \tag{3}$$

By Corollary 5, and since $a_t \le b_t$ for every $t \ge 0$,

$$\mathbf{E}_{\lambda\otimes P}[a_\mu \mathbf{1}_{\{\mu<\nu<\mu+r\delta\}}] \le \mathbf{E}_{\lambda\otimes P}[a_\nu \mathbf{1}_{\{\mu<\nu<\mu+r\delta\}}] + \varepsilon \le \mathbf{E}_{\lambda\otimes P}[b_\nu \mathbf{1}_{\{\mu<\nu<\mu+r\delta\}}] + \varepsilon. \tag{4}$$

Corollary 5 implies in addition that for every $r \in [0, 1]$

$$\mathbf{E}_{\lambda\otimes P}[a_\mu \mathbf{1}_{\{\mu+r\delta<\nu\}}] \le \mathbf{E}_{\lambda\otimes P}[a_{\mu+r\delta} \mathbf{1}_{\{\mu+r\delta<\nu\}}] + \varepsilon. \tag{5}$$

By (2)–(5),

$$v^* - \varepsilon \le \gamma_{\Gamma^*}(\mu, \nu) \le \gamma_\Gamma(\phi, \nu) + 2\varepsilon.$$

Since ν is arbitrary, $\inf_\nu \gamma_\Gamma(\phi, \nu) \ge v^* - 3\varepsilon$.

Consider a second auxiliary stopping game $\Gamma^{**} = (\Omega, \mathcal{A}, P; \mathcal{F}, (a_t^{**}, b_t^{**}, c_t^{**})_{t\ge 0})$, where $a_t^{**} = c_t^{**} = a_t$ and $b_t^{**} = b_t$ for every $t \ge 0$.

A symmetric argument to the one provided above proves that the game Γ^{**} has a value v^{**} and that player 2 has a randomized stopping time ψ that satisfies $\sup_\mu \gamma_\Gamma(\mu, \psi) \le v^{**} + 3\varepsilon$.

Since $c_t^{**} = a_t \le b_t = c_t^*$ for every $t \ge 0$, one has $v^{**} \le v^*$. Since $\sup_\mu \gamma_\Gamma(\mu, \psi) \ge \gamma_\Gamma(\phi, \psi) \ge \inf_\nu \gamma_\Gamma(\phi, \nu)$,

$$v^* \ge v^{**} \ge \sup_\mu \gamma_\Gamma(\mu, \psi) - 3\varepsilon \ge \inf_\nu \gamma_\Gamma(\phi, \nu) - 3\varepsilon \ge v^* - 6\varepsilon.$$

Since ε is arbitrary, $v^* = v^{**}$, so that v^* is the value in randomized stopping times of Γ, and ϕ and ψ are 3ε-optimal randomized stopping times for the two players. The first assertion of the proposition is established.

We now turn to the third assertion of the proposition. If $a_t \le c_t \le b_t$ for every $t \ge 0$, then $\gamma_{\Gamma^{**}}(\mu, \nu) \le \gamma_\Gamma(\mu, \nu) \le \gamma_{\Gamma^*}(\mu, \nu)$ for every pair of stopping times (μ, ν). Hence

$$\begin{aligned} v^{**} &= \sup_\mu \inf_\nu \gamma_{\Gamma^{**}}(\mu, \nu) \le \sup_\mu \inf_\nu \gamma_\Gamma(\mu, \nu) \\ &\le \inf_\nu \sup_\mu \gamma_\Gamma(\mu, \nu) \le \inf_\nu \sup_\mu \gamma_{\Gamma^*}(\mu, \nu) = v^* = v^{**}. \end{aligned}$$

Thus $\sup_\mu \inf_\nu \gamma_\Gamma(\mu, \nu) = \inf_\nu \sup_\mu \gamma_\Gamma(\mu, \nu)$: the value exists and there are ε-optimal stopping times for both players. Moreover, any ε-optimal stopping time for player 1 (resp., player 2) in Γ^* (resp., Γ^{**}) is also ε-optimal in Γ. In particular, if $a_t \le c_t \le b_t$ for every $t \ge 0$, both players have ε-optimal stopping times that are independent of $(c_t)_{t\ge 0}$. □

3.3. Proof of Theorem 3. Define a stopping time τ by

$$\tau = \inf\{t \geq 0, a_t \geq b_t\},$$

where the infimum of an empty set is $+\infty$. Since $(a_t - b_t)_{t\geq 0}$ is progressively measurable with respect to $(\mathcal{F}_t)_{t\geq 0}$, τ is a stopping time (see, e.g., Dellacherie and Meyer (1975, section IV-50)).

We show below that it is optimal for both players to stop at or around time τ (provided the game does not stop before time τ). Hence the problem reduces to the game between times 0 and τ. Since for $t \in [0, \tau[$, $a_t \leq b_t$, Proposition 6 can be applied.

The following notation will be useful in what follows. For a pair of stopping times (μ, ν) and a set $A \in \mathcal{A}$ we define

$$\gamma_\Gamma(\mu, \nu; A) = \mathbf{E}_P[\mathbf{1}_A(a_\mu \mathbf{1}_{\{\mu<\nu\}} + b_\mu \mathbf{1}_{\{\mu>\nu\}} + c_\mu \mathbf{1}_{\{\mu=\nu<+\infty\}})].$$

This is the expected payoff restricted to A. For a pair of randomized stopping times (ϕ, ψ) we define

$$\gamma_\Gamma(\phi, \psi; A) = \int_{[0,1]^2} \gamma_\Gamma(\mu_r, \nu_s; A) dr\ ds,$$

where μ_r and ν_s are the sections of ϕ and ψ, respectively.

Set

$$\begin{aligned} A_0 &= \{\tau = +\infty\}, \\ A_1 &= \{\tau < +\infty\} \cap \{c_\tau \geq a_\tau \geq b_\tau\}, \\ A_2 &= \{\tau < +\infty\} \cap \{a_\tau > c_\tau \geq b_\tau\}, \text{ and} \\ A_3 &= \{\tau < +\infty\} \cap \{a_\tau \geq b_\tau > c_\tau\}. \end{aligned}$$

Observe that (A_0, A_1, A_2, A_3) is an $\mathcal{F}_\tau$-measurable partition of Ω.

Define an $\mathcal{F}_\tau$-measurable function w by

$$w = a_\tau \mathbf{1}_{A_1} + c_\tau \mathbf{1}_{A_2} + b_\tau \mathbf{1}_{A_3}.$$

Define a stopping game $\Gamma^* = (\Omega, \mathcal{A}, P, (\mathcal{F}_t)_{t\geq 0}, (a_t^*, b_t^*, c_t^*)_{t\geq 0})$ by

$$a_t^* = \begin{cases} a_t & t < \tau \\ w & t \geq \tau \end{cases}, \qquad b_t^* = \begin{cases} b_t & t < \tau \\ w & t \geq \tau \end{cases}, \qquad c_t^* = \begin{cases} c_t & t < \tau \\ w & t \geq \tau \end{cases}.$$

That is, the payoff is set to w at and after time τ.

The game Γ^* satisfies the assumptions of Proposition 6 and hence, has a value V in randomized stopping times.

We now prove that V is the value of the game Γ as well. Fix $\varepsilon > 0$. We show only that player 1 has a randomized stopping time ϕ such that $\inf_\nu \gamma_\Gamma(\phi, \nu) \geq V - 7\varepsilon$. An analogous argument shows that player 2 has a randomized stopping time ψ such that $\sup_\mu \gamma_\Gamma(\mu, \psi) \leq V + 7\varepsilon$. Since ε is arbitrary, V is indeed the value in randomized stopping times of Γ.

Assume δ is sufficiently small so that the following conditions hold (by the proofs of Lemma 4 and Proposition 6 such δ exists):

(C1) There exists a stopping time μ such that the randomized stopping time $\phi^* = \mu + r\delta$ is ε-optimal for player 1 in Γ^*.

(C2) $P(\{\mu+\delta<\tau\}) \geq P(\{\mu<\tau\}) - \varepsilon/M$, where $M \in]0,+\infty[$ is a uniform bound of the payoff processes.

(C3) $P(\{|a_t - a_\tau| < \varepsilon, |b_t - b_\tau| < \varepsilon \quad \forall t \in [\tau, \tau+\delta]\}) > 1-\varepsilon$.

We now claim that we can assume without loss of generality that $\mu \leq \tau$. Indeed, assume that $P(\{\mu > \tau\}) > 0$. The set $\{\mu > \tau\}$ is $\mathcal{F}_\tau$-measurable. Define a stopping time $\mu' = \min\{\mu, \tau\}$. We will prove that the randomized stopping time $\phi' = \mu' + r\delta$ is also ε-optimal in Γ^*, which establishes the claim. Given a stopping time ν define a stopping time ν' by $\nu' = \min\{\nu, \tau\}$. By (C1),

$$\begin{aligned} V - \varepsilon &\leq \gamma_{\Gamma^*}(\mu + r\delta, \nu') \\ &= \gamma_{\Gamma^*}(\mu + r\delta, \nu'; \{\mu > \tau\}) + \gamma_{\Gamma^*}(\mu + r\delta, \nu'; \{\mu \leq \tau < \mu + \delta\}) \\ &\quad + \gamma_{\Gamma^*}(\mu + r\delta, \nu'; \{\mu + \delta \geq \tau\}). \end{aligned}$$

On the right-hand side the first term is equal to $\gamma_{\Gamma^*}(\mu' + r\delta, \nu; \{\mu > t\})$, by (C2) the second term is bounded by ε, and the third term is equal to $\gamma_{\Gamma^*}(\mu' + r\delta, \nu; \{\mu + \delta \geq \tau\})$. Therefore, by (C2),

$$\begin{aligned} V - \varepsilon &\leq \gamma_{\Gamma^*}(\mu' + r\delta, \nu; \{\mu > t\}) + \varepsilon + \gamma_{\Gamma^*}(\mu' + r\delta, \nu; \{\mu + \delta \geq \tau\}) \\ &\leq \gamma_{\Gamma^*}(\mu' + r\delta; \nu) + 2\varepsilon, \end{aligned}$$

as desired.

Define a randomized stopping time ϕ as follows:

$$\phi(r, \cdot) = \begin{cases} \mu + r\delta & \{\mu < \tau\} \cup A_0, \\ \tau & \{\mu = \tau\} \cap (A_1 \cup A_2), \\ \mu + r\delta & \{\mu = \tau\} \cap A_3. \end{cases}$$

The randomized stopping times ϕ and ϕ^* differ only over the set $\{\mu = \tau\} \cap (A_1 \cup A_2)$. Since over this set the payoff in Γ^* is w, provided the game terminates after time τ regardless of what the players play, and by (C2),

$$\inf_\nu \gamma_{\Gamma^*}(\phi, \nu) \geq V - 3\varepsilon. \tag{6}$$

Let ν be an arbitrary stopping time. Define a partition (B_0, B_1, B_2) of $[0,1] \times \Omega$ by

$$\begin{aligned} B_0 &= \{\mu + \delta < \tau\} \cup \{\nu < \tau\}, \\ B_1 &= \{\mu < \tau < \mu + \delta\} \cap \{\nu \geq \tau\}, \\ \text{and } B_2 &= \{\mu = \tau\} \cap \{\nu \geq \tau\}. \end{aligned}$$

Over B_0 the game terminates before time τ under (ϕ, ν). In particular,

$$\gamma_\Gamma(\phi, \nu; B_0) = \gamma_{\Gamma^*}(\phi, \nu; B_0). \tag{7}$$

By (C2), $P(B_1) < \varepsilon/M$, so that

$$\gamma_\Gamma(\phi, \nu; B_1) \geq \gamma_{\Gamma^*}(\phi, \nu; B_1) - 2\varepsilon. \tag{8}$$

Over $B_2 \cap A_0$ the game never terminates under (ϕ, ν), so that

$$\gamma_\Gamma(\phi, \nu; B_2 \cap A_0) = \gamma_{\Gamma^*}(\phi, \nu; B_2 \cap A_0) = 0. \tag{9}$$

Over $A_1 \cup A_2$, $\min\{a_\tau, c_\tau\} \geq w$, so that

$$\begin{aligned}\gamma_\Gamma(\phi,\nu; B_2 \cap (A_1 \cup A_2)) &= \mathbf{E}_{\lambda\otimes P}[\mathbf{1}_{B_2\cap(A_1\cup A_2)}(a_\tau \mathbf{1}_{\{\tau<\nu\}} + c_\tau \mathbf{1}_{\{\tau=\nu\}})] \\ &\geq \mathbf{E}_{\lambda\otimes P}[w\mathbf{1}_{\{\tau\leq\nu\}\cap B_2\cap(A_1\cup A_2)}] \\ &= \gamma_{\Gamma^*}(\phi,\nu; B_2 \cap (A_1 \cup A_2)).\end{aligned} \tag{10}$$

Finally, since $\lambda \otimes P(\{\mu + r\delta = \nu\}) = 0$, since $\{\mu = \tau\}$ on B_2, by Corollary 5, since $(c_t)_{t\geq 0}$ is progressively measurable, and since $a_\tau \geq b_\tau = w$ on A_3,

$$\begin{aligned}\gamma_\Gamma(\phi,\nu; B_2 \cap A_3) &= \mathbf{E}_{\lambda\otimes P}[\mathbf{1}_{B_2\cap A_3}(a_{\mu+r\delta}\mathbf{1}_{\{\mu+r\delta<\nu\}} + b_\nu \mathbf{1}_{\{\mu+r\delta>\nu\}} + c_\nu \mathbf{1}_{\{\mu+r\delta=\nu\}})] \\ &= \mathbf{E}_{\lambda\otimes P}[\mathbf{1}_{B_2\cap A_3}(a_{\mu+r\delta}\mathbf{1}_{\{\mu+r\delta<\nu\}} + b_\nu \mathbf{1}_{\{\mu+r\delta>\nu\}})] \\ &\geq \mathbf{E}_{\lambda\otimes P}[\mathbf{1}_{B_2\cap A_3}(a_\tau \mathbf{1}_{\{\mu+r\delta<\nu\}} + b_\tau \mathbf{1}_{\{\mu+r\delta>\nu\}})] - 2\varepsilon \\ &\geq \mathbf{E}_{\lambda\otimes P}[w\mathbf{1}_{B_2\cap A_3}] - 2\varepsilon \\ &= \gamma_{\Gamma^*}(\phi,\nu; B_2 \cap A_3) - 2\varepsilon.\end{aligned} \tag{11}$$

Summing (7)–(11) and using (6) gives us

$$V - 3\varepsilon \leq \gamma_{\Gamma^*}(\phi,\nu) \leq \gamma_\Gamma(\phi,\nu) + 4\varepsilon,$$

as desired.

3.4. On final payoff. Our convention is that the payoff is 0 if no player ever stops. In fact, one can add a final payoff as follows. Let χ be an $\mathcal{A}$-measurable and integrable function. The expected payoff that corresponds to a pair of pure strategies (μ,ν) is

$$\mathbf{E}_P[a_\mu \mathbf{1}_{\{\mu<\nu\}} + b_\nu \mathbf{1}_{\{\mu>\nu\}} + c_\mu \mathbf{1}_{\{\mu=\nu<+\infty\}} + \chi \mathbf{1}_{\{\mu=\nu=+\infty\}}].$$

The expected payoff can be written as

$$\begin{aligned}\mathbf{E}_P[\chi] + \mathbf{E}_P\Big[\Big(a_\mu - \mathbf{E}_P^{\mathcal{F}_\mu}[\chi]\Big)\mathbf{1}_{\{\mu<\nu\}} &+ \Big(b_\nu - \mathbf{E}_P^{\mathcal{F}_\nu}[\chi]\Big)\mathbf{1}_{\{\mu>\nu\}} \\ &+ \Big(c_\mu - \mathbf{E}_P^{\mathcal{F}_\mu}[\chi]\Big)\mathbf{1}_{\{\mu=\nu<+\infty\}}\Big],\end{aligned}$$

where $\mathbf{E}_P^{\mathcal{F}_\mu}[\chi]$ is the conditional expectation of χ given the σ-algebra $\mathcal{F}_\mu$.

Define a process $d_t := \mathbf{E}_P^{\mathcal{F}_t}[\chi]$. Since the filtration satisfies the "usual conditions," $(d_t)_{t\geq 0}$ is a right-continuous martingale (see, e.g., Dellacherie and Meyer (1980, section VI-4) or Lepeltier and Maingueneau (1984, Theorem 4)). Hence we are reduced to the study of the standard stopping game $\Gamma^* = (\Omega, \mathcal{A}, P, (\mathcal{F}_t)_{t\geq 0}, (a_t^*, b_t^*, c_t^*)_{t\geq 0})$ with $a_t^* = b_t - d_t$, $b_t^* = b_t - d_t$, and $c_t^* = c_t - d_t$.

3.5. Right-continuity of the payoff process. For every $s \geq 0$, let $\Gamma[s]$ be the stopping game that starts at time s. Formally, $\Gamma[s]$ is given by $(\Omega, \mathcal{A}, P, (\mathcal{F}'_t, a'_t, b'_t, c'_t)_{t\geq 0})$, where for every $t \geq 0$, $\mathcal{F}'_t = \mathcal{F}_{t+s}$, $a_t = a_{t+s}$, $b_t = b_{t+s}$, and $c_t = c_{t+s}$. Let v_s be the value of $\Gamma[s]$.

The next proposition states that if the payoff processes are right-continuous, the process $(v_t)_{t\geq 0}$ is right-continuous as well.

PROPOSITION 7. *If the processes $(a_t, b_t, c_t)_{t\geq 0}$ are right-continuous, then so is $(v_t)_{t\geq 0}$.*

Proof. For every $t \geq 0$, denote $\tau[t] = \inf\{t \geq s\colon a_s \geq b_s\}$ and define the sets $A_0[t], A_1[t], A_2[t]$, and $A_3[t]$ as in the proof of Theorem 3 with respect to $\tau[t]$. Set

$$w_t = a_{\tau[t]}\mathbf{1}_{A_1[t]} + c_{\tau[t]}\mathbf{A_2[t]} + b_{\tau[t]}\mathbf{1}_{A_3[t]}.$$

Now fix $t \geq 0$. On $\{a_t < b_t\}$, one has $w_t = w_s$ for every $s > t$ sufficiently close to t, so that by Lepeltier and Maingueneau (1984, Theorem 9), the value is right-continuous on this set.

On $\{a_t > c_t > b_t\}$, one has $v_s = c_s$ for every $s \geq t$ sufficiently close to t, and by the right-continuity of $(c_t)_{t\geq 0}$ the same conclusion holds.

On $\{a_t = c_t \geq b_t\}$, one has $\tau[t] = 0$ and $v_t = a_t = c_t$. Moreover, for every $\varepsilon > 0$ and every $s > t$ sufficiently small, one has (i) $a_s > a_t - \varepsilon = v_t - \varepsilon$ and $c_s > c_t - \varepsilon = v_t - \varepsilon$, so that $v_s > v_t - \varepsilon$, and (ii) $b_s < b_t + \varepsilon \leq v_t + \varepsilon$ and $c_s < c_t + \varepsilon = v_t + \varepsilon$, so that $v_s < v_t + \varepsilon$. In particular, $(v_t)_{t\geq 0}$ is right-continuous at t on this set.

A similar argument shows the right-continuity of $(v_t)_{t\geq 0}$ in all of the remaining cases. □

3.6. Noisy stochastic duels. As mentioned in the introduction, the right-continuity of the payoff process can be used to derive, by induction and proper definition of a final payoff, the existence of an equilibrium in a more general class of games, in which (i) each player has to act at most M times, and (ii) the payoff depends on the number of times each player acted, as well as on the exact times in which the players acted. That is, the game is given by a filtration $(\mathcal{F}_t)_{t\geq 0}$ and, for every $0 \leq n, m \leq M$, a right-continuous process $u_{m,n}(t_1, \dots, t_m, t'_1, \dots, t'_n)$ that is defined whenever $t_1 < t_2 < \cdots < t_m$ and $t'_1 < t'_2 < \cdots < t'_n$, and such that $u_{m,n}(t_1, \dots, t_m, t'_1, \dots, t'_n)$ is $\mathcal{F}_{\max\{t_m, t'_n\}}$-measurable. If player 1 acts at times $t_1 < \cdots < t_m$ and player 2 acts at times $t'_1 < \cdots < t'_n$, with $0 \leq m, n \leq M$, the payoff is $u_{m,n}(t_1, \dots, t_m, t'_1, \dots, t'_n)$. This implies, in particular, that every noisy stochastic duel in which each player is endowed with finitely many bullets, the payoff is 1 if player 1 hits player 2, the payoff is –1 if player 2 hits player 1, and the accuracy process is right-continuous, admits a value.

Details are standard and omitted.

Acknowledgment. We thank the anonymous referee, whose comments improved the presentation.

REFERENCES

M. Alario-Nazaret, J. P. Lepeltier, and B. Marchal (1982), *Dynkin games*, in Stochastic Differential Systems (Bad Honnef, 1982), Lecture Notes in Control and Inform. Sci. 43, Springer-Verlag, Berlin, pp. 23–32.

R. Bellman and M. A. Girshick (1949), *An Extension of Results on Duels with Two Opponents, One Bullet Each, Silent Guns, Equal Accuracy*, Rand Publication D-403, Rand Corp., Santa Monica, CA.

J. M. Bismut (1977), *Sur un problème de Dynkin*, Z. Warscheinlichkeitstheorie und Verw. Gebiete, 39, pp. 31–53.

D. Blackwell (1949), *The Noisy Duel, One Bullet Each, Arbitrary Nonmonotone Accuracy*, Rand Publication RM-131, Rand Corp., Santa Monica, CA.

M. J. Brennan and E. S. Schwartz (1977), *Convertible bonds: Valuation and optimal strategies for call and conversion*, J. Finance, 32, pp. 1699–1715.

C. Dellacherie and P.-A. Meyer (1980), *Probabilités et potentiel*, Chapitres V à VIII, Théorie des Martingales, Hermann, Paris (in French); *Probabilities and Potential. B. Theory of Martingales*, North–Holland Math. Stud. 72, North–Holland, Amsterdam, 1982 (in English).

C. Dellacherie and P.-A. Meyer (1975), *Probabilités et potentiel*, Chapitres I à IV, Hermann, Paris (in French); *Probabilities and Potential*, North–Holland Math. Stud. 29, North–Holland, Amsterdam, New York, 1978 (in English).

E. B. Dynkin (1969), *Game variant of a problem on optimal stopping*, Soviet Math. Dokl., 10, pp. 270–274.

D. Fudenberg and J. Tirole (1991), *Game Theory*, MIT Press, Cambridge, MA.

P. Ghemawat and B. Nalebuff (1985), *Exit*, Rand J. Econom., 16, pp. 184–194.

K. Hendricks, A. Weiss, and C. Wilson (1988), *The war of attrition in continuous time with complete information*, Internat. Econom. Rev., 29, pp. 663–680.

J. Kallsen and C. Kühn (2004), *Pricing derivatives of American and game type in incomplete markets*, Finance Stoch., 8, pp. 261–284.

S. Karlin (1959), *Mathematical Methods and Theory in Games, Programming and Economics Vol.* II*: The Theory of Infinite Glames*, Addison-Wesley, Reading, MA.

Y. Kifer (2000), *Game options*, Finance Stoch., 4, pp. 443–463.

C. Kühn and A. E. Kyprianou (2003), *Pricing Israeli Options: A Pathwise Approach*, preprint, Department of Mathematics, University of Utrecht, Utrecht, The Netherlands.

A. E. Kyprianou (2004), *Some calculations for Israeli options*, Finance Stoch., 8, pp. 73–86.

J.-P. Lepeltier and M. A. Maingueneau (1984), *Le jeu de Dynkin en théorie générale sans l'hypothèse de Mokobodski*, Stochastics, 13, pp. 25–44.

J. Maynard-Smith (1974), *The theory of games and the evolution of animal conflicts*, J. Theoret. Biol., 47, pp. 209–221.

J. J. McConnell and E. S. Schwartz (1986), *LYON taming*, J. Finance, 41, pp. 561–576.

R. C. Merton (1973), *Theory of rational option pricing*, Bell J. Econom. and Management Sci., 4, pp. 141–183.

J. Neveu (1975), *Discrete-Parameter Martingales*, North–Holland Math. Library 10, North–Holland, Amsterdam.

T. Radzik and T. E. S. Raghavan (1994), *Handbook of Game Theory with Economic Applications*, Vol. 2, Handbooks in Econom. 11, R. J. Aumann and H. S. Duels, eds., North–Holland, Amsterdam, pp. 761–768.

D. Rosenberg, E. Solan, and N. Vieille (2001), *Stopping games with randomized strategies*, Probab. Theory Related Fields, 119, pp. 433–451.

L. S. Shapley (1951), *The Noisy Duel: Existence of a Value in the Singular Core*, Rand Publication RM-641, Rand Corp., Santa Monica, CA.

L. Stettner (1984), *On closedness of general zero-sum stopping game*, Bull. Polish Acad. Sci. Math., 32, pp. 351–361.

N. Touzi and N. Vieille (2002), *Continuous-time Dynkin games with mixed strategies*, SIAM J. Control Optim., 41, pp. 1073–1088.

M. Yasuda (1985), *On a randomized strategy in Neveu's stopping problem*, Stochastic Process. Appl., 21, pp. 159–166.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 1923–1952

EFFICIENT SOLUTION OF OPTIMAL CONTROL PROBLEMS USING HYBRID SYSTEMS*

MIREILLE BROUCKE†, MARIA DOMENICA DI BENEDETTO‡, STEFANO DI GENNARO§, AND ALBERTO SANGIOVANNI-VINCENTELLI¶

Abstract. We consider the synthesis of optimal controls for continuous feedback systems by recasting the problem to a hybrid optimal control problem: *synthesize optimal enabling conditions for switching between locations in which the control is constant.* An algorithmic solution is obtained by translating the hybrid automaton to a finite automaton using a bisimulation and formulating a dynamic programming problem with extra conditions to ensure non-Zenoness of trajectories. We show that the discrete value function converges to the viscosity solution of the Hamilton–Jacobi–Bellman equation as a discretization parameter tends to zero.

Key words. optimal control, hybrid systems, bisimulation

AMS subject classifications. 49L20, 49L25, 65N99

DOI. 10.1137/S0363012900383090

1. Introduction.

The goal of this paper is the development of a computationally appealing technique for synthesizing optimal controls for continuous feedback systems $\dot{x} = f(x, u)$ by recasting the problem as a hybrid optimal control problem. The hybrid problem is obtained by approximating the control set $U \subset \mathbb{R}^m$ by a finite set $\Sigma \subset U$ and defining vector fields for the locations of the hybrid system of the form $f(x, \sigma)$, $\sigma \in \Sigma$; that is, the control is constant in each location. The hybrid control problem is to synthesize enabling conditions such that a target set $\Omega_f \subset \Omega$ is reached while a hybrid cost function is minimized, for each initial condition in a specified set $\Omega \subset \mathbb{R}^n$.

Casting the problem as a hybrid control problem is not necessarily a simplification because, while algorithmic approaches for solving the controller synthesis problem for specific classes of hybrid systems have appeared [33, 52], no general, efficient algorithm is available. To be able to solve the (nonlinear) hybrid optimal control problem, we must exploit some additional property. We have a feasible and appealing approach if we can translate the problem to an equivalent discrete problem, which abstracts completely the continuous behavior. This translation is possible if we can construct a finite *bisimulation* defined on the hybrid state space, that is, an equivalence relation that induces a partition in each hybrid automaton location that is consistent with the continuous dynamics of that location. A finite bisimulation can be constructed using the geometric approach reported in [10], based on the following key assumption: $n-1$ *local (on* Ω*) first integrals can be expressed analytically for each vector field* $f(x, \sigma)$, $\sigma \in \Sigma$. This assumption is imposed in the transient phase of a feedback system's

*Received by the editors December 20, 2000; accepted for publication (in revised form) June 7, 2004; published electronically April 14, 2005.
http://www.siam.org/journals/sicon/43-6/38309.html

†Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada M5S 3G4 (broucke@control.utoronto.ca).

‡Dip. di Ingegneria Elettrica, Università di L'Aquila, Poggio di Roio, 67040 L'Aquila, Italy, and Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 (marika@eecs.berkeley.edu).

§Dip. di Ingegneria Elettrica, Università di L'Aquila, Poggio di Roio, 67040 L'Aquila, Italy (digennar@dis.uniroma1.it).

¶Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 (alberto@eecs.berkeley.edu).

response, when the vector field is nonvanishing and local first integrals always exist, although finding closed form expressions for them is not always easy or possible. Also, the assumption that the partition be a bisimulation is sufficient but not necessary for the overall approach.

If the assumption is met, then we can transform the hybrid system to a quotient system associated with the finite bisimulation, which is a finite automaton. The control problem posed on the finite automaton is to synthesize a discrete supervisor, providing a switching rule between automaton locations, that minimizes a discrete cost function approximating the original cost function, for each initial discrete state. We provide a dynamic programming solution to this problem, with extra constraints to ensure non-Zenoness of the closed-loop trajectories. By imposing non-Zeno conditions on the synthesis we obtain piecewise constant controls with a finite number of discontinuities in bounded time.

The discrete value function depends on the discretizations of U and of Ω using the bisimulation. We quantify these discretizations by parameters δ and δ_Q. The main theoretical contribution is to show that as $\delta, \delta_Q \to 0$, the discrete value function converges to the unique viscosity solution of the Hamilton–Jacobi–Bellman (HJB) equation.

There is a similarity between our approach to optimal control and *regular synthesis*, introduced in [8], in the sense that both restrict the class of controls to a set that has some desired property and both use a finite partition to define switching behavior. For linear systems, the results on regular synthesis are centered on the bang-bang principle [38], stating that a sufficient class of optimal controls is piecewise constant. If U is a convex polyhedron, then the number of discontinuities of the control is bounded. There is no hope that general bang-bang results are available due to Fuller's example [21, 29]. Nevertheless, in many applications the optimal control is a piecewise continuous function, and methods of regular synthesis of such controls are worth investigating. Our paper focuses on piecewise constant controls and provides a constructive approach to obtaining a cell decomposition, in the spirit of regular synthesis, by using a finite bisimulation, which further allows us to formulate the synthesis problem on its quotient system—a finite automaton.

The idea of using a time abstract model formed by partitioning the continuous state space has been pursued in a number of papers recently. Stiver, Antsaklis, and Lemmon [48] and Yang, Lemmon, and Antsaklis [53] used a partition of the state space to convert a hybrid model to a discrete event system (DES). This enabled them to apply controller synthesis for DESs to synthesize a supervisor. While our approach is related to this methodology, it differs in that we provide conditions for obtaining the partition. In [41] hybrid systems consisting of a linear time-invariant system and a discrete controller that has access to a quantized version of the linear system's output is considered. The quantization results in a rectangular partition of the state space. This approach suffers from spurious solutions that must be trimmed from the automaton behavior. Hybrid optimal control problems have been studied in papers by Witsenhausen [51], Branicky, Borkar, and Mitter [9], and Bensoussan and Menaldi [6]. The first two studies concentrated on problems of well-posedness, necessary conditions, and existence of optimal solutions but did not provide algorithmic solutions. Bensoussan and Menaldi considered a more general model than ours that included continuous dynamics with a measurable control input and a discrete part with impulsive control. Control switches can be autonomous or controlled and may have time delays. They characterized the viscosity solution of a dynamic programming problem

on their model. They constructed open-loop controls, whereas we obtain feedback controls, and they did not consider the numerical implementation.

There has recently been significant progress in developing numerical methods that incorporate geometric invariants of a dynamical or control system. In particular, in the area of geometric mechanics, numerical integrators have been developed that preserve the Hamiltonian, Lie group symmetries, and other integrals of motion [19, 27, 25, 28, 40]. See [36] for an overview of problems where geometric structure is exploited in numerical methods. Our work represents the first general methodology in which geometric invariants are explicitly considered in the numerical solution of optimal control problems. The geometric structure that is present in the optimal control problem is encoded in the bisimulation partition. In effect, the two-step procedure of a time discretization followed by the state discretization via finite element methods that together lead to a fixed point formulation of the approximate solution of a continuous time optimal control problem is circumvented. Instead an exact representation of the time evolution of the system is encoded in the finite element partition, enabling a simplified and more efficient formulation.

The paper is organized as follows. In section 2 we state the optimal control problem, while in section 3 the associated hybrid system is given. In section 4 we review how the bisimulation is constructed. Section 5 formulates the proposed solution using bisimulation and dynamic programming. In section 6 we prove the main theoretical result. In section 7 we present an algorithmic solution of the dynamic programming problem including a formal justification of the algorithm's optimality. In section 8 we give two simple examples. Section 9 summarizes our findings.

2. Optimal control problem. *Notation.* First, we introduce notation. $\mathbf{1}(\cdot)$ is the indicator function. $\mathrm{cl}(A)$ denotes the closure of set A. $\|\cdot\|$ denotes the Euclidean norm. Let $C^1(\mathbb{R}^n)$ and $\mathcal{X}(\mathbb{R}^n)$ denote the sets of continuously differentiable real-valued functions and smooth vector fields on $\mathbb{R}^n$, respectively. $\phi_t(x_0, \mu)$ denotes the trajectory of $\dot{x} = f(x, \mu)$ starting from x_0 and using control $\mu(\cdot)$.

Let U be a compact subset of $\mathbb{R}^m$; Ω an open, bounded, connected subset of $\mathbb{R}^n$; and Ω_f a compact subset of Ω. Define $\mathcal{U}_m$ to be the set of measurable functions mapping $\mathbb{R}^+$ to U. We define the minimum hitting time $T : \mathbb{R}^n \times \mathcal{U}_m \to \mathbb{R}^+$ by

$$T(x,\mu) := \begin{cases} \infty & \text{if } \{t \mid \phi_t(x,\mu) \in \Omega_f\} = \emptyset, \\ \min\{t \mid \phi_t(x,\mu) \in \Omega_f\} & \text{otherwise.} \end{cases} \tag{2.1}$$

A control $\mu \in \mathcal{U}_m$ specified on $[0, T]$ is *admissible* for $x \in \Omega$ if $\phi_t(x, \mu) \in \Omega$ for all $t \in [0, T]$. The set of admissible controls for x is denoted $\mathcal{U}_x$. Let

$$\mathcal{R} := \{x \in \Omega \mid \exists \mu \in \mathcal{U}_x.\ T(x,\mu) < \infty\}.$$

We consider the following stationary optimal control problem. Given $y \in \Omega$,

$$\text{minimize} \qquad J(y,\mu) = \int_0^{T(y,\mu)} L(x(t),\mu(t))dt + h(x(T(y,\mu))) \tag{2.2}$$

$$\text{subject to} \qquad \dot{x} = f(x,\mu), \qquad \text{a.e. } t \in [0, T(y,\mu)], \tag{2.3}$$

$$x(0) = y \tag{2.4}$$

among all admissible controls $\mu \in \mathcal{U}_y$. $J : \mathbb{R}^n \times \mathcal{U}_m \to \mathbb{R}$ is the *cost-to-go* function, $h : \mathbb{R}^n \to \mathbb{R}$ is the *terminal cost*, and $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the *instantaneous cost*. At $T(y, \mu)$ the terminal cost $h(x(T(y, \mu)))$ is incurred and the dynamics are stopped. The control objective is to reach Ω_f from $y \in \Omega$ with minimum cost.

Assumption 2.1.

(1) $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ satisfies $\|f(x', u') - f(x, u)\| \le L_f\big[\|x' - x\| + \|u' - u\|\big]$ for some $L_f > 0$. Let M_f be the upper bound of $\|f(x, u)\|$ on $\Omega \times U$.

(2) $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ satisfies $|L(x', u') - L(x, u)| \le L_L\big[\|x' - x\| + \|u' - u\|\big]$ and $1 \le L(x, u) \le M_L$, $x \in \Omega$, $u \in U$, for some $L_L, M_L > 0$.

(3) $h : \mathbb{R}^n \to \mathbb{R}$ satisfies $|h(x') - h(x)| \le L_h\|x' - x\|$ for some $L_h > 0$, and $h(x) \ge 0$ for all $x \in \Omega$. Let M_h be the upper bound of $|h(x)|$ on Ω.

Remark 2.1. These assumptions ensure existence of solutions to (2.3) and uniqueness of the trajectories $\phi_t(x, \mu)$. Weaker assumptions are possible; see [4].

The *value function* or optimal cost-to-go function $V : \mathbb{R}^n \to \mathbb{R}$ is given by

$$V(y) = \inf_{\mu \in \mathcal{U}_y} J(y, \mu)$$

for $y \in \Omega \setminus \Omega_f$ and by $V(y) = h(y)$ for $y \in \Omega_f$. A control μ is called *ϵ-optimal* for x if $J(x, \mu) \le V(x) + \epsilon$.

It is well known [20] that V satisfies the HJB equation

$$-\inf_{u \in U}\Big\{L(x, u) + \frac{\partial V}{\partial x} f(x, u)\Big\} = 0 \tag{2.5}$$

at each point of $\mathcal{R}$ at which it is differentiable. The HJB equation is an infinitesimal version of the equivalent *dynamic programming principle* (DPP), which says that

$$V(x) = \inf_{\mu \in \mathcal{U}_x}\left\{\int_0^t L(\phi_s(x, \mu), \mu(s))ds + V(\phi_t(x, \mu))\right\}, \quad x \in \Omega \setminus \Omega_f,$$

$$V(x) = h(x), \qquad x \in \Omega_f.$$

The subject of assiduous effort has been that the HJB equation may not have a C^1 solution. This gap in the theory was closed by the introduction of the concept of viscosity solution [32, 14], which can be shown to provide the unique solution of (2.5) without any differentiability assumption. In particular, a bounded uniformly continuous function V is called a *viscosity solution* of HJB provided, for each $\psi \in C^1(\mathbb{R}^n)$, the following hold:

(i) if $V - \psi$ attains a local maximum at $x_0 \in \mathbb{R}^n$, then

$$-\inf_{u \in U}\Big\{L(x_0, u) + \frac{\partial \psi}{\partial x}(x_0) f(x_0, u)\Big\} \le 0;$$

(ii) if $V - \psi$ attains a local minimum at $x_1 \in \mathbb{R}^n$, then

$$-\inf_{u \in U}\Big\{L(x_1, u) + \frac{\partial \psi}{\partial x}(x_1) f(x_1, u)\Big\} \ge 0.$$

Assumption 2.2. For every $\epsilon > 0$ and $x \in \mathcal{R}$, there exists $N_\epsilon > 0$ and an admissible piecewise constant ϵ-optimal control μ having at most N_ϵ discontinuities and such that $\phi_t(x, \mu)$ is transverse to $\partial\Omega_f$.

The transversality assumption implies that the viscosity solution is continuous at the boundary of the target set, a result needed in proving uniform continuity of V

over a finite horizon. The assumption can be replaced by a small-time controllability condition. For a treatment of small-time controllability and compatibility of the terminal cost with respect to continuity of the value function, see [4]. The finite switching assumption holds under mild assumptions, such as Lipschitz continuity of the vector field and cost functions, and is based on approximating measurable functions by piecewise constant functions.

3. Hybrid system. The approach we propose for solving the continuous optimal control problem first requires a mapping to a hybrid system and, second, employs a bisimulation of the hybrid system to formulate a dynamic programming problem on the quotient system. In this section we define the hybrid system. First, we discretize U by defining a finite set $\Sigma_\delta \subset U$ which has a mesh size

$$\delta := \sup_{u \in U} \min_{\sigma \in \Sigma_\delta} \|u - \sigma\|.$$

We define the hybrid automaton $H := (\Sigma \times \mathbb{R}^n, \Sigma_\delta, D, E_h, G)$ with the following components.

State set. $\Sigma \times \mathbb{R}^n$ consists of the finite set $\Sigma = \Sigma_\delta \cup \{\sigma_f\}$ of control locations and n continuous variables $x \in \mathbb{R}^n$. σ_f is a terminal location when the optimal control problem is stopped and the target set is reached. The controller for σ_f may, for instance, be a linear feedback designed using the linearization of the system.

Events. Σ_δ is a finite set of control events.

Vector fields. $D : \Sigma \to \mathcal{X}(\mathbb{R}^n)$ is a function assigning an autonomous vector field to each location. We use the notation $D(\sigma) = f_\sigma$.

Control switches. $E_h \subset \Sigma \times \Sigma$ is a set of control switches. $e = (\sigma, \sigma')$ is a directed edge between a source location σ and a target location σ'. If $E_h(\sigma)$ denotes the set of edges that can be enabled at $\sigma \in \Sigma$, then $E_h(\sigma) := \{(\sigma, \sigma') \mid \sigma' \in \Sigma \setminus \sigma\}$ for $\sigma \in \Sigma_\delta$ and $E_h(\sigma_f) = \emptyset$. Thus, from a source location not equal to σ_f, there is an edge to every other location (but not itself), while location σ_f has no outgoing edges.

Enabling conditions. $G : E_h \to \{g_e\}_{e \in E_h}$ is a function assigning to each edge an enabling (or guard) condition $g \subset \mathbb{R}^n$. We use the notation $G(e) = g_e$. The optimal enabling conditions are unknown and must be synthesized.

3.1. Semantics. A state is a pair (σ, x), $\sigma \in \Sigma$ and $x \in \mathbb{R}^n$. In location $\sigma \in \Sigma_\delta$ the continuous state evolves according to the vector field $f(x, \sigma)$. In location σ_f, the vector field is $\dot{x} = f(x, \mu_f)$, where μ_f is the (not necessarily constant) control of the terminal location. Trajectories of H evolve in *steps* of two types. A *σ-step* is a binary relation $\xrightarrow{\sigma} \subset (\Sigma \times \mathbb{R}^n) \times (\Sigma \times \mathbb{R}^n)$, and we write $(\sigma, x) \xrightarrow{\sigma'} (\sigma', x')$ iff (1) $e = (\sigma, \sigma') \in E_h$, (2) $x \in g_e$, and (3) $x = x'$. A *t-step* is a binary relation $\xrightarrow{t} \subset (\Sigma \times \mathbb{R}^n) \times (\Sigma \times \mathbb{R}^n)$, and we write $(\sigma, x) \xrightarrow{t} (\sigma', x')$ iff (1) $\sigma = \sigma'$ and (2) for some $t \geq 0$, $x' = \phi_t(x, \sigma)$, where $\dot{\phi}_t(x) = f(\phi_t(x, \sigma), \sigma)$. Enabling conditions are *forced* in that an edge is taken instantaneously and as soon as it is enabled.

Example 3.1. Consider the time optimal control problem for the system

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= u. \end{aligned} \tag{3.1}$$

Suppose $\Omega = (-1, 1) \times (-1, 1)$ and $\Omega_f = \overline{B}_\epsilon(0)$, the closed epsilon ball centered at 0. The cost-to-go function is $J(x, \mu) = \int_0^{T(x,\mu)} dt$ and $U = \{u : |u| \leq 1\}$. We select

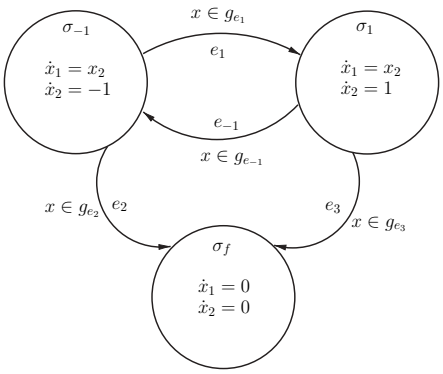


FIG. 3.1. *Hybrid automaton for time optimal control of a double integrator system.*

$\Sigma_\delta = \{-1, 1\}$, so that $\delta = 1$. The hybrid system is show in Figure 3.1. The state set is $\{\sigma_{-1} = -1, \sigma_1 = 1, \sigma_f\} \times \mathbb{R}^2$. $g_{e_{-1}}$ and g_{e_1} are unknown and must be synthesized, while $g_{e_2} = g_{e_3} = \Omega_f$.

4. Bisimulation. Let λ represent an arbitrary time interval. A *bisimulation* of H is an equivalence relation $\simeq \subset (\Sigma_\delta \times \mathbb{R}^n) \times (\Sigma_\delta \times \mathbb{R}^n)$ such that for all states $p_1, p_2 \in \Sigma_\delta \times \mathbb{R}^n$, if $p_1 \simeq p_2$ and $\beta \in \Sigma_\delta \cup \{\lambda\}$, then if $p_1 \xrightarrow{\beta} p_1'$, there exists p_2' such that $p_2 \xrightarrow{\beta} p_2'$ and $p_1' \simeq p_2'$. See Figure 4.1. Intuitively, a bisimulation is an equivalence relation defining a partition on the hybrid state space that preserves reachability over σ-steps and time steps. However, the definition leaves ambiguity about how the partition should be obtained. Alur and Dill [1] gave a construction for timed automata that was based on the first integrals of the continuous dynamics and on the syntax of the enabling and reset conditions. Their approach was first generalized in [10]. Time evolution of the original system is modeled as untimed transitions from equivalence class to equivalence class in the quotient system associated with the bisimulation. Transitions between locations of the hybrid automaton appear also as transitions in the quotient system. Thus if there is a finite number of equivalence classes of $\simeq$, then a finite transition system or finite automaton is obtained which gives a time abstract model of the original system, with reachability properties exactly preserved. For a more thorough discussion of results on bisimulations for hybrid systems, see [26, 2].

We declare the set of "interesting" equivalence classes of $\simeq$, which is assumed to be finite and is denoted Q, to be those that intersect $\Sigma_\delta \times \mathrm{cl}(\Omega)$. For each $q \in Q$ we define a distinguished point $(\sigma, \xi) \in q$, and we use the notation $q = [(\sigma, \xi)]$.

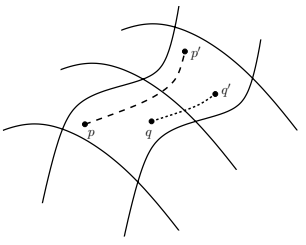


FIG. 4.1. *Illustration of the definition of bisimulation.*

We define a mesh size on Q by

$$\delta_Q = \max_{q \in Q} \sup_{(\sigma,x),(\sigma,y) \in q} \|x - y\|.$$

For each $q = [(\sigma, \xi)] \in Q$ we associate the duration τ_q, the maximum time to traverse q using constant control σ. That is,

$$\tau_q = \sup_{(\sigma,x),(\sigma,y) \in q} \{t \mid y = \phi_t(x, \sigma)\}.$$

This extra data associated with the bisimulation is required in our problem to obtain approximations of the cost functions L and h.

4.1. Geometric construction. We review our method for obtaining finite bisimulations [10] which relies on the following assumptions on the vector fields on Ω.

Assumption 4.1.

(1) For each $\sigma \in \Sigma_\delta$, there exist $n-1$ C^1 functions $\gamma_i^\sigma : \Omega \to \mathbb{R}$, $i = 1, \ldots, n-1$, whose time derivative along solutions of $\dot{x} = f(x, \sigma)$ in Ω is zero.

(2) There exists $m_f > 0$ such that $\|f(x, u)\| \geq m_f$ for all $x \in \mathrm{cl}(\Omega)$, $u \in U$.

Remark 4.1. There is an uncontested view promulgated by Poincaré that differential equations possessing a complete set of first integrals, i.e., completely integrable systems, are the exception rather than the norm. This has led to some confusion about when one can or cannot find first integrals for non-Hamiltonian systems. The primary source of confusion comes from the multiple meanings of the term *integrability*. It appears as a Liouville integrability (the version alluded to by Poincaré and further developed by Arnold [3]), local integrability, algebraic integrability, etc. A type of integrability suitable for non-Hamiltonian systems was proposed by Chavarriga et al. [13] with the terminology *weak integrability* (to contrast with *strong integrability* in the sense of Liouville). Weak integrability is meant to capture that many systems do not exhibit complex behavior such as chaos, even if they are not Hamiltonian.

Let $\dot{x} = f(x)$ be a differential equation with domain of definition $\mathcal{D} \subset \mathbb{R}^n$, and let $\mathcal{O}$ be a set of orbits of the system such that $\mathcal{D} \setminus \mathcal{O}$ is open. Following [13], we say a C^1 function $\gamma : \mathcal{D} \setminus \mathcal{O} \to \mathbb{R}$ is a *weak first integral* of the system $\dot{x} = f(x)$ if γ is constant on each solution of the system contained in $\mathcal{D} \setminus \mathcal{O}$ and γ is nonconstant on any open subset of $\mathcal{D} \setminus \mathcal{O}$. A system is said to be *weakly integrable* if it has $n-1$ functionally independent (on $\mathcal{D} \setminus \mathcal{O}$) weak first integrals.

The relaxation of the requirement that the first integral be a differentiable function on the entire domain of the differential equation means that, for instance, all linear systems are weakly integrable [11], whereas only the Hamiltonian linear systems (centers and saddles in the case of second order linear systems) are integrable in the strong sense. Assumption 4.1(1) is a weak integrability assumption.

There are many methods for finding first integrals, including Lie group symmetry analysis [7, 37], Lax pairs, Painleve analysis, and the Frobenius theorem, among others [16, 22, 46, 49]. A general reference and overview of the methods can be found in [24]. The best-known result for symbolic computation of first integrals is the Prelle–Singer procedure [39]. Reduce and Macsyma implementations of the Prelle–Singer procedure are described in [34, 46], while an implementation in higher dimensions is described in [35]. Algorithms for finding polynomial first integrals are described in [43, 44].

A bisimulation of $\Sigma_\delta \times \mathbb{R}^n$ is found by first constructing partitions for each location of H such that reachability properties are preserved over time steps. In section 5 we describe how to accomodate σ-steps in the quotient system. To obtain a partition

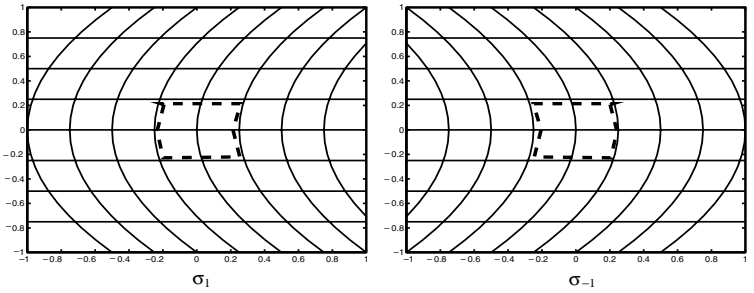


FIG. 4.2. *Partitions for states* σ_1 *and* σ_{-1} *of the hybrid automaton of Figure* 3.1.

consistent with the dynamics of location $\sigma \in \Sigma_\delta$ we use the level sets of the $n-1$ first integrals $\gamma_i^\sigma(x) = y_i^\sigma$, $i = 1, \ldots, n-1$, to bound the flow in $n-1$ independent directions, thus obtaining tubes of trajectories with a rectangular cross section. Next, the level sets of a submersion $\gamma_n^\sigma = y_n^\sigma$ that is transverse to the flow of $\dot{x} = f(x, \sigma)$ are used to divide the tube of trajectories into boxes, so that $(y_1^\sigma, \ldots, y_n^\sigma)$ form a set of Euclidean coordinates $\gamma^\sigma : \Omega \to [-1, 1]^n$ on Ω. That is, we assume that the level sets of γ_i^σ *foliate* the set Ω (see [30] for background on foliations) and, by appropriate scaling, their level values lie between -1 and 1 on Ω. We *discretize* the foliations associated with each γ_i^σ by selecting a finite number of level values. More precisely, fix $k \in Z^+$ and let $\Delta = \frac{1}{2^k}$. Define

$$C_k = \{0, \pm\Delta, \pm 2\Delta, \ldots, \pm 1\}. \tag{4.1}$$

Each $y_i^\sigma = c$ for $c \in C_k$, $i = 1, \ldots, n$, defines a hyperplane in $\mathbb{R}^n$ denoted $\tilde{W}_{i,c}^\sigma$ and a submanifold $W_{i,c}^\sigma = (\gamma^\sigma)^{-1}(\tilde{W}_{i,c}^\sigma)$. The collection of submanifolds for $\sigma \in \Sigma_\delta$ is

$$\mathcal{W}_k^\sigma = \{W_{i,c}^\sigma \mid c \in C_k, i \in \{1, \ldots, n\}\}. \tag{4.2}$$

$\Omega \backslash \mathcal{W}_k^\sigma$ is the union of $2^{n(k+1)}$ disjoint open sets $\mathcal{V}_k^\sigma = \{V_j^\sigma\}$. We define an equivalence relation $\simeq^e$ on $\mathbb{R}^n$ as follows. $y \simeq^e y'$ iff

(1) $y \notin [-1, 1]^n$ iff $y' \notin [-1, 1]^n$, and

(2) if $y, y' \in [-1, 1]^n$, then for each $i = 1, \ldots, n$, $y_i \in (c, c + \Delta)$ iff $y_i' \in (c, c + \Delta)$, and $y_i = c$ iff $y_i' = c$ for all $c \in C_k$.

We define the equivalence relation $\simeq$ on $\Sigma_\delta \times \mathbb{R}^n$ as follows. $(\sigma, x) \simeq (\sigma', x')$ iff (1) $\sigma = \sigma'$ and (2) $\gamma^\sigma(x) \simeq^e \gamma^\sigma(x')$.

Remark 4.2. A consequence of this construction is that if any trajectory of H passing through $q \in Q$ spends zero time in it, then $\tau_q = 0$.

Example 4.1. Continuing Example 3.1, a first integral for vector field $\dot{x}_1 = x_2$, $\dot{x}_2 = 1$ is $x_1 - \frac{1}{2}x_2^2 = c_1$, $c_1 \in \mathbb{R}$. For $\dot{x}_1 = x_2$, $\dot{x}_2 = -1$ a first integral is $x_1 + \frac{1}{2}x_2^2 = c_2$, $c_2 \in \mathbb{R}$. We select a transverse foliation for each vector field, given by $x_2 = c_3$. Partitions for locations σ_1 and σ_{-1} and $\Omega = (-1, 1) \times (-1, 1)$ are shown in Figure 4.2. The equivalence classes of $\simeq$ are pairs consisting of a control event in Σ_δ and of the interiors of regions, open line segments and curves forming the boundaries of two regions, and the points at the corners of regions. $\tau = 0$ for the segments transverse to the flow and the corner points. $\tau = \Delta$ for the interiors of regions and segments tangential to the flow, where $\Delta = .25$ in Figure 4.2.

5. Discrete problem. In this section we transform the hybrid optimal synthesis problem to a dynamic programming problem on a nondeterministic finite automaton.

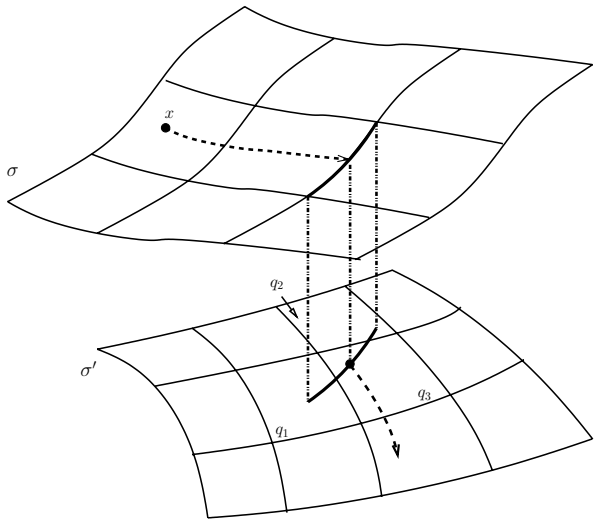


FIG. 5.1. *Partitions for states σ and σ' of a hybrid automaton, and the resulting nondeterminism in A.*

Consider the class of nondeterministic finite automata with cost structure represented by the tuple

$$A = (Q, \Sigma_\delta, E, \hat{L}, \hat{h}).$$

Q is the finite state set, as defined above, and Σ_δ is the set of control events as before. $E \subseteq Q \times Q$ is the transition relation encoding t-steps and σ-steps of H. $(q, q') \in E$, where $q = [(\sigma, \xi)]$ and $q' = [(\sigma', \xi')]$ if either (a) $\sigma = \sigma'$, there exists $x \in \Omega$ such that $(\sigma, x) \in q$, and there exists $\tau > 0$ such that for all $t \in [0, \tau]$, $(\sigma, \phi_t(x, \sigma)) \in q$ and $(\sigma, \phi_{\tau+\epsilon}(x, \sigma)) \in q'$ for arbitrarily small $\epsilon > 0$; (b) $\sigma = \sigma'$, there exists $x \in \Omega$ such that $(\sigma, x) \in q$, and there exists $\tau > 0$ such that for all $t \in [0, \tau)$, $(\sigma, \phi_t(x, \sigma)) \in q$ and $(\sigma, \phi_\tau(x, \sigma)) \in q'$; or (c) $\sigma \neq \sigma'$ and there exists $x \in \Omega$ such that $(\sigma, x) \in q$ and $(\sigma', x) \in q'$. Cases (a) and (b) say that from a point in q, q' is the first state (different from q) reached after following the flow of $f(x, \sigma)$ for some time. Case (c) says that an edge exists between q and q' if their projections to $\mathbb{R}^n$ have nonempty intersection.

Remark 5.1. The requirement that there be an edge from q to q' if their projections to $\mathbb{R}^n$ have nonempty intersection is illustrated in Figure 5.1. We have partitions for controls σ and σ', respectively. In the partition for σ, suppose a trajectory starting at x flows in time using control σ until state q of A is reached, at which time the control is set to σ'. The possible states of A that can be reached from q are q_1, q_2, q_3, and the one-dimensional equivalence classes between them. Hence, edges corresponding to these possible futures for the trajectory must be included in the definition of A. A consequence is that multiple trajectories of A can be defined starting from an initial state. One can think of this construction as overapproximating the identity map in terms of the equivalence classes of $\simeq$. This is the source of nondeterminacy of A.

Let $e = (q, q')$ with $q = [(\sigma, \xi)]$ and $q' = [(\sigma', \xi')]$. $\hat{L} : E \to \mathbb{R}$ is the *discrete instantaneous cost* given by

$$\hat{L}(e) := \begin{cases} \tau_q L(\xi, \sigma) & \text{if } \sigma = \sigma', \\ 0 & \text{if } \sigma \neq \sigma'. \end{cases} \tag{5.1}$$

$\hat{h} : Q \to \mathbb{R}$ is the *discrete terminal cost* given by

$$\hat{h}(q) := h(\xi).$$

The domain of $\hat{h}$ can be extended to Ω, with a slight abuse of notation, by

$$\hat{h}(x) := \hat{h}(q), \tag{5.2}$$

where $q = \arg\min_{q'}\{\|x - \xi'\| \mid q' = [(\sigma', \xi')]\}$. Finally, Q_f is the target set given by the overapproximation of Ω_f,

$$Q_f = \{q \in Q \mid \exists x \in \Omega_f \text{ s.t. } (\sigma, x) \in q\}. \tag{5.3}$$

5.1. Semantics. A transition or *step* of A from $q = [(\sigma, \xi)] \in Q$ to $q' = [(\sigma', \xi')] \in Q$ is denoted $q \overset{\sigma'}{\to} q'$. If $\sigma \neq \sigma'$, the transition is referred to as a *control switch*; otherwise, it is referred to as a *time step*. If $E(q)$ is the set of edges that can be enabled from $q \in Q$, then for $\sigma \in \Sigma_\delta$,

$$E_\sigma(q) = \{e \in E(q) \mid e = (q, q'), q = [(\sigma, \xi)], q' = [(\sigma', \xi')]\}.$$

If $|E_\sigma(q)| > 1$, then we say that $e \in E_\sigma(q)$ is *unobservable* in the sense that when control event σ is issued, it is unknown which edge among $E_\sigma(q)$ is taken. If $\sigma = \sigma'$, then $|E_\sigma(q)| = 1$, by the uniqueness of solutions of ODEs and by the definition of bisimulation.

A *control policy* $c : Q \to \Sigma_\delta$ is a map assigning a control event to each state; $c(q) = \sigma$ is the control event issued when the state is at q. A *trajectory* π of A over c is a sequence $\pi = q_0 \overset{\sigma_1}{\to} q_1 \overset{\sigma_2}{\to} q_2 \overset{\sigma_3}{\to} \cdots$, $q_i \in Q$. A trajectory is *non-Zeno* if between any two nonzero duration time steps there is a finite number of control switches. Note that this definition is slightly different from the traditional definition of non-Zeno trajectories of H [26], in which it is assumed that time steps always have a nonzero duration. Here zero-duration time steps can occur. Let $\Pi_c(q)$ be the set of trajectories starting at q and applying control policy c, and let $\tilde{\Pi}_c(q)$ be the set of trajectories starting at q, applying control policy c, and eventually reaching Q_f. If for every $q \in Q$, $\pi \in \Pi_c(q)$ is non-Zeno, then we say c is an *admissible control policy*. The set of all admissible control policies for A is denoted $\mathcal{C}$.

A control policy c is said to have a *loop* if A has a trajectory $q_0 \overset{c(q_0)}{\to} q_1 \overset{c(q_1)}{\to} \cdots \overset{c(q_{m-1})}{\to} q_m = q_0$, $q_i \in Q$. A control policy has a *Zeno loop* if it has a loop made up of control switches and/or zero duration time steps (i.e., $\tau_q = 0$) only.

LEMMA 5.1. *A control policy c is admissible iff it has no Zeno loops.*

Proof. First we show that a nondeterministic automaton with non-Zeno trajectories has a control policy without Zeno loops. Suppose not. Then a trajectory starting on a state belonging to the loop can take infinitely many steps around the loop before taking a nonzero duration time step. Such a trajectory must necessarily include a control switch (since a zero duration time step is always followed either by a nonzero duration time step or a control switch). Since this control switch occurs infinitely often in a finite time interval, the trajectory is Zeno, a contradiction.

Second, we show that a control policy without Zeno loops implies non-Zeno trajectories. Suppose not. Consider a Zeno trajectory that takes an infinite number of control switches in some finite time interval. Because there are a finite number of states in Q, by the Dirichlet principle [31], one of the states must be repeated in the sequence of states visited during the infinite number of control switches. Note that

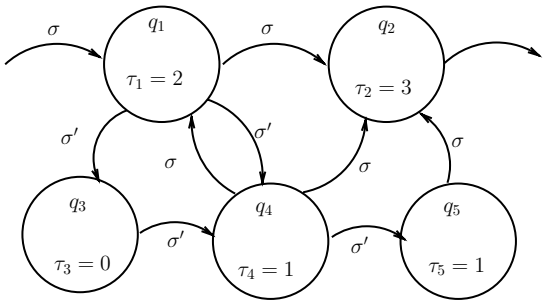


FIG. 5.2. *Fragment of automaton with a zero duration time step.*

this sequence can include zero duration time steps. This implies the existence of a loop in the control policy. Now we argue this loop is Zeno.

First, by Remark 4.2, if $\tau_q = 0$, then all trajectories spend zero time in q. Second, if $\tau_q > 0$, then there exists $\overline{\tau}_q > 0$ such that all trajectories spend at least $\overline{\tau}_q$ time in q. This follows from the boundedness of f and the bisimulation construction (trajectories cannot move between two level sets of γ_n^σ in arbitrarily small time). Since there is a finite number of states in Q, there exists $\overline{\tau} > 0$, the minimum time spent by any trajectory in a state $q \in Q$ with $\tau_q > 0$. The result is that Zeno trajectories only arise by an infinite number of control switches in a zero duration time interval. Hence, we have shown that the loop consists of control switches and/or zero duration time steps only, i.e., it is a Zeno loop. □

Example 5.1. Consider the automaton in Figure 5.2. Suppose that we define a control policy $c(q_1) = \sigma'$, $c(q_3) = \sigma'$, $c(q_4) = \sigma$, and $c(q_5) = \sigma$. Starting at q_1 two possible trajectories are $q_1 \xrightarrow{\sigma'} q_3 \xrightarrow{\sigma'} q_4 \xrightarrow{\sigma} q_2$, or $q_1 \xrightarrow{\sigma'} q_3 \xrightarrow{\sigma'} q_4 \xrightarrow{\sigma} q_1$. The first trajectory has a zero duration time step. The control is inadmissible since the second trajectory has a Zeno loop.

5.2. Dynamic programming. We formulate the dynamic programming problem on A. This involves defining a cost-to-go function and a value function that minimizes it over control policies suitable for nondeterministic automata.

Let $\pi = q_0 \xrightarrow{\sigma_1} q_1 \to \cdots \to q_{N-1} \xrightarrow{\sigma_N} q_N$, where $q_i = [(\sigma_i, \xi_i)]$ and π takes the sequence of edges $e_1 e_2 \dots e_N$. We define a *discrete cost-to-go* $\hat{J} : Q \times \mathcal{C} \to \mathbb{R}$ by

$$\hat{J}(q,c) = \begin{cases} \max\limits_{\pi \in \tilde{\Pi}_c(q)} \left\{ \sum\limits_{j=1}^{N_\pi} \hat{L}(e_j) + \hat{h}(q_{N_\pi}) \right\} & \text{if } \ \Pi_c(q) = \tilde{\Pi}_c(q), \\ \infty & \text{otherwise,} \end{cases}$$

where $N_\pi = \min\{j \geq 0 \mid q_j \in Q_f\}$. We take the maximum over $\tilde{\Pi}_c(q)$ because of the nondeterminacy of A: it is uncertain which among the (multiple) trajectories allowed by c will be taken so we must assume the worst-case situation. The *discrete value function* $\hat{V} : Q \to \mathbb{R}$ is

$$\hat{V}(q) = \min_{c \in \mathcal{C}} \hat{J}(q,c)$$

for $q \in Q \backslash Q_f$ and $\hat{V}(q) = \hat{h}(q)$ for $q \in Q_f$. We show in Proposition 5.2 that $\hat{V}$ satisfies a DPP that takes into account the nondeterminacy of A and ensures that optimal control policies are admissible. Let $\mathcal{A}_q$ be the set of control assignments $c(q) \in \Sigma_\delta$ at q such that c is admissible.

PROPOSITION 5.2. *$\hat{V}$ satisfies*

$$\hat{V}(q) = \min_{c(q)\in\mathcal{A}_q}\Big\{\max_{e=(q,q')\in E_{c(q)}(q)}\big\{\hat{L}(e)+\hat{V}(q')\big\}\Big\}, \quad q\in Q\setminus Q_f, \tag{5.4}$$

$$\hat{V}(q) = \hat{h}(q), \qquad q\in Q_f. \tag{5.5}$$

Proof. Fix $q\in Q$. By definition of $\hat{J}$,

$$\hat{J}(q,c) = \max_{e=(q,q')\in E_{c(q)}(q)}\{\hat{L}(e)+\hat{J}(q',c)\}. \tag{5.6}$$

By definition of $\hat{V}$,

$$\hat{J}(q,c) \geq \max_{e=(q,q')\in E_{c(q)}(q)}\{\hat{L}(e)+\hat{V}(q')\}.$$

Since $c(q)\in\mathcal{A}_q$ is arbitrary,

$$\hat{V}(q) \geq \min_{c(q)\in\mathcal{A}_q}\left\{\max_{e=(q,q')\in E_{c(q)}(q)}\{\hat{L}(e)+\hat{V}(q')\}\right\}.$$

To prove the reverse inequality suppose, by way of contradiction, there exists $\sigma'\in\Sigma_\delta$ such that

$$\hat{V}(q) > \max_{e=(q,q')\in E_{\sigma'}(q)}\{\hat{L}(e)+\hat{V}(q')\} := \hat{L}(e)+\hat{V}(\overline{q}). \tag{5.7}$$

Suppose an optimal admissible policy for $\overline{q}$ is $\overline{c}$. Define $c=\overline{c}$ on $Q\setminus\{q\}$ and $c(q)=\sigma'$. Then $\hat{J}(q,c)=\hat{L}(e)+\hat{V}(\overline{q})<\hat{V}(q)$. This gives rise to a contradiction if we can show c is admissible. Suppose not. Then there exists a loop of control switches and zero duration time steps containing q. Either the loop includes $\overline{q}$, implying $\hat{V}(\overline{q})=\hat{V}(q)$, which contradicts hypothesis (5.7), or the loop includes some other q' such that $(q,q')\in E_{\sigma'}(q)$, implying $\hat{V}(q')=\hat{V}(q)$. But $\hat{V}(\overline{q})\geq\hat{V}(q')$ since $\overline{q}$ gives the worst-case cost over edges with label σ'. This again contradicts hypothesis (5.7). ∎

5.3. Synthesis of g_e. The synthesis of enabling conditions or *hybrid controller synthesis* is typically a postprocessing step of a backward reachability analysis (see, for example, [52]). This situation prevails here as well: (5.4) and (5.5) describe a backward analysis to construct an optimal policy $c\in\mathcal{C}$. Once c is known, the enabling conditions of H are extracted as follows.

Consider each $e=(\sigma,\sigma')\in E$ of H with $\sigma\neq\sigma'$. There are two cases. If $\sigma'\neq\sigma_f$, then $g_e=\big\{x\mid(\sigma,x)\in q, q\in Q, c(q)=\sigma'\big\}$. That is, if the control policy designates switching from $q\in Q$ with label σ to $q'\in Q$ with label σ', then the corresponding enabling condition in H includes the projection to $\mathbb{R}^n$ of q. The second case when $\sigma'=\sigma_f$ is for edges going to the terminal location of H. Then $g_e=\big\{x\mid(\sigma,x)\in q, q\in Q_f\big\}$.

6. Main result. We will prove that $\hat{V}$ converges to V, the viscosity solution of the HJB equation, as $\delta_Q,\delta\to 0$. We make use of a filtration of control sets $\Sigma_k\equiv\Sigma_{\delta_k}$ corresponding to a sequence $\delta_k\to 0$ as $k\to\infty$ in such a manner that $\Sigma_k\subset\Sigma_{k+1}$. Considering (4.2), we define a filtration of families of submanifolds such that $\mathcal{W}_k^\sigma\subset\mathcal{W}_{k+1}^\sigma$ for each $\sigma\in\Sigma_k$.

The proof proceeds in three steps. In the first step we restrict the class of controls to piecewise constant functions whose constant intervals are a function of the state. In particular, the control is constant on equivalence classes of $\simeq$. As δ_k tends to zero this class of piecewise constant controls well approximates ϵ-optimal controls. The Arzela–Ascoli theorem is invoked to show that the limit of a sequence of approximations V_k of the value function using the aforementioned controls is a continuous function V_*. Using techniques of [4], V_* is shown to be the unique viscosity solution of HJB. In the second step we introduce the discrete approximations of L and h. The discrete approximation of h is a one-time error, while the error between L and $\hat{L}$ is shown to be $O(\delta_k^2)$ per interval τ. Since the number of intervals is $O(1/\delta_k)$, the error is $O(\delta_k)$. In the last step we introduce the discrete states Q. The error introduced at each control switch by the nondeterminacy of A is $O(\delta_k)$ and since there are a fixed number of control switches as $\delta_k \to 0$, this error can be made arbitrarily small.

Step 1*: Piecewise constant controls.* In the first step we define a class of piecewise constant functions that depend on the state and show that the value function which minimizes the cost-to-go over this class converges to the viscosity solution of HJB as $\delta_k \to 0$. The techniques of this step are based on those by Bardi and Capuzzo-Dolcetta [4] and are related to those in [12].

We consider the optimal control problem (2.2)–(2.4) when the set of admissible controls is $\mathcal{U}_k^1$, piecewise constant functions consisting of finite sequences of control events $\sigma \in \Sigma_k$, where each σ is applied for a time $\tau(\sigma, x)$ and the trajectory remains in Ω. Let $(\sigma, x) \in q$ for $q \in Q$ and define $\tau(\sigma, x)$ to be the minimum of the time it takes the trajectory starting at x and using control $\sigma \in \Sigma_k$ to reach (ta) $\partial\Omega_f$ or (tb) some x' such that $(\sigma, x') \notin q$. If a trajectory is at x_i at the start of the $(i+1)$th step, then the control σ_{i+1} is applied for time $\tau_{i+1} := \tau(\sigma_{i+1}, x_i)$ and $x_{i+1} = \phi_{\tau_{i+1}}(x_i, \sigma_{i+1})$. Thus U_k^1 is a class of piecewise constant controls whose constant intervals are based on the state partition induced by $\simeq$ (in contrast with a partition of the time interval): the control can only change values on the boundary of equivalence classes.

Let

$$\mathcal{R}_k^1 := \{x \in \Omega \mid \exists \mu \in \mathcal{U}_k^1 \cdot T(x, \mu) < \infty\}.$$

We define the cost-to-go function $J_k^1 : \Omega \times \mathcal{U}_k^1 \to \mathbb{R}$ as follows. For $x \in \Omega$ and $\mu = \sigma_1 \sigma_2 \ldots \in \mathcal{U}_k^1$, if $T(x, \mu) < \infty$, then

$$J_k^1(x, \mu) = \sum_{j=1}^{N} \int_0^{\tau(\sigma_j, x_{j-1})} L(\phi_s(x_{j-1}, \sigma_j), \sigma_j) ds + h(x_N),$$

where $N = \min\{j \geq 0 \mid x_j \in \Omega_f\}$. $J_k^1(x, \mu) = \infty$ otherwise. We define the value function $V_k^1 : \mathbb{R}^n \to \mathbb{R}$ as follows. For $x \in \Omega \setminus \Omega_f$,

$$V_k^1(x) = \inf_{\mu \in \mathcal{U}_k^1} J_k^1(x, \mu), \tag{6.1}$$

and for $x \in \Omega_f$, $V_k^1(x) = h(x)$. The following result is proved using standard arguments from dynamic programming [20].

PROPOSITION 6.1. *V_k^1 satisfies, for all $x \in \mathcal{R}_k^1$,*

$$V_k^1(x) = \min_{\sigma \in \Sigma_k} \left\{ \int_0^{\tau(\sigma, x)} L(\phi_s(x, \sigma), \sigma) ds + V_k^1(\phi_{\tau(\sigma, x)}(x, \sigma)) \right\}. \tag{6.2}$$

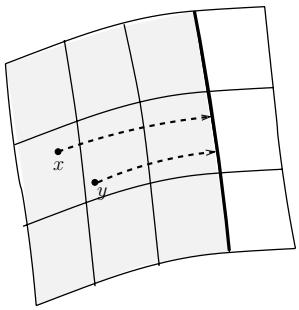


FIG. 6.1. *The shaded region is M^σ_{c-} while the bold curve on its boundary is M^σ_c.*

We would like to show that V^1_k is uniformly bounded and locally uniformly continuous. Considering uniform continuity of V^1_k, let C_k be as in (4.1) and γ^σ_n the submersion whose level sets are transverse to the flow of $\dot{x} = f(x, \sigma)$. Referring to Figure 6.1, for each $\sigma \in \Sigma_k$ and for each fixed $c \in C_k$ we define the regions in $\mathbb{R}^n$

$$M^\sigma_c := \{x \mid \gamma^\sigma_n(x) = c\},$$
$$M^\sigma_{c-} := \{x \mid \gamma^\sigma_n(x) \in (-1, c)\};$$

that is, M_{c-} is the strip of points belonging to a level set of γ^σ_n whose level value is between -1 and c.

Remark 6.1.

(a) If $x, y \in M^\sigma_{c-}$ for some $c \in C_k$ and $\tau(\sigma, x)$ and $\tau(\sigma, y)$ are defined using (tb), then $|\tau(\sigma, x) - \tau(\sigma, y)| \to 0$ and $\|\phi_{\tau(\sigma,x)}(x, \sigma) - \phi_{\tau(\sigma,y)}(y, \sigma)\| \to 0$ as $\|x - y\| \to 0$ in M^σ_{c-}, since M^σ_c is a smooth submanifold. See Figure 6.1. For the details, see [20, Theorem 6.1, pp. 91–94]. If instead $\tau(\sigma, x)$ and $\tau(\sigma, y)$ are defined using (ta) and σ is an ϵ-optimal control for x, then by Assumption 2.2 the same results hold.

(b) For each $x \in \cup_k \mathcal{R}^1_k$ and $\epsilon > 0$ there exists $m \in \mathbb{Z}^+$ and $\mu \in \mathcal{U}^1_m$ such that μ is an ϵ-optimal control for x w.r.t. V^1_k with at most N_ϵ discontinuities and such that $\phi_t(x, \mu)$ is transverse to $\partial\Omega_f$. This follows from Assumption 2.2, $V^1_k(x) \geq V(x)$, and the fact that we can well approximate an ϵ-optimal control for V by a control in $\mathcal{U}^1_m$ for large enough m.

The following lemma shows that V^1_k is locally uniformly continuous.

LEMMA 6.2. *For each $y \in \cup_k \mathcal{R}^1_k$ and $\epsilon > 0$, there exists $m_\epsilon \in \mathbb{Z}^+$ and $\eta_\epsilon > 0$ such that $|V^1_k(x) - V^1_k(y)| < 2\epsilon$ for all $|x - y| < \eta_\epsilon$ with $x \in \mathcal{R}^1_k$ and for all $k > m_\epsilon$.*

Proof. Fix $y \in \cup_k \mathcal{R}^1_k$. By Remark 6.1(b) there exists $m_1 > 0$ and $\mu \in \mathcal{U}^1_{m_1}$ such that μ is an ϵ-optimal control for y satisfying Assumption 2.2. Let $x \in \mathcal{R}^1_{m_1}$. Then $V^1_k(x) - V^1_k(y) \leq J^1_k(x, \mu_x) - J^1_k(y, \mu) + \epsilon$ for any $\mu_x \in \mathcal{U}^1_{m_1}$ and $k > m_1$. If we can show that for fixed y and μ there exists $\mu_x \in \mathcal{U}^1_{m_1}$ such that

$$J^1_k(x, \mu_x) - J^1_k(y, \mu) < \epsilon \tag{6.3}$$

for all $x \in \mathcal{R}^1_{m_1}$ sufficiently close to y, then $V^1_k(x) - V^1_k(y) \leq 2\epsilon$ for all $k \geq m_1$.

Conversely, by Remark 6.1(b) there exists $m_2 > 0$ and $\mu_x \in \mathcal{U}^1_{m_2}$ such that μ_x is an ϵ-optimal control for x satisfying Assumptions 2.2. Then $V^1_k(y) - V^1_k(x) \leq J^1_k(y, \mu) - J^1_k(x, \mu_x) + \epsilon$ for any $\mu \in \mathcal{U}^1_{m_2}$ and $k > m_2$. If we can show that for fixed y there exists $\mu \in \mathcal{U}^1_{m_2}$ such that

$$J^1_k(y, \mu) - J^1_k(x, \mu_x) < \epsilon \tag{6.4}$$

for all $x \in \mathcal{R}^1_{m_2}$ sufficiently close to y, then $V^1_k(x) - V^1_k(y) \geq -2\epsilon$ for all $k \geq m_2$. The result follows by letting $m_\epsilon = \min\{m_1, m_2\}$. Thus, we must show (6.3) and (6.4).

Consider first (6.3). Let $\mu = \sigma_1\sigma_2 \ldots \in \mathcal{U}^1_k$ be an ϵ-optimal control for y such that $y_N \in \partial\Omega_f$. By redefining indices, we can associate with μ the open-loop control $\tilde{\mu} = (\sigma_1, \tau_1)(\sigma_2, \tau_2)\ldots$, where τ_i is the time σ_i is applied. We claim there exists $\tilde{\mu}^x = (\sigma_1, \tau^x_1)(\sigma_2, \tau^x_2)\ldots$ such that as $x \to y$, (a) $x_j \to y_j$, (b) $\tau^x_j \to \tau_j$, and (c) $x_N \in \partial\Omega_f$. Let $T_k = \max_i \tau_i$. Then we have

$$\begin{aligned} J^1_k(x, \tilde{\mu}^x) - J^1_k(y, \tilde{\mu}) &\leq \sum_{j=1}^{N} \int_0^{\tau_j} |L(\phi_s(x_{j-1}, \sigma_j), \sigma_j) - L(\phi_s(y_{j-1}, \sigma_j), \sigma_j)| ds \\ &\quad + \sum_{j=1}^{N} \left| \int_{\tau_j}^{\tau^x_j} L(\phi_s(x_{j-1}, \sigma_j), \sigma_j) ds \right| + |h(y_N) - h(x_N)| \\ &\leq L_L T_k \exp(L_f T_k) \sum_{j=1}^{N} \|x_{j-1} - y_{j-1}\| \\ &\quad + M_L \sum_{j=1}^{N} |\tau^x_j - \tau_j| + L_h |x_N - y_N|. \end{aligned}$$

By the previous claim the right-hand side (r.h.s.) can be made less than ϵ. Thus, we need only show there exists $\tilde{\mu}^x = (\sigma_1, \tau^x_1)(\sigma_2, \tau^x_2)\ldots$, which satisfies the claim, and $\mu^x \in \mathcal{U}^1_k$ can be reconstructed from it, based on the discrete states in Q visited by $\phi_t(x, \tilde{\mu}^x)$.

We argue by induction. Suppose (a)–(c) hold at $j-1$. We show they hold at j. We need only consider the case when $y_{j-1} \in M^{\sigma_j}_{c-}$ and $y_j \in M^{\sigma_j}_c$ for some $c \in C_k$; that is, y_{j-1} lies upstream of y_j (trajectories flow in the increasing $\gamma^{\sigma_j}_n$ direction), while y_j lies on the boundary of an equivalence class where the control is allowed to switch values. For x_{j-1} sufficiently close to y_{j-1}, $x_{j-1} \in M^{\sigma_j}_{c-}$. By Remark 6.1(a) there exists τ^x_j such that $x_j = \phi_{\tau^x_j}(x_{j-1}, \sigma_j) \in M^{\sigma_j}_c$ and $\tau^x_j \to \tau_j$ and $x_j \to y_j$ as $x_{j-1} \to y_{j-1}$. The case $y_{j-1} \in M^{\sigma_j}_{c-}$ and $y_j \in \partial\Omega_f$ follows in the same way from Assumption 2.2. Proving (6.4) follows along the same lines as the proof for (6.3). ∎

To show boundedness of V^1_k, let $T(x) := \inf_{\mu \in \mathcal{U}^1_k} T(x, \mu)$. In light of Assumption 2.1(2), we have that for all $x \in \mathbb{R}^n$, $|V^1_k(x)| \leq T(x) \cdot M_L + M_h$. Consider the set $K_a := \{x \in \mathcal{R}^1_k \mid T(x) < a\}$. Then $|V^1_k(x)| \leq a \cdot M_L + M_h$ for all $x \in K_a$.

We have shown that on each $K_a \subseteq \mathbb{R}^n$, $\{V^1_k\}$ forms a family of equibounded, locally equicontinuous functions. It follows by the Arzela–Ascoli theorem [42] that along some subsequence k_n, $V^1_{k_n}$ converges to a continuous function V_*. The proof of the following result closely follows [4].

PROPOSITION 6.3. *V_* is the unique viscosity solution of HJB.*

Proof. We show that V_* solves HJB in the viscosity sense. Let $\psi \in C^1(\mathbb{R}^n)$ and suppose $x_0 \in \Omega$ is a strict local maximum for $V_* - \psi$. There exists a closed ball B centered at x_0 such that $(V_* - \psi)(x_0) > (V_* - \psi)(x)$ for all $x \in B$. Let $x_{0\delta_k}$ be a maximum point for $V^1_k - \psi$ over B. Since $V^1_k \to V_*$ locally uniformly it follows that $x_{0\delta_k} \to x_0$ as $\delta_k \to 0$. Then, for any $\sigma \in \Sigma_k$, the point $\phi_\tau(x_{0\delta_k}, \sigma)$ is in B (using boundedness of f), for sufficiently small δ_k and $0 \leq \tau \leq \tau(x_{0\delta_k}, \sigma)$, since $\tau(x_{0\delta_k}, \sigma) \to 0$ as $\delta_k \to 0$. Therefore,

$$V^1_k(x_{0\delta_k}) - \psi(x_{0\delta_k}) \geq V^1_k(\phi_\tau(x_{0\delta_k}, \sigma)) - \psi(\phi_\tau(x_{0\delta_k}, \sigma)).$$

Considering (6.2), we have

$$0 = -\min_{\sigma\in\Sigma_k}\left\{V_k^1(\phi_\tau(x_{0\delta_k},\sigma)) - V_k^1(x_{0\delta_k}) + \int_0^\tau L(\phi_s(x_{0\delta_k},\sigma),\sigma)ds\right\}$$
$$\geq -\min_{\sigma\in\Sigma_k}\left\{\psi(\phi_\tau(x_{0\delta_k},\sigma)) - \psi(x_{0\delta_k}) + \int_0^\tau L(\phi_s(x_{0\delta_k},\sigma),\sigma)ds\right\}.$$

Since $\psi \in C^1(\mathbb{R}^n)$, we have by the mean value theorem,

$$0 \geq -\min_{\sigma\in\Sigma_k}\left\{\frac{\partial\psi}{\partial x}(y)\cdot\int_0^\tau f(\phi_s(x_{0\delta_k},\sigma),\sigma)ds + \int_0^\tau L(\phi_s(x_{0\delta_k},\sigma),\sigma)ds\right\},$$

where $y = \alpha x_{0\delta_k} + (1-\alpha)\phi_\tau(x_{0\delta_k},\sigma)$ for some $\alpha \in (0,1)$. Dividing by $\tau > 0$ on each side and taking the limit as $\delta_k \to 0$, we have $V_k^1 \to V_*$, $x_{0\delta_k} \to x_0$, $\tau \to 0$, and $y \to x_{0\delta_k}$. By the fundamental theorem of calculus, the continuity of f and L, and the uniform continuity in u of the expression in brackets, we obtain

$$0 \geq -\inf_{u\in U}\left\{\frac{\partial\psi}{\partial x}(x_0)\cdot f(x_0,u) + L(x_0,u)\right\}.$$

This confirms part (i) of the viscosity solution definition. Part (ii) is proved in an analogous manner. □

Step 2: Approximate cost functions and overapproximation of Ω_f by Q_f. In this step we define a class of piecewise constant controls, denoted $\mathcal{U}_k^2$, nearly the same as $\mathcal{U}_k^1$, to accommodate that trajectories terminate at Q_f, not Ω_f, and we replace the cost functions L and h by approximations L^2 and $\hat{h}$, respectively. We define $\mathcal{U}_k^2 \subset \mathcal{U}_k^1$ to be the class of piecewise continuous controls whose constant time intervals $\tau(\sigma,x)$ are determined by the equivalence classes of $\simeq$ but not $\partial\Omega_f$. That is, case (ta) in the definition of $\tau(\sigma,x)$ in Step 1 is omitted. Next we define an approximate instantaneous cost $L^2 : \Omega\times\Sigma_k \to \mathbb{R}$ by

$$L^2(x,\sigma) := \hat{L}(e), \tag{6.5}$$

where $(\sigma,x) \in q$ and $e = (q,q')$ represents the time step. For $x \in \Omega$ and $\mu = \sigma_1\sigma_2\ldots \in \mathcal{U}_k^2$, if $T(x,\mu) < \infty$, the cost-to-go function $J_k^2 : \Omega\times\mathcal{U}_k^2 \to \mathbb{R}$ is

$$J_k^2(x,\mu) = \sum_{j=1}^N L^2(x_{j-1},\sigma_j) + \hat{h}(x_N),$$

where $N = \min\{j \geq 0 \mid x_j \in Q_f\}$. In other words, J_k^2 is a worst-case cost over a set of trajectories starting at x that visit the same sequence of equivalence classes of $\simeq$, and it is a worst-case cost w.r.t. J_k^1 because $\int_0^{\tau(\sigma,x)} L(\phi_s(x,\sigma),\sigma)ds \leq L^2(x,\sigma)$.

We define a value function $V_k^2 : \mathbb{R}^n \to \mathbb{R}$ as follows. For $x \in \Omega\setminus Q_f$,

$$V_k^2(x) = \inf_{\mu\in\mathcal{U}_k^1} J_k^2(x,\mu), \tag{6.6}$$

and for $x \in Q_f$, $V_k^2(x) = \hat{h}(x)$. For $x \in \Omega$ such that $V_k^2(x) < \infty$, V_k^2 satisfies the DPP

$$V_k^2(x) = \min_{\sigma\in\Sigma_k}\left\{L^2(x,\sigma) + V_k^2(\phi_{\tau(\sigma,x)}(x,\sigma))\right\}.$$

The proof is along the same lines as that of Proposition 5.2.

The following facts are useful for the subsequent result. The first lemma says that τ_q is order δ_k. The second lemma says that given two times τ and τ' that two trajectories spend, respectively, in the same equivalence class, $|\tau - \tau'|$ is order δ_k^2.

LEMMA 6.4. *If $\delta_k < \frac{m_f}{L_f}$, then for all $q \in Q$,*

$$\tau_q \leq \frac{\delta_k}{m_f - L_f \delta_k}. \tag{6.7}$$

Proof. Let $q \in Q$. Fix $x \in \Omega$ and $\sigma \in \Sigma_k$ such that $(\sigma, x) \in q$. We know $\|\phi_{\tau(\sigma,x)}(x,\sigma) - x\| \leq \delta_k$. We have

$$\begin{aligned}
\delta_k \geq \|\phi_{\tau(\sigma,x)} - x\| &= \left\| \int_0^{\tau(\sigma,x)} f(\phi_s(x,\sigma),\sigma) ds \right\| \\
&\geq \left\| \int_0^{\tau(\sigma,x)} f(x,\sigma) ds \right\| - \left\| \int_0^{\tau(\sigma,x)} [f(\phi_s(x,\sigma),\sigma) - f(x,\sigma)] ds \right\| \\
&\geq \tau(\sigma,x)\|f(x,\sigma)\| - \tau(\sigma,x) L_f \delta_k,
\end{aligned}$$

where in the last step we use the fact that $\|\phi_s(x,\sigma) - x\| \leq \delta_k$. Therefore,

$$\tau(\sigma,x) \leq \frac{\delta_k}{\|f(x,\sigma)\| - L_f \delta_k}.$$

Using Assumption 4.1(2) and taking the sup over all $\tau(\sigma,x)$ for q, the result follows. ∎

LEMMA 6.5. *Let $x, x' \in M_c^\sigma$ for some $c \in C_k$ and $\sigma \in \Sigma_k$ such that $\|x - x'\| \leq \delta_k$. Let τ, τ' be times such that $\phi_\tau(x,\sigma), \phi_{\tau'}(x',\sigma) \in M_{c+\Delta}^\sigma$. Then $|\tau - \tau'| \leq c_\gamma \tau \delta_k$ for some $c_\gamma > 0$.*

Proof. We have

$$\int_0^\tau \frac{d}{ds}(\gamma_n^\sigma(\phi_s(x,\sigma)))ds = \int_0^{\tau'} \frac{d}{ds}(\gamma_n^\sigma(\phi_s(x',\sigma)))ds.$$

Let $f = f(\phi_s(x,\sigma),\sigma)$, $f' = f(\phi_s(x',\sigma),\sigma)$, $d\gamma = \frac{d\gamma_n^\sigma(z)}{dz}|_{z=\phi_s(x,\sigma)}$, and $d\gamma' = \frac{d\gamma_n^\sigma(z)}{dz}|_{z=\phi_s(x',\sigma)}$. Then, rearranging terms,

$$\int_0^\tau (f' \cdot d\gamma')ds - \int_0^\tau (f \cdot d\gamma)ds = \int_{\tau'}^\tau (f' \cdot d\gamma')ds.$$

Let L_1 be the Lipschitz constant of $f \cdot d\gamma$ (using the fact that γ_n^σ is smooth). Then

$$\int_{\tau'}^\tau f' \cdot d\gamma' \leq L_1 \tau \|x - x'\| \leq L_1 \tau \delta_k.$$

Since γ_n^σ defines a transversal foliation to vector field $f(\cdot,\sigma)$, $f \cdot d\gamma > 0$. Let $c = \min_{s\in[\tau,\tau']}\{f' \cdot d\gamma'\} > 0$. Letting $c_\gamma = \frac{L_1}{c}$ we obtain the result. ∎

Remark 6.2. If $\mu \in \mathcal{U}_k^1$ is an ϵ-optimal control for x and the first time the trajectory $\phi_t(x,\mu)$ reaches Ω_f (Q_f) is T (T^2), then $T - T^2 \to 0$ and $|\phi_T(x,\mu) - \phi_{T^2}(x,\mu)| \to 0$ as $k \to \infty$. This follows from the fact that the distance between Ω_f and Q_f tends to zero as $k \to \infty$.

We denote by $\mu^2 \in \mathcal{U}_k^2$ the restriction of μ to $[0, T^2]$. Note that if the length of μ is $|\mu| = N$, then $|\mu_2| := N^2 \leq N$. Then we have the following result.

PROPOSITION 6.6. *Let $k_0 \in \mathbb{Z}^+$ be arbitrary, $x \in \mathcal{R}_{k_0}^1$, and $\mu \in \mathcal{U}_{k_0}^1$ be an ϵ-optimal control for x. Then $|J_k^1(x,\mu) - J_k^2(x,\mu^2)| \to 0$ as $k \to \infty$.*

Proof. Suppose $\mu = (\sigma_1, \tau_1)\dots(\sigma_N, \tau_N)$ and $\mu^2 = (\sigma_1, \tau_1)\dots(\sigma_{N^2}, \tau_{N^2})$, where $N^2 \leq N$. Thus, $N - N^2$ additional steps are required to reach $\partial\Omega_f$ after reaching Q_f. Then we have

$$\begin{aligned}\left|J_k^1(x,\mu) - J_k^2(x,\mu^2)\right| \leq \Bigg|&\sum_{j=1}^{N}\left[\int_0^{\tau(\sigma_j, x_{j-1})} L(\phi_s(x_{j-1},\sigma_j),\sigma_j)ds\right] + h(x_N) \\ &- \sum_{j=1}^{N^2}\left[\tau_{q_{j-1}} L(\xi_{j-1},\sigma_j)\right] - \hat{h}(x_{N^2})\Bigg|,\end{aligned}$$

where $(\sigma_j, x_{j-1}) \in q_{j-1}$ and $q_{j-1} = [(\sigma_j, \xi_{j-1})]$. There exists ξ_{N^2} such that $\hat{h}(x_{N^2}) = h(\xi_{N^2})$ and $\|x_{N^2} - \xi_{N^2}\| \leq \delta_k$. Also, using the mean value theorem, there exists $\tilde{t}_{j-1}$ with $\tilde{x}_{j-1} = \phi_{\tilde{t}_{j-1}}(x_{j-1}, \sigma_j)$ and $\|\tilde{x}_{j-1} - \xi_{j-1}\| \leq \delta_k$ such that

$$\begin{aligned}&\left|J_k^1(x,\mu) - J_k^2(x,\mu^2)\right| \\ &\quad\leq \sum_{j=1}^{N^2}\left|\tau(\sigma_j, x_{j-1})L(\tilde{x}_{j-1},\sigma_j) - \tau_{q_{j-1}}L(\xi_{j-1},\sigma_j)\right| \\ &\qquad + \left|\sum_{j=N^2+1}^{N}\left[\int_0^{\tau(\sigma_j,x_{j-1})} L(\phi_s(x_{j-1},\sigma_j),\sigma_j)ds\right]\right| + \left|h(x_N) - \hat{h}(x_{N^2})\right| \\ &\quad\leq \sum_{j=1}^{N^2}\tau_{q_{j-1}}L_L\delta_k + \sum_{j=1}^{N^2}[\tau_{q_{j-1}} - \tau(\sigma_j, x_{j-1})]L(\tilde{x}_{j-1},\sigma_j) \\ &\qquad + (T - T^2)M_L + L_h\|x_N - x_{N^2}\| + L_h\delta_k.\end{aligned}$$

The last three terms on the r.h.s. go to zero as $k \to \infty$ because of Remark 6.2 and since $\delta_k \to 0$. Using Lemma 6.4 the first summation decreases linearly as δ_k. Call the second summation on the r.h.s. B. Splitting B into sums over control switches and time steps, we have

$$\begin{aligned}B &\leq M_L\sum_{j=1}^{N^2}[\tau_{q_{j-1}} - \tau(\sigma_j, x_{j-1})]\mathbf{1}(\sigma_j = \sigma_{j-1}) \\ &\quad + M_L\sum_{j=1}^{N^2}[\tau_{q_{j-1}} - \tau(\sigma_j, x_{j-1})]\mathbf{1}(\sigma_j \neq \sigma_{j-1}) \\ &\leq M_L\sum_{j=1}^{N^2}c_{j-1}\tau_{q_{j-1}}\delta_k + M_L\sum_{j=1}^{N^2}\tau_{q_{j-1}}\mathbf{1}(\sigma_j \neq \sigma_{j-1})\end{aligned}$$

for some $c_{j-1} \in \mathbb{R}$. In the second line we used Lemma 6.5 and the fact that $\tau_{q_{j-1}} \geq \tau(\sigma_j, x_{j-1})$. Using Lemma 6.4 the first summation on the r.h.s. decreases linearly as δ_k. The second term on the r.h.s. goes to zero since, by Assumption 2.2, μ has a fixed number of control switches for all $k \geq k_0$. ∎

Step 3: Discrete states and nondeterminacy. In the last step we compare the value function $V_k^2(x)$ with the discrete value function $\hat{V}$ defined on A. The difference between the two is that trajectories defined over $\mathcal{U}_k^2$ do not include jumps, while trajectories whose time abstract versions are accepted by A can have jumps due to the nondeterminacy of A. Nevertheless, as $k \to \infty$ this discrepancy can be made negligible and we show that the difference between V_k^2 and $\hat{V}$ can be made arbitrarily small.

First we extend the domain of $\hat{V}(q)$, with an abuse of notation, by defining

$$\hat{V}_k(x) := \min_{\sigma \in \Sigma_k} \{\hat{V}(q) \mid (\sigma, x) \in q\}.$$

Also let $\hat{\mathcal{R}}_k = \{x \in \Omega \mid \hat{V}_k(x) < \infty\}$ and $\hat{\mathcal{R}} = \cup_k \hat{\mathcal{R}}_k$.

Remark 6.3.

(a) For each $x \in \cup_k \mathcal{R}_k^1$ and $\epsilon > 0$ there exists $m \in \mathbb{Z}^+$ and $\mu \in \mathcal{U}_m^2$ such that μ is an ϵ-optimal control for x w.r.t. V_k^2 with at most N_ϵ discontinuities. This follows from Remark 6.1(b) and the fact that trajectories in $\mathcal{U}_m^2$ are merely truncations of trajectories in $\mathcal{U}_m^1$.

(b) $\hat{\mathcal{R}} \subset \cup_k \mathcal{R}_k^1$, but the converse is not true, in general.

(c) If μ is an ϵ-optimal control for x w.r.t. V_k^2, then we can assume $\phi_t(x, \mu)$ does not self-intersect, for if it did we could find $\tilde{\mu}$, also ϵ-optimal, which eliminates loops in $\phi_t(x, \mu)$.

(d) $\|x - x'\| \to 0$ as $k \to \infty$ for all x, x', σ, $\sigma' \neq \sigma$ such that $([(\sigma, x)], [(\sigma', x')]) \in E$.

(e) For all $x \in \hat{R}$, $\hat{V}_k(x) \geq V_k^2(x)$. This follows because the argument of the max in the definition of $\hat{J}(q, c)$ is equal to $J_k^2(x, \mu)$, where c is a control policy as defined in section 5.1 with $c(q) = \sigma_1$, $(\sigma_1, x) \in q$, and $\mu = (\sigma_1, \tau_1) \ldots$ is the piecewise continuous control that corresponds to following policy c starting at x. Thus, $J_k^2(x, \mu)$ is the cost for a particular trajectory in $\tilde{\Pi}_c(q)$ which has no jumps at the control switches. Then we have $\hat{J}(q, c) \geq J_k^2(x, \mu)$, since $\hat{J}(q, c)$ maximizes over all trajectories in $\tilde{\Pi}_c(q)$.

PROPOSITION 6.7. *For all $x \in \hat{\mathcal{R}}$, $|\hat{V}_k(x) - V_k^2(x)| \to 0$ as $k \to \infty$.*

Proof. Fix $\epsilon > 0$ and $x \in \hat{\mathcal{R}}$. By Remark 6.3(a) there exists $m > 0$ and an ϵ-optimal control $\mu \in \mathcal{U}_m^2$ for x w.r.t. V_m^2. Denote $\mu = ((\sigma_1, \tau_1) \ldots (\sigma_N, \tau_N))$, where τ_i is the time σ_i is applied. Let c be any control policy on Q that is generated using δ_k and C_k, for $k \geq m$. Then, using Remark 6.3(e),

$$0 \leq \hat{V}_k(x) - V_k^2(x) \leq \hat{J}_k(q, c) - J_k^2(x, \mu) + \epsilon,$$

where $q = [(\sigma_1, x)]$. If we can show there exists $\overline{k} \geq m$ such that for $k > \overline{k}$, there exists a policy $\overline{c}$ such that

$$\hat{J}_k(q, \overline{c}) - J_k^2(x, \mu) < \epsilon,$$

then the result follows.

By Remark 6.3(d) and the transversality of $\phi_t(x, \mu)$ with the level sets of γ_n, we can find $\overline{k} \geq m$ such that for $k > \overline{k}$, there exists a family of (both continuous and discontinuous) trajectories Ψ_k starting at x with the following properties:

(1) $\phi_t(x, \mu) \in \Psi_k$.

(2) $\phi \in \Psi_k$ is defined over a control $\tilde{\mu} = ((\sigma_1, \tilde{\tau}_1), \ldots, (\sigma_N, \tilde{\tau}_N)) \in \mathcal{U}_k^2$ with the same sequence of control values as μ.

(3) $\phi \in \Psi_k$ switches controls on the same (transversal) submanifolds as $\phi_t(x, \mu)$ and reaches Q_f.

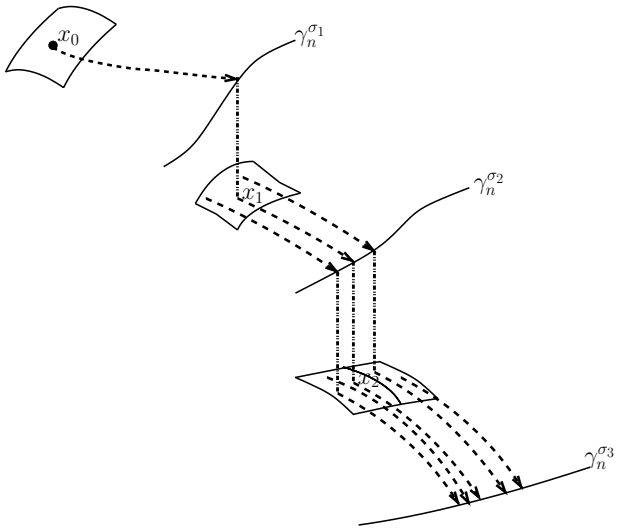


FIG. 6.2. *The family of trajectories* Ψ_k *in the proof of Proposition* 6.7.

(4) If $x_j^- = \phi_{\tau_j}(x_{j-1}, \sigma_j)$, then x_j, the initial condition of the next step, satisfies $([(\sigma_j, x_j^-)], (\sigma_{j+1}, x_j)]) \in E$. Thus, the trajectories of Ψ_k include jumps at the control switches modeling the nondeterminacy of A.

(5) If $\phi \in \Psi_k$ intersects $q \in Q$ in $\mathbb{R}^n$ at the jth step, then that is the only step where it intersects q. Also, all other $\phi' \in \Psi_k$ that intersect q do so at the jth step only. This requirement can be met for sufficiently large k by the fact that ϕ' has no self-intersections, by the fact that there are a finite number of steps, and by Remark 6.3(d). For if ϕ' has a self-intersection, then since ϕ' approaches $\phi_t(x, \mu)$ as $k \to \infty$, this would imply $\phi_t(x, \mu)$ has a self-intersection, contradicting Remark 6.3(c).

The family Ψ_k includes all trajectories starting at x, using the same sequence of control values as μ, and switching on the same equivalence class boundaries $\phi_t(x, \mu)$. Moreover, the initial condition at the start of each step can be any point in an equivalence class that has a nonempty intersection in $\mathbb{R}^n$ with the equivalence class reached at the end of the previous step. One visualizes a tube of trajectories that fans out with each successive control switch, as depicted in Figure 6.2. By choosing k sufficiently large and by transversality, all these trajectories reach Q_f.

Let $W_k(\phi) = \sum_{j=1}^{N} L^2(x_{j-1}, \sigma_j) + \hat{h}(x_N)$. Observe that for $\phi, \phi' \in \Psi_k$, $|W_k(\phi) - W_k(\phi')| \to 0$ as $k \to \infty$, using Lipschitz continuity of L and h and Remark 6.3(d). We can define a control policy $\overline{c}$ in which $q \in Q$ is assigned a time step if q is not visited by any trajectory in Ψ_k. If $q \in Q$ is visited by some $\phi \in \Psi_k$ in its jth step, then we assign $\overline{c}(q) = \sigma_j$. This gives a well-defined value for c because of Property 4. By construction, A accepts the time abstract trajectory starting at q corresponding to each trajectory of Ψ_k. $\overline{c}$ is admissible because otherwise some time abstract trajectory of A would have a Zeno loop. But a time abstract trajectory of A with a Zeno loop has a corresponding timed trajectory in Ψ_k that violates Property 4 of Ψ_k.

Now we observe that

$$\hat{J}(q, \overline{c}) = \max_{\phi \in \Psi_k} W_k(\phi) := W_k(\overline{\phi}).$$

Thus, $\hat{J}_k(q, \overline{c}) - J_k^2(x, \mu) \leq |W_k(\overline{\phi}) - W_k(\phi(x, \mu))| \to 0$ as $k \to \infty$. ∎

Combining Propositions 6.3, 6.6, and 6.7, we have the next theorem.

THEOREM 6.8. *For all* $x \in \hat{\mathcal{R}}$, $\hat{V}_k(x) \to V(x)$ *as* $k \to \infty$.

7. Implementation. So far we have developed a discrete method for solving an optimal control problem based on hybrid systems and bisimulation. Now we focus on the pragmatic question of how the discretized problem can be efficiently solved. In this section we propose a modification of the Dijkstra algorithm suitable for nondeterministic automata and prove that it is optimal and does not synthesize Zeno loops.

7.1. Motivation. Capuzzo Dolcetta and Evans [12] introduced a method for obtaining approximations of viscosity solutions based on time discretization of the HJB equation. The approximations of the value function correspond to a discrete time optimal control problem, for which an optimal control can be synthesized which is piecewise constant. Finite difference approximations were also introduced in [15] and [47]. In general, the time discretized approximation of the HJB equation is solved by finite element methods. Gonzales and Rofman [23] introduced a discrete approximation by triangulating the domain of the problem, while the admissible control set is approximated by a finite set. Gonzales and Rofman's approach is adapted in several papers, including [18]. The approach of [50] uses the special structure of an optimal control problem to obtain a single-pass algorithm to solve the discrete problem, thus bypassing the expensive iterations of a finite element method. See [45] for a recent adaptation of Tsitsiklis' approach. The essential property needed to find a single pass algorithm is to obtain a partition of the domain so that the cost-to-go value from any equivalence class of the partition is determined from knowledge of the cost-to-go from those equivalence classes with strictly smaller cost-to-go values. We obtain a partition of the domain provided by a bisimulation partition. *The combination of the structure of the bisimulation partition and the requirement of non-Zeno trajectories enables us to reproduce the essential property of* [50]*, so that we obtain a Dijkstra-like algorithmic solution.* Our approach has complexity $O(N \log N)$ if suitable data structures are used, where N is the number of locations of the finite automaton. The number N is, of course, exponential in n, the dimension of the continuous state space.

7.2. Nondeterministic Dijkstra algorithm. The dynamic programming solution (5.4)–(5.5) can be viewed as a shortest path problem on a nondeterministic finite graph subject to all optimal paths satisfying a non-Zeno condition. We propose an algorithm that is a modification of the Dijkstra algorithm for deterministic graphs [17]. First we define the notation. F_n is the set of states that have been assigned a control and are deemed finished at iteration n, while U_n are the unfinished states. At each n, $Q = U_n \cup F_n$. $\Sigma_n(q) \subseteq \Sigma_\delta$ is the set of control events at iteration n that take state q to finished states exclusively. $\tilde{U}_n$ is the set of states for which there exists a control event that can take them to finished states exclusively. $\tilde{V}_n(q)$ is a tentative cost-to-go value at iteration n. B_n is the set of best states among $\tilde{U}_n$.

The nondeterministic Dijkstra (NDD) algorithm first determines $\tilde{U}_n$ by checking if any q in U_n can take a step to states belonging exclusively to F_n. For states belonging to $\tilde{U}_n$, an estimate of the value function $\tilde{V}$ following the prescription of (5.4) is obtained: among the set of control events constituting a step into states in F_n, select the event with the lowest worst-case cost. Next, the algorithm determines B_n, the states with the lowest $\tilde{V}$ among $\tilde{U}_n$, and these are added to F_{n+1}. The iteration counter is incremented until it reaches $N = |Q|$. It is assumed in the following description that initially $\hat{V}(q) = \infty$ and $c(q) = \emptyset$ for all $q \in Q$.

PROCEDURE NDD.

$\mathsf{F}_1 = \mathsf{Q}_\mathsf{f}$; $\mathsf{U}_1 = \mathsf{Q} - \mathsf{Q}_\mathsf{f}$;
for each $\mathsf{q} \in \mathsf{Q}_\mathsf{f}$, $\hat{\mathsf{V}}(\mathsf{q}) = \hat{\mathsf{h}}(\mathsf{q})$;

for $\mathsf{n} = 1$ to N, do
  for each $\mathsf{q} \in \mathsf{U}_\mathsf{n}$,
    $\Sigma_\mathsf{n}(\mathsf{q}) = \{\sigma' \in \Sigma_\delta \mid \text{if } \mathsf{q} \xrightarrow{\sigma'} \mathsf{q}', \text{then } \mathsf{q}' \in \mathsf{F}_\mathsf{n}\}$;
  $\tilde{\mathsf{U}}_\mathsf{n} = \{\mathsf{q} \in \mathsf{U}_\mathsf{n} \mid \Sigma_\mathsf{n}(\mathsf{q}) \neq \emptyset\}$;
  for each $\mathsf{q} \in \tilde{\mathsf{U}}_\mathsf{n}$,
    $\tilde{\mathsf{V}}_\mathsf{n}(\mathsf{q}) = \min_{\sigma' \in \Sigma_\mathsf{n}(\mathsf{q})}\{\max_{\mathsf{e}=(\mathsf{q},\mathsf{q}') \in \mathsf{E}_{\sigma'}(\mathsf{q})}\{\hat{\mathsf{L}}(\mathsf{e}) + \hat{\mathsf{V}}(\mathsf{q}')\}\}$;
  $\mathsf{B}_\mathsf{n} = \operatorname{argmin}_{\mathsf{q} \in \tilde{\mathsf{U}}_\mathsf{n}}\{\tilde{\mathsf{V}}_\mathsf{n}(\mathsf{q})\}$;
  for each $\mathsf{q} \in \mathsf{B}_\mathsf{n}$,
    $\hat{\mathsf{V}}(\mathsf{q}) = \tilde{\mathsf{V}}_\mathsf{n}(\mathsf{q})$;
    $\mathsf{c}(\mathsf{q}) = \operatorname{argmin}_{\sigma' \in \Sigma_\mathsf{n}(\mathsf{q})}\{\max_{\mathsf{e}=(\mathsf{q},\mathsf{q}') \in \mathsf{E}_{\sigma'}(\mathsf{q})}\{\hat{\mathsf{L}}(\mathsf{e}) + \hat{\mathsf{V}}(\mathsf{q}')\}\}$;
  endfor
  $\mathsf{F}_{\mathsf{n}+1} = \mathsf{F}_\mathsf{n} \cup \mathsf{B}_\mathsf{n}$; $\mathsf{U}_{\mathsf{n}+1} = \mathsf{Q} - \mathsf{F}_{\mathsf{n}+1}$;
endfor

7.3. Justification. In this section we prove that the algorithm is *optimal*; that is, it synthesizes a control policy so that each $q \in Q$ reaches Q_f with the best worst-case cost. We observe a few properties of the algorithm. First, if all states of Q can reach Q_f in a nondeterministic sense, then $Q - Q_f = \cup_n B_n$. By nondeterministic sense we mean that for each $q \in Q$, there exists a control policy c such that $\Pi_c(q) = \tilde{\Pi}_c(q)$. Note that if this condition is not met, then it can happen that at some iteration of NDD, $\tilde{U}_n = \emptyset$ but $U_n \neq \emptyset$. Second, as in the deterministic case, the algorithm computes $\hat{V}$ in order of level sets of $\hat{V}$. In particular, $\hat{V}(B_n) \leq \hat{V}(B_{n+1})$. Finally, we need the following property.

LEMMA 7.1. *For all $q \in Q$ that can reach Q_f in a nondeterministic sense and for all $\sigma' \in \Sigma_\delta$,*

$$\hat{V}(q) \leq \max_{e=(q,q') \in E_{\sigma'}(q)} \{\hat{L}(e) + \hat{V}(q')\}.$$

Proof. Fix $q \in Q$ and $\sigma' \in \Sigma_\delta$. There are two cases.

Case 1.

$$\hat{V}(q) \leq \max_{e=(q,q') \in E_{\sigma'}(q)} \{\hat{V}(q')\}.$$

In this case the result is obvious.

Case 2.

$$\hat{V}(q) > \max_{e=(q,q') \in E_{\sigma'}(q)} \{\hat{V}(q')\}. \tag{7.1}$$

By assumption, q belongs to some B_n. Suppose without loss of generality that $q \in B_j$. Together with (7.1) this implies $q' \in F_j$ for all q' such that $q \xrightarrow{\sigma'} q'$. This, in turn, means that $\sigma' \in \Sigma_j(q)$ and according to the algorithm

$$\hat{V}(q) = \tilde{V}_j(q) \leq \max_{e=(q,q') \in E_{\sigma'}(q)} \{\hat{L}(e) + \hat{V}(q')\},$$

which proves the result. ∎

THEOREM 7.2. *Algorithm NDD is optimal and synthesizes a control policy with no Zeno loops.*

Proof. First we prove optimality. Let $V(q)$ be the optimal (best worst-case) cost-to-go for $q \in Q$ and $\overline{Q} = \{q \in Q \mid V(q) < \hat{V}(q)\}$. Let $l(\pi_q)$ be the number of edges taken by the shortest optimal (best worst-case) trajectory π_q from q. Define $\overline{q} = \arg\min_{q \in \overline{Q}} \{l(\pi_q)\}$. Suppose that the best worst-case trajectory starting at $\overline{q}$ is $\pi_{\overline{q}} = \overline{q} \xrightarrow{\sigma'} \overline{\overline{q}} \to \cdots$. We showed in the previous lemma that

$$\hat{V}(\overline{q}) \leq \max_{e=(\overline{q},q') \in E_{\sigma'}(\overline{q})} \{\hat{L}(e) + \hat{V}(q')\} = \hat{L}(e) + \hat{V}(\overline{\overline{q}}).$$

Since $\pi_{\overline{q}}$ is the best worst-case trajectory from $\overline{q}$ and by the optimality of $V(\overline{q})$,

$$V(\overline{q}) = \max_{e=(\overline{q},q') \in E_{\sigma'}(\overline{q})} \{\hat{L}(e) + V(q')\} = \hat{L}(e) + V(\overline{\overline{q}}).$$

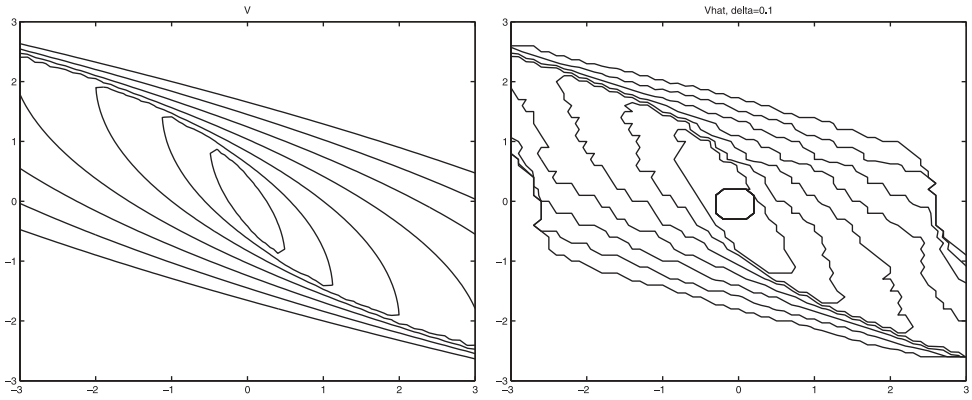
Since $\pi_{\overline{q}}$ is the shortest best worst-case trajectory, we know that $\overline{\overline{q}} \notin \overline{Q}$, so $V(\overline{\overline{q}}) = \hat{V}(\overline{\overline{q}})$. This implies $\hat{V}(\overline{q}) \leq \hat{L}(e) + V(\overline{\overline{q}}) = V(\overline{q})$, a contradiction.

To prove that the algorithm synthesizes a policy with no Zeno loops, we argue by induction. The claim is obviously true for F_1. Suppose that the states of F_n have been assigned controls forming no Zeno loops. Consider F_{n+1}. Each state of B_n takes either a time step or a control switch to F_n, so there cannot be a Zeno loop in B_n. The only possibility is for some $q \in B_n$ to close a Zeno loop with states in F_n. This implies there exists a control assignment that allows an edge from F_n to q to be taken, but this is not allowed by NDD. Thus, F_{n+1} has no Zeno loops. ∎

8. Examples. We consider two simple examples where the solution of the optimal control problem is known in order to illustrate the correctness of the method. The software that generates the optimal enabling conditions is broken into two programs, one that generates the automaton given the information about the bisimulation and one that runs the algorithm NDD. The first program takes as input the control values Σ_δ and the level values of γ_i^σ, $i = 1, \ldots, n$, $\sigma \in \Sigma_\delta$, defining the bisimulation. The functions γ_i^σ, $\hat{L}$, and $\hat{h}$ are compiled with the executable. A data structure that associates to each location of the finite automaton the lower and upper level values of each γ_i^σ allows time steps to be encoded symbolically, namely, by sorting nodes with equal upper and lower first integral level values in ascending order of γ_n^σ level values. The edges of the finite automaton that correspond to σ-steps are generated numerically by evaluating γ_i^σ for $i = 1, \ldots, n$ and each $\sigma \in \Sigma_\delta$ and thereby determining which equivalence classes overlap for each pair (l, l') of locations. In our implementation the grid of sample points is $\{x \in \Omega, \gamma_i \in C^k\}$ in order to correlate with the mesh size of the bisimulation partition. This numerical step can also be performed symbolically if the functions γ_i^σ are polynomials using a quantifier elimination algorithm [5]. However, the quantifier elimination step is expensive, and for approximate solutions it suffices to use a numerical approach.

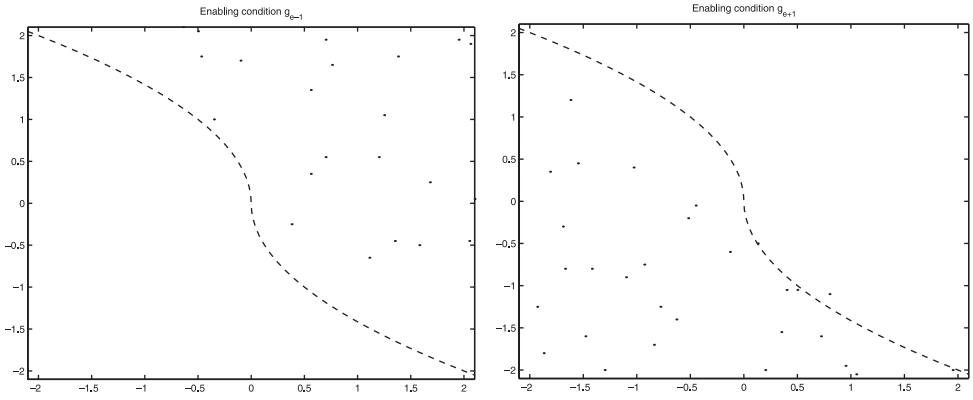
First we apply our method to Examples 3.1 and 4.1. The bang-bang solution obtained using Pontryagin's maximum principle is well known to involve a single switching curve. The continuous value function V is shown in Figure 8.1(a).

The results of algorithm NDD are shown in Figures 8.1(b) and 8.2. In Figure 8.2 the dashed line is the smooth switching curve for the continuous problem. The black dots identify equivalence classes where NDD assigns a control switch. Considering $g_{e_{-1}}$ we see that the boundary of the enabling condition in the upper left corner is a



(a) V (b) $\hat{V}$ for $\Delta = 0.1$.

FIG. 8.1. *Continuous and discrete value functions for double integrator.*



(a) g_{e-1} (b) g_{e1}

FIG. 8.2. *Enabling conditions.*

jagged approximation using equivalence classes of the smooth switching curve. Initial conditions in the upper left corner just inside the enabling condition must switch to a control of $u = -1$; otherwise the trajectory will increase in the x_2 direction and not reach the target. Initial conditions in the upper left corner just outside the enabling condition must allow time to pass until they reach the enabling condition, for if they switched to $u = -1$ they would be unable to reach the target. Hence the upper left boundary of the enabling condition is crisp. The lower right side of the enabling condition which has islands of time steps shows the effect of the nondeterminacy of automaton A. These additional time steps occur because it can be less expensive to take a time step than to incur the cost of the worst-case control switch. Indeed consider an initial condition in Figure 8.2(a) which lies in an equivalence class that takes a time step but should take a control switch according to the continuous optimal

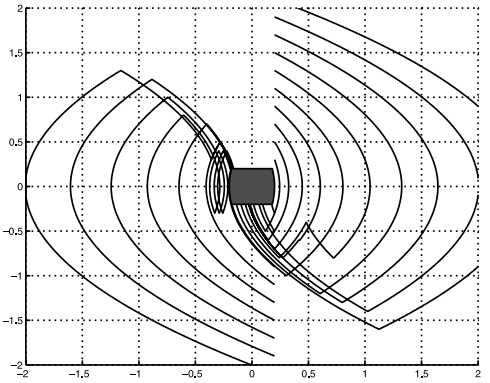


FIG. 8.3. *Trajectories of the closed-loop system.*

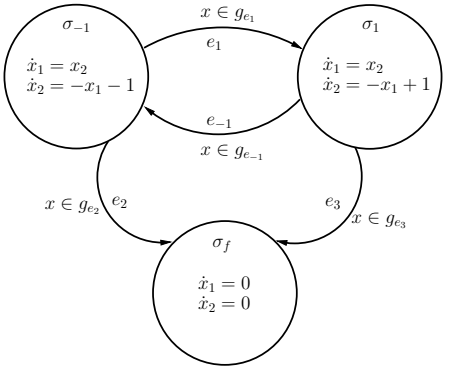


FIG. 8.4. *Hybrid automaton for example* 2.

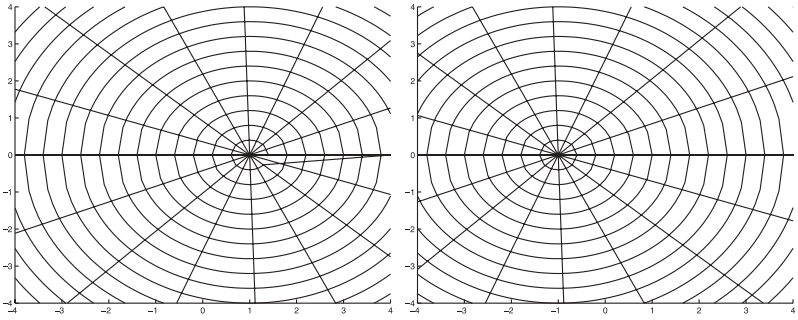
control. Such a point will move up and to the left before it takes a control switch. By moving slightly closer to the target, the worst-case cost-to-go incurred in a control switch is reduced. Notice that all such initial conditions eventually take a control switch. This phenomenon of extra time steps is a function of the mesh size δ: as δ decreases there are fewer extra time steps. Finally we note that the two enabling conditions have an empty intersection, as expected to ensure non-Zeno trajectories.

Figure 8.3 shows trajectories of the closed-loop system using the controller synthesized by NDD. The central shaded region is an enlarged target set.

Next we consider the time optimal control problem for the system

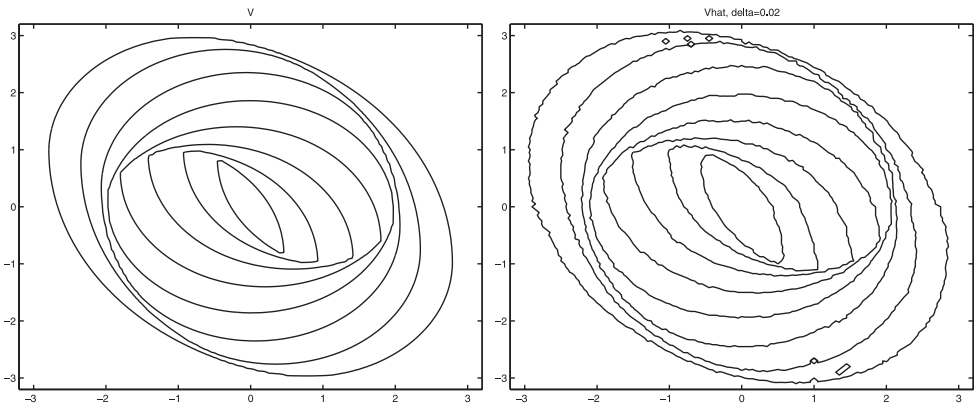
$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -x_1 + u. \end{aligned} \tag{8.1}$$

Suppose $\Omega = (-1,1) \times (-1,1)$ and $\Omega_f = \overline{B}_\epsilon(0)$, the closed epsilon ball centered at 0. The cost-to-go function is $J(x,\mu) = \int_0^{T(x,\mu)} dt$ and $U = \{u : |u| \leq 1\}$. We select $\Sigma_\delta = \{-1, 1\}$ so that $\delta = 1$. The hybrid system is show in Figure 8.4. The state set is $\{\sigma_{-1} = -1, \sigma_1 = 1, \sigma_f\} \times \mathbb{R}^2$. $g_{e_{-1}}$ and g_{e_1} are unknown and must be synthesized, while $g_{e_2} = g_{e_3} = \Omega_f$. A first integral for (8.1) is $\sqrt{(x_1 - u)^2 + x_2^2} = c_1$, where $u = \pm 1$. The transverse foliation is chosen to be defined by the function $\arctan(\frac{x_2}{x_1 - u}) = c_2$. Partitions for locations σ_1 and σ_{-1} are shown in Figure 8.5. The results of algorithm NDD are shown in Figures 8.6(b) and 8.7. In Figure 8.7 the dashed line is the



(a) σ_1 (b) σ_{-1}

FIG. 8.5. *Partitions for states* σ_1 *and* σ_{-1} *of the hybrid automaton of Figure* 8.5.



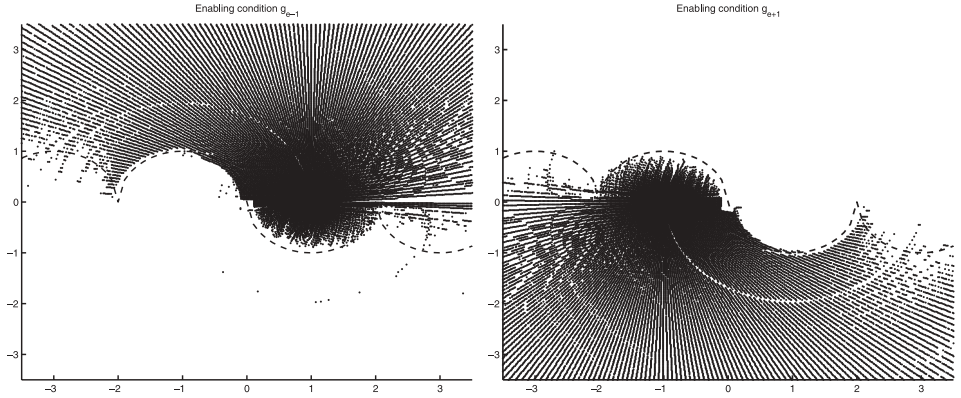
(a) V (b) $\hat{V}$ for $\Delta = 0.02$.

FIG. 8.6. *Continuous and discrete value functions for example* 2.

switching curve for the continuous problem. As in the previous example the black dots identify equivalence classes where NDD assigns a control switch. Figure 8.8 shows trajectories of the closed-loop system using the controller synthesized by NDD. An enlarged target set is at the origin.

Remark 8.1. From these examples we observe that our method is best suited to problems when there are relatively few control switches, as each control switch incurs an error of order δ. Also the method is suited to problems where bang-bang controls are used. The method has advantages in situations where a fine time discretization of the vector field is needed for standard finite element methods. We do not require time discretization because of the particular choice of grid, which captures time evolution exactly. Finally, because the method requires computation of weak first integrals, only systems for which first integrals are computable in closed form are considered.

Table 8.1 shows the computation times for the two examples as a function of δ. The automaton size and the time in seconds to generate it appear in the second and



(a) $g_{e_{-1}}$ (b) g_{e_1}

FIG. 8.7. *Enabling conditions for example* 2.

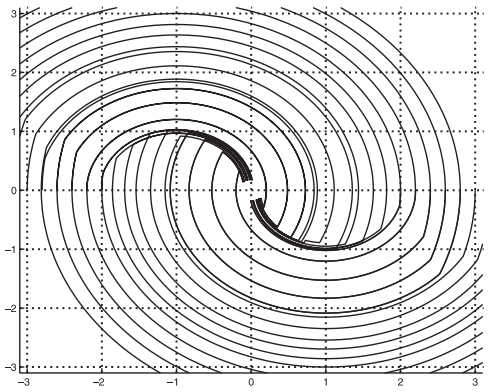


FIG. 8.8. *Trajectories of the closed-loop system for example* 2.

TABLE 8.1

Example	δ	N	Automaton	I/O	NDD	Finished
1	.2	7200	.11	4.57	.02	3274
1	.1	28800	.46	5.69	.08	12835
1	.05	115200	1.97	9.09	.38	51490
1	.025	460800	7.83	22.91	1.77	205624
2	.2	1920	.03	4.42	.01	1920
2	.1	7560	.09	4.78	.07	7560
2	.05	30240	.39	6.13	.28	30240
2	.025	120960	1.55	11.04	1.3	189000
2	.0125	482880	7.99	35.29	7.12	482880

third columns. We report the time for file I/O, which otherwise would dominate the computation times. The time in seconds to run NDD and the size of the set F_n of finished states appear in the last two columns. Note that not all nodes are finished in the first example because the regions of $\mathbb{R}^2$ that are partitioned in the two locations do not overlap perfectly, resulting in the nonexistence of a trajectory that can reach the origin starting in a subset of the nonoverlapping areas.

9. Conclusion. In this paper we have developed a methodology for the synthesis of optimal controls based on hybrid systems and bisimulations. The idea is to translate the optimal control problem to a switching problem on a hybrid system whose locations describe the dynamics when the control is constant. When the vector fields for each location of the hybrid automaton have local first integrals which can be expressed analytically we are able to define a finite bisimulation using the approach of [10]. From the finite bisimulation we obtain a (time abstract) finite automaton on which a dynamic programming problem can be formulated that can be solved efficiently. We proposed an efficient single-pass algorithm to solve this dynamic programming problem and demonstrated its correctness on two simple examples.

REFERENCES

[1] R. Alur and D. Dill, *A theory of timed automata*, Theoret. Comput. Sci., 126 (1994), pp. 183–235.
[2] R. Alur, T. Henzinger, G. Lafferriere, and G. Pappas, *Discrete abstractions of hybrid systems*, Proc. IEEE, 88 (2000), pp. 971–984.
[3] V.I. Arnold, *Mathematical Methods in Classical Mechanics*, Springer-Verlag, New York, 1997.
[4] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.
[5] S. Basu, R. Pollack, and M.-F. Roy, *On the combinatorial and algebraic complexity of quantifier elimination*, J. ACM, 43 (1996), pp. 1002–1045.
[6] A. Bensoussan and J.L. Menaldi, *Hybrid control and dynamic programming*, Dynam. Contin. Discrete Impuls. Systems, 3 (1997), pp. 395–442.
[7] G. Bluman and S. Anco, *Symmetry and Integration Methods for Differential Equations*, Springer-Verlag, New York, 2002.
[8] V.G. Boltyanskii, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control, 4 (1966), pp. 326–361.
[9] M. Branicky, V. Borkar, and S. Mitter, *A unified framework for hybrid control: Model and optimal control theory*, IEEE Trans. Automat. Control, 43 (1998), pp. 31–45.
[10] M. Broucke, *A geometric approach to bisimulation and verification of hybrid systems*, in Hybrid Systems: Computation and Control, F. Vaandrager and J. van Schuppen, eds., Lecture Notes in Comput. Sci. 1569, Springer-Verlag, New York, 1999, pp. 61–75.
[11] M. Broucke, *Reachability analysis for hybrid systems with linear dynamics*, in Proceedings of the Fifteenth International Symposium, Mathematical Theory of Networks and Systems (MTNS '02), Notre Dame, IN, 2002.
[12] I. Capuzzo Dolcetta and L.C. Evans, *Optimal switching for ordinary differential equations*, SIAM J. Control Optim., 22 (1984), pp. 143–161.
[13] J. Chavarriga, H. Giacomini, J. Gine, and J. Llibre, *On the integrability of two-dimensional flows*, J. Differential Equations, 157 (1999), pp. 163–182.
[14] M. Crandall and P. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[15] M.G. Crandall and P.L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.
[16] P.E. Crouch and R.L. Grossman, *The explicit computation of integration algorithms and first integrals for ordinary differential equations with polynomial coefficients using trees*, in Proceedings of International System Symposium on Symbolic and Algebraic Computation (ISSAC '92), P.S. Wang, ed., ACM, New York, 1992, pp. 89–94.
[17] E.W. Dijkstra, *A note on two problems in connection with graphs*, Numer. Math., 1 (1959), pp. 269–271.

[18] M. Falcone, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 1–13.
[19] K. Feng and D.-L. Wang, *Dynamical Systems and Geometric Construction of Algorithms*, in Computational Mathematics in China, Contemp. Math. 163, Z. Shi and C. Yang, eds., AMS, Providence, RI, 1994.
[20] W.H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
[21] A.T. Fuller, *Study of an optimum non-linear control system*, J. Electron. Control, 15 (1963), pp. 63–71.
[22] J. Goldman, *Integrals of multinomial systems of ordinary differential equations*, J. Pure Appl. Algebra, 45 (1987), pp. 225–240.
[23] R. Gonzalez and E. Rofman, *On deterministic control problems: An approximation procedure for the optimal cost* I: *The stationary problem*, SIAM J. Control Optim., 23 (1985), pp. 242–266.
[24] A. Goriely, *Integrability and Nonintegrability of Dynamical Systems*, World Scientific, River Edge, NJ, 2001.
[25] E. Hairer, *Geometric integration of ordinary differential equations on manifolds*, BIT, 41 (2001), pp. 996–1007.
[26] T. Henzinger, *Hybrid automata with finite bisimulations*, in Proceedings of the 22nd ICALP: Automata, Languages and Programming, Lecture Notes in Comput. Sci. 944, Springer-Verlag, New York, 1995, pp. 324–335.
[27] W. Hereman, *Symbolic software for Lie symmetry analysis*, in CRC Handbook of Lie Group Analysis of Differential Equations, N.H. Ibragimov, ed., CRC Press, Boca Raton, FL, 1995.
[28] A. Iserles and A. Zanna, *Preserving algebraic invariants with Runge-Kutta methods*, J. Comput. Appl. Math., 125 (2000), pp. 69–81.
[29] I. Kupka, *The ubiquity of the Fuller phenomenon*, in Nonlinear Controllability and Optimal Control, H.J. Sussmann, ed., Marcel Dekker, New York, 1990.
[30] H.B. Lawson, *The Qualitative Theory of Foliations*, in Regional Conf. Ser. Math. 27, AMS, Providence, RI, 1977.
[31] W. LeVeque, *Fundamentals of Number Theory*, Addison-Wesley, Reading, MA, 1977.
[32] P.L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.
[33] O. Maler, A. Pnueli, and J. Sifakis, *On the synthesis of discrete controllers for timed systems*, in Proceedings of STACS '95, E.W. Mayr and C. Puech, eds., Lecture Notes in Comput. Sci. 900, Springer-Verlag, New York, 1995, pp. 229–242.
[34] Y.-K. Man, *Computing closed form solutions of first order ODE's using the Prelle-Singer procedure*, J. Symbolic Comput., 16 (1993), p. 423.
[35] Y.K. Man, *First integrals of autonomous systems of differential equations and the Prelle-Singer procedure*, J. Phys. A, 27 (1994), pp. L329–L332.
[36] R. McLachlan and C. Scovel, *A survey of open problems in symplectic integration*, in Integration Algorithms and Classical Mechanics, J.E. Marsden, G. Patrick, and W. Shadwick, eds., AMS, Providence, RI, 1996, pp. 151–180.
[37] P. Olver, *Application of Lie Groups to Differential Equations*, Springer-Verlag, New York, 1986.
[38] L.S. Pontryagin, V. Boltiansky, A. Gamkrelitze, and E. Mischenko, *The Mathematical Theory of Optimal Processes*, Interscience Publishers, New York, 1962.
[39] M.J. Prelle and M.F. Singer, *Elementary first integrals of differential equations*, Trans. Amer. Math. Soc., 279 (1983), pp. 215–229.
[40] G. Quispel and H. Capel, *Solving ODE's numerically while preserving a first integral*, Phys. Lett. A, 218 (1996), pp. 223–228.
[41] J. Raisch, *Controllability and observability of simple hybrid control systems-FDLTI plants with symbolic measurements and quantized control inputs*, in International Conference on Control '94, vol. 1, IEE, London, 1994, pp. 595–600.
[42] H. Royden, *Real Analysis*, 3rd ed., Prentice–Hall, Englewood Cliffs, NJ, 1988.
[43] F. Schwarz, *A REDUCE package for determining first integrals of autonomous systems of ordinary differential equations*, Comput. Phys. Comm., 39 (1986), pp. 285–296.
[44] F. Schwarz, *An algorithm for determining polynomial first integrals of autonomous systems of ordinary differential equations*, J. Symbolic Comput., 1 (1985), pp. 229–233.
[45] J.A. Sethian and A. Vladimirsky, *Ordered upwind methods for static Hamilton–Jacobi equations: Theory and algorithms*, SIAM J. Numer. Anal., 41 (2003), pp. 325–363.
[46] W.Y. Sit, *On Goldman's algorithm for solving first-order multinomial autonomous systems*, in Proceedings of Applied Algebra, Algebraic Algorithms and Error-Correcting Codes, 6th International Conference, T. Mora, ed., Springer-Verlag, New York, 1989, pp. 386–395.

[47] P.E. Souganidis, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, J. Differential Equations, 59 (1985), pp. 1–43.

[48] J. Stiver, P. Antsaklis, and M. Lemmon, *A logical DES approach to the design of hybrid control systems*, Math. Comput. Modelling, 23 (1996), pp. 55–76.

[49] J.-M. Strelcyn and S. Wojciechowski, *A method of finding first integrals for three-dimensional dynamical systems*, Phys. Lett. A, 133 (1988), pp. 207–212.

[50] J.N. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), pp. 1528–1538.

[51] H.S. Witsenhausen, *A class of hybrid-state continuous-time dynamic systems*, IEEE Trans. Automat. Control, 11 (1966), pp. 161–167.

[52] H. Wong-Toi, *The synthesis of controllers for linear hybrid automata*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 4607–4612.

[53] X. Yang, M. Lemmon, and P. Antsaklis, *On the supremal controllable sublanguage in the discrete-event model of nondeterministic hybrid control systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 2098–2103.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 1953–1971

BOUNDARY CONTROL OF THE LINEARIZED GINZBURG–LANDAU MODEL OF VORTEX SHEDDING*

OLE MORTEN AAMO†, ANDREY SMYSHLYAEV‡, AND MIROSLAV KRSTIĆ‡

Abstract. In this paper, we continue the development of state feedback boundary control laws based on the backstepping methodology, for the stabilization of unstable, parabolic partial differential equations. We consider the linearized Ginzburg–Landau equation, which models, for instance, vortex shedding in bluff body flows. Asymptotic stabilization is achieved by means of boundary control via state feedback in the form of an integral operator. The kernel of the operator is shown to be twice continuously differentiable, and a series approximation for its solution is given. Under certain conditions on the parameters of the Ginzburg–Landau equation, compatible with vortex shedding modelling on a semi-infinite domain, the kernel is shown to have compact support, resulting in partial state feedback. Simulations are provided in order to demonstrate the performance of the controller. In summary, the paper extends previous work in two ways: (1) it deals with two coupled partial differential equations, and (2) under certain circumstances handles equations defined on a semi-infinite domain.

Key words. partial differential equations, boundary control, stabilization, flow control

AMS subject classifications. 35B37, 35B65, 93D15

DOI. 10.1137/S036301290342601X

1. Introduction.

In this paper, we continue the development of state feedback boundary control laws based on the backstepping methodology [6], for the stabilization of unstable, parabolic partial differential equations [3, 2, 10, 14]. We consider the linearized Ginzburg–Landau equation given by

$$\frac{\partial A(\breve{x},t)}{\partial t} = a_1 \frac{\partial^2 A(\breve{x},t)}{\partial \breve{x}^2} + a_2(\breve{x}) \frac{\partial A(\breve{x},t)}{\partial \breve{x}} + a_3(\breve{x}) A(\breve{x},t) \tag{1}$$

for $\breve{x} \in (0, x_d)$, with boundary conditions

$$A(0,t) = u(t), \tag{2}$$

$$A(x_d,t) = 0, \tag{3}$$

and where $A : [0, x_d] \times \mathbb{R}_+ \to \mathbb{C}$, $a_2 \in C^2([0, x_d]; \mathbb{C})$, $a_3 \in C^1([0, x_d]; \mathbb{C})$, $a_1 \in \mathbb{C}$, $x_d > 0$, and $u : \mathbb{R}_+ \to \mathbb{C}$ is the control input. a_1 is assumed to have strictly positive real part. In order to achieve asymptotic stabilization of the equilibrium at $A \equiv 0$, backstepping is applied resulting in a boundary control law that essentially cuts the term $a_3(\breve{x}) A(\breve{x},t)$ from (1). The result extends the work of [10, 14] in two ways: (1) it deals with two coupled partial differential equations, and (2) under certain circumstances handles equations defined on a semi-infinite domain ($x_d \to \infty$). The theory is supplemented with a case study involving control of vortex shedding in bluff

*Received by the editors April 10, 2003; accepted for publication (in revised form) August 12, 2004; published electronically April 14, 2005. This work was supported by National Science Foundation grant CMS-0329662 and the Norwegian Research Council.
http://www.siam.org/journals/sicon/43-6/42601.html

†Department of Engineering Cybernetics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway (aamo@itk.ntnu.no).

‡Department of Mechanical and Aerospace Engineering, University of California at San Diego, La Jolla, CA 92093-0411 (smysh@mechanics.ucsd.edu, krstic@ucsd.edu).

body flows. Controllers for this problem have previously been designed for finite-dimensional approximations of (1) [7, 8, 1, 9]. In [12, 13], it was shown numerically that the Ginzburg–Landau model for Reynolds numbers close to R_c can be stabilized using proportional feedback from a single measurement downstream of the cylinder to local forcing at the location of the cylinder. In [5], an optimal solution to a boundary control problem formulated for a stationary Ginzburg–Landau model of superconductivity defined on a bounded domain was shown to exist, and the optimality system of equations was solved by employing the finite element method.

The paper is organized as follows. In section 2, equation (1) is rewritten in terms of real variables and coefficients, and the problem statement is given. The main result is stated in section 3. In section 4 partial differential equations governing the feedback kernel are derived, and in section 5 they are transformed to corresponding integral equations. We find a unique solution to the integral equations in section 6, and show that the solution also yields the unique feedback kernel. Stability properties of the chosen target system are established in section 7. In section 9, the results are applied to a model of vortex shedding behind a bluff body immersed in a moving fluid, and it is shown that stabilizing feedback kernels that have compact support can be found even when the domain is semi-infinite. Concluding remarks are offered in section 10.

2. Problem statement. We now rewrite (1) to obtain two coupled partial differential equations in real variables and coefficients by defining

$$\rho(x,t) = \Re(B(x,t)) = \frac{1}{2}\left(B(x,t) + \bar{B}(x,t)\right), \tag{4}$$

$$\iota(x,t) = \Im(B(x,t)) = \frac{1}{2i}\left(B(x,t) - \bar{B}(x,t)\right), \tag{5}$$

where

$$x = \frac{x_d - \breve{x}}{x_d} \text{ and } B(x,t) = A(\breve{x},t)\exp\left(\frac{1}{2a_1}\int_0^{\breve{x}} a_2(\tau)\,d\tau\right). \tag{6}$$

i denotes the imaginary unit, and $\bar{}$ denotes complex conjugation. Equation (1) becomes

$$\rho_t = a_R\rho_{xx} + b_R(x)\rho - a_I\iota_{xx} - b_I(x)\iota, \tag{7}$$

$$\iota_t = a_I\rho_{xx} + b_I(x)\rho + a_R\iota_{xx} + b_R(x)\iota \tag{8}$$

for $x \in (0,1)$, with boundary conditions

$$\rho(0,t) = 0,\ \iota(0,t) = 0, \tag{9}$$

$$\rho(1,t) = u_R(t),\ \iota(1,t) = u_I(t), \tag{10}$$

and where

$$a_R \triangleq \frac{1}{x_d^2}\Re(a_1),\ a_I \triangleq \frac{1}{x_d^2}\Im(a_1), \tag{11}$$

$$b_R(x) \triangleq \Re\left(a_3(\breve{x}) - \frac{1}{2}a_2'(\breve{x}) - \frac{1}{4a_1}a_2^2(\breve{x})\right), \tag{12}$$

$$b_I(x) \triangleq \Im\left(a_3(\breve{x}) - \frac{1}{2}a_2'(\breve{x}) - \frac{1}{4a_1}a_2^2(\breve{x})\right). \tag{13}$$

Notice that transformation (6) serves two purposes: it normalizes the domain, and removes the convective term in (1). The control input will be in the form

$$u_R(t) = \int_0^1 \left[k(1,y)\,\rho(y,t) + k_c(1,y)\,\iota(y,t)\right] dy, \tag{14}$$

$$u_I(t) = \int_0^1 \left[-k_c(1,y)\,\rho(y,t) + k(1,y)\,\iota(y,t)\right] dy. \tag{15}$$

It is the objective of this paper to find stabilizing feedback gain kernels k and k_c.

3. Main result. Our main result states well posedness and stability properties of system (7)–(10) in closed loop with (14)–(15). The proof of the theorem is given in section 8, following intermediate results.

THEOREM 1 (main result). *There exist feedback gain kernels, $k(1,\cdot), k_c(1,\cdot) \in C^2(0,1)$, such that for arbitrary initial data $\rho_0, \iota_0 \in L_\infty(0,1)$, system (7)–(10) in closed loop with (14)–(15) has a unique classical solution $\rho, \iota \in C^{2,1}((0,1)\times(0,\infty))$. The solution satisfies*

$$\|(\rho,\iota)\|_{H_1} \le M \|(\rho_0,\iota_0)\|_{H_1} e^{-ct}, \tag{16}$$

where $M > 0$ and c is a prescribable positive constant.

The basic idea of the control design is to show that the dynamics of the *original* system (7)–(10), with an appropriate choice for (14)–(15), is equivalent to the dynamics of a *target* system that has a specified structure, but whose zero-solution can be assigned desired stability properties by the choice of coefficients. This is achieved by finding an invertible coordinate transformation that transforms solutions of the original system into solutions of the target system, establishing equivalence of norms of solutions of the two systems. The transformation is found as the unique solution to a hyperbolic partial differential equation, as stated in the next section. The proof of existence and uniqueness of solutions to this equation offers a constructive procedure to compute the transformation.

4. Derivation of PDE for the kernels. We want to find a coordinate transformation

$$\tilde\rho(x,t) = \rho(x,t) - \int_0^x \left[k(x,y)\,\rho(y,t) + k_c(x,y)\,\iota(y,t)\right] dy, \tag{17}$$

$$\tilde\iota(x,t) = \iota(x,t) - \int_0^x \left[-k_c(x,y)\,\rho(y,t) + k(x,y)\,\iota(y,t)\right] dy, \tag{18}$$

transforming (7)–(10) into the exponentially stable system (under appropriate conditions on $f_R(x)$ and $f_I(x)$ that are stated in section 7)

$$\tilde\rho_t = a_R\tilde\rho_{xx} + f_R(x)\,\tilde\rho - a_I\tilde\iota_{xx} - f_I(x)\,\tilde\iota, \tag{19}$$

$$\tilde\iota_t = a_I\tilde\rho_{xx} + f_I(x)\,\tilde\rho + a_R\tilde\iota_{xx} + f_R(x)\,\tilde\iota \tag{20}$$

for $x \in (0,1)$, with boundary conditions

$$\tilde\rho(0,t) = 0,\ \tilde\iota(0,t) = 0,\ \tilde\rho(1,t) = 0,\ \tilde\iota(1,t) = 0, \tag{21}$$

and where $f_R, f_I \in C^1([0,1])$. The skew-symmetric form of (17)–(18) is postulated from the skew-symmetric form of (7)–(8). Notice that once the kernels, $k(x,y)$ and

$k_c(x,y)$, have been found, setting $x=1$ in (17)–(18) and using (21) yields the boundary control law (10) with (14)–(15).

LEMMA 2. *If the pair of kernels, $k(x,y)$ and $k_c(x,y)$, satisfy the partial differential equation*

$$k_{xx} = k_{yy} + \beta(x,y)k + \beta_c(x,y)k_c, \tag{22}$$

$$k_{c,xx} = k_{c,yy} - \beta_c(x,y)k + \beta(x,y)k_c \tag{23}$$

for $(x,y) \in \mathcal{T} = \{x,y : 0 < y < x < 1\}$, with boundary conditions

$$k(x,x) = -\frac{1}{2}\int_0^x \beta(\gamma,\gamma)d\gamma, \tag{24}$$

$$k_c(x,x) = \frac{1}{2}\int_0^x \beta_c(\gamma,\gamma)d\gamma, \tag{25}$$

$$k(x,0) = 0, \tag{26}$$

$$k_c(x,0) = 0, \tag{27}$$

where

$$\beta(x,y) = \left[a_R\left(b_R(y) - f_R(x)\right) + a_I\left(b_I(y) - f_I(x)\right)\right] / \left(a_R^2 + a_I^2\right), \tag{28}$$

$$\beta_c(x,y) = \left[a_R\left(b_I(y) - f_I(x)\right) - a_I\left(b_R(y) - f_R(x)\right)\right] / \left(a_R^2 + a_I^2\right), \tag{29}$$

and if (ρ,ι) satisfies (7)–(10) with (14)–(15), then $(\tilde{\rho},\tilde{\iota})$ satisfies (19)–(21).

Proof. Differentiating (17) with respect to time and inserting (7)–(8) we have

$$\begin{aligned}\tilde{\rho}_t(x,t) = {} & a_R\rho_{xx} + b_R(x)\rho - a_I\iota_{xx} - b_I(x)\iota \\ & - \int_0^x \Big[k(x,y)\left(a_R\rho_{xx} + b_R(y)\rho - a_I\iota_{xx} - b_I(y)\iota\right) \\ & \qquad - k_c(x,y)\left(a_I\rho_{xx} + b_I(y)\rho + a_R\iota_{xx} + b_R(y)\iota\right)\Big]dy.\end{aligned} \tag{30}$$

Integrating (30) by parts, and using (9), (26)–(27), and (17)–(18) yield

$$\begin{aligned}\tilde{\rho}_t(x,t) = {} & a_R\tilde{\rho}_{xx}(x,t) + a_R\frac{\partial^2}{\partial x^2}\int_0^x \left[k(x,y)\rho(y,t) + k_c(x,y)\iota(y,t)\right]dy \\ & + b_R(x)\tilde{\rho}(x,t) + b_R(x)\int_0^x \left[k(x,y)\rho(y,t) + k_c(x,y)\iota(y,t)\right]dy \\ & - a_I\tilde{\iota}_{xx}(x,t) - a_I\frac{\partial^2}{\partial x^2}\int_0^x \left[-k_c(x,y)\rho(y,t) + k(x,y)\iota(y,t)\right]dy \\ & - b_I(x)\tilde{\iota}(x,t) - b_I(x)\int_0^x \left[-k_c(x,y)\rho(y,t) + k(x,y)\iota(y,t)\right]dy \\ & - k(x,x)a_R\rho_x(x,t) + k(x,x)a_I\iota_x(x,t) - k_c(x,x)a_I\rho_x(x,t) - k_c(x,x)a_R\iota_x(x,t) \\ & + k_y(x,x)a_R\rho(x,t) - k_y(x,x)a_I\iota(x,t) + k_{c,y}(x,x)a_I\rho(x,t) + k_{c,y}(x,x)a_R\iota(x,t) \\ & - \int_0^x \left[k_{yy}(x,y)\left(a_R\rho(y,t) - a_I\iota(y,t)\right) + k(x,y)\left(b_R(y)\rho(y,t) - b_I(y)\iota(y,t)\right)\right. \\ & \left. + \; k_{c,yy}(x,y)\left(a_I\rho(y,t) + a_R\iota(y,t)\right) + k(x,y)\left(b_I(y)\rho(y,t) + b_R(y)\iota(y,t)\right)\right]dy.\end{aligned} \tag{31}$$

Applying the relation

$$\begin{aligned}\frac{\partial^2}{\partial x^2}\int_0^x \kappa(x,y)\upsilon(y,t)\,dy = {} & \int_0^x \kappa_{xx}(x,y)\upsilon(y,t)\,dy + \kappa_x(x,x)\upsilon(x,t) \\ & + \; \upsilon(x,t)\frac{d\kappa(x,x)}{dx} + \kappa(x,x)\upsilon_x(x,t)\end{aligned} \tag{32}$$

to appropriate terms in (31), and using (17)–(18) again, we get

$$\begin{aligned}(33)\quad \tilde{\rho}_t(x,t) &= a_R\tilde{\rho}_{xx}(x,t) - a_I\tilde{\iota}_{xx}(x,t) + b_R(x)\,\tilde{\rho}(x,t) - b_I(x)\,\tilde{\iota}(x,t)\\ &\quad + 2a_R\frac{dk(x,x)}{dx}\tilde{\rho}(x,t) + 2a_R\frac{dk_c(x,x)}{dx}\tilde{\iota}(x,t) + 2a_I\frac{dk_c(x,x)}{dx}\tilde{\rho}(x,t)\\ &\quad - 2a_I\frac{dk(x,x)}{dx}\tilde{\iota}(x,t) + \int_0^x R(x,y)\,\rho(y,t)\,dy + \int_0^x I(x,y)\,\iota(y,t)\,dy,\end{aligned}$$

where

$$\begin{aligned}(34)\quad R(x,y) &= a_R\left(k_{xx}(x,y) - k_{yy}(x,y)\right) + a_I\left(k_{c,xx}(x,y) - k_{c,yy}(x,y)\right)\\ &\quad + \left(2a_R\frac{dk(x,x)}{dx} + 2a_I\frac{dk_c(x,x)}{dx} + b_R(x) - b_R(y)\right)k(x,y)\\ &\quad + \left(-2a_R\frac{dk_c(x,x)}{dx} + 2a_I\frac{dk(x,x)}{dx} + b_I(x) - b_I(y)\right)k_c(x,y)\end{aligned}$$

and

$$\begin{aligned}(35)\quad I(x,y) &= -a_I\left(k_{xx}(x,y) - k_{yy}(x,y)\right) + a_R\left(k_{c,xx}(x,y) - k_{c,yy}(x,y)\right)\\ &\quad + \left(2a_R\frac{dk_c(x,x)}{dx} - 2a_I\frac{dk(x,x)}{dx} - b_I(x) + b_I(y)\right)k(x,y)\\ &\quad + \left(2a_R\frac{dk(x,x)}{dx} + 2a_I\frac{dk_c(x,x)}{dx} + b_R(x) - b_R(y)\right)k_c(x,y).\end{aligned}$$

Substituting (22)–(23) and (24)–(25) into (34)–(35) yields

$$\begin{aligned}(36)\quad &R(x,y)\\ &= \left(a_R\beta(x,y) - a_I\beta_c(x,y) - \left(a_R\beta(x,x) - a_I\beta_c(x,x)\right) + b_R(x) - b_R(y)\right)k(x,y)\\ &+ \left(a_R\beta_c(x,y) + a_I\beta(x,y) - \left(a_R\beta_c(x,x) + a_I\beta(x,x)\right) + b_I(x) - b_I(y)\right)k_c(x,y),\end{aligned}$$

and

$$\begin{aligned}(37)\quad &I(x,y)\\ &= \left(-\left(a_R\beta_c(x,y) + a_I\beta(x,y)\right) + a_R\beta_c(x,x) + a_I\beta(x,x) - b_I(x) + b_I(y)\right)k(x,y)\\ &+ \left(a_R\beta(x,y) - a_I\beta_c(x,y) - \left(a_R\beta(x,x) - a_I\beta_c(x,x)\right) + b_R(x) - b_R(y)\right)k_c(x,y).\end{aligned}$$

From (28)–(29), we see that

$$(38)\qquad a_R\beta(x,y) - a_I\beta_c(x,y) = b_R(y) - f_R(x)$$

and

$$(39)\qquad a_R\beta_c(x,y) + a_I\beta(x,y) = b_I(y) - f_I(x),$$

so it follows that

$$(40)\qquad R(x,y) = I(x,y) \equiv 0.$$

In view of (40), and using (24)–(25), (33) becomes

$$\begin{aligned}(41)\quad \tilde{\rho}_t(x,t) =& a_R\tilde{\rho}_{xx}(x,t) - a_I\tilde{\iota}_{xx}(x,t)\\ &+ \left(-\left(a_R\beta(x,x) - a_I\beta_c(x,x)\right) + b_R(x)\right)\tilde{\rho}(x,t)\\ &+ \left(a_R\beta_c(x,x) + a_I\beta(x,x) - b_I(x)\right)\tilde{\iota}(x,t).\end{aligned}$$

Equation (19) now follows by substituting (38)–(39) into (41). The boundary conditions (21) follow by setting $x = 0$ and $x = 1$ in (17)–(18), and using (9)–(10) and (14)–(15). Equation (20) follows similarly by starting from the time derivative of (18). □

5. Converting the PDE into an integral equation. In the following lemma, (22)–(27) is converted into an integral equation, that is suitable for analysis by a fixed point method to establish existence and uniqueness of solutions.

LEMMA 3. *Any pair of kernels, $k(x,y)$ and $k_c(x,y)$, satisfying* (22)–(27)*, also satisfy the integral equation*

$$\begin{aligned}(42)\quad G(\xi,\eta) =& -\frac{1}{4}\int_\eta^\xi b(\tau,0)\,d\tau\\ &+ \frac{1}{4}\int_\eta^\xi\int_0^\eta b(\tau,s)\,G(\tau,s)\,ds d\tau + \frac{1}{4}\int_\eta^\xi\int_0^\eta b_c(\tau,s)\,G_c(\tau,s)\,ds d\tau,\end{aligned}$$

$$\begin{aligned}(43)\quad G_c(\xi,\eta) =& \frac{1}{4}\int_\eta^\xi b_c(\tau,0)\,d\tau\\ &- \frac{1}{4}\int_\eta^\xi\int_0^\eta b_c(\tau,s)\,G(\tau,s)\,ds d\tau + \frac{1}{4}\int_\eta^\xi\int_0^\eta b(\tau,s)\,G_c(\tau,s)\,ds d\tau,\end{aligned}$$

where

$$(44)\qquad \xi = x + y,\ \eta = x - y,$$

$$(45)\qquad G(\xi,\eta) = k\left(\frac{\xi+\eta}{2},\frac{\xi-\eta}{2}\right),\ G_c(\xi,\eta) = k_c\left(\frac{\xi+\eta}{2},\frac{\xi-\eta}{2}\right),$$

$$(46)\qquad b(\xi,\eta) = \beta\left(\frac{\xi+\eta}{2},\frac{\xi-\eta}{2}\right),\ b_c(\xi,\eta) = \beta_c\left(\frac{\xi+\eta}{2},\frac{\xi-\eta}{2}\right).$$

Proof. Using the relations

$$(47)\qquad k_x(x,y) = G_\xi\frac{\partial\xi}{\partial x} + G_\eta\frac{\partial\eta}{\partial x} = G_\xi(\xi,\eta) + G_\eta(\xi,\eta),$$

$$(48)\qquad k_y(x,y) = G_\xi\frac{\partial\xi}{\partial y} + G_\eta\frac{\partial\eta}{\partial y} = G_\xi(\xi,\eta) - G_\eta(\xi,\eta),$$

$$\begin{aligned}k_{xx}(x,y) &= \frac{\partial}{\partial\xi}\left(G_\xi(\xi,\eta) + G_\eta(\xi,\eta)\right)\frac{\partial\xi}{\partial x} + \frac{\partial}{\partial\eta}\left(G_\xi(\xi,\eta) + G_\eta(\xi,\eta)\right)\frac{\partial\eta}{\partial x}\\ &= G_{\xi\xi}(\xi,\eta) + 2G_{\eta\xi}(\xi,\eta) + G_{\eta\eta}(\xi,\eta),\end{aligned}\qquad(49)$$

$$\begin{aligned}k_{yy}(x,y) &= \frac{\partial}{\partial\xi}\left(G_\xi(\xi,\eta) - G_\eta(\xi,\eta)\right)\frac{\partial\xi}{\partial y} + \frac{\partial}{\partial\eta}\left(G_\xi(\xi,\eta) - G_\eta(\xi,\eta)\right)\frac{\partial\eta}{\partial y}\\ &= G_{\xi\xi}(\xi,\eta) - 2G_{\eta\xi}(\xi,\eta) + G_{\eta\eta}(\xi,\eta),\end{aligned}\qquad(50)$$

(22)–(23) with boundary conditions (24)–(27) are transformed to

$$G_{\eta\xi}(\xi,\eta)=\frac{1}{4}\left[b(\xi,\eta)G(\xi,\eta)+b_c(\xi,\eta)G_c(\xi,\eta)\right], \tag{51}$$

$$G_{c,\eta\xi}(\xi,\eta)=\frac{1}{4}\left[-b_c(\xi,\eta)G(\xi,\eta)+b(\xi,\eta)G_c(\xi,\eta)\right], \tag{52}$$

$$G(\xi,0)=-\frac{1}{4}\int_0^{\xi}b(\tau,0)\,d\tau, \tag{53}$$

$$G_c(\xi,0)=\frac{1}{4}\int_0^{\xi}b_c(\tau,0)\,d\tau, \tag{54}$$

$$G(\xi,\xi)=0, \tag{55}$$

$$G_c(\xi,\xi)=0. \tag{56}$$

Integrating (51) and (52) with respect to η from 0 to η, we obtain

$$\begin{aligned}G_\xi(\xi,\eta)-G_\xi(\xi,0)=&\frac{1}{4}\int_0^{\eta}b(\xi,s)G(\xi,s)\,ds\\&+\frac{1}{4}\int_0^{\eta}b_c(\xi,s)G_c(\xi,s)\,ds,\end{aligned} \tag{57}$$

$$\begin{aligned}G_{c,\xi}(\xi,\eta)-G_{c,\xi}(\xi,0)=&-\frac{1}{4}\int_0^{\eta}b_c(\xi,s)G(\xi,s)\,ds\\&+\frac{1}{4}\int_0^{\eta}b(\xi,s)G_c(\xi,s)\,ds.\end{aligned} \tag{58}$$

Integrating (57) and (58) with respect to ξ from η to ξ, and using (53)–(56) we obtain (42)–(43). □

6. Analysis of the integral equation.

THEOREM 4. *Equation* (22)–(23) *with boundary conditions* (24)–(27) *has a unique* $C^2(\overline{T})$ *solution satisfying*

$$|k(x,y)|\le Me^{2Mx}, \tag{59}$$

$$|k_c(x,y)|\le Me^{2Mx}, \tag{60}$$

where M *depends only on* a_1, $a_2(\cdot)$, $a_3(\cdot)$, $f_R(\cdot)$, $f_I(\cdot)$, *and is given in* (65).

Proof. Set

$$G_0(\xi,\eta)=-\frac{1}{4}\int_\eta^{\xi}b(\tau,0)\,d\tau, \tag{61}$$

$$G_{n+1}(\xi,\eta)=\frac{1}{4}\int_\eta^{\xi}\int_0^{\eta}b(\tau,s)G_n(\tau,s)\,ds d\tau+\frac{1}{4}\int_\eta^{\xi}\int_0^{\eta}b_c(\tau,s)G_{c,n}(\tau,s)\,ds d\tau, \tag{62}$$

$$G_{c,0}(\xi,\eta)=\frac{1}{4}\int_\eta^{\xi}b_c(\tau,0)\,d\tau, \tag{63}$$

and

$$G_{c,n+1}=-\frac{1}{4}\int_\eta^{\xi}\int_0^{\eta}b_c(\tau,s)G_n(\tau,s)\,ds d\tau+\frac{1}{4}\int_\eta^{\xi}\int_0^{\eta}b(\tau,s)G_{c,n}(\tau,s)\,ds d\tau. \tag{64}$$

Denote

$$B = \sup_{(\xi,\eta)\in\mathcal{T}_1} |b(\xi,\eta)|,\ B_c = \sup_{(\xi,\eta)\in\mathcal{T}_1} |b_c(\xi,\eta)|,\ M = \max\{B, B_c\}, \tag{65}$$

where $\mathcal{T}_1 \triangleq \{\xi,\eta : 0 < \xi < 2, 0 < \eta < \min(\xi, 2-\xi)\}$. For $G_0(\xi,\eta)$ and $G_{c,0}(\xi,\eta)$ we have

$$|G_0(\xi,\eta)| \le \frac{1}{4}\int_\eta^\xi |b(\tau,0)|\,d\tau \le \frac{1}{4}B(\xi-\eta) \le \frac{B}{2}, \tag{66}$$

$$|G_{c,0}(\xi,\eta)| \le \frac{1}{4}\int_\eta^\xi |b_c(\tau,0)|\,d\tau \le \frac{1}{4}B_c(\xi-\eta) \le \frac{B_c}{2}, \tag{67}$$

where we have used the fact that $0 < \xi - \eta < 2$. Suppose that

$$|G_n(\xi,\eta)| \le MK^n\frac{(\xi+\eta)^n}{n!}, \tag{68}$$

$$|G_{c,n}(\xi,\eta)| \le MK^n\frac{(\xi+\eta)^n}{n!}, \tag{69}$$

where $K > 0$ is a constant that will be determined later. Clearly, (68)–(69) hold for $n = 0$. Noting that

$$\begin{aligned}
\int_\eta^\xi\int_0^\eta |G_n(\tau,s)|\,ds d\tau &\le \frac{MK^n}{n!}\int_\eta^\xi\int_0^\eta (\tau+s)^n\,ds d\tau \\
&= \frac{MK^n}{(n+1)!}\int_\eta^\xi \left[(\tau+\eta)^{n+1} - \tau^{n+1}\right]d\tau \\
&\le \frac{MK^n}{(n+1)!}\int_\eta^\xi (\tau+\eta)^{n+1}\,d\tau \\
&\le \frac{MK^n}{(n+1)!}\int_\eta^\xi (\xi+\eta)^{n+1}\,d\tau \\
&\le 2MK^n\frac{(\xi+\eta)^{n+1}}{(n+1)!}, \qquad (70)
\end{aligned}$$

we obtain from (62) that

$$|G_{n+1}(\xi,\eta)| \le \frac{1}{2}M(B+B_c)K^n\frac{(\xi+\eta)^{n+1}}{(n+1)!}, \tag{71}$$

and from (64) that $|G_{c,n+1}(\xi,\eta)|$ satisfies the same bound (71). Therefore, setting $K = M$, we obtain

$$|G_{n+1}(\xi,\eta)| \le MK^{n+1}\frac{(\xi+\eta)^{n+1}}{(n+1)!}, \tag{72}$$

$$|G_{c,n+1}(\xi,\eta)| \le MK^{n+1}\frac{(\xi+\eta)^{n+1}}{(n+1)!}. \tag{73}$$

Thus, (68) and (69) are proved by induction, and the series

$$G(\xi,\eta) = \sum_{n=0}^\infty G_n(\xi,\eta) \text{ and } G_c(\xi,\eta) = \sum_{n=0}^\infty G_{c,n}(\xi,\eta) \tag{74}$$

converge uniformly in $\overline{\mathcal{T}}_1$, and is a solution of (42)–(43). G and G_c are $C^2\left(\overline{\mathcal{T}}_1\right)$ since b and b_c are $C^1\left(\overline{\mathcal{T}}_1\right)$. The bounds (59)–(60) follow from (68)–(69), (74), and the fact that $K = M$. It can be shown by the method of successive approximations that if $(G_1, G_{c,1})$ and $(G_2, G_{c,2})$ are two different solutions of (42)–(43), the resulting homogeneous integral equation for $(G, G_c) = (G_1 - G_2, G_{c,1} - G_{c,2})$ has a unique solution $(G, G_c) = 0$, which proves that the solution (74) is unique. We can check that (74) satisfies (51)–(56) by direct substitution. Equations (51)–(56) have a unique solution by Lemma 3. □

Exponential stability of the target system (19)–(21) in the L_2 and H_1 norms is proved in the next section. In order to be able to imply stability of the closed loop system (7)–(10) from that result, we need to establish equivalence of norms of (ρ, ι) and $(\tilde{\rho}, \tilde{\iota})$ in L_2 and H_1. This is done by proving that transformation (17)–(18) is invertible. The inverse transformation has the form

$$\rho(x,t) = \tilde{\rho}(x,t) - \int_0^x \left[l(x,y)\,\tilde{\rho}(y,t) + l_c(x,y)\,\tilde{\iota}(y,t)\right] dy, \tag{75}$$

$$\iota(x,t) = \tilde{\iota}(x,t) - \int_0^x \left[-l_c(x,y)\,\tilde{\rho}(y,t) + l(x,y)\,\tilde{\iota}(y,t)\right] dy. \tag{76}$$

The following result holds for the kernels $l(x,y)$ and $l_c(x,y)$ of transformation (75)–(76).

THEOREM 5. *If the pair of kernels, $l(x,y)$ and $l_c(x,y)$, satisfy the partial differential equation*

$$l_{xx} = l_{yy} - \beta(y,x)l - \beta_c(y,x)l_c, \tag{77}$$

$$l_{c,xx} = l_{c,yy} + \beta_c(y,x)l - \beta(y,x)l_c, \tag{78}$$

with boundary conditions

$$l(x,x) = \frac{1}{2}\int_0^x \beta(\gamma,\gamma)d\gamma, \tag{79}$$

$$l_c(x,x) = -\frac{1}{2}\int_0^x \beta_c(\gamma,\gamma)d\gamma, \tag{80}$$

$$l(x,0) = 0, \tag{81}$$

$$l_c(x,0) = 0, \tag{82}$$

and if $(\tilde{\rho}, \tilde{\iota})$ satisfies (19)–(21), *then (ρ, ι) satisfies* (7)–(10) *with* (14)–(15). *System* (77)–(82) *has a unique $C^2\left(\overline{\mathcal{T}}\right)$ solution satisfying*

$$|l(x,y)| \le Me^{2Mx}, \tag{83}$$

$$|l_c(x,y)| \le Me^{2Mx}, \tag{84}$$

where M is given in (65).

Proof. The proof is similar to those of Lemmas 2 and 3, and Theorem 4. □

7. Stability analysis.

THEOREM 6. *Suppose $c > 0$, and select $f_R(x)$ and $f_I(x)$ such that*

$$\sup_{x\in[0,1]} \left(f_R(x) + \frac{1}{2}\left|f_I'(x)\right|\right) \le -\frac{1}{2}c. \tag{85}$$

Then the solution $(\tilde{\rho},\tilde{\iota})\equiv(0,0)$ *of system* (19)–(21) *is exponentially stable in the* $L_2(0,1)$ *and* $H_1(0,1)$ *norms.*

COROLLARY 7. *Suppose* $c>0$*, and set* $f_R(x)=-c$ *and* $f_I(x)\equiv 0$*. Then the solution* $(\tilde{\rho},\tilde{\iota})\equiv(0,0)$ *of system* (19)–(21) *is exponentially stable in the* $L_2(0,1)$ *and* $H_1(0,1)$ *norms.*

Proof. Consider the function

$$E(t)=\frac{1}{2}\int_0^1\left(\tilde{\rho}(x,t)^2+\tilde{\iota}(x,t)^2\right)dx. \tag{86}$$

Its time derivative along solutions of system (19)–(21) is

$$\begin{aligned}\dot{E}(t)&=\int_0^1\left[\tilde{\rho}\left(a_R\tilde{\rho}_{xx}+f_R(x)\tilde{\rho}-a_I\tilde{\iota}_{xx}-f_I(x)\tilde{\iota}\right)\right.\\&\qquad\left.+\tilde{\iota}\left(a_I\tilde{\rho}_{xx}+f_I(x)\tilde{\rho}+a_R\tilde{\iota}_{xx}+f_R(x)\tilde{\iota}\right)\right]dx\\&=\int_0^1\left(\tilde{\rho}\left(a_R\tilde{\rho}_{xx}+f_R(x)\tilde{\rho}-a_I\tilde{\iota}_{xx}\right)+\tilde{\iota}\left(a_I\tilde{\rho}_{xx}+a_R\tilde{\iota}_{xx}+f_R(x)\tilde{\iota}\right)\right)dx\\&=-\int_0^1 a_R\left(\tilde{\rho}_x^2+\tilde{\iota}_x^2\right)dx+\int_0^1 f_R(x)\left(\tilde{\rho}^2+\tilde{\iota}^2\right)dx+a_I\int_0^1\left(\tilde{\rho}_x\tilde{\iota}_x-\tilde{\iota}_x\tilde{\rho}_x\right)dx\\&\le\int_0^1 f_R(x)\left(\tilde{\rho}^2+\tilde{\iota}^2\right)dx.\end{aligned} \tag{87}$$

So, from (85), and the comparison principle, we have

$$E(t)\le E(0)e^{-ct}\text{ for }t\ge 0. \tag{88}$$

Set

$$V(t)=\frac{1}{2}\int\left(\tilde{\rho}_x^2(x,t)+\tilde{\iota}_x^2(x,t)\right)dx. \tag{89}$$

The time derivative of $V(t)$ along solutions of system (19)–(21) is

$$\begin{aligned}\dot{V}(t)&=\int_0^1\left(\tilde{\rho}_x\tilde{\rho}_{xt}+\tilde{\iota}_x\tilde{\iota}_{xt}\right)dx\\&=-\int_0^1\left(\tilde{\rho}_{xx}\tilde{\rho}_t+\tilde{\iota}_{xx}\tilde{\iota}_t\right)dx\\&=-\int_0^1\left[\tilde{\rho}_{xx}\left(a_R\tilde{\rho}_{xx}+f_R(x)\tilde{\rho}-a_I\tilde{\iota}_{xx}-f_I(x)\tilde{\iota}\right)\right.\\&\qquad\left.+\tilde{\iota}_{xx}\left(a_I\tilde{\rho}_{xx}+f_I(x)\tilde{\rho}+a_R\tilde{\iota}_{xx}+f_R(x)\tilde{\iota}\right)\right]dx\\&=-a_R\int_0^1\left(\tilde{\rho}_{xx}^2+\tilde{\iota}_{xx}^2\right)dx+\int_0^1 f_R(x)\left(\tilde{\rho}_x^2+\tilde{\iota}_x^2\right)dx\\&\quad+\int_0^1 f_I'(x)\left(\tilde{\iota}_x\tilde{\rho}-\tilde{\rho}_x\tilde{\iota}\right)dx-\frac{1}{2}\int_0^1 f_R''(x)\left(\tilde{\rho}^2+\tilde{\iota}^2\right)dx\\&\le\int_0^1 f_R(x)\left(\tilde{\rho}_x^2+\tilde{\iota}_x^2\right)dx+\int_0^1 f_I'(x)\left(\tilde{\iota}_x\tilde{\rho}-\tilde{\rho}_x\tilde{\iota}\right)dx-\frac{1}{2}\int_0^1 f_R''(x)\left(\tilde{\rho}^2+\tilde{\iota}^2\right)dx\\&\le\int_0^1\left(f_R(x)+\frac{1}{2}\left|f_I'(x)\right|\right)\left(\tilde{\rho}_x^2+\tilde{\iota}_x^2\right)dx+\frac{1}{2}\int_0^1\left(\left|f_I'(x)\right|-f_R''(x)\right)\left(\tilde{\rho}^2+\tilde{\iota}^2\right)dx\\&\le\int_0^1\left(f_R(x)+\frac{1}{2}\left|f_I'(x)\right|\right)\left(\tilde{\rho}_x^2+\tilde{\iota}_x^2\right)dx+\frac{1}{2}c_2\int_0^1\left(\tilde{\rho}^2+\tilde{\iota}^2\right)dx,\end{aligned}$$

$$\le-\frac{c}{2}V(t)+c_2E(0)e^{-ct}, \tag{90}$$

where we have used (85) and defined

$$c_2 \triangleq \max\left\{\sup_{x\in[0,1]}\left(\left|f_I'(x)\right| - f_R''(x)\right), 0\right\}. \tag{91}$$

From the comparison principle, we get

$$V(t) \le \left(V(0) + 2\frac{c_2}{c}E(0)\right)e^{-\frac{c}{2}t} - 2\frac{c_2}{c}E(0)\,e^{-ct}, \tag{92}$$

so we obtain

$$V(t) \le \left(V(0) + \frac{2c_2}{c}E(0)\right)e^{-\frac{c}{2}t} \text{ for } t \ge 0. \tag{93}$$

Since (Poincaré inequality)

$$E(t) \le \frac{1}{2}V(t), \tag{94}$$

we get

$$V(t) \le c_3 V(0)\,e^{-\frac{c}{2}t} \text{ for } t \ge 0, \tag{95}$$

with $c_3 = 1 + c_2/c$. □

8. Proof of Theorem 1. From Theorem 6, $(\tilde{\rho}, \tilde{\iota}) = 0$ is exponentially stable in the L_2 and H_1 norms. Since Theorems 4 and 5 establish equivalence of norms of (ρ, ι) and $(\tilde{\rho}, \tilde{\iota})$ in L_2 and H_1, the stability statements of Theorem 6 also hold for the solution $(\rho, \iota) \equiv (0, 0)$ of system (7)–(8). From standard results for uniformly parabolic[1] equations (see, for instance, [4]), it follows that system (19)–(20), with Dirichlet boundary conditions (21) and initial data $\tilde{\rho}_0, \tilde{\iota}_0 \in L_\infty(0,1)$, has a unique classical solution $\tilde{\rho}, \tilde{\iota} \in C^{2,1}((0,1)\times(0,\infty))$. The smoothness properties of k, k_c, l, and l_c stated in Theorems 4 and 5 then provide well posedness of system (7)–(10) in closed loop with (14)–(15).

9. Application to a model of vortex shedding. The objective of this section is to provide a numerical demonstration of our results applied to a fluid flow control problem. An interesting feature of the system we study in this example, is that it is defined on an infinite domain ($x_d \to \infty$), yet, we obtain feedback gain kernels which have compact support.

9.1. The model. In flows past submerged obstacles, the phenomenon of vortex shedding occurs provided the Reynolds number is sufficiently large. A popular prototype model flow for studying vortex shedding, is the flow past a two-dimensional circular cylinder, as sketched in Figure 1. The vortices, which are alternatively shed from the upper and lower sides of the cylinder, induce an undesirable periodic force that acts on the cylinder. The dynamics of the cylinder wake, often referred to as the von Kármán vortex street, is governed by the Navier–Stokes equation, however, in [13], a simplified model was suggested in terms of the Ginzburg–Landau equation

$$\frac{\partial A}{\partial t} = a_1\frac{\partial^2 A}{\partial \breve{x}^2} + a_2(\breve{x})\frac{\partial A}{\partial \breve{x}} + a_3(\breve{x})A + a_4|A|^2 A + \delta(\breve{x})u, \tag{96}$$

[1]System (19)–(20) is uniformly parabolic in $(0,1)$, with module of parabolicity a_R.

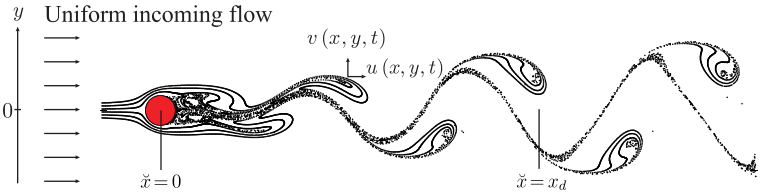


FIG. 1. *Vortex shedding from a cylinder visualized by passive tracer particles.*

where $\breve{x} \in \mathbb{R}$, $A : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{C}$, $a_1, a_4 \in \mathbb{C}$, and $a_2, a_3 : \mathbb{R} \to \mathbb{C}$. δ denotes the Dirac distribution and $u : \mathbb{R}_+ \to \mathbb{C}$ is the control input. Thus, actuation is in the form of local forcing at $\breve{x} = 0$, which is the location of the cylinder. The boundary conditions are $A(\pm\infty, t) = 0$, that is, homogeneous Dirichlet boundary conditions. $A(x, t)$ may represent any physical variable (velocities (u, v) or pressure p), or derivations thereof, along the centerline $y = 0$, see Figure 1. The choice will have an impact on the performance of the Ginzburg–Landau model, and associating A with the transverse fluctuating velocity $v(x, y = 0, t)$ seems to be a particularly good choice [11]. In order to implement the scheme in practice, transfer functions between $A(0)$ and the physical actuation, and the physical sensing and $A(x)$, would have to be determined, either experimentally or computationally. The physical actuation could for instance be micro/synthetic jet actuators distributed on the cylinder surface. Numerical values for the coefficients in (96) were determined from experiments in [13], and are reproduced in Appendix A.

We now simplify this problem to fit into the framework of the previous analysis. We linearize around the zero-solution, discard the upstream subsystem by replacing the local forcing at $\breve{x} = 0$ with boundary input at this location, and truncate the downstream subsystem at some $x_d > 0$. The resulting system is of the form (1)–(3), defined on the interval $[0, x_d]$. We justify the truncation of the system by noting that the upstream subsystem (the region to the left of the cylinder in Figure 1) is approximately uniform flow, whereas the downstream subsystem (the region to the right of the cylinder in Figure 1) can be approximated to any desired level of accuracy by selecting x_d sufficiently large. We are now in a position to apply our results, and we will do so for different choices of x_d.

In this numerical example, we set the Reynolds[2] number to $R = 50$, which corresponds to supercritical flow for which vortex shedding will occur in the uncontrolled case. For this choice of Reynolds number, the numerical coefficients of (7)–(8) derived from the coefficients given in Appendix A are $a_R = 0.156/x_d^2$, $a_I = 0$, and $b_R(x)$ and $b_I(x)$ are plotted in Figure 2 for $x_d = 2.5$, $x_d = 5$, and $x_d = 7.5$.

9.2. Feedback kernels. In terms of the feedback gain kernels, $k(1, x)$ and $k_c(1, x)$, the boundary feedback (2) is given by

$$u(t) = \int_0^{x_d} \frac{1}{x_d} \left(k\left(1, \frac{x_d - \breve{x}}{x_d}\right) - ik_c\left(1, \frac{x_d - \breve{x}}{x_d}\right)\right) \times \exp\left(\frac{1}{2a_1}\int_0^{\breve{x}} a_2(\tau)\, d\tau\right) A(\breve{x}, t)\, d\breve{x}. \tag{97}$$

[2]The Reynolds number for flow past a circular cylinder is usually defined as $R = \rho U_\infty D/\mu$, where U_∞ is the free stream velocity, D is the cylinder diameter, and ρ and μ are density and viscosity of the fluid, respectively. Vortex shedding occurs when $R > 47$.

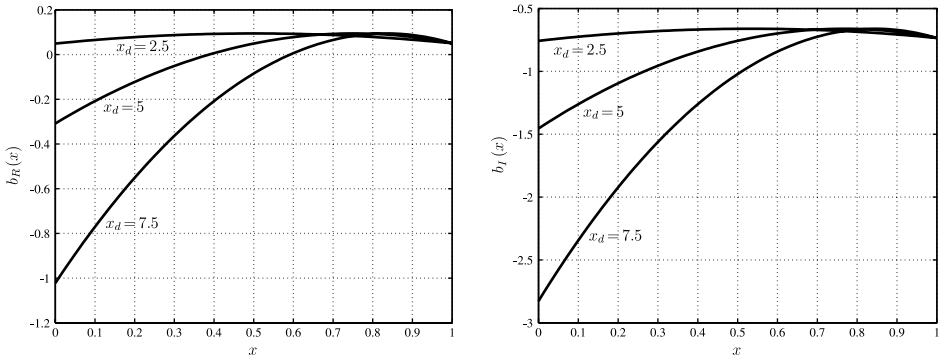


FIG. 2. $b_R(x)$ *and* $b_I(x)$ *for* $x_d = 2.5$, $x_d = 5$, *and* $x_d = 7.5$.

Thus, the feedback gain kernel for the original system (1)–(3) is complex-valued, and given by

$$k_u(\breve{x}) = \frac{1}{x_d}\left(k\left(1, \frac{x_d - \breve{x}}{x_d}\right) - ik_c\left(1, \frac{x_d - \breve{x}}{x_d}\right)\right)\exp\left(\frac{1}{2a_1}\int_0^{\breve{x}} a_2(\tau)\, d\tau\right). \tag{98}$$

Setting $f_R(x) = -0.2$ and $f_I(x) = 0$, exponential stability is assured by Corollary 7, and the stabilizing feedback gain kernel (98) can be calculated numerically using formulas (61)–(64), (74), (45)–(46), and (98). Figure 3 shows the feedback gain kernel (98) for $x_d = 2.5$, $x_d = 5$, and $x_d = 7.5$. It is clear that the feedback gain kernels grow rather rapidly with increasing x_d, which is an undesirable feature since we want to make x_d large in order to minimize the effect of truncating the downstream subsystem. The increase can be seen in connection with Figure 2, which shows that the absolute values of the differences $b_R(x) - f_R(x)$ and $b_I(x) - f_I(x)$ increase with increasing x_d. In other words, the control effort needed to change the dynamics of system (7)–(10) into that of (19)–(21) increases with the degree to which the two systems differ. Therefore, the functions $f_R(x)$ and $f_I(x)$ must be chosen more intelligently than the simple case of setting them constant. Theorem 6 allows some flexibility in choosing $f_R(x)$ and $f_I(x)$ within the constraints of (85). In order to postpone choosing x_d, we study f_R, f_I, b_R, and b_I as functions of $\breve{x}$ rather than x in the following. This is convenient since f_R, f_I, b_R, and b_I are invariant of x_d when treated as functions of $\breve{x}$. Recall that when x_d is chosen, the two domains are related by $x = (x_d - \breve{x})/x_d$. We propose to choose $f_R(\breve{x})$ and $f_I(\breve{x})$ as close to $b_R(\breve{x})$ and $b_I(\breve{x})$ as possible, without violating the conditions of Theorem 6, which we now write

$$\sup_{\breve{x}}\left(f_R(\breve{x}) + \frac{1}{2}\left|f_I'(\breve{x})\right|\right) \le -\frac{1}{2}c. \tag{99}$$

Towards that end, we first set them equal, that is $f_R(\breve{x}) = b_R(\breve{x})$ and $f_I(\breve{x}) = b_I(\breve{x})$, and plot (99) along with $-\frac{1}{2}c = -0.2$. The result is shown in Figure 4, for $\breve{x} \in [0, 20]$. The figure shows that the conditions for stability are already satisfied, without control, for $\breve{x} \in [x_s, 20]$ (in fact, the stability conditions are satisfied for $\breve{x} \in [x_s, \infty)$), which means that it suffices to alter $f_R(\breve{x})$ and $f_I(\breve{x})$ in $[0, x_s)$ in order to satisfy (99). Thus,

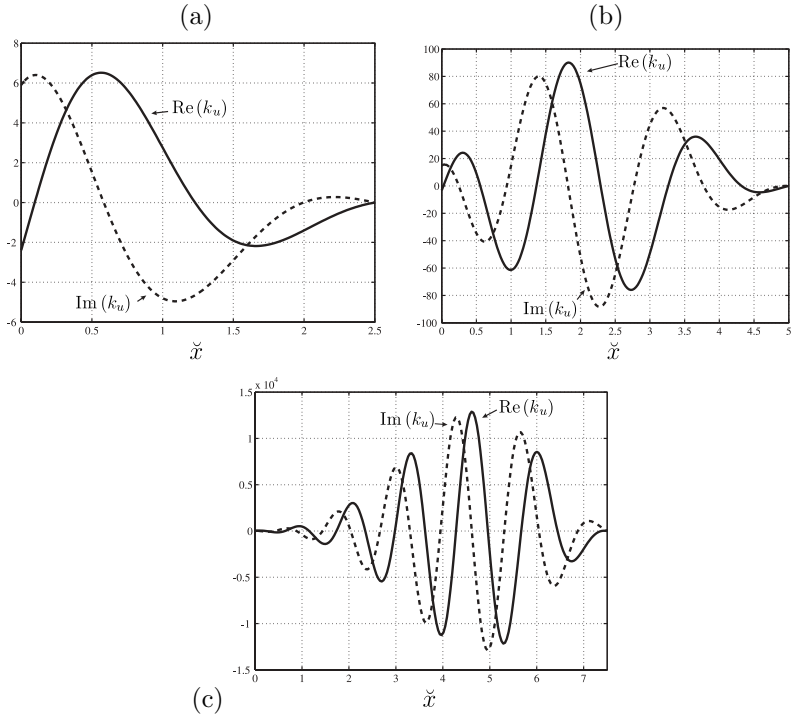


FIG. 3. *Feedback kernel* (98) *for* (a) $x_d = 2.5$, (b) $x_d = 5$, *and* (c) $x_d = 7,5$. $f_R(x) = -0.2$, *and* $f_I(x) = 0$.

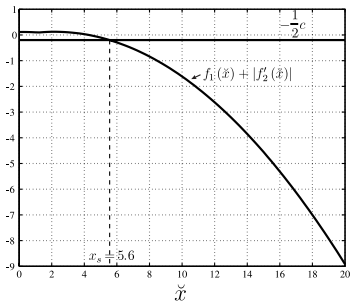


FIG. 4. *The stability criterion* (99), *when* $f_R(\breve{x}) = b_R(\breve{x})$ *and* $f_I(\breve{x}) = b_I(\breve{x})$, *is satisfied for* $\breve{x} \geq x_s$.

we set[3]

$$f_R(\breve{x}) = \begin{cases} -\frac{1}{2}c - \frac{1}{2}\left|b_I'(\breve{x})\right| & \text{for } 0 \leq x < x_s, \\ b_R(\breve{x}) & \text{for } \breve{x} \geq x_s, \end{cases} \tag{100}$$

$$f_I(\breve{x}) = b_I(\breve{x}) \text{ for all } \breve{x}. \tag{101}$$

With these choices of $f_R(\breve{x})$ and $f_I(\breve{x})$, we calculate numerically the stabilizing

[3]By the choice of x_s, $f_R(\breve{x})$ is continuous. In this example, we ignore the fact that our choice of $f_R(\breve{x})$ may not be C^1, although this can easily be achieved by smoothing $f_R(\breve{x})$ in a small neighborhood of x_s.

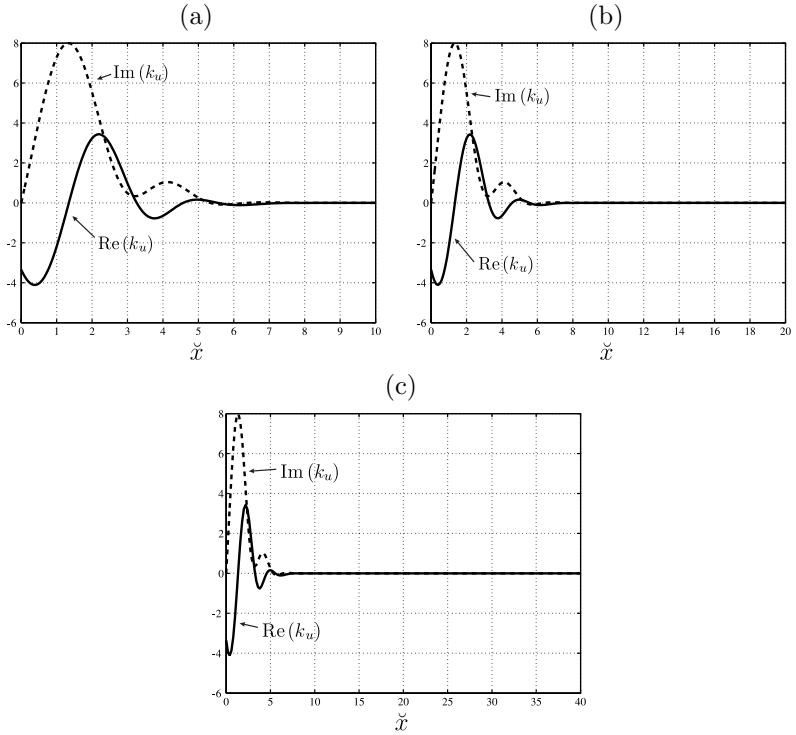


FIG. 5. *Feedback kernels* (98) *for* $f_R(\breve{x})$ *and* $f_I(\breve{x})$ *as defined in* (100)–(101)*, and for* $x_d = 10$ (a), 20 (b), *and* 40 (c).

feedback gain kernel (98) for $x_d = 10$, $x_d = 20$, and $x_d = 40$. Figure 5 shows the result. As expected, the feedback gain kernels look similar, and in particular, they appear to be zero for $\breve{x}$ larger than approximately 7.5. In fact, they are identical and have compact support, as stated formally in the next theorem.

THEOREM 8. *Given* $c > 0$*, suppose there exists* $x_s \in (0, \infty)$ *such that*

$$b_R(\breve{x}) + \frac{1}{2}\left|b_I'(\breve{x})\right| \leq -\frac{1}{2}c \quad \text{for } \breve{x} \geq x_s. \tag{102}$$

Then $f_R(\breve{x})$ *and* $f_I(\breve{x})$*, satisfying* (99)*, can be chosen such that* $f_R(\breve{x}) = b_R(\breve{x})$ *and* $f_I(\breve{x}) = b_I(\breve{x})$ *for* $\breve{x} \in [x_s, x_d]$*. The resulting stabilizing feedback gain kernel* (98) *has compact support contained in* $[0, 2x_s]$*. Moreover, all choices of* $x_d \geq 2x_s$ *will produce the same stabilizing feedback gain kernel* (98) *in* $[0, 2x_s]$*.*

Proof. The existence of $f_R(\breve{x})$ and $f_I(\breve{x})$ satisfying the criterion for stability (99) follows trivially from (102). To prove that the kernel has support contained in $[0, 2x_s]$, we show that it is identically zero outside this interval. We have that

$$b(\xi, 0) = \beta(x, x) = 0, \tag{103}$$

$$b_c(\xi, 0) = \beta_c(x, x) = 0 \tag{104}$$

for

$$\xi \in \left[0, 2\left(1 - \frac{x_s}{x_d}\right)\right). \tag{105}$$

It follows that

$$G_0(\xi,\eta)=0,\ G_{c,0}(\xi,\eta)=0 \text{ for } (\xi,\eta)\in\mathcal{T}_1,\ \xi\le 2\left(1-\frac{x_s}{x_d}\right). \tag{106}$$

Now, suppose that

$$G_n(\xi,\eta)=0,\ G_{c,n}(\xi,\eta)=0 \text{ for } (\xi,\eta)\in\mathcal{T}_1,\ \xi\le 2\left(1-\frac{x_s}{x_d}\right). \tag{107}$$

From (62) and (64), we get

$$G_{n+1}(\xi,\eta)=0\ G_{c,n+1}(\xi,\eta)=0,\ \text{for } (\xi,\eta)\in\mathcal{T}_1,\ \xi\le 2\left(1-\frac{x_s}{x_d}\right). \tag{108}$$

Thus, (107) is proved by induction, and

$$\left.\begin{array}{c} k(1,y)=G(1+y,1-y)=0 \\ k_c(1,y)=G_c(1+y,1-y)=0 \end{array}\right\} \text{ for } 0\le y\le 2\left(1-\frac{x_s}{x_d}\right)-1. \tag{109}$$

Therefore, $k_u(\breve{x})=0$ for

$$0\le\frac{x_d-\breve{x}}{x_d}\le 2\left(1-\frac{x_s}{x_d}\right)-1, \tag{110}$$

which yields

$$x_d\ge\breve{x}\ge 2x_s. \tag{111}$$

In order to prove the last part of the theorem, we need to show that for any $x_{d,1},x_{d,2}\in[2x_s,\infty)$,

$$\left.\begin{array}{c} \frac{1}{x_{d,1}}k_{x_{d,1}}\left(1,\frac{x_{d,1}-\breve{x}}{x_{d,1}}\right)=\frac{1}{x_{d,2}}k_{x_{d,2}}\left(1,\frac{x_{d,2}-\breve{x}}{x_{d,2}}\right) \\ \frac{1}{x_{d,1}}k_{c,x_{d,1}}\left(1,\frac{x_{d,1}-\breve{x}}{x_{d,1}}\right)=\frac{1}{x_{d,2}}k_{c,x_{d,2}}\left(1,\frac{x_{d,2}-\breve{x}}{x_{d,2}}\right) \end{array}\right\} \text{ for } \breve{x}\in[0,2x_s], \tag{112}$$

where the additional subscripts, $x_{d,1}$ and $x_{d,2}$, on variables identify the domain of the problem from which they stem. We have

$$k_{x_d}\left(1,\frac{x_d-\breve{x}}{x_d}\right)=G_{x_d}\left(2-\frac{1}{x_d}\breve{x},\frac{1}{x_d}\breve{x}\right). \tag{113}$$

From (103)–(105) it follows that

$$x_dG_{0,x_d}\left(2-\frac{1}{x_d}\breve{x},\frac{1}{x_d}\breve{x}\right)=-\frac{1}{4}\int_{\breve{x}}^{2x_s}b_{x_d}\left(2-\frac{1}{x_d}\tau,0\right)d\tau \tag{114}$$

for $x_d\ge 2x_s$. From the definition of b, we have that

$$x_{d,2}^2b_{x_{d,1}}\left(2-\frac{1}{x_{d,1}}\tau,0\right)=x_{d,1}^2b_{x_{d,2}}\left(2-\frac{1}{x_{d,2}}\tau,0\right) \tag{115}$$

for $\tau\in[0,2x_s]$. It follows from (114)–(115) that

$$x_{d,2}G_{0,x_{d,1}}\left(2-\frac{1}{x_{d,1}}\breve{x},\frac{1}{x_{d,1}}\breve{x}\right)=x_{d,1}G_{0,x_{d,2}}\left(2-\frac{1}{x_{d,2}}\breve{x},\frac{1}{x_{d,2}}\breve{x}\right). \tag{116}$$

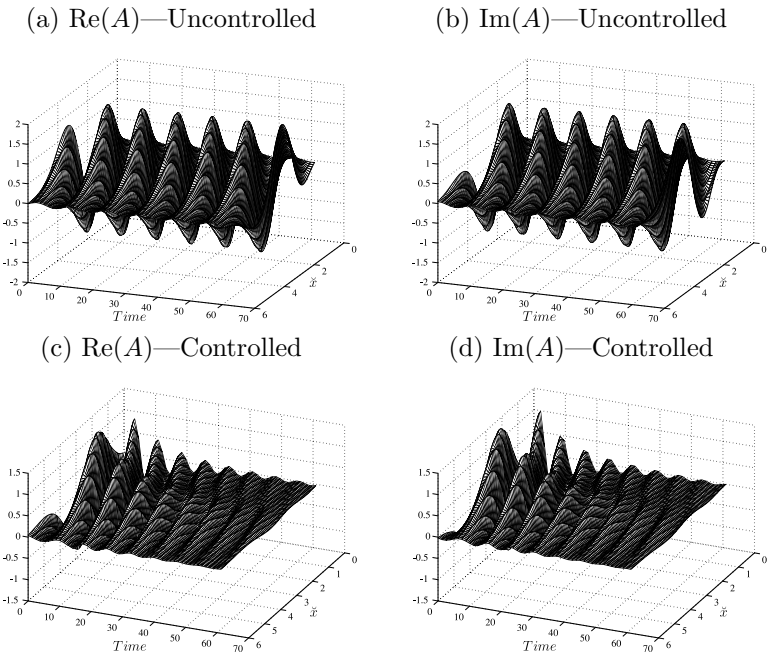


FIG. 6. *Comparison of the uncontrolled and controlled cases. For clarity, only a part of the computational domain is shown.*

Similar arguments for $G_{c,0}$, G_n, and $G_{c,n}$ yield

$$x_{d,2}G_{x_{d,1}}\left(2-\frac{1}{x_{d,1}}\breve{x},\frac{1}{x_{d,1}}\breve{x}\right)=x_{d,1}G_{x_{d,2}}\left(2-\frac{1}{x_{d,2}}\breve{x},\frac{1}{x_{d,2}}\breve{x}\right), \tag{117}$$

which in turn gives (112). □

The significance of Theorem 8 is that it guarantees stabilization of the system evolving on an infinite domain by solving the stabilization problem on a finite domain. The procedure for verifying the conditions of the theorem was demonstrated above, but for clarity we repeat it in the following remark.

Remark 9. The key to applying Theorem 8 is being able to find an x_s that satisfies (102). This is most easily done by inspecting a graph as the one shown in Figure 4. Once x_s is found, a possible choice of $f_R(\breve{x})$ and $f_I(\breve{x})$ is given in (100)–(101). Other choices are possible, and in particular, care should be taken to ensure necessary smoothness properties of $f_R(\breve{x})$. Also, note that the estimate for the support of (98) is not tight, as suggested by Figures 4 and 5, which indicate that $k_u(\breve{x})$ is supported on approximately $[0, 7.5]$ while $2x_s = 11.2$. Theorem 8 states that $[0, 7.5] \subseteq [0, 2x_s]$, which is true.

9.3. Numerical simulations. For completeness, we include numerical simulations of the controlled and uncontrolled systems. The simulations have been performed by discretizing (1) on the domain $\breve{x} \in [0, 15]$, using finite differences on a grid of 200 nodes. To make the simulation study more interesting, the nonlinear term in (96) is accounted for in the simulations. Figures 6(a) and 6(b) show the real and imaginary parts of $A(\breve{x}, t)$ for the uncontrolled case. The system is in a periodic state reminiscent of vortex shedding. Figures 6(c) and 6(d) show the real and imaginary parts of

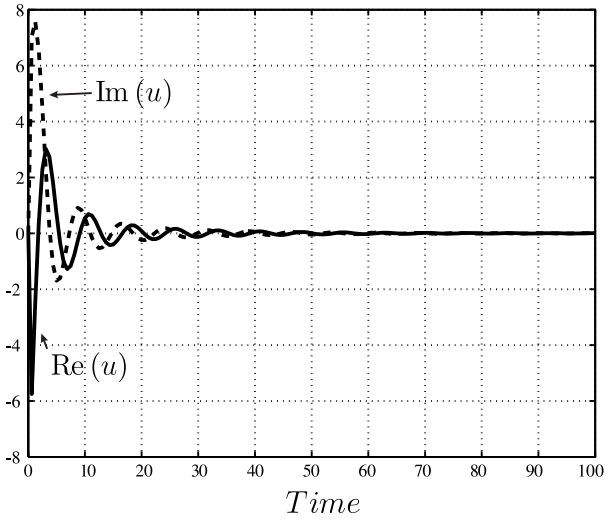


FIG. 7. *Control effort,* $u(t)$.

$A(\breve{x},t)$ for the controlled case. The figures show that $A(\breve{x},t)$ is effectively driven to zero by the control. Figure 7 shows the control effort.

10. Conclusions. This paper extends previous work in two ways: (1) it deals with two coupled partial differential equations, and (2) under certain circumstances handles equations defined on a semi-infinite domain. For the linearized Ginzburg–Landau equation, asymptotic stabilization is achieved by means of boundary control via state feedback in the form of an integral operator. The kernel of the operator is shown to be twice continuously differentiable, and a series approximation for its solution is given. Under certain conditions (given in (102)) on the parameters of the Ginzburg–Landau equation, compatible with vortex shedding modelling on a semi-infinite domain, the kernel is shown to have compact support, resulting in partial state feedback. Simulations are provided in order to demonstrate the performance of the controller.

Appendix A. Coefficients for the Ginzburg–Landau equation. The numerical coefficients below are taken from [13, Appendix A]

$$
\begin{aligned}
R_c &= 47,\\
x^t &= 1.183 - 0.031i,\\
\omega_0^t &= 0.690 + 0.080i + (-0.00159 + 0.00447i)(R - R_c),\\
k_0^t &= 1.452 - 0.844i + (0.00341 + 0.011i)(R - R_c),\\
\omega_{kk}^t &= -0.292i,\\
\omega_{xx}^t &= 0.108 - 0.057i,\\
k_x^t &= 0.164 - 0.006i,\\
\omega_0(\breve{x}) &= \omega_0^t + \frac{1}{2}\omega_{xx}^t\left(\breve{x} - x^t\right)^2,\\
k_0(\breve{x}) &= k_0^t + k_x^t\left(\breve{x} - x^t\right),
\end{aligned}
$$

$$
\begin{aligned}
a_1 &= \frac{1}{2} i \omega_{kk}^t, \\
a_2(\breve{x}) &= \omega_{kk}^t k_0(\breve{x}), \\
a_3(\breve{x}) &= -\left(\omega_0(\breve{x}) + \frac{1}{2}\omega_{kk}^t k_0^2(\breve{x})\right) i, \\
a_4 &= -0.0225 + 0.0671 i.
\end{aligned}
$$

Acknowledgment. We thank Professor Peter Monkewitz for helpful explanations on relationships between Navier–Stokes and Ginzburg–Landau models of vortex shedding and on implementability of GL-based controllers on NS simulations or experiments.

REFERENCES

[1] O. M. Aamo and M. Krstić, *Global stabilization of a nonlinear Ginzburg–Landau model of vortex shedding*, Eur. J. Control, 10 (2004), pp. 105–116.

[2] A. Balogh and M. Krstić, *Infinite dimensional backstepping-style feedback transformations for a heat equation with an arbitrary level of instability*, Eur. J. Control, 8 (2002), pp. 165–176.

[3] D. M. Bosković, M. Krstić, and W. J. Liu, *Boundary control of an unstable heat equation via measurement of domain-averaged temperature*, IEEE Trans. Automat. Control, 46 (2001), pp. 2022–2028.

[4] A. Friedman, *Partial Differential Equations of Parabolic Type*, Robert E. Krieger, 1983.

[5] M. D. Gunzburger, L. S. Hou, and S. S. Ravindran, *Analysis and approximation of optimal control problems for a simplified Ginzburg–Landau model of superconductivity*, Numer. Math., 77 (1997), pp. 243–268.

[6] M. Krstić, I. Kanellakopoulos, and P. Kokotović, *Nonlinear and Adaptive Control Design*, Wiley, New York, 1995.

[7] E. Lauga and T. R. Bewley, *H_∞ control of linear global instability in models of non-parallel wakes*, in Proceedings of the Second International Symposium on Turbulence and Shear Flow Phenomena, Stockholm, Sweden, 2001.

[8] E. Lauga and T. R. Bewley, *The decay of stabilizability with Reynolds number in a linear model of spatially developing flows*, R. Soc. Proc. Ser. A Math. Phys. Eng. Sci., 459 (2003), pp. 2077–2095.

[9] E. Lauga and T. R. Bewley, *Performance of a linear robust control strategy on a nonlinear model of spatially-developing flows*, J. Fluid Mech., 512 (2004), pp. 343–374.

[10] W. Liu, *Boundary feedback stabilization of an unstable heat equation*, SIAM J. Control Optim., 42 (2003), pp. 1033–1043.

[11] P. A. Monkewitz, *private communication*.

[12] D. S. Park, D. M. Ladd, and E. W. Hendricks, *Feedback control of a global mode in spatially developing flows*, Phys. Lett. A, 182 (1993), pp. 244–248.

[13] K. Roussopoulos and P. A. Monkewitz, *Nonlinear modelling of vortex shedding control in cylinder wakes*, Phys. D, 97 (1996), pp. 264–273.

[14] A. Smyshlyaev and M. Krstić, *Closed-form boundary state feedbacks for a class of 1-d partial integro-differential equations*, IEEE Trans. Automat. Control, 49 (2004), pp. 2185–2202.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 1972–1986

MULTIRATE PERIODIC SYSTEMS AND CONSTRAINED ANALYTIC FUNCTION INTERPOLATION PROBLEMS*

LI CHAI† AND LI QIU‡

Abstract. Multirate periodic systems and some related constrained analytic function interpolation problems are studied in this paper. After showing how to convert a general multirate periodic system to an equivalent linear time invariant (LTI) system with a structural constraint, we formulate some analytic function interpolation problems with such a constraint that can find various applications in the study of multirate and periodic systems. Both the solvability conditions and characterization of all solutions are presented to these constrained interpolation problems.

Key words. multirate systems, periodic systems, Nevanlinna–Pick interpolation

AMS subject classifications. 93C55, 93C35, 30E10, 91A28, 41A05

DOI. 10.1137/S0363012903430056

1. Introduction.

Periodic and multirate systems are finding more and more applications in control, signal processing, communication, econometrics, and numerical mathematics. There are several reasons for this.

• In signal processing, the use of periodic and multirate systems can often lead to the reduction of the required transmission rate, storage space, or computational complexity for a given task, depending on the application [24].

• In large-scale multivariable digital systems, often it is unrealistic, or sometimes impossible, to sample all physical signals uniformly at one single rate. In such situations, one is forced to use multirate sampling.

• Periodic and multirate systems can often achieve objectives that cannot be achieved by single rate systems. Examples include gain margin improvement and simultaneous stabilization [14].

The study of periodic systems can be traced back to [10]. Examples of more recent studies are [2, 15]. The study of multirate systems goes back to the late 1950s. A renewal of research on multirate systems has occurred since 1980 within the signal processing, communication, and control communities. The driving force for studying multirate systems in signal processing comes from the need for sampling rate conversion, subband coding, and their ability to generate wavelets. Multirate signal processing is now one of the most vibrant areas of research within the signal processing community; see the recent book [24] and references therein. In the communication community, blind identification and equalization call for the use of multirate sampling [23]. In the control community, two groups of research stand out: (i) using multirate control to achieve something that otherwise cannot be achieved by single rate control (see, for example, [14]) and (ii) optimal design of multirate controllers [7, 18].

*Received by the editors June 16, 2003; accepted for publication (in revised form) June 4, 2004; published electronically April 14, 2005. This work was supported in part by the Hong Kong Research Grants Council under grants HKUST6054/99E and HKUST6171/02E and in part by the National Natural Science Foundation of China under grant 60304011.

http://www.siam.org/journals/sicon/43-6/43005.html

†Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China. Current address: Department of Automation, Hangzhou Dianzi University, Hangzhou, 310018, China (lchai@hziee.edu.cn).

‡Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China (eeqiu@ee.ust.hk).

A standard technique for treating periodic and multirate systems is called lifting in control theory [7, 14] and blocking in signal processing [15, 24]. First, we establish a setup of multirate periodic (MP) systems, which cover many familiar systems as special cases. Using the technique of lifting, an MP system can be converted to an equivalent linear time invariant (LTI) system satisfying a causality constraint that requires the feedthrough term to be block lower triangular. Motivated by this fact we propose and then study the problem of analytic function interpolation with an additional constraint that requires the value of the interpolating function at the origin to be block lower triangular [4, 5]. These constrained analytic function interpolation problems play the same role for multirate systems as their unconstrained counterparts do for single rate systems.

Analytic function interpolation problems have a very rich history in mathematics, and there has been a large volume of literature on this subject; see the recent books [1, 11, 13]. Many successful approaches have been proposed to solve the analytic function interpolation problems since the theory was first proposed at the beginning of the last century. In particular, Sarason [21] encompassed different classical interpolation problems in a representation theorem of operators commuting with special contractions, which was later developed to a general framework on commutant lifting theorem [11, 22]. On the other hand, using the realization method from the system theory, Ball, Gohberg, and Rodman [1] present another systematic way to deal with the interpolation of rational matrix functions. Recently, Foias et al. [12] combined the commutant lifting theorem from operator theory and state-space method from system theory to provide a unified approach for a more general setup of the problems, where they used the concept of operator-valued functions with operator arguments.

The increasing research interest on analytic function interpolation theory is also partly due to its wide applications in a variety of engineering problems such as those in control, circuit theory, and digital filter design [4, 8, 13]. The Nevanlinna–Pick (NP) interpolation theory was first brought into system theory by Youla and Saito, who gave a circuit theoretical proof of the Pick criterion [28]. In the early stage of the development of $\mathcal{H}_\infty$ control theory, the analytic function interpolation theory played a fundamental role [25]. A detailed review of this connection can be found in [13]. Recently, some new methods in high-resolution spectral estimation have been presented based on the NP interpolation with degree constraints [3]. The NP interpolation and Carathéodory–Fejér (CF) interpolation problems are also used extensively in robust model validation and identification [5, 6, 16].

In this paper, we propose a general model of multirate and periodic systems which covers single rate periodic systems and many other multirate systems in the literature as special cases. We also propose and solve some constrained analytic function interpolation problems that play the same role in multirate and periodic systems as the unconstrained counterparts do in single rate systems. That is, our results can be applied directly to multirate and periodic systems for $\mathcal{H}_\infty$ control, robust model validation and identification, etc. We present the necessary and sufficient solvability conditions and the parametrization of all solutions explicitly. The interpolation and distance problems involving analytic function with such structural constraints were first discussed in [13], but explicit solutions to the problem considered in this paper were not given there.

The notation used in this paper is standard. The real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively. The open unit disc of the complex plane is denoted

by $\mathbb{D}$. Let

$$p = \begin{bmatrix} p_1 & \cdots & p_l \end{bmatrix} \text{ and } q = \begin{bmatrix} q_1 & \cdots & q_l \end{bmatrix} \tag{1.1}$$

be two vectors, where p_i and q_i, $i = 1, \ldots, l$, are nonnegative integers. Denote

$$|p| = \sum_{i=1}^{l} p_i \quad \text{and} \quad |q| = \sum_{i=1}^{l} q_i. \tag{1.2}$$

For $k = 0, \ldots, l$, define

$$\Pi_k(p) = \operatorname{diag}(0_{p_1}, \ldots, 0_{p_k}, I_{p_{k+1}}, \ldots, I_{p_l}), \tag{1.3}$$

$$\Pi_k(q) = \operatorname{diag}(0_{q_1}, \ldots, 0_{q_k}, I_{q_{k+1}}, \ldots, I_{q_l}), \tag{1.4}$$

where 0_n denote the $n \times n$ zero matrix and I_n the $n \times n$ unit matrix. Note that $\Pi_0(p) = I_{|p|}$, $\Pi_l(p) = 0_{|p|}$, $\Pi_0(q) = I_{|q|}$, and $\Pi_l(q) = 0_{|q|}$. The set of $|q| \times |p|$ matrices is denoted by $\mathbb{C}^{|q|\times|p|}$, and every such matrix is assumed to have an underlining partition so that its ijth block is q_i by p_j. Hence we have

$$\mathbb{C}^{|q|\times|p|} := \left\{ \begin{bmatrix} M_{11} & \cdots & M_{1l} \\ \vdots & \ddots & \vdots \\ M_{l1} & \cdots & M_{ll} \end{bmatrix} : M_{ij} \in \mathbb{C}^{q_i \times p_j} \right\}.$$

Note that the entry M_{ij} is "empty" if $q_i = 0$ or $p_j = 0$. The block lower triangular subset of $\mathbb{C}^{|q|\times|p|}$, denoted by $\mathcal{T}(q,p)$, consists of all matrices with $M_{ij} = 0, i < j$, and the strictly block lower triangular subset, $\mathcal{T}_s(q,p)$, consists of all matrices with $M_{ij} = 0, i \le j$. Let $\mathcal{H}_\infty^{|q|\times|p|}$ denote the Hardy class of all uniformly bounded analytic functions on $\mathbb{D}$ with values in $\mathbb{C}^{|q|\times|p|}$. For any $G(\lambda) \in \mathcal{H}_\infty^{|q|\times|p|}$, there exist $G_0, G_1, \ldots \in \mathbb{C}^{|q|\times|p|}$ such that $G(\lambda) = \sum_{m=0}^{\infty} \lambda^m G_m$ for $\lambda \in \mathbb{D}$.

2. Multirate periodic systems. To introduce the general setup of multirate periodic systems, we need the concept of signals with time-varying dimensions. A signal with time-varying dimensions is defined as

$$x = \{\ldots, x(-2), x(-1), |x(0), x(1), x(2), \ldots\},$$

where $x(k) \in \mathbb{R}^{p(k)}$ and $p(k)$ is a nonnegative integer for any k. Here the vertical line indicates the position of time zero. Note that when $p(k) = 0$, $x(k) \in \mathbb{R}^{p(k)}$ means that $x(k)$ is always equal to 0. If $p(k)$ is periodic with period l, i.e., $p(k+l) = p(k)$ for any k, we call x a signal with l-periodically time-varying dimensions. Define the l-step shift operator S^l as

$$\begin{aligned} &S^l\{\ldots, x(-1), |x(0), x(1), \ldots\} \\ &= \{\ldots, x(-l-1), |x(-l), x(-l+1), \ldots\}. \end{aligned} \tag{2.1}$$

Denote P_k as the truncation operator, i.e.,

$$P_k\{\ldots, x(k-1), x(k), x(k+1), \ldots\} = \{\ldots, x(k-1), x(k), 0, \ldots\}. \tag{2.2}$$

Consider the system G_{mp}, shown in Figure 2.1, where the input u with $u(k) \in \mathbb{R}^{p(k)}$ and output y with $y(k) \in \mathbb{R}^{q(k)}$ are signals with l-periodically time-varying dimensions; that is, $p(k)$ and $q(k)$ are periodic with period l. Assume that G_{mp} satisfy the following properties:

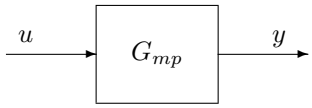


FIG. 2.1. *The multirate periodic system.*

(1) Linearity. The system G_{mp} is a linear operator.
(2) Periodicity. G_{mp} satisfies $G_{mp}S^l = S^l G_{mp}$, where S^l is given by (2.1).
(3) Causality. G_{mp} satisfies $P_k G_{mp}(I - P_k) = 0$, where P_k is given by (2.2).

In this paper, we focus on the systems that satisfy linear, periodic, and causal properties defined above. We call them multirate periodic (MP) systems. The general class of MP systems defined here covers many familiar classes of systems as special cases.

If $p(k) = p_1$ and $q(k) = q_1$ for all $k \in \mathbb{Z}$, then an MP system is a usual l-periodic system, for which there is a vast literature [2]. The multirate feature arises when $p(k)$ and $q(k)$ are truly time-varying. Let l be a multiple of m and n. If

$$u(k) \in \begin{cases} \mathbb{R}^{p_1} & \text{if } m|k \\ \{0\} & \text{otherwise} \end{cases} \quad \text{and} \quad y(k) \in \begin{cases} \mathbb{R}^{q_1} & \text{if } n|k \\ \{0\} & \text{otherwise} \end{cases},$$

then such an MP system is a dual rate system considered in [17]. Let l be a multiple of integers $m_i, i = 1, \ldots, s$, and $n_j, j = 1, \ldots, t$. If

$$u(k) = \begin{bmatrix} u_1(k) \\ \vdots \\ u_s(k) \end{bmatrix} \quad \text{and} \quad y(k) = \begin{bmatrix} y_1(k) \\ \vdots \\ y_t(k) \end{bmatrix},$$

where

$$u_i(k) \in \begin{cases} \mathbb{R}^{p_i} & \text{if } m_i|k \\ \{0\} & \text{otherwise} \end{cases} \quad \text{and} \quad y_j(k) \in \begin{cases} \mathbb{R}^{q_j} & \text{if } n_j|k \\ \{0\} & \text{otherwise} \end{cases},$$

then such an MP system becomes a general multirate system with uniform synchronized but different sampling in each input or output channel [7, 18, 20, 26]. The study of periodic and multirate systems in such a generality as indicated above, however, has never been done before.

Remark 1. One advantage of modelling a multirate system as a periodic system with periodically time-varying input-output spaces is that it better relates the present study to the rich theory on the usual periodic systems, as surveyed in [2]. Other advantages are its generality (it allows for nonuniform and asynchronous sampling) and its convenience (the treatments using this model take similar forms to those using other models, such as the one discussed in [7, 18]).

A standard method for the analysis of MP systems is to use lifting or blocking. For the MP system shown in Figure 2.1, define a lifting operator L_l on $\bigoplus_{k=-\infty}^{\infty} \mathbb{R}^{p(k)}$ by

$$L_l : \{\cdots | u(0), u(1), \cdots\} \mapsto \left\{ \cdots \middle| \begin{bmatrix} u(0) \\ u(1) \\ \vdots \\ u(l-1) \end{bmatrix}, \begin{bmatrix} u(l) \\ u(l+1) \\ \vdots \\ u(2l-1) \end{bmatrix}, \cdots \right\}.$$

L_l on $\bigoplus_{k=-\infty}^{\infty} \mathbb{R}^{q(k)}$ is defined similarly. Then the lifted system $G = L_l G_{mp} L_l^{-1}$ is an LTI system in the sense that $GS^1 = S^1 G$, where S^1 is the unit shift. Hence it has transfer function in the λ-transform ($\lambda = \frac{1}{z}$)

$$G(\lambda) = \begin{bmatrix} G_{11}(\lambda) & \cdots & G_{1l}(\lambda) \\ \vdots & \ddots & \vdots \\ G_{l1}(\lambda) & \cdots & G_{ll}(\lambda) \end{bmatrix}.$$

The LTI system G is not an arbitrary LTI system. Instead, its direct feedthrough term $G(0)$ is subject to a constraint that results from the causality of G_{mp} :

$$G_{ij}(0) = 0_{q(i)\times p(j)} \quad \text{for } 1 \le i < j \le l;$$

i.e., $G(0)$ is a block lower triangular matrix. Therefore the causality constraint can be represented by

$$G(0) \in \mathcal{T}(q, p),$$

where $p = [\ p(1) \quad \cdots \quad p(l)\ ]$ and $q = [\ q(1) \quad \cdots \quad q(l)\ ]$. Notice that the form of the causality here is simpler than that in [7, 18] due to the new form of the model.

3. Constrained analytic interpolation problems. In this section, we will present some constrained analytic function interpolation problems, which can be viewed as a multirate version of the standard interpolation problems. These constrained interpolation problems have various applications in MP systems as do their unconstrained counterparts in single rate systems. We first present a general case: a constrained tangential NP interpolation problem. Some more useful special cases are then formulated for convenience. In the following sections, we always assume that p, q, $|p|$, and $|q|$ are defined by (1.1)–(1.2).

Problem 1 (constrained tangential NP interpolation). Given $U \in \mathbb{C}^{|p|\times n}$, $Y \in \mathbb{C}^{|q|\times n}$, and $Z \in \mathbb{C}^{n\times n}$ with spectral radius $\rho(Z) < 1$, find (if possible) a function $G(\lambda) = \sum_{m=0}^{\infty} G_m \lambda^m \in \mathcal{H}_\infty^{|q|\times|p|}$ such that

(i) $\|G\|_\infty \le 1$,

(ii) $\sum_{m=0}^{\infty} G_m U Z^m = Y$,

(iii) $G(0) \in \mathcal{T}(q, p)$.

Roughly speaking, the integer n in the problem determines the number of interpolation conditions.

We can also formulate the following interpolation problems with block lower triangular constraints, which are special cases of Problem 1.

Problem 2 (constrained classical NP interpolation). Given a set of complex numbers $\lambda_\alpha \in \mathbb{D}$ along with matrices $U_\alpha \in \mathbb{C}^{|p|\times n}$ and $Y_\alpha \in \mathbb{C}^{|q|\times n}$ for $\alpha = 1, \ldots, s$, find (if possible) a function $G \in \mathcal{H}_\infty^{|q|\times|p|}$ such that

(i) $\|G\|_\infty \le 1$,

(ii) $G(\lambda_\alpha) U_\alpha = Y_\alpha$ for $\alpha = 1, \ldots, s$,

(iii) $G(0) \in \mathcal{T}(q, p)$.

Problem 3 (constrained CF interpolation). Given $U_\beta \in \mathbb{C}^{|p|\times n}$ and $Y_\beta \in \mathbb{C}^{|q|\times n}$, $\beta = 0, \ldots, r-1$, find (if possible) a function $G(\lambda) = \sum_{m=0}^{\infty} G_m \lambda^m \in \mathcal{H}_\infty^{|q|\times|p|}$ such that

(i) $\|G\|_\infty \le 1$,

(ii) $$\begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_{r-1} \end{bmatrix} = \begin{bmatrix} G_0 & 0 & \cdots & 0 \\ G_1 & G_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ G_{r-1} & G_{r-2} & \cdots & G_0 \end{bmatrix} \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ U_{r-1} \end{bmatrix},$$

(iii) $G(0) \in \mathcal{T}(q,p)$.

Problem 4 (constrained simultaneous NP and CF interpolation). Given $U_{1,\beta} \in \mathbb{C}^{|p|\times n}$, $Y_{1,\beta} \in \mathbb{C}^{|q|\times n}$ for $j = 0,\dots,r-1$, and $U_\alpha \in \mathbb{C}^{|p|\times n}$, $Y_\alpha \in \mathbb{C}^{|q|\times n}$, and $\lambda_\alpha \in \mathbb{D}$ for $\alpha = 2,\dots,s$, find (if possible) a function $G(\lambda) = \sum_{m=0}^{\infty} G_m \lambda^m \in \mathcal{H}_\infty^{|q|\times|p|}$ such that

(i) $\|G\|_\infty \le 1$,

(ii) $$\begin{bmatrix} Y_{1,0} \\ Y_{1,1} \\ \vdots \\ Y_{1,r-1} \end{bmatrix} = \begin{bmatrix} G_0 & 0 & \cdots & 0 \\ G_1 & G_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ G_{r-1} & G_{r-2} & \cdots & G_0 \end{bmatrix} \begin{bmatrix} U_{1,0} \\ U_{1,1} \\ \vdots \\ U_{1,r-1} \end{bmatrix},$$

(iii) $G(\lambda_\alpha)U_\alpha = Y_\alpha$ for $\alpha = 2,\dots,s$,

(iv) $G(0) \in \mathcal{T}(q,p)$.

Before giving the necessary and sufficient conditions of the above constrained analytic function interpolation problems, we need a result on matrix positive completion.

The matrix positive completion problem is as follows [9]: For a block matrix $B = [B_{ij}]_{i,j=1}^n$, given $B_{ij}, |j-i| \le m$, satisfying $B_{ij} = B_{ji}^*$, find the remaining matrices $B_{ij}, |j-i| > m$, such that the block matrix B is positive definite. The matrix positive completion problem was first proposed by Dym and Gohberg [9], who gave the following result.

LEMMA 3.1. *The matrix positive completion problem has a solution if and only if*

$$\begin{bmatrix} B_{ii} & \cdots & B_{i(i+m)} \\ \vdots & & \vdots \\ B_{(i+m)i} & \cdots & B_{(i+m)(i+m)} \end{bmatrix} \ge 0, \quad i = 1,\dots,n-m. \tag{3.1}$$

Reference [27] gave a detailed discussion of this problem and presented an explicit description of the set of all solutions via a linear fractional map whose coefficients are given in terms of the original data. However, Lemma 3.1 is sufficient for our purpose. We are now in a position to give the main result of this section.

THEOREM 3.2. *There exists a solution to Problem* 1 *if and only if*

$$Q - \tilde{Q} + Y^*\Pi_k(q)Y - U^*\Pi_k(p)U \ge 0 \tag{3.2}$$

for all $k = 1,\dots,l$, *where* Q *and* $\tilde{Q}$ *are, respectively, the unique solutions of Lyapunov equations*

$$Q = Z^*QZ + U^*U, \tag{3.3}$$

$$\tilde{Q} = Z^*\tilde{Q}Z + Y^*Y. \tag{3.4}$$

Here $\Pi_k(p)$ *and* $\Pi_k(q)$ *are defined in* (1.3)–(1.4).

Proof. The structural constraint on the interpolation function G can be viewed as an additional interpolation condition,

$$G(0)\,I = T,$$

for some $T \in \mathcal{T}(q,p)$. Set $\lambda_0 = 0$, $U_0 = I$, and $Y_0 = T$. By the solvability condition of the standard NP interpolation problem [12], the constrained interpolation problem has a solution if and only if there exists $T \in \mathcal{T}(q,p)$ such that

$$Q_a - \tilde{Q}_a \geq 0, \tag{3.5}$$

where Q_a and $\tilde{Q}_a$ satisfy

$$Q_a = \begin{bmatrix} \lambda_0 I & 0 \\ 0 & Z \end{bmatrix}^* Q_a \begin{bmatrix} \lambda_0 I & 0 \\ 0 & Z \end{bmatrix} + \begin{bmatrix} I \\ U^* \end{bmatrix} \begin{bmatrix} I & U \end{bmatrix}, \tag{3.6}$$

$$\tilde{Q}_a = \begin{bmatrix} \lambda_0 I & 0 \\ 0 & Z \end{bmatrix}^* \tilde{Q}_a \begin{bmatrix} \lambda_0 I & 0 \\ 0 & Z \end{bmatrix} + \begin{bmatrix} T^* \\ Y^* \end{bmatrix} \begin{bmatrix} T & Y \end{bmatrix}. \tag{3.7}$$

It is easy to see from (3.6)–(3.7) that

$$Q_a = \begin{bmatrix} I & U \\ U^* & Q \end{bmatrix} \text{ and } \tilde{Q}_a = \begin{bmatrix} T^*T & T^*Y \\ Y^*T & \tilde{Q} \end{bmatrix}.$$

Substituting Q_a and $\tilde{Q}_a$ into the inequality (3.5), we have

$$\begin{bmatrix} I - T^*T & U - T^*Y \\ U^* - Y^*T & Q - \tilde{Q} \end{bmatrix} \geq 0. \tag{3.8}$$

The left-hand side of (3.8) can be rewritten as

$$\begin{bmatrix} I & U \\ U^* & Q - \tilde{Q} + Y^*Y \end{bmatrix} - \begin{bmatrix} T^* \\ Y^* \end{bmatrix} \begin{bmatrix} T & Y \end{bmatrix}.$$

By the Schur complement, (3.8) is equivalent to

$$\begin{bmatrix} I & U & T^* \\ U^* & Q - \tilde{Q} + Y^*Y & Y^* \\ T & Y & I \end{bmatrix} \geq 0. \tag{3.9}$$

Therefore, the constrained NP interpolation problem has a solution if and only if (3.9) holds for a block lower triangular matrix T. This is a matrix positive completion problem. By Lemma 3.1, there is a block lower triangular T such that (3.9) holds if and only if

$$\begin{bmatrix} \Pi_k(p) & \Pi_k(p)U & 0 \\ U^*\Pi_k(p) & Q - \tilde{Q} + Y^*Y & Y^*[I_q - \Pi_k(q)] \\ 0 & [I_q - \Pi_k(q)]Y & I_q - \Pi_k(q) \end{bmatrix} \geq 0 \tag{3.10}$$

for $k = 0, \ldots, l$. Using the Schur complement twice, we can easily show that (3.10) is equivalent to

$$Q - \tilde{Q} + Y^*\Pi_k(q)Y - U^*\Pi_k(p)U \geq 0 \tag{3.11}$$

for $k = 0, \ldots, l$. We claim that inequality (3.11) when $k = l$ implies the case when $k = 0$. In fact, when $k = 0$, inequality (3.11) gives

$$Q - \tilde{Q} + Y^*Y - U^*U \geq 0. \tag{3.12}$$

Note that inequality (3.12) is equivalent to

$$Z^*(Q-\tilde{Q})Z \geq 0.$$

When $k=l$, inequality (3.11) gives

$$Q-\tilde{Q} \geq 0. \tag{3.13}$$

It is obvious that inequality (3.13) implies (3.12). □

Remark 2. If there is no constraint, then we have $l=1$. In this case, the condition in Theorem 3.2 becomes $Q-\tilde{Q} \geq 0$, which is a well-known result in the literature [1, 11, 12].

Remark 3. To verify the condition in Theorem 3.2, two Lyapunov equations, (3.3) and (3.4), can be combined into one:

$$Q-\tilde{Q} = Z^*(Q-\tilde{Q})Z + U^*U - Y^*Y.$$

However, Q and $\tilde{Q}$ will be used in the next section.

Q and $\tilde{Q}$ can be given directly from the original data in some special cases. We end this section by providing the explicit formula for these special cases.

COROLLARY 3.3. *There exists a solution to Problem* 2 *if and only if*

$$\left[\frac{U_\alpha^*U_\beta - Y_\alpha^*Y_\beta}{1-\lambda_\alpha^*\lambda_\beta} + Y_\alpha^*\Pi_k(q)Y_\beta - U_\alpha^*\Pi_k(p)U_\beta\right]_{\alpha,\beta=1}^{s} \geq 0 \tag{3.14}$$

for $k=1,\ldots,l$.

Proof. Note that Problem 2 can be viewed as a special case of Problem 1 with

$$\begin{aligned} Z &= \operatorname{diag}(\lambda_1 I_n, \ldots, \lambda_s I_n),\\ U &= [\ U_1 \quad \cdots \quad U_s\ ],\\ Y &= [\ Y_1 \quad \cdots \quad Y_s\ ]. \end{aligned}$$

Then it is easy to check that

$$Q = \left[\frac{U_\alpha^*U_\beta}{1-\bar{\lambda}_\alpha\lambda_\beta}\right]_{\alpha,\beta=1}^{s} \quad \text{and} \quad \tilde{Q} = \left[\frac{Y_\alpha^*Y_\beta}{1-\bar{\lambda}_\alpha\lambda_\beta}\right]_{\alpha,\beta=1}^{s}$$

are the solution of the Lyapunov equations (3.3) and (3.4), respectively. The result then follows from Theorem 3.2 directly. □

For $V = [\ V_0 \quad \cdots \quad V_{r-1}\ ]$, we use T_V to denote a corresponding lower Toeplitz matrix

$$T_V := \begin{bmatrix} V_0 & 0 & \cdots & 0\\ V_1 & V_0 & \ddots & \\ \vdots & \vdots & \ddots & 0\\ V_{r-1} & V_{r-2} & \cdots & V_0 \end{bmatrix}. \tag{3.15}$$

COROLLARY 3.4. *For the data of Problem* 3*, denote*

$$\begin{aligned} U &= [\ U_0 \quad \cdots \quad U_{r-1}\ ],\\ Y &= [\ Y_0 \quad \cdots \quad Y_{r-1}\ ]. \end{aligned}$$

Then there exists a solution to Problem 3 *if and only if*

$$T_U^* \begin{bmatrix} I_{p(r-1)} & 0 \\ 0 & I_p - \Pi_k(p) \end{bmatrix} T_U - T_Y^* \begin{bmatrix} I_{q(r-1)} & 0 \\ 0 & I_q - \Pi_k(q) \end{bmatrix} T_Y \geq 0 \tag{3.16}$$

for all $k = 1, \ldots, l$.

Proof. Note that Problem 3 can be viewed as a special case of Problem 1 with U, Y, and

$$Z = \begin{bmatrix} 0 & I_n & & 0 \\ & 0 & \ddots & \\ & & \ddots & I_n \\ & & & 0 \end{bmatrix}_{rn \times rn}.$$

Hence Q can be computed by

$$\begin{aligned} Q &= \sum_{m=0}^{\infty} Z^{*m} U^* U Z^m = \sum_{m=0}^{r-1} Z^{*m} U^* U Z^m \\ &= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ U_0^* \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & U_0 \end{bmatrix} + \cdots + \begin{bmatrix} U_0^* \\ \vdots \\ U_{r-1}^* \end{bmatrix} \begin{bmatrix} U_0 & \cdots & U_{r-1} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & \cdots & U_0^* \\ \vdots & \vdots & \iddots & \vdots \\ 0 & U_0^* & \cdots & U_{r-2}^* \\ U_0^* & U_1^* & \cdots & U_{r-1}^* \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & U_0 \\ 0 & \cdots & U_0 & U_1 \\ \vdots & \iddots & \vdots & \vdots \\ U_0 & \cdots & U_{r-2} & U_{r-1} \end{bmatrix}. \end{aligned}$$

Similarly, we have

$$\tilde{Q} = \begin{bmatrix} 0 & 0 & \cdots & Y_0^* \\ \vdots & \vdots & \iddots & \vdots \\ 0 & Y_0^* & \cdots & Y_{r-2}^* \\ Y_0^* & Y_1^* & \cdots & Y_{r-1}^* \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & Y_0 \\ 0 & \cdots & Y_0 & Y_1 \\ \vdots & \iddots & \vdots & \vdots \\ Y_0 & \cdots & Y_{r-2} & Y_{r-1} \end{bmatrix}.$$

Condition (3.2) then becomes

$$Q - \tilde{Q} + Y^* \Pi_k(q) Y - U^* \Pi_k(p) U \geq 0. \tag{3.17}$$

By pre- and postmultiplying inequality (3.17) by

$$\begin{bmatrix} 0 & \cdots & I_n \\ \vdots & \iddots & \vdots \\ I_n & \cdots & 0 \end{bmatrix}_{rn \times rn},$$

we obtain another equivalent condition

$$T_U^* T_U - T_Y^* T_Y + \begin{bmatrix} Y_{r-1}^* \\ \vdots \\ Y_0^* \end{bmatrix} \Pi_k(q) \begin{bmatrix} Y_{r-1} & \cdots & Y_0 \end{bmatrix} - \begin{bmatrix} U_{r-1}^* \\ \vdots \\ U_0^* \end{bmatrix} \Pi_k(p) \begin{bmatrix} U_{r-1} & \cdots & U_0 \end{bmatrix} \geq 0.$$

This is exactly (3.16) after some simple algebraic manipulations. ∎

COROLLARY 3.5. *For the data of Problem* 4*, denote*

$$U_1 = \begin{bmatrix} U_{10} & \cdots & U_{1(r-1)} \end{bmatrix},$$
$$Y_1 = \begin{bmatrix} Y_{10} & \cdots & Y_{1(r-1)} \end{bmatrix}.$$

Then there exists a solution to Problem 4 *if and only if*

$$\begin{bmatrix} A_{11k} & A_{21k}^* \\ A_{21k} & A_{22k} \end{bmatrix} \geq 0 \tag{3.18}$$

for all $k = 1, \ldots, l$*, where*

$$A_{11k} = T_{U_1}^* \begin{bmatrix} I_{p(r-1)} & 0 \\ 0 & I_p - \Pi_k(p) \end{bmatrix} T_{U_1} - T_{Y_1}^* \begin{bmatrix} I_{q(r-1)} & 0 \\ 0 & I_q - \Pi_k(q) \end{bmatrix} T_{Y_1},$$

$$A_{21k} = \begin{bmatrix} \bar{\lambda}_2^{r-1} U_2^* & \cdots & \bar{\lambda}_2 U_2^* & U_2^* \\ \vdots & \cdots & \vdots & \vdots \\ \bar{\lambda}_s^{r-1} U_s^* & \cdots & \bar{\lambda}_s U_s^* & U_s^* \end{bmatrix} \begin{bmatrix} I_{p(r-1)} & 0 \\ 0 & I_p - \Pi_k(p) \end{bmatrix} T_{U_1} - \begin{bmatrix} \bar{\lambda}_2^{r-1} Y_2^* & \cdots & \bar{\lambda}_2 Y_2^* & Y_2^* \\ \vdots & \cdots & \vdots & \vdots \\ \bar{\lambda}_s^{r-1} Y_s^* & \cdots & \bar{\lambda}_s Y_s^* & Y_s^* \end{bmatrix} \begin{bmatrix} I_{q(r-1)} & 0 \\ 0 & I_q - \Pi_k(q) \end{bmatrix} T_{Y_1},$$

$$A_{22k} = \left[\frac{U_\alpha^* U_\beta - Y_\alpha^* Y_\beta}{1 - \bar{\lambda}_\alpha \lambda_\beta} - U_\alpha^* \Pi_k(p) U_\beta + Y_\alpha^* \Pi_k(q) Y_\beta \right]_{\alpha,\beta=2}^{s}.$$

Proof. Note that Problem 4 can be viewed as a special case of Problem 1 with U, Y, and Z, where

$$U = \begin{bmatrix} U_1 & U_2 & \cdots & U_s \end{bmatrix},$$
$$Y = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_s \end{bmatrix},$$
$$Z = \operatorname{diag}(Z_1, \lambda_2 I_n, \ldots, \lambda_s I_n),$$

$$Z_1 = \begin{bmatrix} 0 & I_n & & 0 \\ & 0 & \ddots & \\ & & \ddots & I_n \\ & & & 0 \end{bmatrix}_{rn \times rn}.$$

Some simple algebraic manipulations show that

$$Q = \begin{bmatrix} Q_{11} & Q_{21}^* \\ Q_{21} & Q_{22} \end{bmatrix} \quad \text{and} \quad \tilde{Q} = \begin{bmatrix} \tilde{Q}_{11} & \tilde{Q}_{21}^* \\ \tilde{Q}_{21} & \tilde{Q}_{22} \end{bmatrix}$$

satisfy the Lyapunov equations (3.3) and (3.4), respectively, where

$$Q_{11} = \begin{bmatrix} 0 & \cdots & 0 & U_{10}^* \\ 0 & \cdots & U_{10}^* & U_{11}^* \\ \vdots & \cdots & \vdots & \vdots \\ U_{10}^* & \cdots & U_{1(r-2)}^* & U_{1(r-1)}^* \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & U_{10} \\ 0 & \cdots & U_{10} & U_{11} \\ \vdots & \cdots & \vdots & \vdots \\ U_{10} & \cdots & U_{1(r-2)} & U_{1(r-1)} \end{bmatrix},$$

$$Q_{21} = \begin{bmatrix} \bar{\lambda}_2^{r-1} U_2^* & \cdots & \bar{\lambda}_2 U_2^* & U_2^* \\ \bar{\lambda}_3^{r-1} U_3^* & \cdots & \bar{\lambda}_3 U_3^* & U_3^* \\ \vdots & \cdots & \vdots & \vdots \\ \bar{\lambda}_s^{r-1} U_s^* & \cdots & \bar{\lambda}_s U_s^* & U_s^* \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & U_{10} \\ 0 & \cdots & U_{10} & U_{11} \\ \vdots & \cdots & \vdots & \vdots \\ U_{10} & \cdots & U_{1(r-2)} & U_{1(r-1)} \end{bmatrix},$$

$$Q_{22} = \left[\frac{U_\alpha^* U_\beta}{1 - \bar{\lambda}_\alpha \lambda_\beta} \right]_{\alpha,\beta=2}^{s},$$

$$\tilde{Q}_{11} = \begin{bmatrix} 0 & \cdots & 0 & Y_{10}^* \\ 0 & \cdots & Y_{10}^* & Y_{11}^* \\ \vdots & \cdots & \vdots & \vdots \\ Y_{10}^* & \cdots & Y_{1(r-2)}^* & Y_{1(r-1)}^* \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & Y_{10} \\ 0 & \cdots & Y_{10} & Y_{11} \\ \vdots & \cdots & \vdots & \vdots \\ Y_{10} & \cdots & Y_{1(r-2)} & Y_{1(r-1)} \end{bmatrix},$$

$$\tilde{Q}_{21} = \begin{bmatrix} \bar{\lambda}_2^{r-1} Y_2^* & \cdots & \bar{\lambda}_2 Y_2^* & Y_2^* \\ \bar{\lambda}_3^{r-1} Y_3^* & \cdots & \bar{\lambda}_3 Y_3^* & Y_3^* \\ \vdots & \cdots & \vdots & \vdots \\ \bar{\lambda}_s^{r-1} Y_s^* & \cdots & \bar{\lambda}_s Y_s^* & Y_s^* \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & Y_{10} \\ 0 & \cdots & Y_{10} & Y_{11} \\ \vdots & \cdots & \vdots & \vdots \\ Y_{10} & \cdots & Y_{1(r-2)} & Y_{1(r-1)} \end{bmatrix},$$

$$\tilde{Q}_{22} = \left[\frac{Y_\alpha^* Y_\beta}{1 - \bar{\lambda}_\alpha \lambda_\beta} \right]_{\alpha,\beta=2}^{s}.$$

Condition (3.2) then becomes

$$Q - \tilde{Q} + Y^* \Pi_k(q) Y - U^* \Pi_k(p) U \geq 0. \tag{3.19}$$

By pre- and postmultiplying inequality (3.19) by

$$\begin{bmatrix} 0 & \cdots & I_n & \\ \vdots & \iddots & \vdots & 0 \\ I_n & \cdots & 0 & \\ & 0 & & I_{(s-1)n} \end{bmatrix},$$

we obtain condition (3.18) after some direct operator manipulations. □

Remark 4. Corollaries 3.3 and 3.4 can be directly used for robust model validation of multirate systems following the method for LTI systems studied in [6, 16].

4. Parametrization of all solutions. In this section, we characterize all solutions G to Problem 1 when the solvability condition is satisfied. We will consider only the generic case when $Q - \tilde{Q} > 0$. The unlikely case when $Q - \tilde{Q}$ is singular is technically more involved.

Since the characterization for the unconstrained case has been given in [12], our strategy in solving the constrained problem is then to choose, if possible, from this characterization all those solutions that satisfy the structural constraint. The same notation is used as in previous sections and more notation is needed. Given an operator Δ and two operator matrices

$$\Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \quad \text{and} \quad \Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix},$$

the linear fractional transformation associated with Λ and Δ is denoted by

$$\mathcal{F}(\Lambda, \Delta) = \Lambda_{11} + \Lambda_{12}\Delta(I - \Lambda_{22}\Delta)^{-1}\Lambda_{21},$$

and the star product of Λ and Γ is defined as

$$\Lambda \star \Gamma = \begin{bmatrix} \Lambda_{11} + \Lambda_{12}\Gamma_{11}(I - \Lambda_{22}\Gamma_{11})^{-1}\Lambda_{21} & \Lambda_{12}(I - \Gamma_{11}\Lambda_{22})^{-1}\Gamma_{12} \\ \Gamma_{21}(I - \Lambda_{22}\Gamma_{11})^{-1}\Lambda_{21} & \Gamma_{21}(I - \Lambda_{22}\Gamma_{11})^{-1}\Lambda_{22}\Gamma_{12} + \Gamma_{22} \end{bmatrix}.$$

Here we assume that the operator manipulations are all compatible. With these definitions, we have

$$\mathcal{F}(\Lambda, \mathcal{F}(\Gamma, \Delta)) = \mathcal{F}(\Lambda \star \Gamma, \Delta).$$

The following lemma from [19] will be used later.

LEMMA 4.1. *For $M \in \mathbb{C}^{|q|\times|p|}$, the following statements are equivalent:*

(1) *There exists $T \in \mathcal{T}(q,p)$ such that $\|M + T\| \le 1$.*

(2) *There exists*

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$$

with $P_{11} \in \mathcal{T}(q,p)$, $P_{12} \in \mathcal{T}(q,q)$ invertible, $P_{21} \in \mathcal{T}(p,p)$ invertible, and $P_{22} \in \mathcal{T}_s(p,q)$ such that

$$\begin{bmatrix} M + P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$$

is unitary.

A way to find P from M was given in [19]. Recall that an operator valued function Θ is said to be two-sided inner if Θ is an inner function and $\Theta(e^{jw})$ is almost everywhere unitary. For U, Y, and Z in Problem 1, assume that $Q - \tilde{Q} > 0$, where Q and $\tilde{Q}$ are defined by (3.3) and (3.4), respectively. By [12, Theorem III 7.2], there exist matrices $C \in \mathbb{C}^{n\times|p|}$ and $E \in \mathbb{C}^{|p|\times|p|}$ such that the state space model $\{Z, C, U, E\}$ is controllable and observable and the transfer function

$$\Theta(\lambda) := E + \lambda U(I - \lambda Z)^{-1}C \tag{4.1}$$

is two-sided inner in $\mathcal{H}^{|p|\times|p|}$. It follows from QR factorization that there is a special $\Theta(\lambda)$ such that $E^* \in \mathcal{T}(p,p)$ and (4.1) holds. By Cholesky factorization, there exist $N \in \mathcal{T}(q,q)$ and $S \in \mathcal{T}(p,p)$ such that

$$N^*N = [I + Y(Q - \tilde{Q})^{-1}Y^*]^{-1}, \tag{4.2}$$

$$S^*S = [I + C^*\tilde{Q}(Q - \tilde{Q})^{-1}QC]^{-1}. \tag{4.3}$$

Let $A_0 = (Q - Z^*\tilde{Q}Z)^{-1}Z^*(Q - \tilde{Q})$. It is shown in [12, Proposition V 1.7] that A_0 is stable. Define

$$\Phi(\lambda) = \begin{bmatrix} \Phi_{11}(\lambda) & \Phi_{12}(\lambda) \\ \Phi_{21}(\lambda) & \Phi_{22}(\lambda) \end{bmatrix},$$

where

$$\begin{aligned}
\Phi_{11}(\lambda) &= Y(I - \lambda A_0)^{-1}(Q - Z^*\tilde{Q}Z)^{-1}U^*, \\
\Phi_{12}(\lambda) &= N^{-1} - \lambda Y A_0(I - \lambda A_0)^{-1}(Q - \tilde{Q})^{-1}Y^*N^{-1}, \\
\Phi_{21}(\lambda) &= S\Theta^*(\lambda) - S^{-1}C^*Q(I - \lambda A_0)^{-1}(Q - \tilde{Q})^{-1}\tilde{Q}C\Theta^*(\lambda), \\
\Phi_{22}(\lambda) &= -\lambda S^{-1}C^*Q(I - \lambda A_0)^{-1}(Q - \tilde{Q})^{-1}Y^*N^{-1}.
\end{aligned}$$

The set of all $G(\lambda)$ solving the unconstrained interpolation problem is then given by

$$G(\lambda) = \mathcal{F}(\Phi(\lambda), V(\lambda)),$$

where $V(\lambda)$ is a contractive analytic function in $\mathcal{H}_\infty^{|q|\times|p|}$. Obviously, the set of all solutions to the constrained interpolation problem is

$$\{G(\lambda) = \mathcal{F}(\Phi(\lambda), V(\lambda)) : G(0) \in \mathcal{T}(q,p)\}. \tag{4.4}$$

It is easy to check that

$$\begin{aligned}
\Phi_{11}(0) &= Y(Q - Z^*\tilde{Q}Z)^{-1}U^*, \\
\Phi_{12}(0) &= N^{-1}, \\
\Phi_{21}(0) &= SE^* - S^{-1}C^*Q(Q - \tilde{Q})^{-1}\tilde{Q}CE^* \\
&= S[I - C^*Q(Q - \tilde{Q})^{-1}\tilde{Q}C]E^* \\
&= S^{-1}E^*, \\
\Phi_{22}(0) &= 0.
\end{aligned}$$

Now assume condition (3.2) in Theorem 3.2 is satisfied. Then there is a contractive analytic function $V(\lambda)$ in $\mathcal{H}_\infty^{|q|\times|p|}$ such that

$$G(0) = \Phi_{11}(0) + N^{-1}V(0)S^{-1}E^* \in \mathcal{T}(q,p).$$

That is,

$$\| - N\Phi_{11}(0)(S^{-1}E^*)^{-1} + N\hat{G}(0)(S^{-1}E^*)^{-1}\| \le 1.$$

By Lemma 4.1, there exists

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$$

with $P_{11} \in \mathcal{T}(q,p)$, $P_{12} \in \mathcal{T}(q,q)$ invertible, $P_{21} \in \mathcal{T}(p,p)$ invertible, and $P_{22} \in \mathcal{T}_s(p,q)$ such that

$$B := \begin{bmatrix} -N\Phi_{11}(0)(S^{-1}E^*)^{-1} + P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$$

is unitary. Define $\Psi = \Phi \star B$. It is easy to check that $\Psi_{11} \in \mathcal{T}(q,p)$, $\Psi_{12} \in \mathcal{T}(q,q)$, $\Psi_{21} \in \mathcal{T}(p,p)$, $\Psi_{22} \in \mathcal{T}_s(p,q)$, and both Ψ_{12} and Ψ_{21} are invertible. By setting a bijective map $R = \mathcal{F}(B,V)$, we have

$$G(\lambda) = \mathcal{F}(\Phi, V) = \mathcal{F}(\Phi, \mathcal{F}(B,R)) = \mathcal{F}(\Phi \star B, R) = \mathcal{F}(\Psi, R).$$

Note that $G(0) \in \mathcal{T}(q,p)$ if and only if $R(0) \in \mathcal{T}(q,p)$. Hence the set (4.4) can be rewritten as

$$\{G(\lambda) = \mathcal{F}(\Psi, R) : R \in \mathcal{H}_\infty^{|q|\times|p|} \text{ with } R(0) \in \mathcal{T}(q,p) \text{ and } \|R\| \le 1\}.$$

This gives us the main result of this section, the following theorem.

THEOREM 4.2. *For Problem* 1*, assume that* $Q - \tilde{Q} > 0$ *and the solvability condition* (3.2) *holds. Then the set of all interpolants* $G(\lambda)$ *is given by*

$$G(\lambda) = \mathcal{F}(\Psi(\lambda), R(\lambda)),$$

where R *is a contractive analytic function with* $R(0) \in \mathcal{T}(q,p)$.

5. Conclusion. In this paper, we study the MP systems and some related analytic function interpolation problems. We show that each MP system has an equivalent LTI system with a causality constraint which can be represented by a set of block lower triangular matrices. We then study some analytic function interpolation problems with such a constraint. The necessary and sufficient solvability conditions are given using the result of the positive matrix completion problem. Finally, all the solutions are presented in terms of linear fractional transformation.

REFERENCES

[1] J. A. BALL, I. GOHBERG, AND L. RODMAN, *Interpolation of Rational Matrix Functions*, Oper. Theory Adv. Appl. 45, Birkhäuser, Basel, Switzerland, 1990.

[2] S. BITTANTI AND P. COLANERI, *Invariant representations of discrete-time periodic systems*, Automatica J. IFAC, 36 (2000), pp. 1777–1793.

[3] C. I. BYRNES, T. T. GEORGIOU, AND A. LINDQUIST, *A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint*, IEEE Trans. Automat. Control, 46 (2001), pp. 822–839.

[4] L. CHAI AND L. QIU, *Multirate systems and related interpolation problems*, in Perspectives in Robust Control, Lecture Notes in Control and Inform. Sci. 268, R. Moheimani, ed., Springer-Verlag, London, 2001, pp. 29–40.

[5] L. CHAI AND L. QIU, *Model validation of multirate systems from time-domain experimental data*, IEEE Trans. Automat. Control, 47 (2002), pp. 346–351.

[6] J. CHEN AND G. GU, *Control-oriented System Identification: An* $\mathcal{H}_\infty$ *Approach*, John Wiley and Sons, New York, 2000.

[7] T. CHEN AND L. QIU, $\mathcal{H}_\infty$ *design of general multirate sampled-data control systems*, Automatica J. IFAC, 30 (1994), pp. 1139–1152.

[8] P. DELSARTE, Y. GENIN, AND Y. KAMP, *On the role of the Nevanlinna-Pick problem in circuit and system theory*, Internat. J. Circuit Theory Appl., 9 (1981), pp. 177–187.

[9] H. DYM AND I. GOHBERG, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36 (1981), pp. 1–24.

[10] G. FLOQUET, *Sur les equations differentielles linearies a cofficients periodicques*, Annales de l'Ecole Normale Superieur, 12 (1883), pp. 47–89.

[11] C. FOIAS AND A. E. FRAZHO, *The Commutant Lifting Approach to Interpolation Problems*, Oper. Theory Adv. Appl. 44, Birkhäuser, Basel, Switzerland, 1990.

[12] C. FOIAS, A. E. FRAZHO, I. GOHBERG, AND M. A. KAASHOEK, *Metric Constrained Interpolation, Commutant Lifting and Systems*, Oper. Theory Adv. Appl. 100, Birkhäuser, Basel, Switzerland, 1998.

[13] J. W. Helton, J. A. Ball, C. R. Johnson, and J. N. Palmer, *Operator Theory, Analytic Functions, Matrices, and Electrical Engineering*, CBMS Reg. Conf. Ser. Math., AMS, Providence, RI, 1987.

[14] P. P. Khargonekar, K. Poolla, and A. Tannenbaum, *Robust control of linear time-invariant plants using periodic compensation*, IEEE Trans. Automat. Control, 30 (1985), pp. 1088–1096.

[15] R. A. Meyer and C. S. Burrus, *A unified analysis of multirate and periodically time-varying digital filters*, IEEE Trans. Circuits and Systems, 22 (1975), pp. 162–168.

[16] K. Poolla, P. P. Khargonekar, A. Tikku, J. Krause, and K. M. Nagpal, *A time-domain approach to model validation*, IEEE Trans. Automat. Control, 39 (1994), pp. 1088–1096.

[17] L. Qiu and T. Chen, *$\mathcal{H}_2$-optimal design of multirate sampled-data systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 2506–2511.

[18] L. Qiu and T. Chen, *Multirate sampled-data systems: All $\mathcal{H}_\infty$ suboptimal controllers and the minimum entropy controllers*, IEEE Trans. Automat. Control, 44 (1999), pp. 537–550.

[19] L. Qiu and T. Chen, *Unitary dilation approach to contractive matrix completion*, Linear Algebra Appl., 379 (2004), pp. 345–352.

[20] L. Qiu and K. Tan, *Direct state space solution of multirate sampled-data $\mathcal{H}_2$ optimal control*, Automatica J. IFAC, 34 (1998), pp. 234–245.

[21] D. Sarason, *Generalized interpolation in $\mathcal{H}_\infty$*, Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.

[22] Sz.-Nagy and C. Foias, *Harmonic Analysis of Operators on Hilbert Space*, North–Holland, Amsterdam-London, American Elsevier, New York, Akadémiai Kiadó, Budapest, 1970.

[23] L. Tong, G. Xu, and T. Kailath, *Blind identification and equalization based on second-order statistics: A time domain approach*, IEEE Trans. Signal Process., 42 (1994), pp. 2242–2256.

[24] P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[25] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

[26] P. G. Voulgaris, M. A. Dahleh, and L. Valavani, *$\mathcal{H}_\infty$ and $\mathcal{H}_2$ optimal controllers for periodic and multirate systems*, Automatica J. IFAC, 30 (1994), pp. 251–263.

[27] H. J. Woerdeman, *Strictly contractive and positive completions for block matrices*, Linear Algebra Appl., 136 (1990), pp. 63–105.

[28] D. C. Youla and M. Saito, *Interpolation with positive-real functions*, J. Franklin Inst., 284 (1967), pp. 77–108.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 1987–1999

OPTIMALITY OF STASIS AND SMALL SWITCHING CYCLES IN PLANAR SYSTEMS WITH TWO-VALUED CONTROLS*

STEWART D. JOHNSON†

Abstract. For planar ($x \in \mathbb{R}^2$) control systems $x' = f(x, \mu)$ with a two-valued control $\mu \in \{\mu_1, \mu_2\}$ we consider small cycles created by rapidly switching between the two control values. A two-cycle is a periodic cycle on which the periodic control changes value twice in one period. Stasis is a relaxed (i.e., probabilistic) control μ_{r} with a fixed point (called a stasis point) x_{r} where $0 = f(x_{\mathrm{r}}, \mu_{\mathrm{r}}) = \int f(x_{\mathrm{r}}, \mu)\, d\mu_{\mathrm{r}}$. Generically, stasis points can be approximated by two-cycles, and every two-cycle must contain a stasis point. Also generically, stasis points form curves (called stasis curves). The unit metric on these curves parameterize families of small two-cycles near the stasis curve. Under general conditions, we show that average performance is differentiable with respect to this parameterization, and thus necessary conditions for optimality of small two-cycles versus stasis can be explicitly calculated.

Key words. optimal control, optimal periodic control, relaxed control

AMS subject classifications. 37N35, 49N25, 49N20, 49K15, 49K30

DOI. 10.1137/S036301290342231X

1. Approximating optimal control.

The standard control problem involves optimizing (maximizing or minimizing) a performance function

$$Q = \int_0^T H(x, u, t)\, dt$$

over some set of *admissible controls* $u(t) \in \mathcal{U}$ subject to constraint by a differential equation

$$x' = F(x, u, t) \tag{1}$$

and possibly constrained by a variety of endpoint conditions on T, $x(0)$, and $x(T)$. Excellent introductions to control theory include [3, 10, 13].

Optimal periodic control [4, 6, 7] involves optimizing average performance

$$Q = \frac{1}{T} \int_0^T H(x, u, t)\, dt$$

over periodic trajectories of period T under a differential equation (1). This constrains $x(t) = x(t + T)$ and $u(t) = u(t + T)$ for all t, and in particular $x(0) = x(T)$.

The introduction of relaxed controls [2, 4, 6, 7, 12] raises several questions. First, when can a relaxed control be approximated by an admissible control? Certainly not all can be; an example appears later. Second, in the case where the relaxed control can be approximated, in what way does the performance depend on the approximation? What happens to performance as we take closer approximations to a relaxed control? We may expect continuity because performance is defined by an integral, but can

*Received by the editors February 4, 2003; accepted for publication (in revised form) May 23, 2004; published electronically April 14, 2005.
http://www.siam.org/journals/sicon/43-6/42231.html
†Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267 (sjohnson@williams.edu).

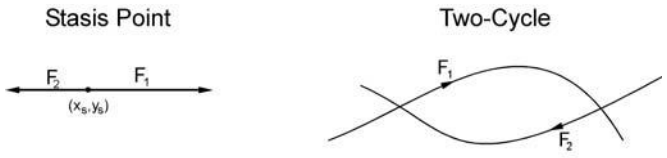


FIG. 1. *Stasis and two-cycles.*

we expect some form of smoothness near the "boundaries" of the set of admissible controls?

This paper contains an extensive analysis of these questions in a specific setting. We consider periodic optimization of $\mathbb{R}^2$ systems with control functions that switch between two fixed values. Such situations arise naturally in bang-bang controls [8, 9, 11].

We consider relaxed controls that produce fixed trajectories at antiparallel points called *stasis points* and consider periodic approximations that switch between values twice in one period, called *two-cycles* (see Figure 1). Under reasonable assumptions we show that stasis points are generically approximable by two-cycles, and that every two-cycle (or every n-cycle for that matter) must contain a stasis point. We show that stasis points generically form curves that parameterize the family of approximating two-cycles as two-parameter family. Performance is a differentiable function of this parameterization. In fact as the parameters converge to the relaxed control, performance has a finite one-sided derivative. This main result on differentiability is proven using a standardizing diffeomorphism and technical calculation.

The next section of this paper contains the basic constructs and topological results. Section 3 examines performance in this setting and ends with a statement of the main theorem. In section 4 we introduce a standardizing diffeomorphism and prove some basic analytic results in the standardized framework. Section 5 analyzes the family of two-cycles in the standard framework, and section 6 is a technical and very calculational proof of the main theorem. Finally, section 7 contains an application to a problem studied by Gilbert [5].

2. Stasis and switching two-cycles. Consider a controlled planar flow[1]

$$\begin{aligned} x' &= f(x, y, \mu), \\ y' &= g(x, y, \mu) \end{aligned} \tag{2}$$

with $(x, y) \in \mathcal{D} \subset \mathbb{R}^2$ and $\mu \in \mathbb{R}^n$. We restrict ourselves to two distinct values of the control parameter, μ_1, μ_2, and piecewise constant control functions $\mu(t)$ that switch between the two values. Taking $f_i(x, y) = f(x, y, \mu_i)$, $g_i(x, y) = g(x, y, \mu_i)$, this amounts to studying the pair of systems

$$\begin{aligned} x' &= f_1(x, y), & x' &= f_2(x, y), \\ y' &= g_1(x, y), & y' &= g_2(x, y). \end{aligned} \tag{3}$$

Trajectories in system (2) are composed of segments of trajectories from these two systems, as in Figure 2. Specifically, $x(t)$, $y(t)$, $\mu(t)$ is a trajectory for (2) iff $x(t)$, $y(t)$ is a trajectory for $x' = f_i(x, y)$, $y' = g_i(x, y)$ from (3) on the closure of any time interval where $\mu(t) = \mu_i$.

[1]This work is specific to planar geometry, and notation reflecting this bias is favored.

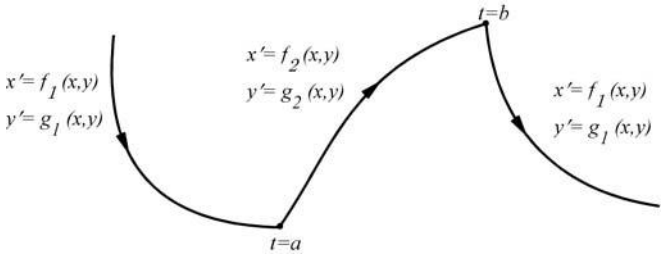


FIG. 2. *A switching trajectory.*

A *cycle* of length $T > 0$ is a nonstationary trajectory $x(t), y(t), \mu(t)$ defined for $[0, T]$ with $x(T) = x(0)$, $y(T) = y(0)$, and $\mu(t + T) = \mu(t)$. A cycle is *switching* if μ is discontinuous on $[0, T]$ and *nonswitching* if μ is constant. The cycle is a *two-cycle* if μ has exactly two discontinuities on $[0, T]$.

Without relaxed controls, a fixed trajectory can occur only at a fixed point in one of the two component systems. But if we allow relaxed controls, trajectories can be fixed at places where the two flows are antiparallel. Specifically, writing

$$\mathbf{F}_i = \begin{pmatrix} f_i \\ g_i \end{pmatrix},$$

a point $(x_s, y_s) \in \mathcal{D} \subset \mathbb{R}^2$ is a *stasis point* if (x_s, y_s) is not fixed in either system and, for some $\kappa > 0$,

$$\mathbf{F}_1(x_s, y_s) = -\kappa\, \mathbf{F}_2(x_s, y_s). \tag{4}$$

A trajectory can be fixed at a stasis point by a relaxed control

$$\mu = \begin{cases} 1 & \text{with probability } \frac{1}{\kappa+1}, \\ 2 & \text{with probability } \frac{\kappa}{\kappa+1}\ . \end{cases} \tag{5}$$

In the classic theory for smooth flows, a cycle in a simply connected domain must orbit a fixed point. Similarly, a switching cycle in a simply connected domain must orbit either a stasis or a fixed point.

LEMMA 1. *If $N \subset \mathcal{D}$ is a simply connected open set containing a switching cycle, then N contains a stasis point or a fixed point.*

Proof. Suppose N contains no stasis or fixed points. Let

$$\mathbf{U} = \frac{\mathbf{F}_1}{|\mathbf{F}_1|} + \frac{\mathbf{F}_2}{|\mathbf{F}_2|}\ .$$

This vector field is continuous and nonzero on simply connected N and is therefore parallel to $\mathbf{U} = \nabla H$ for some continuous surface $H : N \mapsto \mathbb{R}$. Let $x(t), y(t)$ be a trajectory under $\mathbf{F}_1$. Then

$$\begin{aligned} \begin{pmatrix} x' \\ y' \end{pmatrix} \cdot \mathbf{U} &\propto \mathbf{F}_1 \cdot \left(\frac{\mathbf{F}_1}{|\mathbf{F}_1|} + \frac{\mathbf{F}_2}{|\mathbf{F}_2|} \right) \\ &= |\mathbf{F}_1| + \frac{\mathbf{F}_1 \cdot \mathbf{F}_2}{|\mathbf{F}_2|} \\ &\geq 0 \end{aligned}$$

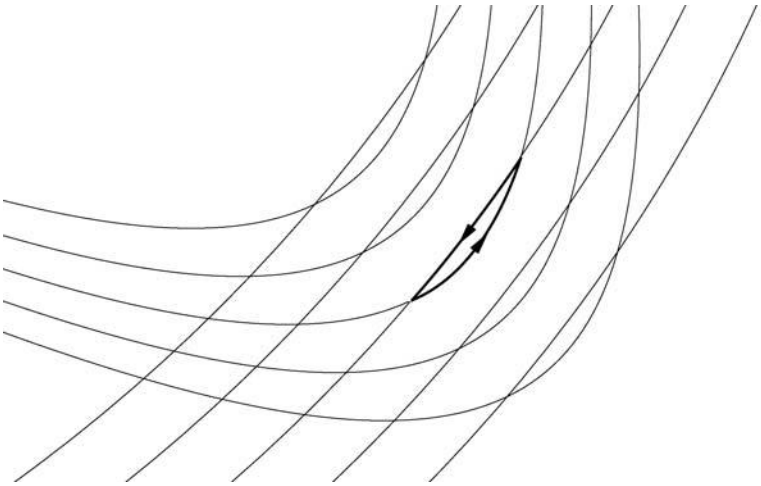


FIG. 3. *Small switching cycles.*

with equality only at a stasis point. This similarly holds for $\mathbf{F}_2$, and so values of $H(x(t), y(t))$ are strictly increasing on any switching trajectory, prohibiting a cycle. □

A switching cycle without stasis points (but with a fixed point) can be constructed from, for example, an inward spiral and a source.

The next lemma establishes generic existence of switching cycles near stasis points. The idea is that near a typical point, a smooth flow is approximated to the second order by a flow of constant curvature. The curvature of the two flows being different at a stasis point creates small cycles, as in Figure 3.

Difference in curvature near a stasis point is calculated as a change in flux of one vector field across trajectories of the other. That is, the flux of field $\mathbf{F}_1$ across $\mathbf{F}_2$ is simply the cross product:[2]

$$\mathbf{F}_1 \otimes \mathbf{F}_2 = \mathbf{F}_1 \cdot \mathbf{F}_2^{\perp} = -\mathbf{F}_1^{\perp} \cdot \mathbf{F}_2.$$

This value is zero at a stasis point. Small cycles exist when this value changes sign as a trajectory passes through a stasis point. We say a stasis point (x_s, y_s) is *regular* if

$$\frac{d}{dt}\, \mathbf{F}_1 \otimes \mathbf{F}_2\big|_{x_s, y_s} \neq 0,$$

where the derivative is taken along either flow. We have the following.

LEMMA 2. *Every neighborhood of a regular stasis point contains a switching cycle.*

The proof of this lemma follows by the construction of delta-cycles in section 5. Note that stasis points without switching cycles can be constructed with flows of equal curvature, as in Figure 4. Thus there are (nonregular) stasis points where there is no periodic approximation.

Stasis points are zero-level curves of the surface $\mathbf{F}_1 \otimes \mathbf{F}_2$, and regularity at a stasis point implies it is not a critical point of that surface. Hence we have the following lemma.

[2]Define the perpendicular operator as a clockwise right angle rotation $(x, y)^{\perp} = (y, -x)$.

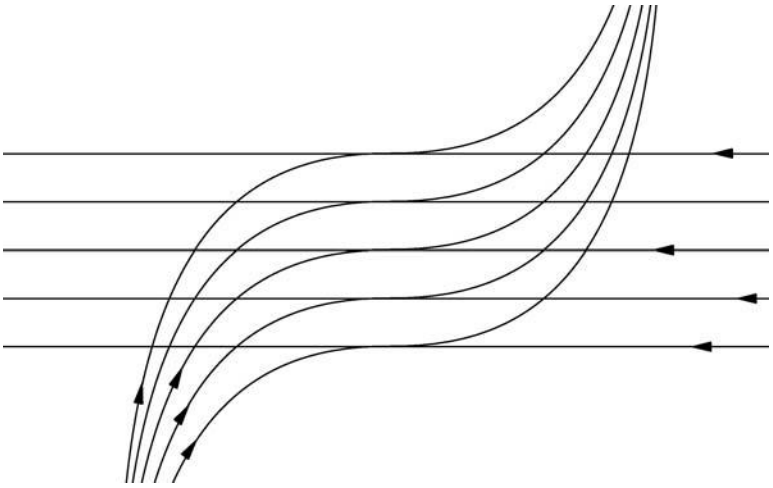


FIG. 4. *Stasis without cycles.*

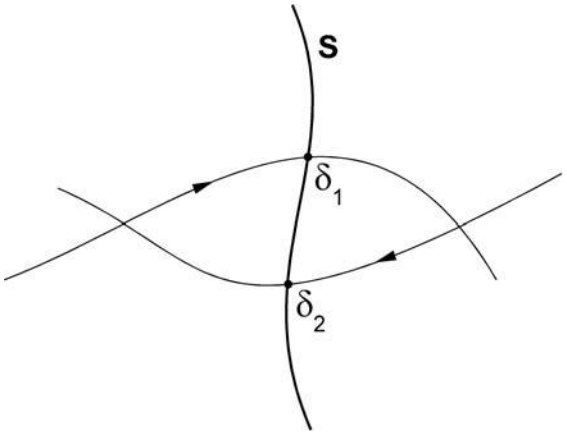


FIG. 5. *Two-cycle through* $\mathbf{S}(\delta_1)$ *and* $\mathbf{S}(\delta_2)$.

LEMMA 3. *For any regular stasis point, there exists a neighborhood in which the set of stasis points forms a smooth curve.*

Given a regular stasis point x_s, y_s, we can unit parameterize the stasis curve near (x_s, y_s) as $\mathbf{S}(\delta)$ with

$$\mathbf{S}(0) = (x_s, y_s). \tag{6}$$

Lemma 3 implies that $\mathbf{S}(\delta)$ is well defined in a neighborhood of $\delta = 0$. If ϵ is sufficiently small, pairs δ_1, δ_2 with $\epsilon > \delta_1 > \delta_2 > -\epsilon$ will uniquely determine a two-cycle that crosses the stasis line exactly twice, at δ_1 and δ_2, as in Figure 5.

We organize the subscripts so that for the two-cycle corresponding to $\delta_1 > \delta_2$, the trajectory segment under $\mathbf{F}_1$ will contain $\mathbf{S}(\delta_1)$ and the segment under $\mathbf{F}_2$ will contain $\mathbf{S}(\delta_2)$. This is accomplished by choosing the direction of parameterization so that

$$\mathbf{S}'(0)^{\perp} \cdot \nabla \left(\mathbf{F}_1 \otimes \mathbf{F}_2\right)\big|_{\mathbf{S}(0)} < 0. \tag{7}$$

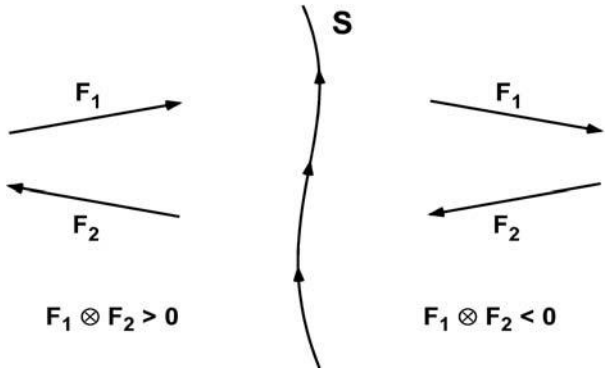


FIG. 6. *Direction of parameterization.*

That is, as we traverse $\mathbf{S}(\delta)$ in the positive direction, the value of $\mathbf{F}_1 \otimes \mathbf{F}_2$ is positive to the left and negative to the right, as in Figure 6.

By choosing the appropriate direction for time we are also free to assume, without loss of generality,

$$\mathbf{S}'(0)^{\perp} \cdot \mathbf{F}_1 > 0, \tag{8}$$

and this is consistent with Figures 5 and 6.

3. Performance of small cycles. The performance of small cycles is measured as the average performance over the cycle [4, 6, 7]. Thus if $x(t), y(t), \mu(t)$ is periodic of period T, and H is the payoff function, then performance is

$$Q = \frac{1}{T}\int_0^T H(x, y, \mu)dt.$$

With focus restricted to two values μ_1, μ_2 of the control parameter, we take $H_1(x, y) = H(x, y, \mu_1)$ and $H_2(x, y) = H(x, y, \mu_2)$. Thus if we have a two-cycle with one period defined by

$$u(t) = \begin{cases} \mu_1 & \text{for} \quad a \le t < b, \\ \mu_2 & \text{for} \quad b \le t < c, \end{cases}$$

we have the performance

$$Q = \frac{\int_a^b H_1(x, y)dt + \int_b^c H_2(x, y)dt}{c - a}. \tag{9}$$

For small ϵ, and $\epsilon > \delta_1 > \delta_2 > -\epsilon$, define $Q(\delta_1, \delta_2)$ as the performance of the two-cycle corresponding to δ_1, δ_2.

At a regular stasis point x_s, y_s we calculate the stasis performance from the relaxed control (5) and the relation (4):

$$\begin{aligned} Q^*(x_s, y_s) &= \frac{H_1(x_s, y_s) + \kappa\, H_2(x_s, y_s)}{\kappa + 1} \\ &= \frac{H_1(x_s, y_s)\,|\mathbf{F}_2(x_s, y_s)| + H_2(x_s, y_s)\,|\mathbf{F}_1(x_s, y_s)|}{|\mathbf{F}_1(x_s, y_s)| + |\mathbf{F}_2(x_s, y_s)|}. \end{aligned} \tag{10}$$

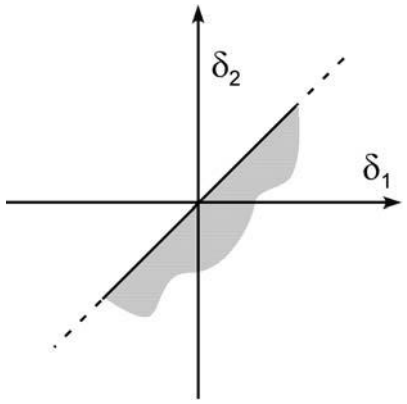


FIG. 7. *Half-closed neighborhood.*

Given a two-cycle, let D be the distance between the switching points $(x(a), y(a))$ and $(x(b), y(b))$, and rewrite (9) as

$$Q(\delta_1, \delta_2) = \frac{\int_a^b H_1(x,y)dt + \int_b^c H_2(x,y)dt}{c-a} = \frac{\frac{D}{c-b}\ \frac{1}{b-a}\ \int_a^b H_1(x,y)dt + \frac{D}{b-a}\ \frac{1}{c-b}\ \int_b^c H_2(x,y)dt}{\frac{D}{c-b} + \frac{D}{b-a}}.$$

Comparing this last expression to (10) and using the approximations

$$\begin{aligned} |\mathbf{F}_1(x_s, y_s)| &\approx \frac{D}{b-a}, \\ |\mathbf{F}_2(x_s, y_s)| &\approx \frac{D}{c-b}, \\ H_1(x_s, y_s) &\approx \frac{1}{b-a}\ \int_a^b H_1(x,y)dt, \\ H_2(x_s, y_s) &\approx \frac{1}{c-b}\ \int_b^c H_2(x,y)dt, \end{aligned}$$

it follows that

$$\lim_{\substack{\delta_1, \delta_2 \to \delta \\ -\epsilon < \delta_1 < \delta_2 < \epsilon}} Q(\delta_1, \delta_2) = Q^*(\mathbf{S}(\delta)).$$

Hence performance at stasis is approximated by performance on small two-cycles. Taking $Q(\delta, \delta) = Q^*(\mathbf{S}(\delta))$ defines $Q(\delta_1, \delta_2)$ as a continuous function for $\epsilon > \delta_1 \geq \delta_2 > -\epsilon$.

In general, there is some half-closed set in the δ_1, δ_2 plane containing $\delta_1 = \delta_2$ and contained in $\delta_1 \geq \delta_2$ near $(0, 0)$ (as in Figure 7) on which $Q(\delta_1, \delta_2)$ is defined and continuous. The following is the main theorem.

THEOREM 4. *If $\mathbf{S}(\delta)$ is a regular stasis curve in a neighborhood of $\delta = 0$, then $Q(\delta_1, \delta_2)$ is C^1 on the half-closed set $-\epsilon < \delta_1 \leq \delta_2 < \epsilon$ for some $\epsilon > 0$.*

Here we use the convention that a function is differentiable at a closed boundary if the one-sided derivatives exist and are continuous. The usefulness of this theorem

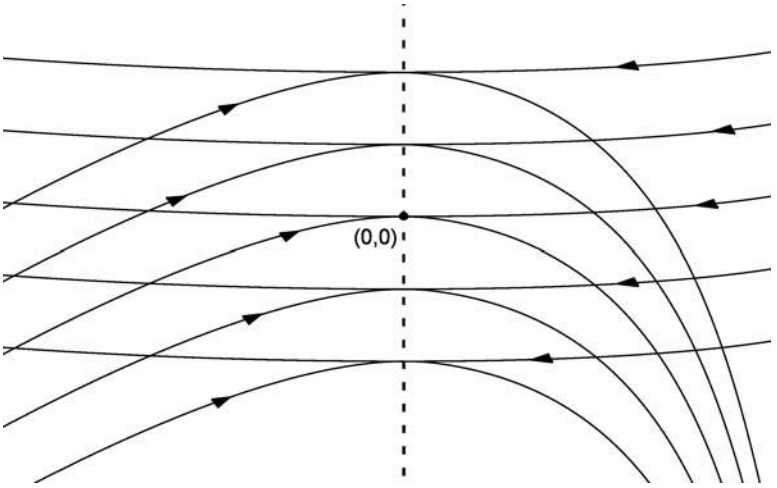


FIG. 8. *Rectification near stasis.*

lies in being able to calculate the partial derivatives and check for optimality. We demonstrate how to compute these partials in a direct computational proof of the theorem.

4. Rectification. We prove Theorem 4 by direct computation in a standardized reference frame. This section defines the standardizing diffeomorphism, of which several choices are possible. For example, we could rectify $\mathbf{F}_1$ by applying a diffeomorphism Φ that conjugates it to the constant flow $\widetilde{\mathbf{F}}_1 = \frac{d\Phi}{d(x,y)} \circ \mathbf{F}_1 \circ \Phi^{-1} = (1,0)$ (see [1]). After some experimentation, we find that applying a diffeomorphism that makes the two flows perpendicular to a vertical stasis line creates an efficient reference frame (see Figure 8). Specifically, on the neighborhood

$$R_\epsilon = \{-\epsilon < u < \epsilon,\ -\epsilon^2 < v < \epsilon^2\}, \tag{11}$$

our standard system is a pair of flows $\widetilde{\mathbf{F}}_i,\ i = 1, 2$, of the form

$$\begin{aligned} u' &= \kappa_i + \zeta_i u + \rho_i u^2 + \eta_i v + O(\epsilon^3), \\ v' &= \gamma_i u + \lambda_i u^2 + O(\epsilon^3) \\ &\text{with } \ \kappa_1\kappa_2 < 0 \ \text{ and } \ \kappa_1\gamma_2 - \kappa_2\gamma_1 \neq 0. \end{aligned} \tag{12}$$

Any C^3 system with a vertical stasis curve $u = 0$ and both flows horizontal through this curve must be of this form. By switching indices and/or the temporal direction we can, without loss of generality, assume

$$\kappa_1 > 0, \qquad \kappa_2 < 0, \qquad \kappa_1\gamma_2 - \kappa_2\gamma_1 < 0, \tag{13}$$

and this would be consistent with assumptions (7), (8) and Figures 5, 6.

LEMMA 5. *Any C^3 pair $\mathbf{F}_1, \mathbf{F}_2$ at a regular stasis point x_s, y_s with stasis curve $\mathbf{S}(\delta)$ can be rectified to the standard form* (12) *in a neighborhood R_ϵ for sufficiently small ϵ.*

Proof. In fact, this can be done in several ways. For example, since diffeomorphisms preserve stasis points, any nonsingular diffeomorphism

$$\Phi = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix}$$

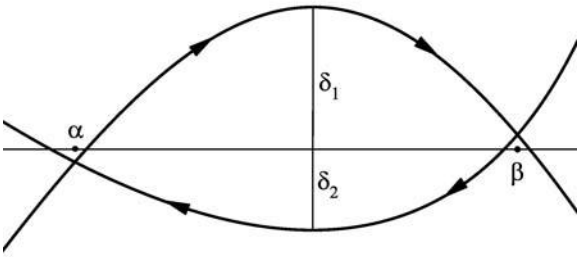


FIG. 9. *Delta-cycle.*

with $u = \mathbf{F}_1 \otimes \mathbf{F}_2$ will produce a conjugated system $\widetilde{\mathbf{F}}_1$, $\widetilde{\mathbf{F}}_2$ with $u = 0$ as the stasis line. Furthermore, taking v as any nontrivial solution to $v_x f_i + v_y g_i = 0$ for either $i = 1, 2$ will produce a conjugated system

$$\begin{pmatrix} \widetilde{f}_1 & \widetilde{f}_2 \\ \widetilde{g}_1 & \widetilde{g}_2 \end{pmatrix} = \begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix} \begin{pmatrix} f_1 & f_2 \\ g_1 & g_2 \end{pmatrix}$$

that flows horizontally through $u = 0$. The flows are antiparallel on $u = 0$, and so $\widetilde{g}_i = v_x f_i + v_y g_i = 0$ for $u = 0$ for both $i = 1, 2$. We can even preserve the parameterization of the stasis curve by imposing the boundary values $v(\mathbf{S}(\delta)) = \delta$.

Since diffeomorphisms preserve regular stasis points, we have $\kappa_1 \kappa_2 < 0$ and

$$\frac{d}{dt}\, \widetilde{\mathbf{F}}_1 \otimes \widetilde{\mathbf{F}}_2 \,\big|_{0,0} = (\kappa_1 \gamma_2 - \kappa_2 \gamma_1)\ \frac{du}{dt}\bigg|_{u=0} = (\kappa_1 \gamma_2 - \kappa_2 \gamma_1)\, \kappa_i,$$

implying $\kappa_1 \gamma_2 - \kappa_2 \gamma_1 \neq 0$. □

It is thus sufficient to prove Theorem 4 for systems of form (12). The following lemma, which follows from direct computation, will be useful.

LEMMA 6. *For $\epsilon > 0$, $\kappa_i \neq 0$, the differential equation*

$$v'(u) = \frac{\gamma u + \lambda u^2 + O(\epsilon^3)}{\kappa + \zeta u + \rho u^2 + \eta v + O(\epsilon^3)}$$

with initial value $v(0) = \delta$ for $\delta^2 < \epsilon$ and $|u| < \epsilon$ has approximate solution

$$v(u) = \delta + \frac{\gamma}{2\kappa} u^2 + \left(\frac{\lambda\kappa - \gamma\zeta}{3\kappa^2} \right) u^3 + O(\epsilon^4).$$

5. Delta-cycles. For our standard system (12) in a neighborhood R_ϵ (defined (11)), and assuming $\kappa_1 \gamma_2 - \kappa_2 \gamma_1 < 0$, we construct the family of two-cycles in the form of *delta-cycles* as depicted in Figure 9. For δ_1,δ_2 values near zero, consider the trajectory $u_i(t), v_i(t)$ in system (12) starting at $(0, \delta_i)$. Applying Lemma 6 to the implicit functions $v_1(u)$ and $v_2(u)$ yields

$$\begin{aligned} v_1(u) - v_2(u) = (\delta_1 - \delta_2) &- \left(\frac{\kappa_1 \gamma_2 - \kappa_2 \gamma_1}{2\kappa_1 \kappa_2} \right) u^2 \\ &+ \left(\frac{\lambda_1 \kappa_1 - \gamma_1 \zeta_1}{3\kappa_1^2} - \frac{\lambda_2 \kappa_2 - \gamma_2 \zeta_2}{3\kappa_2^2} \right) u^3 + O(\epsilon^4). \end{aligned}$$

The assumptions $\kappa_1\gamma_2 - \kappa_2\gamma_1 < 0$ and $\kappa_1\kappa_2 < 0$ imply that for $\delta_1 > \delta_2$ sufficiently small, this expression will vanish at points $u = \alpha, \beta$ for $\alpha < 0 < \beta$, which can be explicitly approximated. For small values of A the cubic $A - Bx^2 + Cx^3 + O(A^2)$ has roots

$$x = \pm\sqrt{\frac{A}{B}} + \frac{AC}{2B^2} + O(A^{3/2}).$$

Hence,

$$\begin{aligned}\alpha = &-\sqrt{\frac{2\kappa_1\kappa_2}{\kappa_1\gamma_2 - \kappa_2\gamma_1}}\,(\delta_1 - \delta_2)^{1/2} \\ &-\frac{2}{3}\frac{\kappa_2^2\kappa_1\lambda_1 + \kappa_2^2\gamma_1\zeta_1 + \kappa_1^2\kappa_2 - \kappa_1^2\gamma_2\zeta_2}{(\kappa_1\gamma_2 - \kappa_2\gamma_1)}\,(\delta_1 - \delta_2) \\ &+ O(\epsilon^3),\end{aligned} \tag{14}$$

$$\begin{aligned}\beta = &\sqrt{\frac{2\kappa_1\kappa_2}{\kappa_1\gamma_2 - \kappa_2\gamma_1}}\,(\delta_1 - \delta_2)^{1/2} \\ &-\frac{2}{3}\frac{\kappa_2^2\kappa_1\lambda_1 + \kappa_2^2\gamma_1\zeta_1 + \kappa_1^2\kappa_2 - \kappa_1^2\gamma_2\zeta_2}{(\kappa_1\gamma_2 - \kappa_2\gamma_1)}\,(\delta_1 - \delta_2) \\ &+ O(\epsilon^3).\end{aligned} \tag{15}$$

This construction demonstrates the existence of two-cycles near a regular stasis point for sufficiently small $\delta_1 < \delta_2$, which proves Lemma 2.

6. Proof of Theorem 4. We are now prepared to prove Theorem 4 by direct calculation in the form of the following lemma.

LEMMA 7. *For systems ($i = 1, 2$)*

$$\begin{aligned}u' &= \kappa_i + \zeta_i u + \rho_i u^2 + \eta_i v + O(\epsilon^3), \\ v' &= \gamma_i u + \lambda_i u^2 + O(\epsilon^3)\end{aligned} \tag{16}$$

with $\kappa_1\kappa_2 < 0$, $\kappa_1\gamma_2 - \kappa_2\gamma_1 < 0$, and associated costs

$$H_i(u, v) = A_i + B_i u + C_i u^2 + D_i v + O(\epsilon^3),$$

for $\delta_1 \geq \delta_2$ sufficiently small we have

$$\begin{aligned}Q(\delta_1, \delta_2) = &\frac{\kappa_1 A_2 - \kappa_2 A_1}{\kappa_1 - \kappa_2} \\ &+ (A_1 - A_2)(\delta_1 - \delta_2)\frac{\left(\gamma_1\eta_1\kappa_2^2 - \gamma_2\eta_2\kappa_1^2 + 2\kappa_2^2\kappa_1\rho_1 - 2\kappa_1^2\kappa_2\rho_2\right)}{3(\kappa_1 - \kappa_2)^2(\kappa_1\gamma_2 - \kappa_2\gamma_1)} \\ &+ (A_1 - A_2)(\delta_1 - \delta_2)\frac{2\kappa_1\kappa_2(\zeta_2\kappa_1 - \zeta_1\kappa_2)(\lambda_2\kappa_1 - \lambda_1\kappa_2 + \gamma_2\zeta_1 - \gamma_1\zeta_2)}{3(\kappa_1 - \kappa_2)^2(\kappa_1\gamma_2 - \kappa_2\gamma_1)^2} \\ &+ (A_1 - A_2)(\delta_1\eta_1\kappa_2 - \delta_2\eta_2\kappa_1)\frac{1}{(\kappa_1 - \kappa_2)^2} \\ &+ (B_1\gamma_2 - B_2\gamma_1)(\delta_1 - \delta_2)\frac{\kappa_2\kappa_1(\zeta_1\kappa_2 - \zeta_2\kappa_1)}{(\kappa_1 - \kappa_2)(\kappa_1\gamma_2 - \kappa_2\gamma_1)^2}\end{aligned}$$

$$
\begin{aligned}
&+ \left(B_1\kappa_2 - B_2\kappa_1\right)\left(\delta_1 - \delta_2\right)\frac{\kappa_2\kappa_1\left(\kappa_1\lambda_2 - \kappa_2\lambda_1\right)}{\left(\kappa_1 - \kappa_2\right)\left(\kappa_1\gamma_2 - \kappa_2\gamma_1\right)^2}\\
&- \left(C_1\kappa_2 - C_2\kappa_1\right)\left(\delta_1 - \delta_2\right)\frac{\kappa_2\kappa_1}{3\left(\kappa_1 - \kappa_2\right)\left(\kappa_1\gamma_2 - \kappa_2\gamma_1\right)}\\
&- \left(D_1\gamma_1\kappa_2^2 - D_2\gamma_2\kappa_1^2\right)\left(\delta_1 - \delta_2\right)\frac{1}{3\left(\kappa_1 - \kappa_2\right)\left(\kappa_1\gamma_2 - \kappa_2\gamma_1\right)}\\
&- \left(D_1\delta_1\kappa_2 - D_2\delta_2\kappa_1\right)\frac{1}{\kappa_1 - \kappa_2}\\
&+ O\left(\left|\delta_1\right|^{3/2}, \left|\delta_2\right|^{3/2}\right).
\end{aligned}
$$

Proof. Given cost functions H_1 and H_2 and using the implicit functions $v_1(u)$ and $v_2(u)$, we have from (9)

$$
Q(\delta_1, \delta_2) = \frac{\int_\alpha^\beta \frac{H_1(u, v_1(u))}{f_1(u, v_1(u))}\, du + \int_\alpha^\beta \frac{H_2(u, v_2(u))}{f_2(u, v_2(u))}\, du}{\int_\alpha^\beta \frac{1}{f_1(u, v_1(u))}\, du + \int_\alpha^\beta \frac{1}{f_2(u, v_2(u))}\, du}. \tag{17}
$$

Using Lemma 6 we expand the integrands

$$
\begin{aligned}
\frac{H_i(u, v_i(u))}{f_i(u, v_i(u))} &= \frac{A_i + B_i u + C_i u^2 + D_i v_i(u)}{\kappa_i + \zeta_i u + \rho_i u^2 + \eta_i v_i(u)}\\
&= \frac{A_i}{\kappa_i} + \left(\frac{D_i\kappa_i - A_i\eta_i}{\kappa_i^2}\right)\delta_i + \left(\frac{B_i\kappa_i - \zeta_i A_i}{\kappa_i^2}\right)u\\
&\quad + \left(\frac{2C_i\kappa_i^2 + D_i\gamma_i\kappa_i - 2A_i\rho_i\kappa_i - A_i\eta_i\gamma_i - 2\zeta_i\kappa_i B_i + 2\zeta_i^2 A_i}{2\kappa_i^3}\right)u^2\\
&\quad + O(\epsilon^3).
\end{aligned} \tag{18}
$$

Hence the integrands in (17) are essentially quadratic. With α, β from (14), (15) we compute for general quadratics with coefficients a_i, b_i,

$$
\begin{aligned}
&\frac{\int_\alpha^\beta \left(a_1 + a_2 u + a_3 u^2 + O(\epsilon^3)\right)\, du}{\int_\alpha^\beta \left(b_1 + b_2 u + b_3 u^2 + O(\epsilon^3)\right)\, du}\\
&\quad = \frac{a_1}{b_1} + \frac{2\left(\kappa_2^2\kappa_1\lambda_1 - \zeta_1\gamma_1\kappa_2^2 - \kappa_1^2\kappa_2\lambda_2 + \kappa_1^2\gamma_2\zeta_2\right)}{3b_1^2\left(\gamma_2\kappa_1 - \gamma_1\kappa_2\right)^2}\left(a_2b_1 - a_1b_2\right)\left(\delta_1 - \delta_2\right)\\
&\qquad + \frac{2\kappa_2\kappa_1}{3b_1^2\left(\gamma_2\kappa_1 - \gamma_1\kappa_2\right)}\left(b_1a_3 - a_1b_3\right)\left(\delta_1 - \delta_2\right) + O(\epsilon^3).
\end{aligned}
$$

The theorem follows from substituting a_i, b_i using (18):

$$
\begin{aligned}
a_1 &= \frac{1}{\kappa_1^2}\left(A_1\kappa_1 + \delta_1 D_1\kappa_1 - \delta_1 A_1\eta_1\right)\\
&\quad - \frac{1}{\kappa_2^2}\left(A_2\kappa_2 + \delta_2 D_2\kappa_2 - \delta_2 A_2\eta_2\right),\\
a_2 &= \frac{1}{\kappa_1^2}\left(B_1\kappa_1 - A_1\zeta_1\right) - \tfrac{1}{\kappa_2^2}\left(B_2\kappa_2 - A_2\zeta_2\right),
\end{aligned}
$$

$$a_3 = \frac{1}{2\kappa_1^3}\left(A_1(2\zeta_1^2 - 2\rho_1\kappa_1 - \eta_1\gamma_1) - 2B_1\zeta_1\kappa_1 + 2C_1\kappa_1^2 + D_1\gamma_1\kappa_1\right)$$
$$- \frac{1}{2\kappa_2^3}\left(A_2(2\zeta_2^2 - 2\rho_2\kappa_2 - \eta_2\gamma_2) - 2B_2\zeta_2\kappa_2 + 2C_2\kappa_2^2 + D_2\gamma_2\kappa_2\right),$$
$$b_1 = \frac{1}{\kappa_1^2}(\kappa_1 - \delta_1\eta_1) - \frac{1}{\kappa_2^2}(\kappa_2 - \delta_2\eta_2),$$
$$b_2 = -\frac{\zeta_1}{\kappa_1^2} + \frac{\zeta_2}{\kappa_2^2},$$
$$b_3 = \frac{1}{2\kappa_1^3}(2\zeta_1^2 - 2\rho_1\kappa_1 - \eta_1\gamma_1) - \frac{1}{2\kappa_2^3}(2\zeta_2^2 - 2\rho_2\kappa_2 - \eta_2\gamma_2). \qquad \square$$

7. Example: Vehicle cruise. Suppose you can switch between the two following systems:

$$\text{COASTING:} \quad y' = -f(y).$$
$$\text{ENGAGED:} \quad y' = M - g(y) - f(y).$$

Here, we can think of y as velocity and $f(y)$ as friction. As would arise in bang-bang controls, we consider M the maximum thrust we can apply, and $g(y)$ the cost or discount from applying that thrust at speed y. We impose the restriction that y must stay close to a prescribed average velocity R. Define u as the difference between the actual and target velocities, $u = y - R$, and let v be position $v' = u$. With $F(u) = f(u + R)$ and $G(u) = g(u + R)$ we have the pair of systems

$$\begin{array}{ll} u' = -F(u), & u' = M - G(u) - F(u), \\ v' = u, & v' = u\,. \end{array}$$

This is in standard form (12), has $\lambda_i = \eta_i = 0$, $\gamma_i = 1$, and

$$\begin{aligned} \kappa_1 &= -F(0), \\ \zeta_1 &= -F'(0), \\ \rho_1 &= -\frac{1}{2}F''(0), \\ \kappa_2 &= M - F(0) - G(0), \\ \zeta_2 &= -F'(0) - G'(0), \\ \rho_2 &= -\frac{1}{2}F''(0) - \frac{1}{2}G''(0). \end{aligned}$$

Taking performance as minimizing the average thrust discount,

$$Q = \frac{1}{T}\int_0^T G(u(t))\,dt,$$

we take $A_1 = B_1 = C_1 = D_1 = A_2 = B_2 = 0$, $C_2 = G'(0)$, and $D_2 = \frac{1}{2}G''(0)$, yielding

$$Q(\delta_1, \delta_2) = -\frac{1}{3}(\delta_2 - \delta_1)\,(M - F(0) - G(0))\ \frac{d}{du}\left(G'(u)\left(\tfrac{F(u)}{M-G(u)}\right)^2\right)_{u=0}.$$

On the other hand, if performance is measured as average thrust, we define

$$I(t) = \begin{cases} 0 & \text{if COASTING,} \\ 1 & \text{if ENGAGED} \end{cases}$$

and take

$$Q = \frac{1}{T}\int_0^T M\ I(t)\ dt.$$

Taking $A_1 = B_1 = C_1 = D_1 = B_2 = C_2 = D_2 = 0$ and $A_2 = M$ we get

$$Q(\delta_1, \delta_2)$$
$$= \frac{(\delta_1 - \delta_2)F(0)(M - F(0) - G(0))}{3(M - G(0))^4}\left((M - G(0))\frac{d^2}{du^2}\big(F(u)(M - G(u))\big)_{u=0} + 2\left(G'(0)\right)^2 F(0)\right).$$

The coefficient $\frac{F(0)(M-F(0)-G(0))}{3(M-G(0))^4}$ and the expression $M - G(0)$ are always positive (if the vehicle can move at all), and so optimality is indicated by the second derivative of $F(u)(M - G(u))$. If

$$\frac{d^2}{du^2}\big(F(u)(M - G(u))\big)_{u=0} > -\frac{2\left(G'(0)\right)^2 F(0)}{M - G(0)},$$

then stasis is never optimal. In particular if G is constant, $\frac{d^2F}{du^2}(0) > 0$ implies that stasis is not optimal. This yields a general and quite powerful converse to Theorem 8 in [5].

REFERENCES

[1] V. I. Arnol'd, *Ordinary Differential Equations*, Springer-Verlag, Berlin, 1992.

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

[3] S. Barnett, *Introduction to Mathematical Control Theory*, Oxford University Press, Oxford, 1975.

[4] F. Colonius, *Optimal Periodic Control*, Lecture Notes in Math. 1313, Springer-Verlag, Berlin, 1988.

[5] E. G. Gilbert, *Vehicle cruise: Improved fuel economy by periodic control*, Automatica, 12 (1976), pp. 159–166.

[6] E. G. Gilbert, *Optimal periodic control: A general theory of necessary conditions*, SIAM J. Control Optim., 15 (1977), pp. 717–746.

[7] G. Guardabassi, A. Locatelli, and S. Rinaldi, *Status of periodic optimization of dynamical systems*, J. Optim. Theory Appl., 14 (1974), pp. 1–20.

[8] J. Guckenheimer and S. Johnson, *Planar hybrid systems*, Hybrid Systems II, Lecture Notes in Comput. Sci. 999, Springer-Verlag, Berlin, 1995, pp. 202–225.

[9] S. Johnson, *Simple hybrid systems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 1655–1665.

[10] F. L. Lewis and V. L. Syrmos, *Optimal Control*, John Wiley and Sons, New York, 1995.

[11] J. J. O'Donnell, *Bounds on limit cycles in two-dimensional bang-bang control systems with an almost time-optimal switching curve*, IEEE Trans. Automat. Control, 9 (1964), pp. 448–457.

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

[13] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control*, W. B. Saunders, Philadelphia, London, Toronto, 1969.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2000–2008

ABSOLUTE STABILITY CRITERIA FOR A GENERALIZED LUR'E PROBLEM WITH DELAY IN THE FEEDBACK*

A. A. ZEVIN† AND M. A. PINSKY‡

Abstract. A nonautonomous linear system controlled by a nonlinear sector-restricted feedback with a time-varying delay is considered. Delay-independent sufficient conditions for absolute stability and instability (expressed in the transfer function of the linear part and the sector bounds) are established. For a system with an exponentially stable linear part, an upper bound for the Lyapunov exponent is found. It is shown that if the transfer function is sign-constant, asymptotic stability of the system with the margin-linear feedback guarantees absolute stability of the considered system; thus, such systems satisfy the known Aizerman conjecture. They include, in particular, a closed-loop system consisting of any number of time-varying first order links and feedback with arbitrary delay. Under some additional condition (which is certainly true for a time-invariant linear block), the obtained stability criterion is precise. The approach employed in the proofs is based on a direct analysis of the corresponding Volterra equation which contains only the transfer function of the linear block and, therefore, embraces a wide range of control systems. As an example, a second order system is considered; it is shown that here the obtained stability bound is reached for a linear feedback with a periodic delay function.

Key words. Lur'e problem, time-varying system, delay in feedback, absolute stability condition, Aizerman class

AMS subject classification. 93D10

DOI. 10.1137/S0363012903437599

1. Introduction.

The classical Lur'e problem is to find conditions for absolute stability of a control system consisting of a linear block and a nonlinear feedback contained within a prescribed sector [1]. Over the last few decades there has appeared an extensive literature devoted to the problem and its generalization. Most of the known results are obtained by the frequency domain or Lyapunov function methods and relate to systems with a time-invariant or periodic linear block (e.g., [2, 3, 4, 5, 6, 7, 8, 9, 10]). The Lyapunov method enables us, in principle, to tackle arbitrary time-varying systems; however, finding the Lyapunov function for such systems is, generally, a difficult problem.

In [11] sufficient stability conditions for the Lur'e problem that are equally applied to time-invariant and time-varying systems were found. The results are based on a direct analysis of the corresponding integral Volterra equation regarding the input of the nonlinear block $\sigma(t)$. In the current paper we extend this approach to systems with delay in the feedback. Namely, we assume that the corresponding output is of the form $\varphi = \varphi(\sigma(t-\tau(t)), t)$ where the function $\tau(t)$ is piecewise continuous, nonnegative, and bounded for $t \in [0, \infty)$. The corresponding integral equation becomes

$$\sigma(t) = f(t) + \int_0^t w(t,s)\varphi(\sigma(s-\tau(s)), s)\,\mathrm{d}s, \tag{1.1}$$

where $w(t,s)$ is the transfer function of the linear block. Note that no other information on the linear block is employed, so the last may be described, in particular, by

*Received by the editors November 10, 2003; accepted for publication (in revised form) June 15, 2004; published electronically April 14, 2005.
http://www.siam.org/journals/sicon/43-6/43759.html
†Transmag Research Institute, Academy of Sciences of Ukraine, 49005 Dnepropetrovsk, Piesarzhevsky 5, Ukraine (zevin@npkista.dp.ua).
‡Department of Mathematics, University of Nevada-Reno, Reno, NV 89557 (pinsky@unr.edu).

partial differential or integral equations.

The piecewise continuous scalar valued function $f(t)$ describes oscillation of the corresponding linear system in the absence of the feedback caused by nonzero initial conditions and, perhaps, external perturbation disappearing at infinity. We assume that the linear block is asymptotically stable; thus

$$|f(t)| \to 0 \text{ as } t \to \infty. \tag{1.2}$$

The function $\varphi(\sigma, t)$ belongs to the class $\Phi(K_1, K_2)$, i.e., satisfies the inequality

$$K_1\sigma^2 \le \varphi(\sigma, t)\sigma \le K_2\sigma^2, \quad \sigma \in (-\infty, \infty). \tag{1.3}$$

We assume that with a given initial function $\sigma(t)$ for $t < 0$, the solution $\sigma(t)$ of (1.1) is continuable on $[0, \infty)$.

DEFINITION 1. *System* (1.1) *is called absolutely stable in the class* $\Phi(K_1, K_2)$ *if for any functions* $f(t)$, $\varphi(\sigma, t)$, *satisfying conditions* (1.2), (1.3), *and any piecewise continuous nonnegative bounded for* $t \in [0, \infty)$ *function* $\tau(t)$, *the corresponding solution* $\sigma(t)$ *of* (1.1) *satisfies the condition*

$$|\sigma(t)| \to 0 \text{ as } t \to \infty. \tag{1.4}$$

If condition (1.4) is not fulfilled for some $\varphi(\sigma, t)$, $\tau(t)$, and $f(t)$ from the indicated classes, the system is referred to as unstable.

Putting $\varphi_1(\sigma, t) = \varphi(\sigma, t) - K_1\sigma - K\sigma, K = 0.5(K_2 - K_1)$, and retaining the previous notation, we reduce (1.3) to the form

$$-K\sigma^2 \le \varphi(\sigma, t)\sigma \le K\sigma^2, \quad \sigma \in (-\infty, \infty). \tag{1.5}$$

Thus, we replace the class $\Phi(K_1, K_2)$ by $\Phi(-K, K)$; therewith we assume that the transfer function $w(t, s)$ in (1.1) is changed correspondingly.

In section 2 a value K_* is found such that for $K < K_*$, the system is absolutely stable independent on the delay function $\tau(t)$ (see Theorem 1). If the linear block is exponentially stable and the function $f(t)$ exponentially tends to zero, so does the solution $\sigma(t)$; Theorem 2 provides an upper bound for the corresponding Lyapunov exponent. For a class of linear blocks (including, in particular, the autonomous ones) the value K_0 is found such that the system is unstable in the class $\Phi(-K_0, K_0)$ for any $\tau(t)$ (see Theorem 3).

In section 3 systems with a nonnegative transfer function are considered. It is shown (see Theorem 4) that asymptotic stability for $\varphi(\sigma, t) = K\sigma$ guarantees absolute stability of the system in the class $\Phi(-K, K)$. Thus, such systems for arbitrary delay $\tau(t)$ in the feedback satisfy the Aizerman conjecture [12] (note that the known results of this kind [13, 14, 15] relate to time-invariant systems). Under some additional condition, a precise upper bound for the stability sector is found (see Theorem 5).

In section 4 applications of the obtained results to some systems are discussed. It is shown that a closed-loop system consisting of any number of first order time-varying links and arbitrary delay in the feedback satisfies the Aizerman conjecture in the class $\Phi(-K, K)$. For a second order time-invariant system, delay independent bounds K_* and K_0 are found in explicit forms; the bound K_* is precise because it is reached for some periodic delay $\tau(t)$.

2. Absolute stability and instability criteria. Suppose that the linear block is exponentially stable; then

$$|w(t,s)| \le C\exp[-\Delta(t-s)], \tag{2.1}$$

where the constants C and $\Delta > 0$ are independent of t and s.

Let us put

$$W(t) = \int_0^t w(t,s)|\mathrm{d}s,$$

$$W_+(t_k) = \sup W(t) \quad \text{for } t \ge t_k, \tag{2.2}$$

$$W_\infty = \overline{\lim_{t\to\infty}}\, W(t) = \lim_{t_k\to\infty} W_+(t_k).$$

Here W_∞ is the upper limit of $W(t)$ as $t \to \infty$; it coincides with the conventional limit when the last exists. This is certainly the case when the linear block is time-invariant. In fact, in such a system $w(t,s) = w(t-s)$, so, setting $t - s = z$, we obtain

$$W(t) = \int_0^t w(z)|\mathrm{d}z. \tag{2.3}$$

The function $W(t)$ in (2.3) increases monotonically and, therefore, tends to the limit.

The following theorem establishes a sufficient condition for absolute stability of system (1.1), (1.5).

THEOREM 1. *If*

$$K < K_* = 1/W_\infty, \tag{2.4}$$

the system is absolutely stable in the class $\Phi(-K,K)$.

Proof. Let $\sigma(t)$ be a solution of (1.1). First let us show that for any $t_1 \ge 0$, there exists $t_m \ge t_1$ such that $|\sigma(t_m)| \ge |\sigma(t)|$ for $t \in [t_1,\infty)$. In fact, otherwise there is a sequence $t_1, t_2, \ldots,\ t_k \to \infty$ as $k \to \infty$, such that $|\sigma(t)| \le |\sigma(t_k)|$ for $t \in [t_1, t_k]$. Then from (1.1) and (1.5) we have

(2.5)

$$|\sigma(t_k)| \le R(t_k,t_1) + K\int_{t_1}^{t_k} |w(t_k,s)\|\sigma(s-\tau(s))|\mathrm{d}s \le R(t_k,t_1) + KW(t_k)|\sigma(t_k)|,$$

where

$$R(t_k,t_1) = |f(t_k)| + K\int_0^{t_1} |w(t_k,s)\|\sigma(s-\tau(s))|\mathrm{d}s.$$

Observing that $W(t_k) \le W_+(t_k), W_+(t_k) \to W_\infty, R(t_k,t_1) \to 0$ as $t_k \to \infty$, and, by (2.4), $KW_\infty < 1$, we find that inequality (2.5) cannot hold for large k. The contradiction obtained shows that there exists a sequence $t_m, m = 1,2,\ldots$, such that $t_m \to \infty$ as $m \to \infty$ and $|\sigma(t_m)| \ge |\sigma(t)|$ for $t \in [t_m,\infty)$. Evidently, $|\sigma(t_m)| \ge |\sigma(t_{m+1})| \ge 0$; therefore there exists $\sigma_\infty = \lim|\sigma(t_m)|$ as $t_m \to \infty$. Let us prove that $\sigma_\infty = 0$.

By assumption, $\tau(t) \le h$ for some h. Assuming $t_m - t_i \ge h$, analogously (2.5), we find

(2.6)
$$|\sigma(t_m)| \le R(t_m, t_i) + K \int_{t_i}^{t_m} |w(t_m, s)||\sigma(s - \tau(s))|\mathrm{d}s \le R(t_m, t_i) + KW(t_m)|\sigma(t_i)|.$$

Since the sequence $|\sigma(t_m)|, m = 1, 2, \dots$, is convergent, then for any $\varepsilon > 0$, there exists i such that $|\sigma(t_i| - |\sigma(t_m| < \varepsilon$ for all $m > i$. Therefore, from (2.6) we find

$$|\sigma(t_m)|[1 - KW(t_m)] < R(t_m, t_i) + \varepsilon KW(t_m). \tag{2.7}$$

Since $R(t_m, t_i) < \varepsilon$ for large $t_m - t_i$, $\lim KW_+(t_m) = KW_\infty < 1$ as $t_k \to \infty$, and $W(t_m) \le W_+(t_m)$, then $KW(t_m) < 1$ for large m. Therefore, inequality (2.7) is true only if $|\sigma(t_m)| \to 0$ as $m \to \infty$, i.e., $|\sigma(t)| \to 0$ as $t \to \infty$. □

Suppose that along with (2.1),

$$|f(t)| \le C \exp(-\Delta_1 t) \tag{2.8}$$

for some $\Delta_1 > 0$. In particular, if $f(t)$ is a solution of the linear system in the absence of external perturbations, then $\Delta_1 = \Delta$.

DEFINITION 2. *System* (1.1) *is called absolutely exponentially stable in the class* $\Phi(K_1, K_2)$ *if for any functions* $\varphi(\sigma, t), f(t)$, *satisfying conditions* (1.3), (2.8) *and any piecewise continuous nonnegative bounded for* $t \in [0, \infty)$ *function* $\tau(t)$, *there exists a constant* $\beta > 0$ *such that for some* C, *the corresponding solution* $\sigma(t)$ *of* (1.1) *satisfies the inequality*

$$|\sigma(t)| \le C \exp(-\beta t).$$

The infimum of $\lambda = -\beta$ for which the value C exists is called the Lyapunov exponent of the function $\sigma(t)$ (see, for example, [16]).

The following theorem establishes an upper bound for the Lyapunov exponent of the solutions $\sigma(t)$ of system (1.1), (1.5).

Let β_* be the root of the equation

$$W_\infty(\beta, h) = 1/K, \tag{2.9}$$

where

$$W_\infty(\beta, h) = \overline{\lim_{t\to\infty}} W_+(\beta, h, t), \quad W_+(\beta, h, t) = \int_0^t \exp[\beta(t + h - s)]|w(t, s)|\mathrm{d}s.$$

THEOREM 2. *The Lyapunov exponent of the solution* $\sigma(t)$ *satisfies the inequality*

$$\lambda \le -\beta_*. \tag{2.10}$$

Proof. Setting in (1.1)

$$\sigma(t) = \exp(-\beta t) y(t), \tag{2.11}$$

where $\beta \in (0, \beta_* < \Delta_1)$, we obtain

$$y(t) = \exp(\beta t) f(t) + \exp(\beta t) \int_0^t w(t, s) \varphi[\exp(-\beta s) y(s - \tau(s)), s]\mathrm{d}s \tag{2.12}$$

whence analogously (2.5) we have

$$|y(t_k)| \le R(\beta, t_k, t_1) + KW(\beta, h, t_k)|y(t_k)|. \tag{2.13}$$

Clearly, $W_\infty(\beta, h)$ increases with β, so, in view of (2.9), $KW_\infty(\beta, h) < 1$ for $\beta < \beta_*$. Therefore, analogously to the proof of Theorem 1, we find that $y(t)$ is bounded on $(0, \infty)$, which, along with (2.11), proves the theorem. □

Let us now obtain a condition guaranteeing instability of the system. To this end, we put

$$W^0(t) = \int_0^t w(t, s)\, \mathrm{d}s. \tag{2.14}$$

Suppose there exists

$$W_0 = \lim_{t\to\infty} W^0(t) \ne 0. \tag{2.15}$$

THEOREM 3. *If*

$$K \ge K_0 = 1/|W_0|, \tag{2.16}$$

then system (1.1), (1.5) *is unstable.*

Proof. Let us put

$$f_0(t) = 1 - W^0(t)/W_0, \qquad \varphi_0(\sigma) = K_0\sigma \operatorname{sgn} W_0, \qquad \sigma(t) \equiv 1 \quad \text{for } t < 0. \tag{2.17}$$

In view of (2.15) and (2.17), $|f_0(t)| \to 0$ as $t \to \infty$; by (2.16), $\varphi_0(\sigma) \in \Phi(-K, K)$. By a direct substitution one can check that $\sigma(t) \equiv 1$ is the corresponding solution of (1.1). Since it does not satisfy condition (1.4), the system is unstable.

Let K_b be the value of the constant K such that the system is stable in the class $\Phi(-K, K)$ for $K < K_b$ and unstable for $K \ge K_b$. Then from Theorems 1 and 3 it follows that K_b satisfies the inequality

$$1/W_\infty \le K_b \le 1/|W_0|. \tag{2.18}$$

3. Systems with sign-constant transfer function. Suppose now that the transfer function $w(t, s)$ is sign-constant. Without loss of generality, we assume that

$$w(t, s) \ge 0 \quad \text{for} \quad t \ge s \ge 0 \tag{3.1}$$

because the case $w(t, s) \le 0$ is reduced to (3.1) by substitution: $w_1(t, s) = -w(t, s)$, $\varphi_1(\sigma, t) = -\varphi(\sigma, t)$.

THEOREM 4. *System* (1.1), (3.1) *is absolutely stable in the class* $\Phi(-K, K)$ *if it is stable for* $\varphi = K\sigma(t - \tau(t))$.

Proof. Let $\sigma_0(t)$ be the solution of the equation

$$\sigma_0(t) = f_0(t) + K\int_0^t w(t, s)\sigma_0(s - \tau(s))\, \mathrm{d}s, \tag{3.2}$$

where

$$\sigma_0(t) = |\sigma(t)| \text{ for } t < 0, \qquad f_0(t) = |f(t)| + \exp(-t) \text{ for } t \ge 0. \tag{3.3}$$

Clearly, $\sigma_0(t-\tau(s)) > |\sigma(t-\tau(s))| > 0$ for sufficiently small $t > 0$. Let us show that this inequality cannot break as t increases. In fact, let $\sigma_0(t_1-\tau(t_1)) = \sigma(t_1-\tau(t_1))$ for some t_1; then, subtracting (1.1) from (3.2), we find

(3.4)
$$0 = |f(t_1)| - f(t_1) + \exp(-t_1) + \int_0^{t_1} w(t_1,s)[K\sigma_0(s-\tau(s)) - \varphi(\sigma(s-\tau(s)),s)]\,\mathrm{d}s,$$

which is impossible, because the right-hand side of (3.4) is positive ($K\sigma_0 > \varphi(\sigma)$ for $\sigma_0 > |\sigma|$).

If $\sigma_0(t_1) = -\sigma(t_1)$, then, summing (3.2) and (1.1), we find

$$0 = |f(t_1)| + f(t_1) + \exp(-t_1) + \int_0^{t_1} w(t_1,s)[K\sigma_0(s-\tau(s)) + \varphi(\sigma(s-\tau(s)),s)]\,\mathrm{d}s,$$

where the right-hand side is also positive.

The obtained contradictions show that $\sigma_0(t) > |\sigma(t)|$ for $t > 0$ and, therefore, $\sigma(t) \to 0$ as $t \to \infty$.

Suppose, moreover, that limit (2.15) exists.

THEOREM 5. *For absolute stability of system* (1.1), (3.1), *it is necessary and sufficient that*

$$K < 1/W_\infty. \tag{3.5}$$

In fact, by (3.1), $W(t) = W^0(t), W_\infty = W_0$, so Theorem 5 follows directly from inequality (2.18).

4. Discussion. The obtained results embrace a wide range of control systems with, generally, time-varying linear block and arbitrary delay $\tau(t)$ in the feedback. In accordance with sufficient stability condition (2.4), such a system is absolutely stable in the class $\Phi(-W_\infty^{-1}+\varepsilon, W_\infty^{-1}-\varepsilon)$ where $\varepsilon > 0$ is an arbitrary small value. If the function $f(t)$ in (1.1) exponentially tends to zero, the stability is exponential; the corresponding Lyapunov exponent satisfies inequality (2.10) (see Theorem 2). If limit (2.15) exists, the system is certainly unstable in the wider class $\Phi(-W_0^{-1}, W_0^{-1})$ (see Theorem 3).

Note that these results can be extended to the case when the nonlinearity bounds are time-dependent, i.e.,

$$-K(t)\sigma^2 \le \varphi(\sigma,t)\sigma \le K(t)\sigma^2, \quad K(t) \ge 0. \tag{4.1}$$

In fact, as is clear from the proofs, it is only necessary to replace KW_∞ and KW_0 in the above conditions by

$$\overline{\lim_{t\to\infty}} \int_0^t |K(t)w(t,s)|\mathrm{d}s \quad \text{and} \quad \overline{\lim_{t\to\infty}} \int_0^t K(t)w(t,s)\,\mathrm{d}s, \tag{4.2}$$

respectively.

As is known, the Lur'e problem was first formulated for the system

$$\dot{x} = Ax + b\varphi(cx), \tag{4.3}$$

where $x \in R^n$, and b and c are column and row vectors, respectively. The problem is reduced to (1.1), where $\varphi = \varphi(\sigma)$, $\sigma = cx$, and $w(t,s) = w(t-s)$, because (4.3) is time-invariant.

In 1949 Aizerman conjectured [12] that system (4.3) is absolutely stable in the class $\varphi(\sigma) \in \Phi(K_1, K_2)$, provided that the linear system $\dot{x} = Ax + kbcx$ is stable for any $k \in [K_1, K_2]$. Subsequently, counterexamples showed that this conjecture is, in general, false (the history of the Aizerman conjecture can be found in Gil' [13]). So the problem is to find classes of systems satisfying the Aizerman conjecture. The first result in this direction was obtained by Gil' [13], who proved that if in system (4.3) the transfer function is nonnegative, then its absolute stability in the class $\Phi(0, K)$ is guaranteed by stability of the system $\dot{x} = Ax + Kbcx$. Recently he extended this result to distributed and delay time-invariant systems [14, 15].

In [11] time-variable systems with a nonnegative transfer function were considered via direct analysis of the corresponding Volterra equation. As a result, it was shown that stability of such a system for $\varphi(\sigma, t) = K\sigma$ guarantees absolute stability in the class $\Psi(-K, K)$. Theorem 4 of the present paper extends this result to systems with arbitrary delay $\tau(t)$ in the feedback. Namely, if the transfer function $w(t, s)$ is nonnegative, then for absolute stability of the system in the class $\Phi(-K, K)$, it is necessary and sufficient that it is asymptotically stable for $\varphi = K\sigma(t - \tau(t))$. If in (2.15) the limit W_0 exists (in particular, if the linear block is time-invariant), the precise bound for the stability sector equals $K = 1/W_\infty$ (Theorem 5) for any delay $\tau(t)$. Note that at first sight the invariance of the stability sector on $\tau(t)$ looks surprising; however, this is due to the fact that for $K = 1/W_\infty$ and $f(t)$, determined by (2.17), equation (1.1) admits the "unstable" solution $\sigma(t) \equiv 1$ for any $\tau(t)$.

When analyzing the Lur'e problem, the nonlinearity class $\Phi(K_1, K_2)$ is usually reduced to $\Phi(0, K)$ ($K = K_2 - K_1$) by substitution: $\varphi_1(\sigma, t) = \varphi(\sigma, t) - K_1\sigma$. It can be shown that on the transformation $\varphi_1(\sigma, t) = \varphi(\sigma, t) - c\sigma$, the corresponding transfer function $w(t, s, c)$ increases with c, provided that $w(t, s, c) > 0$ for $t \geq s \geq 0$. So, if for $c = K_1$ ($\Phi = \Phi(0, K)$) condition (3.1) is satisfied, then it certainly holds for $c = 0.5(K_1 + K_2) > K_1$ ($\Phi = \Phi(-K, K)$). The converse is not, in general, true; i.e., condition (3.1), which is valid in the class $\Phi(-K, K)$, may be lost on the transformation to the class $\Phi(0, K)$.

Clearly, on the transformation $\varphi_1(\sigma, t) = \varphi(\sigma, t) - c\sigma$, the lower bound of the stability sector, provided by Theorem 4, becomes $K(c) = -K + 2c$. Thus, the largest stability sector corresponds to the minimal value $c < 0$ for which the transfer function $w(t, s, c)$ is still nonnegative.

Let $[K_1^0, K_2^0]$ be the Hurwitz angle of linear system (1.1) with $\varphi(\sigma, t) = k\sigma$ (i.e., it is asymptotically stable for $K_1^0 < k < K_2^0$ and unstable or not asymptotically stable for $k = K_1^0$ and $k = K_2^0$). If the transfer function is nonnegative and the limit W_0 exists, then from Theorem 5 it follows that $K_2^0 = 1/W_\infty$. Since, by Theorem 4, stability for $\varphi(\sigma, t) = K\sigma$ provides stability in the class $\Phi(-K, K)$, then $K_1^0 \leq -K_2^0$. A more detailed analysis shows that $K_1^0 < -K_2^0$.

Let the linear block be a closed-loop system consisting of n (generally, time-varying) links. Suppose that the individual transfer functions $w_i(t, s), i = 1, \ldots, n$, are sign-constant. For a sign-constant input, the output of each link is sign-constant as well; therefore, so is the transfer function $w(t, s)$ of the entire linear block.

Suppose, in particular, that the links are of the first order; i.e., the linear block is described by the equations

$$\begin{aligned} &\dot{x}_1 + a_1(t)x_i = 0, \\ &\dot{x}_1 + a_i(t)x_i = k_i x_{i-1}, i = 2, \ldots, n. \end{aligned} \tag{4.4}$$

An extensive literature is devoted to an analysis of feasibility of the Aizerman

conjecture in closed-loop systems with the first order time-invariant links and time-invariant feedback $\varphi(\sigma)$. In particular, Bergen and Williams proved [17] that systems of the third order satisfy the Aizerman conjecture. Trukhan extended this result on systems with up to five stable links of the first order [18]. For an arbitrary number of links, the transfer function is positive, so stability in the class $\Phi(0, K)$ follows from the theorem of Gil' [13]. Let us show that the above findings enable us to essentially generalize these results in some respects.

Evidently, the individual transfer function of a link,

$$w_i(t,s) = \exp\left[-\int_s^t a_i(s)\mathrm{d}s\right],\ i = 1, \dots, n, \tag{4.5}$$

is positive; hence, the transfer function of time-varying system (4.4) is positive as well. So, from Theorem 4 it follows that system (4.4) with the feedback $\varphi(\sigma(t - \tau(t)), t)$ is absolutely stable in the class $\Phi(-K, K)$, provided that it is stable for $\varphi = K\sigma(t - \tau(t), t)$. If in (2.15) the limit W_0 exists, then for any prescribed delay $\tau(t)$, the obtained bound for the stability sector coincides with the upper bound of the Hurwitz angle, i.e., $K_* = 1/W_\infty = K_2^0$ (Theorem 5). However, as is mentioned above, the lower bound of the obtained stability sector, $-1/W_\infty$, is larger than K_1^0 (note that the results [17, 18] provide stability of the particular systems in the whole Hurwitz angle).

Consider now the second order system

$$\begin{gathered}\ddot{x}(t) + 2h\dot{x}(t) + (1 + h^2)x(t) + \varphi(x(t - \tau(t)), t) = 0,\\ -Kx^2 \le \varphi(x,t)x \le Kx^2, \quad \varphi(0,t) = 0.\end{gathered} \tag{4.6}$$

Here

$$w(t,s) = \exp[-h(t-s)]\sin(t-s), \qquad W^0(t) = 1/1 + h^2\exp(-ht)[h\sin t - \cos t];$$

thus $W_0 = 1/(1 + h^2)$. Integrating (2.2), we find (see [11]) $W_\infty = [1 + \exp(-\pi h)]/[1 + \exp(-\pi h)]$.

By Theorems 1, 2, and 3, for any $\tau(t)$, system (4.6) is absolutely exponentially stable if $K < K_* = 1/W_\infty = \coth(-\pi h/2)$ and unstable for $K \ge K_0 = 1 + h^2$.

The functions $K_*(h)$ and $K_0(h)$ are plotted in Figure 1; as shown, they approach each other as h increases. This, in particular, testifies that the impact of a delay $\tau(t)$ in the feedback on the system stability decreases with an increase of stability of the linear part.

Note that the obtained delay-independent stability condition is precise, because for $K_* = \coth(-\pi h/2)$, the function $\tau(t)$ can be found such that (4.6) is unstable. In fact, let us put

$$\varphi(x(t - \tau(t)), t) = K_* x(t - \tau(t)), \quad \tau(t) = t \text{ for } t \in [0, \pi), \quad \tau(t + \pi) = \tau(t). \tag{4.7}$$

Setting $x(0) = -1$, $\dot{x}(0) = 0$ and observing that $\varphi(x(t - \tau(t)), t) = K_*$ for $t \in [0, \pi)$, we find that the corresponding solution is

$$x(t) = \exp(-ht)\left(-\frac{K_*}{1 + h^2} - 1\right)(\cos t + h\sin t) + \frac{K_*}{1 + h^2}. \tag{4.8}$$

Setting in (4.8) $K_* = \coth(-\pi h/2)$, we obtain $x(\pi) = 1$, $\dot{x}(\pi) = 0$. Clearly, $\varphi(x(t - \tau(t)), t) = -K_*$ for $t \in [\pi, 2\pi)$, so analogously we find $x(2\pi) = -1$, $\dot{x}(2\pi) = 0$. Thus, the corresponding solution of (4.6) is 2π-periodic; therefore, in accordance with the above definition, the system is unstable.

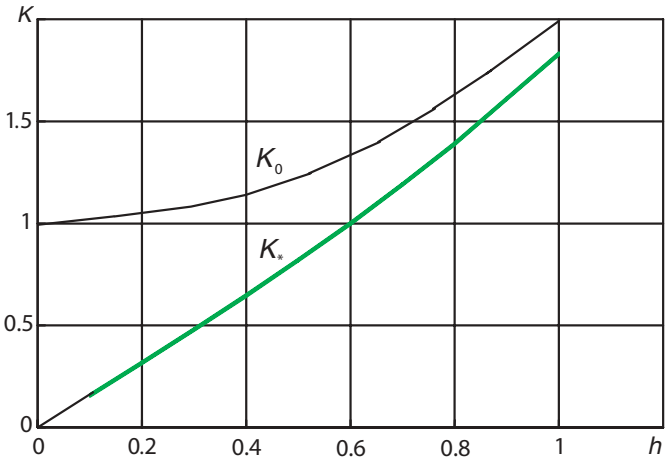


FIG. 1.

Acknowledgment. The authors are grateful to the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] A. I. LUR'E, *Some Nonlinear Problems in the Theory of Automatic Control*, H.M. Stationery Off., London, 1957.
[2] V. M. POPOV, *Absolute stability of nonlinear systems of automatic control*, Autom. Remote Control, 22 (1962), pp. 857–875.
[3] R. E. KALMAN, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Nat. Acad. Sci. USA, 49 (1963), pp. 201–205.
[4] M. A. AIZERMAN AND F. R. GANTMACHER, *Absolute Stability of Regulator Systems*, Holden-Day, San Francisco, London, Amsterdam, 1964.
[5] S. LEFSCHETZ, *Stability of Nonlinear Control Systems*, Math. Sci. Engrg. 13, Academic Press, New York, London, 1965.
[6] E. S. PYATNITSKY, *New studies in absolute stability of automatic control systems. Survey*, Avtomat. i Telemekh., 6 (1968), pp. 5–36 (English translation in Automat. Remote Control).
[7] J. C. WILLEMS, *The Analysis of Feedback Systems*, MIT Press, Cambridge, MA, 1971.
[8] K. S. NARENDRA AND J. F. TAYLOR, *Frequency Domain Criteria for Absolute Stability*, Academic Press, New York, 1973.
[9] M. G. SAFONOV, *Stability and Robustness of Multivariable Feedback Systems*, MIT Press, Cambridge, MA, 1980.
[10] V. A. YAKUBOVICH, *Dichotomy and absolute stability of nonlinear systems with periodically nonstationary linear part, Systems Control Lett.*, 11 (1988), pp. 221–228.
[11] A. A. ZEVIN AND M. A. PINSKY, *A new approach to the Lur'e problem in the theory of absolute stability*, SIAM J. Control Optim., 42 (2003), pp. 1895–1904.
[12] M. A. AIZERMAN, *On a conjecture from absolute stability theory*, Uspekhi Mat. Nauk, 4 (1949), pp. 25–49 (in Russian).
[13] M. I. GIL', *On one class of absolutely stable systems*, Soviet Phys. Dokl., 28 (1983), pp. 811–816.
[14] M. I. GIL', *Stability of Finite- and Infinite-Dimensional Systems*, Kluwer Internat. Ser. Engrg. Comput. Sci. 455, Kluwer Academic Publishers, Boston, 1998.
[15] M. I. GIL', *Boundedness of solutions of nonlinear differential delay equations with positive Green functions and the Aizerman–Myshkis problem*, Nonlinear Anal., 49 (2002), pp. 1065–1078.
[16] E. A. BARBASHIN, *Introduction to the Theory of Stability*, Wolters–Noordhoff, Groningen, The Netherlands, 1970.
[17] A. R. BERGEN AND S. WILLIAMS, *Verification of Aizerman's conjecture for a class of third-order systems*, IRE Trans., AC-7 (1962), pp. 42–46.
[18] N. M. TRUKHAN, *On single-loop systems absolutely stable in the Hurwitzian angle*, Avtomat. i Telemekh., 11 (1966), pp. 5–8 (English translation in Automat. Remote Control).

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2009–2019

A CORRECTED PROOF OF THE STOCHASTIC VERIFICATION THEOREM WITHIN THE FRAMEWORK OF VISCOSITY SOLUTIONS*

FAUSTO GOZZI†, ANDRZEJ ŚWIĘCH‡, AND XUN YU ZHOU§

Abstract. We present a full and corrected proof of the stochastic verification theorem that was first obtained by Zhou, Yong, and Li [SIAM J. Control Optim., 35 (1997), pp. 243–253].

Key words. stochastic optimal control, verification theorem, HJB equation, viscosity solution, superdifferential, Lebesgue point

AMS subject classifications. 93E20, 49L20, 49L25

DOI. 10.1137/S0363012903428184

1. Introduction. The dynamic programming approach is applied to optimal control problems, deterministic and stochastic alike, primarily via the so-called verification theorem. The verification theorem verifies whether a given admissible control is optimal and, more importantly, suggests a way of constructing optimal feedback controls. The classical version of the verification theorem (see, e.g., [4]), as with the dynamic programming at large, has the inherent deficiency of having to assume the smoothness of the value function, whereas it is by now well known that even the simplest optimal control problem may not possess a smooth value function. Therefore, it is imperative, to both the theoretical development and practical applications, that a nonsmooth version of the verification theorem be available. A verification theorem for deterministic optimal control within the framework of one of the nonsmooth analyses—namely, viscosity solutions—was put forward by Zhou in [9]. The stochastic version of the result in [9] was first correctly stated in [10], although with an incorrect proof. Specifically, in a key step in the proof of [10], a test function for the superdifferential was employed and Ito's formula was applied. However, the superdifferential is pointwisely associated with probability sample points, and thus the test function also implicitly depends on the sample points. In this case, Ito's formula would not apply. Another proof was presented in the book [8, Chapter 5, section 5.2]. It fixed the aforementioned problem of the original proof of [10] by incorporating the conditional probability, but unfortunately it still contained some, albeit technically subtle, gaps (see Remark 4.2 below for details). In this paper we give a corrected proof of the result in [10]. The proof presented here is based on arguments of the one in [8] and may at first appear to be very similar. However, it has a few delicate differences that take care of the points that were missing in the proof of [8]. Since the verification theorem is a result of fundamental importance, we think it is worthwhile to present

*Received by the editors May 22, 2003; accepted for publication (in revised form) December 22, 2003; published electronically April 14, 2005.
http://www.siam.org/journals/sicon/43-6/42818.html

†Facoltà di Economia of LUISS, Roma, Italy (fgozzi@luiss.it). This author was supported by the "Progetto di Ateneo" (1999–2001) on "Methods of static and dynamic optimization in Economics and Finance," Università di Roma "La Sapienza."

‡School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (swiech@math.gatech.edu). This author was supported by NSF grant DMS 0098565.

§Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk). This author was supported by RGC earmarked grants CUHK 4234/01E.

its full and detailed proof even though it may repeat some of the material available in the literature.

2. The stochastic optimal control problem. In this section we introduce a class of stochastic optimal control problems that we consider in the paper. We will use the following basic notation throughout the paper. A' will denote the transpose of a matrix A, and $\mathbb{R}^{n\times k}$ and $S^{n\times n}$ will denote, respectively, the set of all $n\times k$ matrices and the set of all $n\times n$ symmetric matrices. Given a probability space $(\Omega,\mathcal{F},P)$ with a filtration $\{\mathcal{F}_t : a\le t\le b\}$ $(-\infty<a<b<+\infty)$, a separable Banach space Z with norm $|\cdot|_Z$, and $1\le p<+\infty$, we define the set $L^p_{\mathcal{F}_t}(a,b;Z)=\{\phi(\cdot)=\{\phi(t,\omega) : a\le t\le b\}|\ \phi(\cdot)$ is an $\mathcal{F}_t$-adapted, Z-valued measurable process on $[a,b]$, and $\mathbb{E}\int_a^b|\phi(t,\omega)|_Z^p dt<+\infty\}$.

Let $T\in(0,+\infty)$ be a finite time horizon and let $(\Omega,\mathcal{F},\mathbb{P})$ be a complete probability space. For any initial data $s\in[0,T]$ and $y\in\mathbb{R}^n$, the state equation of the problem is

$$\text{(1)}\qquad \begin{cases} dx(t)=b(t,x(t),u(t))dt+\sigma(t,x(t),u(t))dW(t),\\ x(s)=y, \end{cases}$$

where W is an m-dimensional standard Brownian motion defined on $(\Omega,\mathcal{F},\mathbb{P})$, and $u(\cdot)$ (the control strategy) is a stochastic process with values in a control space U, adapted to the filtration (augmented by $\mathbb{P}$-null sets) generated by W. The functional to minimize is

$$J(s,y;u(\cdot))=\mathbb{E}\left[\int_s^T f(t,x(t),u(t))dt+h(x(T))\right],$$

where $u(\cdot)$ is a given control strategy and $x(\cdot)$ is the corresponding state trajectory solving (1). The infimum will be taken over a suitable set of admissible controls that will be defined below. Throughout the paper we will always assume that b,σ,f, and h satisfy the following hypothesis. (We use the symbol $|\cdot|$ to denote norms in various spaces.)

Hypothesis 2.1. The functions

$$\begin{aligned} &b:[0,T]\times\mathbb{R}^n\times U\to\mathbb{R}^n, && \sigma:[0,T]\times\mathbb{R}^n\times U\to\mathbb{R}^{n\times m},\\ &f:[0,T]\times\mathbb{R}^n\times U\to\mathbb{R}, && h:\mathbb{R}^n\to\mathbb{R} \end{aligned}$$

are uniformly continuous. Moreover, there exists a constant L such that for functions $\phi=b,\sigma,f,h$ we have

(i) $|\phi(t,x_1,u)-\phi(t,x_2,u)|\le L|x_1-x_2|$ $\forall t\in[0,T]$, $\forall x_1,x_2\in\mathbb{R}^n$, $\forall u\in U$;
(ii) $|\phi(t,0,u)|\le L$.

We will work, as in [8, Chapter 5, section 4], with the so-called weak formulation of control problems. For a given $s\in[0,T]$, we define the set of (weakly) admissible controls $U^w_{ad}[s,T]$ to be the collection of all 5-tuples $(\Omega,\mathcal{F},\mathbb{P},W,u(\cdot))$ satisfying the following conditions:

(i) $(\Omega,\mathcal{F},\mathbb{P})$ is a standard probability space.
(ii) $W=\{W(t) : s\le t\le T\}$ is an m-dimensional standard Brownian motion defined on $(\Omega,\mathcal{F},\mathbb{P})$ over $[s,T]$ (with $W(s)=0$ almost surely), and $\mathcal{F}^s_t$ is the augmented filtration generated by W. Moreover, $\mathcal{F}=\sigma(\cup_t\mathcal{F}^s_t)$ (see [6, p. 285] for details and the definition of augmented filtration).

(iii) $u : [s,T] \times \Omega \to U$ is a U-valued $\mathcal{F}_t^s$-adapted process where U is a given complete separable metric space.

We will write $(\Omega, \mathcal{F}, \mathbb{P}, W, u(\cdot)) \in U_{ad}^w[s,T]$, but occasionally we will write only $u(\cdot) \in U_{ad}^w[s,T]$ if no ambiguity arises.

We remark that in light of Hypothesis 2.1 for every $(s,y) \in [0,T] \times \mathbb{R}^n$ and $(\Omega, \mathcal{F}, \mathbb{P}, W, u(\cdot)) \in U_{ad}^w[s,T]$, there exists a unique solution $x(\cdot)$ of (1) (see, for instance, [8, Theorem 6.16]). It is called the state corresponding to the control strategy $u(\cdot) \in U_{ad}^w[s,T]$, and $(x(\cdot), u(\cdot))$ is called an admissible pair. The objective of the optimal control problem is to minimize the cost functional $J(s,y;u(\cdot))$ (for a given $(s,y) \in [0,T) \times \mathbb{R}^n$) over all $u(\cdot) \in U_{ad}^w[s,T]$. We denote the above problem by $C_{s,y}$ to emphasize the dependence on the initial time s and the initial state y. The value function for the problem is defined as

$$V(s,y) = \inf_{u(\cdot) \in U_{ad}^w[s,T]} J(s,y;u(\cdot)).$$

An admissible pair $(x^*(\cdot), u^*(\cdot))$ is called *optimal* for $C_{s,y}$ if $u^*(\cdot)$ achieves the minimum of $J(s,y;u(\cdot))$ over $U_{ad}^w[s,T]$.

In the dynamic programming approach to the problem, one proves that V is the unique classical solution of an associated Hamilton–Jacobi–Bellman (HJB) equation if $V \in C^{1,2}([0,T] \times \mathbb{R}^n)$. Defining the current value Hamiltonian H_{CV} as

$$H_{CV}(t,x,q,Q;u) = \frac{1}{2}\mathrm{tr}(\sigma(t,x,u)'Q\sigma(t,x,u)) + \langle q, b(t,x,u)\rangle + f(t,x,u)$$

and the infimum value Hamiltonian H_{MIN} as

$$H_{MIN}(t,x,q,Q) = \inf_{u \in U} H_{CV}(t,x,q,Q;u),$$

the HJB equation associated with the family of problems $\{C_{s,y}\}_{(s,y)\in[0,T]\times\mathbb{R}^n}$ is

$$(2) \quad \begin{cases} -v_t(t,x) - H_{MIN}\left(t,x,Dv(t,x),D^2v(t,x)\right) = 0, \ (t,x) \in (0,T) \times \mathbb{R}^n, \\ v(T,x) = h(x), \qquad x \in \mathbb{R}^n, \end{cases}$$

where Dv and D^2v denote, respectively, the gradient and Hessian of v with respect to x.

As a part of the dynamic programming approach, the so-called verification technique plays an important role in testing for the optimality of a given admissible pair and (more importantly) in constructing optimal feedback controls. The classical verification theorem is as follows (see Fleming and Rishel [4, Theorem VI.4.1] and Yong and Zhou [8, Theorem 5.1, p. 268]).

THEOREM 2.2. *Let $W \in C^{1,2}([0,T] \times \mathbb{R}^n)$ be a solution of the HJB equation* (2). *Then we have the following.*

(a) *$W(s,y) \le J(s,y;u(\cdot))$ for any $(s,y) \in [0,T] \times \mathbb{R}^n$ and any $u(\cdot) \in U_{ad}^w[s,T]$.*

(b) *Suppose a given admissible pair $(x^*(\cdot), u^*(\cdot))$ for the problem $C_{s,y}$ satisfies*

$$(3) \quad -W_t(t,x^*(t)) - H_{CV}\left(t,x^*(t),DW(t,x^*(t)),D^2W(t,x^*(t));u^*(t)\right) = 0$$

$\mathbb{P}$-a.s., a.e. $t \in [s,T]$. Then $(x^*(\cdot), u^*(\cdot))$ is an optimal pair for the problem $C_{s,y}$.

Remark 2.3. Equality (3) is equivalent to the more familiar

$$H_{CV}\left(t,x^*(t),DW(t,x^*(t)),D^2W(t,x^*(t));u^*(t)\right) = H_{MIN}\left(t,x^*(t),DW(t,x^*(t)),D^2W(t,x^*(t))\right).$$

However, as is generally well known, the value function is not smooth and the HJB equation may not have a smooth solution at all. The verification theorem, whose proof is presented in this paper, is a nonsmooth version of Theorem 2.2. We refer the reader to [10, 8] for more discussion on this verification theorem and its consequences.

3. Preliminaries: Stochastic processes, Lebesgue points, and viscosity solutions. In this subsection we present some preliminary results needed in the main proof.

3.1. Viscosity solutions and semidifferentials. To make the paper self-contained, we present the definition of viscosity solution we will be using. It is slightly different from the typical and most commonly used definition (see [1, 5, 8]).

First, by a parabolic neighborhood of a point (t_0, x_0), we mean the intersection of an open neighborhood of (t_0, x_0) with the set of all (t, x) such that $t \geq t_0$.

DEFINITION 3.1. *An uppersemicontinuous (respectively, lowersemicontinuous) function v on $(0,T] \times \mathbb{R}^n$ is called a viscosity subsolution (respectively, supersolution) of* (2) *if*

$$v(T,x) \leq (\geq) h(x) \qquad \text{for } x \in \mathbb{R}^n$$

and if

$$-\phi_t(t_0,x_0) - H_{MIN}(t_0,x_0,D\phi(t_0,x_0),D^2\phi(t_0,x_0)) \leq (\geq) 0 \tag{4}$$

whenever $v - \phi$ attains a local maximum (respectively, minimum) at (t_0, x_0) in a parabolic neighborhood of (t_0, x_0) for $\phi \in C^{1,2}((0,T) \times \mathbb{R}^n)$. A function v is called a viscosity solution of (2) *if it is both a viscosity subsolution and a viscosity supersolution of* (2).

The typical definition uses full (two-sided) neighborhoods instead of parabolic (one-sided) ones. However, it is well known that both definitions are equivalent. To see this, suppose that v is a viscosity subsolution in the two-sided sense and that $v - \phi$ has a one-sided local maximum at (t_0, x_0), which can obviously be assumed to be strict. Then it is easy to see that for small $\mu > 0$, the functions $v - \phi - \frac{\mu}{t - t_0}$ have two-sided local maxima at points (t_μ, x_μ) such that $(t_\mu, x_\mu) \to (t_0, x_0)$ as $\mu \to 0$. Therefore from the two-sided definition we get

$$-\phi_t(t_\mu,x_\mu) - H_{MIN}(t_\mu,x_\mu,D\phi(t_\mu,x_\mu),D^2\phi(t_\mu,x_\mu)) \leq -\frac{\mu}{(t-t_\mu)^2} \leq 0$$

and, letting $\mu \to 0$, we obtain (4).

Definition 3.1 is equivalent to the one in which test functions are replaced by parabolic semijets, i.e., second order parabolic sub- and superdifferentials (see [1]).

DEFINITION 3.2. *We say that $(p,q,Q) \in D^{1,2,+}_{t^+,x} v(t,x)$, the second order one-sided parabolic superdifferential of v at (t,x), if, for all $y \in \mathbb{R}^n, s \geq t$,*

$$v(s,y) \leq v(t,x) + p(s-t) + \langle q, y-x\rangle + \frac{1}{2}\langle Q(y-x), y-x\rangle + o(s-t+|y-x|^2).$$

The second order one-sided parabolic subdifferential of v at (t,x), $D^{1,2,-}_{t^+,x} v$ is defined by reversing the inequality, i.e., $D^{1,2,-}_{t^+,x} v(t,x) = -D^{1,2,+}_{t^+,x}(-v)(t,x)$.

The equivalence of Definition 3.1 and the definition in which derivatives of test functions are replaced by elements of super- and subdifferentials are established with the help of a well-known result that we present below and whose proof can be found, for instance, in [8, Chapter 4, Lemma 5.4].

LEMMA 3.3. *Let v be an uppersemicontinuous function on $(0,T)\times\mathbb{R}^n$ and let $(t_0,x_0)\in(0,T)\times\mathbb{R}^n$. Then $(p,q,Q)\in D^{1,2,+}_{t^+,x}v(t_0,x_0)$ if and only if there exists a function $\phi\in C^{1,2}((0,T)\times\mathbb{R}^n)$ such that $v-\phi$ attains a strict global maximum at (t_0,x_0) relative to the set of (t,x) such that $t\geq t_0$ and*

$$(\phi(t_0,x_0),\phi_t(t_0,x_0),D\phi(t_0,x_0),D^2\phi(t_0,x_0))=(v(t_0,x_0),p,q,Q). \tag{5}$$

Moreover, if v has polynomial growth, i.e., if

$$|v(t,x)|\leq C(1+|x|^k)\quad \text{for some } k\geq 1,\ (t,x)\in(0,T)\times\mathbb{R}^n, \tag{6}$$

then ϕ can be chosen so that $\phi,\phi_t,D\phi,D^2\phi$ satisfy (6) *(with possibly different constants C).*

3.2. Lebesgue points. Here we recall and present a few results on Lebesgue points for functions with values in abstract spaces.

DEFINITION 3.4. *Let Z be a Banach space and let $z:[a,b]\to Z$ be a measurable function that is Bochner integrable. We say that t is a right Lebesgue point of z if*

$$\lim_{h\to 0^+}\frac{1}{h}\int_t^{t+h}|z(r)-z(t)|_Z dr=0.$$

We then have the following result (see, for instance, [2, Theorem 9, p. 49]).

LEMMA 3.5. *Let $z:[a,b]\to Z$ be as in Definition* 3.4. *Then the set of right Lebesgue points of z is of full measure in $[a,b]$.*

The following simple lemma ensures that any process in $L^1_{\mathcal{F}_t}(a,b;\mathbb{R}^k)$ for some $k\geq 1$ can be regarded as the Bochner integrable function $z:[a,b]\to L^1(\Omega;\mathbb{R}^k)$.

LEMMA 3.6. *Any $z\in L^1_{\mathcal{F}_t}(a,b;\mathbb{R}^k)$ is Bochner integrable when regarded as a map from $[a,b]$ to $L^1(\Omega;\mathbb{R}^k)$.*

Proof. According to [3, p. 92], if $(\Omega,\mathcal{F},\mathbb{P})$ is a complete probability space, Z is a separable Hilbert space, and $\mathcal{F}$ is countably generated apart from null sets, then the space $L^1(\Omega;Z)$ is separable. Since the filtration generated by a Wiener process is always countably generated (see, for instance, [7, Exercise 4.21, Chapter 1]), the space $L^1(\Omega;\mathbb{R}^k)$ is separable. Since

$$\int_a^b|z(t,\cdot)|_{L^1(\Omega;\mathbb{R}^k)}dt<+\infty,$$

we only need to show that the function z is measurable as a map from $[a,b]$ to $L^1(\Omega;\mathbb{R}^k)$. By the separability of $L^1(\Omega;\mathbb{R}^k)$, the measurability is equivalent to the weak measurability, so it is enough to show that for every function $w\in L^\infty(\Omega;\mathbb{R}^k)$, the map

$$t\to\mathbb{E}(z(t,\omega)\cdot w(\omega))$$

is Lebesgue measurable. But this is obvious from Fubini's theorem, as

$$\mathbb{E}\int_a^b|z(t,\omega)\cdot w(\omega)|dt<+\infty. \qquad \square$$

The following technical result will be used in the proof of the verification theorem.

PROPOSITION 3.7. *Let the standing assumptions of section 2 be satisfied and let* $(s,y) \in [0,T]\times\mathbb{R}^n$. *Given any admissible pair* $(x(\cdot),u(\cdot))$ *for the control problem* $C_{s,y}$, *define the processes*

$$z_1(r) = b(r,x(r),u(r)), \quad z_2(r) = \sigma(r,x(r),u(r))\,\sigma(r,x(r),u(r))'.$$

Then

$$\lim_{h\to 0^+} \mathbb{E}\frac{1}{h}\int_t^{t+h} |z_i(r)-z_i(t)|dr \to 0 \quad as \;\; h\to 0^+, \;\; a.e.\; t\in[s,T], \;\; i=1,2. \tag{7}$$

Proof. It is clear by Hypothesis 2.1 that $z_1, z_2 \in L^1_{\mathcal{F}_t}(s,T;Z)$ with $Z=\mathbb{R}^n$ and $Z=\mathbb{R}^{n\times n}$, respectively. Hence, by Lemma 3.6 they are Bochner integrable when regarded as $L^1(\Omega;Z)$-valued maps. It then follows from Lemma 3.5 that the set of right Lebesgue points of z_1 and z_2, as maps from $[s,T]$ to $L^1(\Omega;Z)$, is of full measure in $[s,T]$, which gives the claim. □

4. Verification theorem. We are going to prove the following theorem.

THEOREM 4.1. *Let* $U\in C((0,T]\times\mathbb{R}^n)$ *be a viscosity subsolution of the HJB* (2) *satisfying* (6) *for some* $k\geq 1$ *and such that* $U(T,x)=h(x)$. *Then we have the following.*

(a) $U(s,y)\leq J(s,y;u(\cdot))$ *for any* $(s,y)\in(0,T]\times\mathbb{R}^n$ *and any* $u(\cdot)\in U^w_{ad}[s,T]$.

(b) *Fix any* $(s,y)\in(0,T)\times\mathbb{R}^n$. *Let* $(x^*(\cdot),u^*(\cdot))$ *be an admissible pair for the problem* $C_{s,y}$. *Suppose that there exists* $(p^*,q^*,Q^*)\in L^2_{\mathcal{F}_t}(s,T;\mathbb{R})\times L^2_{\mathcal{F}_t}(s,T;\mathbb{R}^n)\times L^2_{\mathcal{F}_t}(s,T;S^{m\times n})$ *(where the filtration* $\mathcal{F}_t$ *is* $(\mathcal{F}^s_t)_{t\in[s,T]}$*) such that, for a.e.* $t\in[s,T]$,

$$(p^*(t),q^*(t),Q^*(t))\in D^{1,2,+}_{t^+,x}W(t,x^*(t)),\ \mathbb{P}\text{-}a.s., \tag{8}$$

and

$$\mathbb{E}\int_s^T [p^*(t)+H_{CV}(t,x^*(t),q^*(t),Q^*(t);u^*(t))]\,dt\leq 0. \tag{9}$$

Then $(x^*(\cdot),u^*(\cdot))$ *is an optimal pair for the problem* $C_{s,y}$.

Proof. Part (a) is obvious since $U\leq V$ in view of the standard comparison results for bounded viscosity sub- and supersolutions (see, for instance, [5, Theorem 9.1, Chapter V]). We just have to notice that if U satisfies (6) and K is big enough, then the function

$$U_\delta(t,x) = U(t,x) - \delta e^{K(T-t)}|x|^{k+1}$$

is a viscosity subsolution of (2) for every $\delta>0$ that is bounded from above. Similarly, one can perturb V to a supersolution V_δ since V satisfies (6) with $k=1$ under Hypothesis 2.1 (see [8, p. 178, Proposition 3.1]). The comparison then follows from the comparison for U_δ and V_δ upon letting $\delta\to 0$.

It remains to prove part (b) of the theorem. To simplify the notation, we set

$$f^*(t)=f(t,x^*(t),u^*(t)), \qquad h^*(t)=h(x^*(t)),$$
$$b^*(t)=b(t,x^*(t),u^*(t)), \qquad \sigma^*(t)=\sigma(t,x^*(t),u^*(t)).$$

Fix $t_0\in[s,T]$ such that (8) holds at t_0, and (7) holds at t_0 for $z_1(\cdot)=b^*(\cdot)$ and $z_2(\cdot)=\sigma^*(\cdot)\,\sigma^*(\cdot)'$. The set of such points is of full measure in $[s,T]$ by Proposition 3.7.

We now fix $\omega_0 \in \Omega$ such that the regular conditional probability given $\mathcal{F}_{t_0}^s$, denoted by $\mathbb{P}\left(\cdot|\mathcal{F}_{t_0}^s\right)(\omega_0)$, is well defined (see [6, pp. 84–85] for the definition and properties of conditional expectation and the regular conditional probability). We consider the probability space $\left(\Omega, \mathcal{F}, \mathbb{P}\left(\cdot|\mathcal{F}_{t_0}^s\right)(\omega_0)\right)$. In this probability space the random variables

$$x^*(t_0),\ p^*(t_0),\ q^*(t_0),\ Q^*(t_0)$$

are almost surely constant (see [8, Proposition 2.13, Chapter 1]) and are equal to

$$x^*(t_0,\omega_0),\ p^*(t_0,\omega_0),\ q^*(t_0,\omega_0),\ Q^*(t_0,\omega_0),$$

respectively. Denote $x_0 = x^*(t_0,\omega_0)$. We remark that in this probability space the Brownian motion W is still a standard Brownian motion although now $W(t_0) = W(t_0,\omega_0)$ almost surely. The space is now equipped with a new filtration $(\mathcal{F}_r^{t_0})_{r\in[t_0,T]}$ and the control process $u^*(\cdot)$ is adapted to this new filtration (see [8, p. 179] for more on this). The main point is that for $\mathbb{P}$-a.s., ω_0 the process $x^*(\cdot)$ is a solution of (1) on (t_0,T) in the new probability space $\left(\Omega, \mathcal{F}, \mathbb{P}\left(\cdot|\mathcal{F}_{t_0}^s\right)(\omega_0)\right)$ with the initial condition $x^*(t_0) = x^*(t_0,\omega_0)$.

We will write $\mathbb{E}_{\omega_0}$ to denote the expectation with respect to the measure $\mathbb{P}\left(\cdot|\mathcal{F}_{t_0}^s\right)(\omega_0)$ and will write $L^p_{\mathcal{F}_t,\omega_0}(t_0,T;Z)$ and $L^p_{\omega_0}(\Omega;Z)$ when these spaces are defined with respect to this measure.

Using Lemma 3.3, we now take a test function $\phi \in C^{1,2}((0,T)\times\mathbb{R}^n)$ such that

$$\phi(t,x) > U(t,x) \quad \text{for every } (t,x)\in(0,T)\times\mathbb{R}^n,\ (t,x)\neq(t_0,x_0), \tag{10}$$

$$\begin{aligned}&\left(\phi(t_0,x_0),\phi_t(t_0,x_0),D\phi(t_0,x_0),D^2\phi(t_0,x_0)\right)\\ &= \left(U(t_0,x_0),p^*(t_0,\omega_0),q^*(t_0,\omega_0),Q^*(t_0,\omega_0)\right),\end{aligned} \tag{11}$$

and such that $\phi, \phi_t, D\phi, D^2\phi$ are polynomially bounded; i.e., they satisfy (6). It is important to bear in mind that ϕ is a fixed deterministic function when (t_0,ω_0) is fixed.

We are going to apply Ito's formula to ϕ on the probability space $\left(\Omega, \mathcal{F}, \mathbb{P}\left(\cdot|\mathcal{F}_{t_0}^s\right)(\omega_0)\right)$. To do this, we observe that the processes

$$\phi_t(r,x^*(r)),\quad \langle D\phi(r,x^*(r)),b^*(r)\rangle,\quad \frac{1}{2}\mathrm{tr}\left[\sigma^*(r)'D^2\phi(r,x^*(r))\sigma^*(r)\right]$$

belong to $L^1_{\mathcal{F}_t,\omega_0}(t_0,T;\mathbb{R})$ due to Hypothesis 2.1 on b and σ, to the estimate

$$\mathbb{E}_{\omega_0}\sup_{t_0\le r\le T}|x^*(r)|^l \le K_T(1+|x_0|^l) \quad \text{for every } l\ge 1 \tag{12}$$

(see, for instance, [8, Theorem 6.16, Chapter 1]), and to the polynomial growth of $\phi, \phi_t, D\phi, D^2\phi$. Moreover, by the same argument, the process

$$\sigma^*(r)'\phi_x(r,x^*(r))$$

belongs to $L^2_{\mathcal{F}_t,\omega_0}(t_0,T;\mathbb{R}^n)$. Therefore (see, for instance, [8, Chapter 1, section 5.3]) we can apply Ito's formula to obtain that, for any $h>0$,

$$\begin{aligned}&\phi(t_0+h,x^*(t_0+h)) - \phi(t_0,x^*(t_0))\\ &= \int_{t_0}^{t_0+h}\left[\phi_t(r,x^*(r)) + \langle D\phi(r,x^*(r)),b^*(r)\rangle + \frac{1}{2}\mathrm{tr}\left(\sigma^*(r)'D^2\phi(r,x^*(r))\sigma^*(r)\right)\right]dr\\ &\quad + \int_{t_0}^{t_0+h}\langle D\phi(r,x^*(r)),\sigma^*(r)\,dW(r)\rangle.\end{aligned}$$

In particular, taking the expected value $\mathbb{E}_{\omega_0}$, dividing both sides by h, and using (10), it follows that

$$\begin{aligned}
&\mathbb{E}_{\omega_0}\frac{1}{h}\left[U\left(t_0+h,x^*\left(t_0+h\right)\right)-U\left(t_0,x^*\left(t_0\right)\right)\right]\\
&\leq \mathbb{E}_{\omega_0}\left[\frac{1}{h}\int_{t_0}^{t_0+h}\left[\phi_t\left(r,x^*\left(r\right)\right)+\left\langle D\phi\left(r,x^*\left(r\right)\right),b^*\left(r\right)\right\rangle\right.\right.\\
&\qquad\left.\left.+\frac{1}{2}\mathrm{tr}\left(\sigma^*\left(r\right)'D^2\phi\left(r,x^*\left(r\right)\right)\sigma^*\left(r\right)\right)\right]dr\right].
\end{aligned}\tag{13}$$

We want to pass to the limit as $h\to 0^+$ above. To this end, we treat each term of (13) separately.

(I) First of all, thanks to the continuity of x^* and to the continuity of ϕ_t, we get

$$\frac{1}{h}\int_{t_0}^{t_0+h}\phi_t\left(r,x^*\left(r\right)\right)dr\to\phi_t\left(t_0,x^*\left(t_0\right)\right),\qquad \mathbb{P}\left(\cdot|\mathcal{F}_{t_0}^s\right)\left(\omega_0\right)\text{- a.s.}$$

Now, using estimate (12) and the polynomial growth of ϕ_t, we can apply the dominated convergence theorem to obtain

$$\lim_{h\to 0^+}\mathbb{E}_{\omega_0}\left|\frac{1}{h}\int_{t_0}^{t_0+h}\phi_t\left(r,x^*\left(r\right)\right)dr-\phi_t\left(t_0,x^*\left(t_0\right)\right)\right|=0.$$

(II) With regard to the second term of (13), we have

$$\begin{aligned}
&\mathbb{E}_{\omega_0}\frac{1}{h}\int_{t_0}^{t_0+h}\left\langle D\phi\left(r,x^*\left(r\right)\right),b^*\left(r\right)\right\rangle dr-\left\langle D\phi\left(t_0,x^*\left(t_0\right)\right),b^*\left(t_0\right)\right\rangle\\
&=\mathbb{E}_{\omega_0}\frac{1}{h}\int_{t_0}^{t_0+h}\left\langle D\phi\left(r,x^*\left(r\right)\right)-D\phi\left(t_0,x^*\left(t_0\right)\right),b^*\left(r\right)\right\rangle dr\\
&\quad+\mathbb{E}_{\omega_0}\frac{1}{h}\int_{t_0}^{t_0+h}\left\langle D\phi\left(t_0,x^*\left(t_0\right)\right),b^*\left(r\right)-b^*\left(t_0\right)\right\rangle dr.
\end{aligned}\tag{14}$$

Now

$$\begin{aligned}
&\left|\mathbb{E}_{\omega_0}\frac{1}{h}\int_{t_0}^{t_0+h}\left\langle D\phi\left(r,x^*\left(r\right)\right)-D\phi\left(t_0,x^*\left(t_0\right)\right),b^*\left(r\right)\right\rangle dr\right|\\
&\leq\mathbb{E}_{\omega_0}\frac{1}{h}\int_{t_0}^{t_0+h}\left|D\phi\left(r,x^*\left(r\right)\right)-D\phi\left(t_0,x^*\left(t_0\right)\right)\right|\left|b^*\left(r\right)\right|dr\\
&\leq\mathbb{E}_{\omega_0}\left[\frac{1}{h}\int_{t_0}^{t_0+h}\left|D\phi\left(r,x^*\left(r\right)\right)-D\phi\left(t_0,x^*\left(t_0\right)\right)\right|^2dr\right]^{\frac{1}{2}}\left[\frac{1}{h}\int_{t_0}^{t_0+h}\left|b^*\left(r\right)\right|^2dr\right]^{\frac{1}{2}}\\
&\leq\left[\mathbb{E}_{\omega_0}\frac{1}{h}\int_{t_0}^{t_0+h}\left|D\phi\left(r,x^*\left(r\right)\right)-D\phi\left(t_0,x^*\left(t_0\right)\right)\right|^2dr\right]^{\frac{1}{2}}\left[\mathbb{E}_{\omega_0}\frac{1}{h}\int_{t_0}^{t_0+h}\left|b^*\left(r\right)\right|^2dr\right]^{\frac{1}{2}}.
\end{aligned}$$

Using Hypothesis 2.1 and estimate (12), we have

$$\begin{aligned}
\mathbb{E}_{\omega_0}\frac{1}{h}\int_{t_0}^{t_0+h}\left|b^*\left(r\right)\right|^2dr&\leq\frac{1}{h}\int_{t_0}^{t_0+h}2L^2\left(1+\mathbb{E}_{\omega_0}\left|x^*\left(r\right)\right|^2\right)dr\\
&\leq 2L^2\left(1+K_T\left(1+|x_0|^2\right)\right),
\end{aligned}\tag{15}$$

whereas arguing as in part (I) for ϕ_t, we get

$$\lim_{h\to 0^+} \mathbb{E}_{\omega_0} \frac{1}{h} \int_{t_0}^{t_0+h} |D\phi(r, x^*(r)) - D\phi(t_0, x^*(t_0))|\, dr = 0. \tag{16}$$

From (15) and (16), it follows that the first term of (14) goes to zero as $h \to 0^+$. For the second term of (14), we observe that

$$\begin{aligned} &\left| \mathbb{E}_{\omega_0} \frac{1}{h} \int_{t_0}^{t_0+h} \langle D\phi(t_0, x^*(t_0)), b^*(r) - b^*(t_0) \rangle\, dr \right| \\ &\quad \le |D\phi(t_0, x_0)|\, \mathbb{E}_{\omega_0} \frac{1}{h} \int_{t_0}^{t_0+h} |b^*(r) - b^*(t_0)|\, dr. \end{aligned}$$

However, by the choice of t_0,

$$\begin{aligned} 0 &= \lim_{h\to 0^+} \mathbb{E} \frac{1}{h} \int_{t_0}^{t_0+h} |b^*(r) - b^*(t_0)|\, dr \\ &= \lim_{h\to 0^+} \mathbb{E}\left[\mathbb{E}\left[\frac{1}{h} \int_{t_0}^{t_0+h} |b^*(r) - b^*(t_0)|\, dr \middle| \mathcal{F}_{t_0}^s \right] \right] \\ &= \lim_{h\to 0^+} \mathbb{E}\left[\mathbb{E}_{\omega_0} \frac{1}{h} \int_{t_0}^{t_0+h} |b^*(r) - b^*(t_0)|\, dr \right]. \end{aligned}$$

This implies that $\mathbb{E}_{\omega_0} \frac{1}{h} \int_{t_0}^{t_0+h} |b^*(r) - b^*(t_0)|\, dr$ converges to 0 in $L^1(\Omega; \mathbb{R}^1)$ as $h \to 0^+$. Hence, there is a subsequence $h_l \to 0^+$ so that, for $\mathbb{P}$-a.s. ω_0,

$$\mathbb{E}_{\omega_0} \frac{1}{h} \int_{t_0}^{t_0+h} |b^*(r) - b^*(t_0)|\, dr \to 0.$$

This proves that the second term of (14) goes to zero as $h_l \to 0^+$.

(III) The third term of (13) is treated using the same ideas. Since

$$\operatorname{tr}(ABC) = \operatorname{tr}(BCA) \quad \text{and} \quad \operatorname{tr}(AB) - \operatorname{tr}(AC) = \operatorname{tr} A(B - C),$$

we have

$$\begin{aligned} &\mathbb{E}_{\omega_0}\left[\frac{1}{h} \int_{t_0}^{t_0+h} \left[\frac{1}{2} \operatorname{tr}\left(\sigma^*(r)' D^2\phi(r, x^*(r)) \sigma^*(r) \right) \right.\right. \\ &\qquad \left.\left. - \frac{1}{2} \operatorname{tr}\left(\sigma^*(t_0)' D^2\phi(t_0, x^*(t_0)) \sigma^*(t_0) \right) \right] dr \right] \qquad (17) \\ &= \mathbb{E}_{\omega_0}\left[\frac{1}{h} \int_{t_0}^{t_0+h} \frac{1}{2} \operatorname{tr}\left(\left[D^2\phi(r, x^*(r)) - D^2\phi(t_0, x^*(t_0)) \right] \sigma^*(r) \sigma^*(r)' \right) dr \right] \\ &\quad + \mathbb{E}_{\omega_0}\left[\frac{1}{h} \int_{t_0}^{t_0+h} \frac{1}{2} \operatorname{tr}\left(D^2\phi(t_0, x^*(t_0)) \left[\sigma^*(r) \sigma^*(r)' - \sigma^*(t_0) \sigma^*(t_0)' \right] \right) dr \right]. \end{aligned}$$

Now, employing the same arguments that we used to show that the right-hand side of (14) goes to zero, we obtain that the right-hand side of (17) goes to 0 as some subsequence $h_{l'} \to 0^+$.

Summing up, we have proved that, for any sequence $h \to 0^+$, one can find a subsequence $h_{l''} \to 0^+$ so that

$$\begin{aligned}
&\lim_{h_{l''}\to 0^+} \mathbb{E}_{\omega_0}\left[\frac{1}{h_{l''}}\int_{t_0}^{t_0+h_{l''}}\left[\phi_t\left(r,x^*\left(r\right)\right)+\left\langle D\phi\left(r,x^*\left(r\right)\right),b^*\left(r\right)\right\rangle\right.\right.\\
&\qquad\qquad\left.\left.+\frac{1}{2}\mathrm{tr}\left(\sigma^*\left(r\right)'D^2\phi\left(r,x^*\left(r\right)\right)\sigma^*\left(r\right)\right)\right]dr\right]\\
&\quad=\phi_t\left(t_0,x_0\right)+\left\langle D\phi\left(t_0,x_0\right),b^*\left(t_0\right)\right\rangle+\frac{1}{2}\mathrm{tr}\left(\sigma^*\left(t_0\right)'D^2\phi\left(t_0,x_0\right)\sigma^*\left(t_0\right)\right).
\end{aligned}$$

Consequently, the above holds for any sequence $h \to 0^+$. Passing to the limsup in (13) and using (11), we get that

$$\begin{aligned}
&\limsup_{h\to 0^+}\mathbb{E}_{\omega_0}\frac{1}{h}\left[U\left(t_0+h,x^*\left(t_0+h\right)\right)-U\left(t_0,x^*\left(t_0\right)\right)\right]\\
&\quad\le\phi_t\left(t_0,x_0\right)+\left\langle D\phi\left(t_0,x_0\right),b^*\left(t_0\right)\right\rangle+\frac{1}{2}\mathrm{tr}\left(\sigma^*\left(t_0\right)'D^2\phi\left(t_0,x_0\right)\sigma^*\left(t_0\right)\right)\\
&\quad=p^*\left(t_0,\omega_0\right)+\left\langle q^*\left(t_0,\omega_0\right),b^*\left(t_0,\omega_0\right)\right\rangle+\frac{1}{2}\mathrm{tr}\left(\sigma^*\left(t_0,\omega_0\right)'Q^*\left(t_0,\omega_0\right)\sigma^*\left(t_0,\omega_0\right)\right). \qquad (18)
\end{aligned}$$

Now, using (18) and applying Fatou's lemma, we get

$$\begin{aligned}
&\limsup_{h\to 0^+}\mathbb{E}\left[\frac{1}{h}\left[U\left(t_0+h,x^*\left(t_0+h\right)\right)-U\left(t_0,x^*\left(t_0\right)\right)\right]\right]\\
&\quad=\limsup_{h\to 0^+}\mathbb{E}\left[\mathbb{E}_{\omega_0}\frac{1}{h}\left[U\left(t_0+h,x^*\left(t_0+h\right)\right)-U\left(t_0,x^*\left(t_0\right)\right)\right]\right]\\
&\quad\le\mathbb{E}\left[\limsup_{h\to 0^+}\left[\mathbb{E}_{\omega_0}\frac{1}{h}\left[U\left(t_0+h,x^*\left(t_0+h\right)\right)-U\left(t_0,x^*\left(t_0\right)\right)\right]\right]\right]\\
&\quad\le\mathbb{E}\left[p^*\left(t_0\right)+\left\langle q^*\left(t_0\right),b^*\left(t_0\right)\right\rangle+\frac{1}{2}\mathrm{tr}\left(\sigma^*\left(t_0\right)'Q^*\left(t_0\right)\sigma^*\left(t_0\right)\right)\right] \qquad (19)
\end{aligned}$$

for a.e. $t_0 \in [s,T]$. The rest of the proof goes exactly as in [8, p. 272]. We apply [8, Lemma 5.2, p. 270] to the function $g(t) = \mathbb{E}\,U(t,x^*(t))$ and use (19) and (9) to obtain

$$\mathbb{E}\,U(T,x^*(T)) - U(s,y) \le -\mathbb{E}\int_s^T f^*(t)dt,$$

which, since $U \le V$ and $U(T,x^*(T)) = V(T,x^*(T)) = h(x^*(T))$, yields

$$\mathbb{E}\left[\int_s^T f\left(t,x^*(t),u^*(t)\right)dt + h(x^*(T))\right] \le V(s,y),$$

which means that the control $u^*(\cdot)$ is optimal. □

Remark 4.2. Now we are in a position to discuss the (subtle) gap presented in the proof of the book [8]. On page 271 of [8], equation (5.12) was derived by considering a Lebesgue point t of the integrand in the displayed equation immediately before (5.12). However, the test function ϕ there depends, though implicitly, on the variable t. Hence, it is not clear whether such a t exists in the first place. In this paper, we have fixed this problem via the Lebesgue points of the problem data ($b^*(\cdot)$ and $\sigma^*(\cdot)\sigma^*(\cdot)'$) viewed as $L^1(\Omega;\mathbb{R}^n)$ or $L^1(\Omega;\mathbb{R}^{n\times n})$-valued functions, together with some delicate estimates.

Acknowledgments. Part of this research was carried out when the third author was visiting Scuola Normale Superiore, Italy. The hospitality of the host, Giuseppe Da Prato, is gratefully acknowledged. Moreover, the authors thank the editors and two reviewers for their helpful comments and suggestions on an earlier version of the paper.

REFERENCES

[1] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.
[2] J. Diestel and J. J. Uhl, Jr., *Vector Measures*, Math. Surveys Monogr., 15, AMS, Providence, RI, 1977.
[3] H. Doob, *Measure Theory,* Springer-Verlag, New York, 1994.
[4] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control,* Springer-Verlag, New York, 1975.
[5] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions,* Appl. Math. 25, Springer-Verlag, New York, 1993.
[6] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Grad. Texts in Math. 113, Springer-Verlag, New York, 1999.
[7] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, 3rd ed., Springer-Verlag, New York, 1999.
[8] J. Yong and X. Y. Zhou, *Stochastic Controls. Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.
[9] X. Y. Zhou, *Verification theorems within the framework of viscosity solutions*, J. Math. Anal. Appl., 177 (1993), pp. 208–225.
[10] X. Y. Zhou, J. Yong, and X. Li, *Stochastic verification theorems within the framework of viscosity solutions*, SIAM J. Control Optim., 37 (1997), pp. 243–253.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2020–2035

DIFFERENTIAL GAMES WITH ERGODIC PAYOFF*

MRINAL K. GHOSH† AND K. S. MALLIKARJUNA RAO‡

Abstract. We address a zero-sum differential game with ergodic payoff. We study this problem via the viscosity solutions of an associated Hamilton–Jacobi–Isaacs equation. Under certain condition, we establish the existence of a value and prove certain representation formulae.

Key words. differential games, value, Hamilton–Jacobi–Isaacs equation, viscosity solution

AMS subject classifications. 90D25, 90D26

DOI. 10.1137/S0363012903404511

1. Introduction.

In this article, we consider a general, nonlinear controlled dynamical system

$$\dot{x}(t) = b(x(t), u_1(t), u_2(t)) \tag{1.1}$$

with performance index $r(x(t), u_1(t), u_2(t))$, where u_1, u_2 are controls. Associated to this controlled dynamical system we can pose two kinds of problem—$\mathcal{H}_\infty$ control and the differential game. In $\mathcal{H}_\infty$ control, the performance index is referred to as the output response and u_2 as disturbance. A closed set $\mathcal{T}$ with respect to which the undisturbed system ($u_2 = 0$) is stable and a constant γ are given. The problem is to find a strategy $\alpha = \alpha[u_2]$ such that

$$\int_0^t |r(x(s), \alpha[u_2](s), u_2(s))|^2 \, ds \le \gamma^2 \int_0^t |u_2(s)|^2 \, ds \tag{1.2}$$

for all $t \ge 0$ and all disturbances u_2. If we can find such a strategy, we say that the problem is solvable with disturbance attenuation level γ.

The other problem is the differential game problem. In this case, we call the performance index the running payoff function. There are two controllers or decision makers called players. Player 1 wishes to minimize the running payoff function on finite or infinite time horizon over his control variables $u_1(t)$, whereas Player 2 wishes to maximize the same over his control variables $u_2(t)$. Since the interests of the two players are conflicting, the basic problem is to resolve this conflict by arriving at solution that serves the interests of both players. In other words, we look for a min-max/max-min solution to this problem. For infinite horizon problems, one usually considers two payoff criteria: the discounted payoff criterion and the ergodic or averaged payoff criterion. These two payoff criteria are, in some sense, complementary to each other. The immediate future is far more important than the distant future in the discounted payoff criterion. Quite contrary to this, the finite time behavior of

*Received by the editors August 7, 2003; accepted for publication (in revised form) July 5, 2004; published electronically April 14, 2005. This work was supported in part by the IISc-DRDO Programme on Advanced Engineering Mathematicics, NSF grants ECS-0218207 and ECS-0225448, and the Office of Naval Research Electric Ship Research and Development Consortium.
http://www.siam.org/journals/sicon/43-6/40451.html

†Department of Mathematics, Indian Institute of Science, Bangalore 560 012, India, and Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78712 (mkg@math.iisc.ernet.in).

‡CMI, Université de Provence, 39, Rue F. J. Curie, 13 453 Marseille, France (mallikmpd@yahoo.co.in).

the system is irrelevant in the ergodic payoff criterion. It is the asymptotic behavior of the ergodic payoff that matters. Thus in the ergodic payoff criterion, one looks for some kind of stability or averaging mechanism taking place.

The differential game (in the sense of Elliott–Kalton) with discounted payoff criterion has been studied extensively in the literature; see [1] and the references therein. The basic idea is to show that the lower and upper value functions satisfy the dynamic programming principle (DPP) and thus they are viscosity solutions to corresponding Hamilton–Jacobi–Isaacs (HJI) equations. If the Isaacs minimax principle holds, then by a minimax theorem, one obtains that the differential game has value. This procedures does not seem to be applicable to the differential games with ergodic payoff. Thus in order to study the differential games with ergodic payoff, we need to approach the problem in a different way.

In the traditional approach to differential games, one first establishes the DPP, which in turn leads to the HJI equations. In this article, we follow a reverse approach which was used by Świech [18] to treat a stochastic differential game with a finite horizon payoff criterion (see also [17]). The main idea is to use the integration along the trajectories of the controlled dynamical system to study the HJI equations. Since the HJI equation in general does not admit classical solutions, we need to use the concept of viscosity solutions introduced by Crandall and Lions [4]. We show that if the HJI equation corresponding to ergodic payoff criterion has a viscosity solution, then the scalar quantity appearing in the HJI equation is the ergodic value for the differential game problem under certain stability assumption on the dynamics. Further, under a dissipativity assumption, we show that the HJI equation has a viscosity solution. The novelty of this approach is that it is quite simple and it can be used to prove the DPP.

There is a close connection between $\mathcal{H}_\infty$ control and differential games. A $\mathcal{H}_\infty$ control problem can be viewed as a differential game problem (see [3]). Using this observation, several authors have studied HJI equations and established DPP for the solutions; see [6], [9], [13], [15], [16], and the references therein. In [16] an $\mathcal{H}_\infty$ control problem is considered and studied using the viscosity solution techniques. Some representation formulas are proved for the viscosity solutions of the associated HJI equation. As a consequence, the author obtained the DPP for the viscosity solutions and established the value function to be the minimal viscosity solution under some nonnegativity assumptions and certain stability assumptions. In [9], [13], an analogous problem in the stochastic case is considered. Here the authors first obtained the DPP. The value function is again shown to be the minimal viscosity solution. The results in the deterministic case are obtained by letting the diffusion coefficient be zero. Further in [13], the uniqueness of the viscosity solution is established in a certain class of functions with some growth conditions. Thus the results in these articles are similar to ours. However, the assumptions considered there are different from the assumptions in this article. Note that in the mentioned articles, the ergodic value corresponding to the associated differential game turns out to be zero. Thus the results presented in this article can be seen as more general concerning differential games with ergodic payoff. We now describe our problem.

Let U_i, $i = 1, 2$, be given compact metric spaces. Let $\mathcal{A}_i$, $i = 1, 2$, denote the set of all measurable functions $u_i : [0, \infty) \to U_i$. The set $\mathcal{A}_i$ is called the set of all admissible controls for player i. Consider the d-dimensional controlled dynamical system $x(\cdot)$ described by

$$\left\{ \begin{array}{rcl} \dot{x}(t) & = & b(x(t), u_1(t), u_2(t)),\ t > 0, \\ x(0) & = & x, \end{array} \right. \tag{1.3}$$

where $b : \mathbf{R}^d \times U_1 \times U_2 \to \mathbf{R}^d$ and $u_i(\cdot) \in \mathcal{A}_i$. We assume that

(A1) b is continuous and there exists a constant $C_1 > 0$ such that for all $u_1 \in U_1$ and $u_2 \in U_2$

$$|b(x, u_1, u_2) - b(y, u_1, u_2)| \le C_1|x - y|.$$

Let $r : \mathbf{R}^d \times U_1 \times U_2 \to \mathbf{R}^d$ be the payoff function. We assume that

(A2) r is continuous and there exists a constant $C_2 > 0$ such that for all $u_1 \in U_1$ and $u_2 \in U_2$

$$|r(x, u_1, u_2) - r(y, u_1, u_2)| \le C_2|x - y|.$$

Let Γ denote the set of all maps $\alpha : \mathcal{A}_2 \to \mathcal{A}_1$ that are nonanticipative in the sense that for any $t > 0$ and $u_2, \tilde{u}_2 \in U_2$, $u_2(s) = \tilde{u}_2(s)$ for all $s \le t$ implies $\alpha[u_2](s) = \alpha[\tilde{u}_2](s)$ for all $s \le t$. Similarly, Δ is defined to be the set of all maps from $\mathcal{A}_1$ to $\mathcal{A}_2$ that are nonanticipative.

Let

$$\rho^+(x) := \sup_{\beta\in\Delta} \inf_{u_1(\cdot)\in\mathcal{A}_1} \limsup_{T\to\infty} \frac{1}{T} \int_0^T r(x(t), \beta[u_1](s), u_1(s))\, ds,$$

$$\rho^-(x) := \inf_{\alpha\in\Gamma} \sup_{u_2(\cdot)\in\mathcal{A}_2} \limsup_{T\to\infty} \frac{1}{T} \int_0^T r(x(t), u_2(s), \alpha[u_2](s))\, ds.$$

The functions $\rho^+(x), \rho^-(x)$ are called the upper and lower ergodic value functions associated with the differential game. If $\rho^+(x) = \rho^-(x) = \rho$, a constant for all x, we say that the differential game with ergodic payoff criterion has a value.

The rest of the paper is organized as follows. In section 2, we prove that if the associated HJI equation has a viscosity solution (ρ, w), then the upper and lower values coincide with ρ, and thus the differential game has value. We then prove some more representation formulas for the ergodic value. We also prove DPP for viscosity solution and a partial uniqueness result for viscosity solutions. In section 3, we show the existence of a viscosity solution to the HJI equation in two ways under a suitable assumption. Section 4 contains some concluding remarks.

2. Viscosity solutions and ergodic value. Consider the following HJI equations

$$\rho = \inf_{u_1\in U_1} \sup_{u_2\in U_2} \{b(x, u_1, u_2) \cdot Dw(x) + r(x, u_1, u_2)\},\ x \in \mathbf{R}^d \tag{2.1}$$

and

$$\rho = \sup_{u_2\in U_2} \inf_{u_1\in U_1} \{b(x, u_1, u_2) \cdot Dw(x) + r(x, u_1, u_2)\},\ x \in \mathbf{R}^d. \tag{2.2}$$

Definition 2.1. *A viscosity subsolution of* (2.1) *is a pair* (ρ, w)*, where* ρ *is a real number and* $w(\cdot)$ *is an upper semicontinuous function such that for* $x \in \mathbf{R}^d$ *and a smooth function* ϕ*, we have*

$$\rho \le \inf_{u_1\in U_1} \sup_{u_2\in U_2} \{b(x, u_1, u_2) \cdot D\phi(x) + r(x, u_1, u_2)\}$$

whenever $w-\phi$ has a local maximum at x. A pair (ρ, w) of a real number ρ and a lower semicontinuous function $w(\cdot)$ is said to be a viscosity supersolution of (2.1) *if for $x \in \mathbf{R}^d$ and a smooth function ϕ, we have*

$$\rho \geq \inf_{u_1 \in U_1} \sup_{u_2 \in U_2} \{b(x, u_1, u_2) \cdot D\phi(x) + r(x, u_1, u_2)\}$$

whenever $w-\phi$ has a local minimum at x. A viscosity solution of (2.1) *is a pair (ρ, w) that is both viscosity sub- and supersolution of* (2.1)*. Similarly, a viscosity solution of* (2.2) *is defined.*

We now proceed to prove the main result of this section, which provides estimates for ρ^+ in terms of viscosity sub- and supersolutions of (2.1) and, similarly, for ρ^-, in terms of viscosity sub- and supersolutions of (2.2). We prove this result under the following additional assumption:

(A3) For each $x \in \mathbf{R}^d$, there is a constant $M = M(x) > 0$ such that $|x(t)| < M$ for all $t \geq 0$, where $x(\cdot)$ is the solution of (1.3) under any pair of admissible controls $(u_1(\cdot), u_2(\cdot)) \in \mathcal{A}_1 \times \mathcal{A}_2$.

Remark 2.2. Since for any $t, s \geq 0$,

$$x(t) - x(s) = \int_s^t b(x(\tau), u_1(\tau), u_2(\tau))\, d\tau$$

and $|x(\tau)| \leq M$ by assumption (A3), we can find a constant $C > 0$ such that

$$|x(t) - x(s)| \leq C|t - s|.$$

Thus under assumptions (A1) and (A3), the solutions of (1.3) are globally Lipschitz continuous.

We now state and prove the main result of this section. Throughout the section, we assume (A1)–(A3).

THEOREM 2.3. (i) *Let (ρ, w) be a viscosity subsolution of* (2.1)*. Then*

$$\rho \leq \sup_{\beta \in \Delta} \inf_{u_1(\cdot) \in \mathcal{A}_1} \liminf_{T \to \infty} \frac{1}{T} \int_0^T r(x(s), u_1(s), \beta[u_1](s))\, ds. \tag{2.3}$$

(ii) *Let (ρ, w) be a viscosity supersolution of* (2.1)*. Then*

$$\rho \geq \sup_{\beta \in \Delta} \inf_{u_1(\cdot) \in \mathcal{A}_1} \limsup_{T \to \infty} \frac{1}{T} \int_0^T r(x(s), u_1(s), \beta[u_1](s))\, ds. \tag{2.4}$$

(iii) *Let (ρ, w) be a viscosity subsolution of* (2.2)*. Then*

$$\rho \leq \inf_{\alpha \in \Gamma} \sup_{u_2(\cdot) \in \mathcal{A}_2} \liminf_{T \to \infty} \frac{1}{T} \int_0^T r(x(s), \alpha[u_2](s), u_2(s))\, ds. \tag{2.5}$$

(iv) *Let (ρ, w) be a viscosity supersolution of* (2.2)*. Then*

$$\rho \geq \inf_{\alpha \in \Gamma} \sup_{u_2(\cdot) \in \mathcal{A}_2} \limsup_{T \to \infty} \frac{1}{T} \int_0^T r(x(s), \alpha[u_2](s), u_2(s))\, ds. \tag{2.6}$$

Proof. We prove (iii) and (iv); (i) and (ii) can be proved similarly.

Let (ρ, w) be a viscosity subsolution of (2.2). Assume that w is $C^{1,1}$ (i.e., w is differentiable with bounded and Lipschitz derivatives). Let K be the common

Lipschitz constant associated with w, Dw. Then (ρ, w) satisfies (2.2) in the classical sense. In particular, for any $\epsilon > 0$ and any $x \in \mathbf{R}^d$,

$$\rho - \epsilon < \sup_{u_2 \in U_2} \inf_{u_1 \in U_1} (b(x, u_1, u_2) \cdot Dw(x) + r(x, u_1, u_2)). \tag{2.7}$$

Set

$$\Lambda(x, u_2) = \inf_{u_1 \in U_1} (b(x, u_1, u_2) \cdot Dw(x) + r(x, u_1, u_2)).$$

Then it is easy to note that Λ is uniformly continuous on $\mathbf{R}^d \times U_2$. Since U_2 is separable, we can find a sequence $\{u_2^i\}$ in U_2 and a family of balls $\{B_{r_i}(x_i)\}$ covering $\mathbf{R}^d$ such that

$$\rho - \epsilon < \Lambda(x, u_2^i) \text{ for all } x \in B_{r_i}(x_i) \text{ and } i.$$

Note that here the sequence $\{u_2^i\}$ can be chosen to be finite since U_2 is compact. In that case, the sequence of balls $\{B_{r_i}(x_i)\}$ should be replaced by a finite family of open sets.

Define, $\psi : \mathbf{R}^d \to U_2$ by

$$\psi(x) = u_2^k \text{ if } x \in B_{r_k}(x_k) \setminus \bigcup_{i=1}^{k-1} B_{r_i}(x_i).$$

Then ψ is a Borel map and $\rho - \epsilon < \Lambda(x, \psi(x)) \ \forall x \in \mathbf{R}^d$. We make the following claims.

Claim A. For $x \in \mathbf{R}^d$, $m > 0$, there exists $\beta^m \in \Delta$ such that

$$(\rho - \epsilon)N - C\frac{N}{m} - \int_0^N r(x(s), u_1(s), \beta^m[u_1](s))\, ds \le w(x(N)) - w(x)$$

for any positive integer N, where $x(\cdot)$ is the solution of (1.3) with the initial condition $x(0) = x$ under controls $(u_1(\cdot), \beta^m[u_1](\cdot))$ and C is a constant depending on K, C_1, C_2 but not on x, N, and m.

Claim B. For each $\alpha \in \Gamma$, we can find $\tilde{u}_1(\cdot) \in \mathcal{A}_1$ and $\tilde{u}_2(\cdot) \in \mathcal{A}_2$ such that

$$\beta^m[\tilde{u}_1](\cdot) = \tilde{u}_2(\cdot) \text{ and } \alpha[\tilde{u}_2](\cdot) = \tilde{u}_1(\cdot). \tag{2.8}$$

Assuming the claims to be true, we complete the proof of (2.5). Divide the inequality in Claim A by N, and let $N \to \infty$ to obtain

$$(\rho - \epsilon) \le C\frac{1}{m} + \liminf_{N \to \infty} \frac{1}{N} \int_0^N r(x(s), u_1(s), \beta^m[u_1](s))\, ds. \tag{2.9}$$

Using (2.8) in (2.9), we deduce

$$(\rho - \epsilon) \le C\frac{1}{m} + \inf_{\alpha \in \Gamma} \sup_{u_2(\cdot) \in \mathcal{A}_2} \liminf_{N \to \infty} \frac{1}{N} \int_0^N r(x(s), \alpha[u_2](s), u_2(s))\, ds.$$

Letting $m \to \infty$, we obtain

$$(\rho - \epsilon) \le \inf_{\alpha \in \Gamma} \sup_{u_2(\cdot) \in \mathcal{A}_2} \liminf_{N \to \infty} \frac{1}{N} \int_0^N r(x(s), \alpha[u_2](s), u_2(s))\, ds.$$

We now need to replace the limit along the integers by the limit along any real sequence. For this, choose any sequence $T_n \to \infty$. Then

$$\frac{1}{T_n}\int_o^{T_n} r(x(s),\alpha[u_2](s),u_2(s))\,ds$$
$$= \frac{1}{T_n}\int_0^{[T_n]} r(x(s),\alpha[u_2](s),u_2(s))\,ds + \frac{1}{T_n}\int_{[T_n]}^{T_n} r(x(s),\alpha[u_2](s),u_2(s))\,ds.$$

Using (A3), we note that the second term on the right-hand side of the above equality vanishes as $n \to \infty$. Note also the fact that

$$\left|\frac{1}{T_n}\int_0^{[T_n]} r(x(s),\alpha[u_2](s),u_2(s))\,ds - \frac{1}{[T_n]}\int_0^{[T_n]} r(x(s),\alpha[u_2](s),u_2(s))\,ds\right| \to 0$$

as $n \to \infty$. Thus

$$\lim_{n\to\infty}\frac{1}{T_n}\int_o^{T_n} r(x(s),\alpha[u_2](s),u_2(s))\,ds = \lim_{n\to\infty}\frac{1}{[T_n]}\int_o^{[T_n]} r(x(s),\alpha[u_2](s),u_2(s))\,ds.$$

Since this is true for any sequence (T_n) tending to ∞, we obtain

$$(\rho-\epsilon) \le \inf_{\alpha\in\Gamma}\sup_{u_2(\cdot)\in\mathcal{A}_2}\liminf_{T\to\infty}\frac{1}{T}\int_0^T r(x(s),\alpha[u_2](s),u_2(s))\,ds.$$

This proves (2.5) under the assumption that w is $C^{1,1}$. We now turn to the general case. Let w_ϵ be the sup-convolution of w, i.e.,

$$w_\epsilon(y) = \sup_{|z|\le M+2}\left\{w(z) - \frac{|z-y|^2}{2\epsilon}\right\}.$$

Then w_ϵ converges to w uniformly as $\epsilon \to 0$ on $B_{M+1} := \bar{B}(0, M+1)$, and w_ϵ are Lipschitz continuous and satisfy a.e. on B_{M+1}

$$\rho \le \inf_{u_2\in U_2}\sup_{u_1\in U_1}\{b(y,u_1,u_2)\cdot Dw_\epsilon(y) + r(y,u_1,u_2)\} + \sigma_1(\epsilon)$$

for some modulus σ_1 (see [10], [11]). For each $\delta > 0$, let w_ϵ^δ be a smooth approximation of w_ϵ such that $w_\epsilon^\delta, Dw_\epsilon^\delta$ are smooth and they converge to w_ϵ, Dw_ϵ uniformly on compact sets, respectively, and they all have the same Lipschitz constant. Now, using these facts, we can find another modulus σ_2 such that

$$\rho \le \inf_{u_2\in U_2}\sup_{u_1\in U_1}\{b(y,u_1,u_2)\cdot Dw_\epsilon^\delta(y) + r(y,u_1,u_2)\} + \sigma_1(\epsilon) + \sigma_2(\delta) \tag{2.10}$$

on $B_{M+1/2}$. Note that σ_2 may depend on ϵ and x. Observe that while proving (2.5), we have used the smoothness of w only in B_M. Thus we can use the above arguments with w_ϵ^δ and (2.10) to conclude

$$\rho \le \inf_{\alpha\in\Gamma}\sup_{u_2(\cdot)\in\mathcal{A}_2}\liminf_{T\to\infty}\frac{1}{T}\int_o^T r(x(s),\alpha[u_2](s),u_2(s))\;ds + \sigma_1(\epsilon) + \sigma_2(\delta),$$

where $x(\cdot)$ is the solution of (1.3) with the initial condition $x(0) = x$ under the controls $(\alpha[u_2](\cdot), u_2(\cdot))$. Now letting δ and then ϵ to 0, we obtain (2.5). This completes the proof of part (iii). We now proceed to prove the claims.

Proof of Claim A. Let $t = \frac{1}{m}$. Define

$$u_2^m(s) = \psi(x) \text{ for } s \in [0, t).$$

We extend the definition of $(u_2^m(\cdot), x(\cdot))$ to $[0, (i+1)\,t)$ assuming that it has been defined on $[0, i\,t)$ as follows. Let $x(\cdot)$ be the solution (2.1) with initial value x and controls $(u_1(\cdot), u_2^m(\cdot))$ in the interval $[0, it)$. Set

$$u_2^m(s) = \psi(x((i\,t)^-)) \text{ for } s \in [i\,t, (i+1)\,t).$$

Note that $x((it)^-)$ exists since $X(\cdot)$ is Lipschitz continuous and bounded. This defines $u_2^m(\cdot)$ on $\mathbf{R}$.

Let $x(.)$ be the solution of (1.3) with initial value $x(0) = x$ and controls $(u_1(\cdot), u_2^m(\cdot)\,)$. Then,

$$\begin{aligned}
&w(x((i+1)t)) - w(x(it)) \\
&\quad = \int_{it}^{(i+1)t} Dw(x(s)) \cdot b(x(s), u_1(s), u_2^m(s))\, ds \\
&\quad = \int_{it}^{(i+1)t} (Dw(x(s)) - Dw(x(it))) \cdot b(x(s), u_1(s), u_2^m(s))\, ds \\
&\qquad + \int_{it}^{(i+1)t} Dw(x(it)) \cdot (b(x(s), u_1(s), u_2^m(s)) - b(x(it), u_1(s), u_2^m(s)))\, ds \\
&\qquad + \int_{it}^{(i+1)t} (Dw(x(it)) \cdot b(x(it), u_1(s), u_2^m(s)) + r(x(it), u_1(s), u_2^m(s)))\, ds \\
&\qquad + \int_{it}^{(i+1)t} (r(x(s), u_1(s), u_2^m(s)) - r(x(it), u_1(s), u_2^m(s)))\, ds \\
&\qquad - \int_{it}^{(i+1)t} r(x(s), u_1(s), u_2^m(s))\, ds.
\end{aligned}$$

Note that w, Dw, b are all Lipschitz along the trajectory $x(\cdot)$ and they are bounded by assumptions (A1) and (A3). Using these facts in the above together with the definition of ψ, we obtain,

$$\begin{aligned}
w(x((i+1)t)) - w(x(it)) &\geq -C \int_{it}^{(i+1)t} (s - i\,t)\, ds + (\rho - \epsilon) \int_{it}^{(i+1)t} ds \\
&\quad - \int_{it}^{(i+1)t} r(x(s), u_1(s), u_2^m(s))\, ds \\
&= -C\,t^2 + (\rho - \epsilon)t - \int_{it}^{(i+1)t} r(x(s), u_1(s), u_2^m(s))\, ds
\end{aligned}$$

for a constant $C > 0$ which will depend only on x and other Lipschitz constants. Now define a strategy $\beta^m \in \Delta$ by $\beta^m[u_1](\cdot) = u_2^m(\cdot)$ for $u_1(\cdot) \in \mathcal{A}_1$. Note that $\beta^m[u_1](\cdot)$ on

$[it,(i+1)t)$ depends only on $[0,it)$. Adding these inequalities for $i=0,\dots,Nm-1$, we get the inequality stated in Claim A.

Proof of Claim B. We define such controls inductively. Let $\tilde{u}_2(\cdot)=\psi(x)$ on $[0,t)$. Define $\tilde{u}_1|_{[0,t)}=\alpha[\tilde{u}_2]|_{[0,t)}$. Having known $\tilde{u}_1(\cdot)$ and $\tilde{u}_2(\cdot)$ on $[0,it)$, we define $\tilde{u}_2(\cdot)$ on $[it,(i+1)t)$ by $\tilde{u}_2(s)=\psi(x(it)^-)$, where $x(\cdot)$ satisfies

$$\dot{x}(s)=b(x(s),\tilde{u}_1(s),\tilde{u}_2(s)),\ s\in[0,it)$$

and $x(0)=x$. It is now easy to check (2.8). This completes the proof of Claim B.

We now prove part (iv). Let w be a viscosity supersolution of (2.2) and assume $w\in C^{1,1}$. The proof for general w follows from an argument as in that of (iii). One has for any $\epsilon>0$ and any $x\in\mathbf{R}^d$,

$$\sup_{u_2\in U_2}\inf_{u_1\in U_1}(b(x,u_1,u_2)\cdot Dw(x)+r(x,u_1,u_2))<\rho+\epsilon.$$

Set

$$\Lambda(x,u_1,u_2)=(b(x,u_1,u_2)\cdot Dw(x)+r(x,u_1,u_2)).$$

By the uniform continuity of Λ, we can find a countable family $B_{r_i}(x_i)\times B_{r_i}(u_2^i)$ covering $\mathbf{R}^d$ and a sequence $u_1^i\in U_1$ such that

$$\Lambda(x,u_1^i,u_2)<\rho+\epsilon\quad\forall\ (x,u_2)\in B_{r_i}(x_i)\times B_{r_i}(u_2^i).$$

Define a map $\psi:\mathbf{R}^d\times U_2\to U_1$ by

$$\psi(x,u_2)=u_1^k \text{ if } (x,u_2)\in B_{r_k}(x_k)\times B_{r_k}(u_2^k)\setminus\bigcup_{i=1}^{k-1}B_{r_i}(x_i)\times B_{r_i}(u_2^i).$$

Then ψ is Borel measurable and

$$\Lambda(x,\psi(x,u_2),u_2)<\rho+\epsilon\quad\forall(x,u_2).$$

Claim C. For each integer $m>0$, there exists $\alpha^m\in\Gamma$ such that

$$\int_0^N r(x(s),\alpha^m[u_2](s),u_2(s))\,ds+w(x(N))-w(x)\le(\rho+\epsilon)N+C\frac{N}{m}$$

for all positive integers N and $u_2(\cdot)\in\mathcal{A}_2$, where $x(\cdot)$ is the solution of (1.3) with the initial condition $x(0)=x$ under controls $(\alpha^m[u_2](\cdot),u_2(\cdot))$ and C is a constant independent of N and m.

Assuming that the claim is true, we see, on dividing by N and letting $N\to\infty$,

$$\limsup_{N\to\infty}\frac{1}{N}\int_0^N r(x(s),\alpha^m[u_2](s),u_2(s))\ ds\le(\rho+\epsilon)+\frac{C}{m},$$

which implies

$$\inf_{\alpha\in\Gamma}\sup_{u_2(\cdot)\in U_2}\limsup_{N\to\infty}\frac{1}{N}\int_0^N r(x(s),\alpha[u_2](s),u_2(s))\ ds\le\rho.$$

From this one can deduce (iv).

Proof of Claim C. Let $t = 1/m$. Define $\alpha^m[u_2](s)) = \psi(x, u_2(s))$ for $s \in [0, t)$. Assuming that we have defined $\alpha^m[u_2](\cdot), x(\cdot)$ on $[0, it)$, we extend its definition to $[0, (i+1)t)$ as follows. Let $x(\cdot)$ satisfy (1.3) in $(0, it)$ with the initial condition $x(0) = x$ under the controls $(\alpha^m[u_2](\cdot), u_2(\cdot))$. Then define $\alpha^m[u_2](s) = \psi(x((it)^-, u_2(s)))$ for $s \in [it, (i + 1)t)$. This defines $\alpha^m \in \Gamma$.

Now let $x(\cdot)$ denote the solution of (1.3) with the initial condition $x(0) = x$ under the controls $(\alpha^m[u_2](\cdot), u_2(\cdot))$. Then, for any i, as in Claim A, we can show that

$$w(x((i + 1)t)) - w(x(it)) \le C\, t^2 + (\rho + \epsilon)t - \int_{it}^{(i+1)t} r(x(s), \alpha^m[u_2](s), u_2(s))\, ds.$$

Summing over i from 0 to $Nm - 1$, we obtain Claim C. □

As an immediate consequence of the theorem, we obtain the following comparison principle.

COROLLARY 2.4. *Assume that* (ρ, w), $(\bar\rho, \bar w)$ *are viscosity sub- and supersolutions of* (2.1) *(or* (2.2)*). Then,* $\rho \le \bar\rho$.

Proof. We prove for the case of (2.1). The proof of (2.2) follows similarly. By parts (i) and (ii) of Theorem 2.3, we have

$$\rho \le \sup_{\beta\in\Delta} \inf_{u_1(\cdot)\in\mathcal{A}_1} \liminf_{T\to\infty} \frac{1}{T} \int_0^T r(x(s), u_1(s), \beta[u_1](s))\, ds$$

and

$$\bar\rho \ge \sup_{\beta\in\Delta} \inf_{u_1(\cdot)\in\mathcal{A}_1} \limsup_{T\to\infty} \frac{1}{T} \int_0^T r(x(s), u_1(s), \beta[u_1](s))\, ds.$$

Hence $\rho \le \bar\rho$. □

Remark 2.5. In this corollary, we have not assumed any growth on w and $\bar w$. If w and $\bar w$ are given to be bounded, then one can give a very simple proof of this comparison principle using comparison principle for stationary HJI equations (see [12]).

Note that under assumptions (A1)–(A3), if (2.1) has a viscosity solution (ρ, w), then $\rho = \rho^+$, and if (2.2) has a viscosity solution $(\bar\rho, w)$, then $\rho = \rho^-$, using Theorem 2.3. Thus if the Isaacs minimax condition holds, i.e., for any $x, p \in \mathbf{R}^d$, if we have

$$\inf_{u_2\in U_2} \sup_{u_1\in U_1} \{b(x, u_1, u_2) \cdot p + r(x, u_1, u_2)\} = \sup_{u_1\in U_1} \inf_{u_2\in U_2} \{b(x, u_1, u_2) \cdot p + r(x, u_1, u_2)\},$$

then, using Fan's minimax theorem [8] we can deduce the following result. We omit the details.

THEOREM 2.6. *Assume that the Isaacs minimax condition holds. Assume that* (ρ, w) *is a viscosity solution of* (2.1) *or equivalently of* (2.2)*. Then* $\rho = \rho^+(x) = \rho^-(x)$ *for all* $x \in \mathbf{R}^d$.

By interchanging the roles of taking limits as $T \to \infty$ and taking infimum and supremum over controls in the proof of the Theorem 2.3, we obtain the following result.

THEOREM 2.7. *Let* (ρ, w) *be a viscosity solution of* (2.1)*. Then*

$$\rho = \lim_{T\to\infty} \sup_{\beta\in\Delta} \inf_{u_1(\cdot)\in\mathcal{A}_1} \frac{1}{T} \int_0^T r(x(s), u_1(s), \beta[u_1](s))\, ds.$$

Similarly, if $(\bar{\rho}, \bar{w})$ *is a viscosity solution of* (2.2)*, then*

$$\bar{\rho} = \lim_{T\to\infty} \inf_{\alpha\in\Gamma} \sup_{u_2(\cdot)\in\mathcal{A}_2} \frac{1}{T} \int_0^T r(x(s), \alpha[u_2](s), u_2(s))\, ds.$$

Remark 2.8. Let $w^+(T,x)$ and $w^-(T,x)$ denote the upper and lower value functions of the finite horizon problem with horizon T, dynamics (1.3), payoff function r, and zero terminal cost; i.e., they are defined as follows:

$$w^+(T,x) := \sup_{\beta\in\Delta} \inf_{u_1(\cdot)\in\mathcal{A}_1} \int_0^T r(s, x(s), u_1(s), \beta[u_1](s))\, ds$$

and

$$w^-(T,x) := \inf_{\alpha\in\Gamma} \sup_{u_2(\cdot)\in\mathcal{A}_2} \int_0^T r(s, x(s), \alpha[u_2](s), u_2(s))\, ds,$$

where $x(\cdot)$ is solution of (1.3) with the initial condition $x(0) = x$ under respective controls. Then the conclusion of the above theorem can be restated as

$$\rho = \lim_{T\to\infty} \frac{w^+(T,x)}{T} \text{ and } \bar{\rho} = \lim_{T\to\infty} \frac{w^-(T,x)}{T}.$$

This can be seen as the longtime behavior of viscosity solutions of HJI equations corresponding to differential games on finite horizon. We refer to [2], [14] for the study of longtime behavior of viscosity solutions of Hamilton–Jacobi equations.

We now give another representation formula for ρ in terms of the discounted value of the differential game. Let w_λ denote the upper value of the differential game on an infinite horizon with discount factor $\lambda > 0$, i.e.,

$$w_\lambda(x) = \sup_{\beta\in\Delta} \inf_{u_1(\cdot)\in\mathcal{A}_1} \int_0^\infty e^{-\lambda s} r(x(s), u_1(s), \beta[u_1](s))\, ds;$$

then

$$\rho = \lim_{\lambda\to 0} \lambda w_\lambda(x).$$

An analogous statement holds for the lower value function. This is the content of our next result. We closely follow the arguments in the proof of Theorem 2.3.

THEOREM 2.9. (i) *Let* (ρ, w) *be a viscosity solution of* (2.1)*. Then*

$$\rho = \lim_{\lambda\to 0} \sup_{\beta\in\Delta} \inf_{u_1(\cdot)\in\mathcal{A}_1} \lambda \int_0^\infty e^{-\lambda s} r(x(s), u_1(s), \beta[u_1](s))\, ds.$$

(ii) *Similarly, if* (ρ, w) *is a viscosity solution of* (2.2)*, then*

$$\rho = \lim_{\lambda\to 0} \inf_{\alpha\in\Gamma} \sup_{u_2(\cdot)\in\mathcal{A}_2} \lambda \int_0^\infty e^{-\lambda s} r(x(s), \alpha[u_2](s), u_2(s))\, ds.$$

Proof. We prove only (ii); (i) can be proved in an analogous way. Again we prove this under the additional assumption that w is $C^{1,1}$. The proof of the general case can be done as before.

Fix x. Let $\beta^m \in \Delta$ be as in the proof of Theorem 2.3. Let $u_1(\cdot) \in \mathcal{A}_1$. Let $x(\cdot)$ denote the solution of (1.3) with the initial condition $x(0) = x$ under the controls $(u_1(\cdot), \beta^m[u_1](\cdot))$. Then for a.e. s,

$$\frac{d}{ds}e^{-\lambda s}w(x(s)) = e^{-\lambda s}b(x(s), u_1(s), \bar{u}_2) \cdot Dw(x(s)) - \lambda e^{-\lambda s}w(x(s)).$$

Now following the arguments in the proof of Claim A of Theorem 2.3, we obtain

$$\begin{aligned}
&e^{-\lambda(i+1)t}w(x((i+1)t)) - e^{-\lambda it}w(x(it)) \\
&\qquad = \int_{it}^{(i+1)t} e^{-\lambda s} Dw(x(s)) \cdot b(x(s), u_1(s), \beta^m[u_1](s))\, ds \\
&\qquad \geq -C \int_{it}^{(i+1)t} e^{-\lambda s}(s - i\,t)\, ds + (\rho - \epsilon) \int_{it}^{(i+1)t} e^{-\lambda s}\, ds \\
&\qquad\quad - \int_{it}^{(i+1)t} e^{-\lambda s} r(x(s), u_1(s), \beta^m[u_1](s))\, ds \\
&\qquad \geq -C\, t \frac{1}{\lambda}\left[e^{-\lambda it} - e^{-\lambda(i+1)t}\right] + (\rho - \epsilon)\frac{1}{\lambda}\left[e^{-\lambda it} - e^{-\lambda(i+1)t}\right] \\
&\qquad\quad - \int_{it}^{(i+1)t} e^{-\lambda s} r(x(s), u_1(s), \beta^m[u_1](s))\, ds.
\end{aligned}$$

Adding these inequalities for $i = 0, \ldots, Nm - 1$, and multiplying by λ, we get

$$\begin{aligned}
\lambda e^{-\lambda N} w(x(N)) - \lambda w(x) \geq C\frac{1}{m}[1 - e^{-\lambda N}] + (\rho - \epsilon)[1 - e^{-\lambda N}] \\
- \lambda \int_0^N e^{-\lambda s} r(x(s), u_1(s), \beta^m[u_1](s))\, ds.
\end{aligned}$$

Now letting $N \to \infty$, we obtain

$$\rho - \epsilon + \lambda w(x_0) \leq \lambda \int_0^\infty e^{-\lambda s} r(x(s), u_1(s), \beta^m[u_1](s))\, ds - C\frac{1}{m}.$$

Using (2.8), we get

$$\rho - \epsilon + \lambda w(x_0) \leq \inf_{\alpha \in \Gamma} \sup_{u_2(\cdot) \in \mathcal{A}_2} \lambda \int_0^\infty e^{-\lambda s} r(x(s), \alpha[u_2](s), u_2(s))\, ds.$$

Now taking limit as $\lambda \to 0$ and then $\epsilon \to 0$, we get

$$\rho \leq \liminf_{\lambda \to 0} \inf_{\alpha \in \Gamma} \sup_{u_2(\cdot) \in \mathcal{A}_2} \lambda \int_0^\infty e^{-\lambda s} r(x(s), \alpha[u_2](s), u_2(s))\, ds.$$

Similarly, we can obtain

$$\rho \geq \limsup_{\lambda \to 0} \inf_{\alpha \in \Gamma} \sup_{u_2(\cdot) \in \mathcal{A}_2} \lambda \int_0^\infty e^{-\lambda s} r(x(s), \alpha[u_2](s), u_2(s))\, ds.$$

This completes part (ii). □

Remark 2.10. If (ρ, w) is a viscosity subsolution of (2.2), then note that the following result holds:

$$\rho \le \inf_{\alpha\in\Gamma} \sup_{u_2(\cdot)\in\mathcal{A}_2} \liminf_{\lambda\to 0} \lambda \int_0^\infty e^{-\lambda s} r(x(s), \alpha[u_2](s), u_2(s))\, ds.$$

Similar statements hold for the other cases.

We now present a dynamic programming principle for the viscosity solutions of (2.1) and (2.2).

THEOREM 2.11. (i) *Let (ρ, w) be a viscosity solution of* (2.1). *Then for any $T > 0$,*

$$w(x) = \sup_{\beta\in\Delta} \inf_{u_1(\cdot)\in\mathcal{A}_1} \left[\int_0^T r(x(s), u_1(s), \beta[u_1](s))\, ds + w(x(T)) \right] - \rho T.$$

(ii) *Let (ρ, w) be a viscosity solution of* (2.2). *Then for any $T > 0$,*

$$w(x) = \inf_{\alpha\in\Gamma} \sup_{u_2(\cdot)\in\mathcal{A}_2} \left[\int_0^T r(x(s), \alpha[u_2](s), u_2(s))\, ds + w(x(T)) \right] - \rho T.$$

Proof. We prove (ii); (i) can be proved analogously. Let $T > 0$ and m a positive integer. Take $t = T/m$. As in Claim C, we obtain $\alpha^m(.)$, given ϵ, $u_2(.)$, such that

$$w(x(T)) - w(x) \le -\int_0^T r(x(s), u_1(s), u_2(s))\, ds + (\rho + \epsilon)T - C\frac{T^2}{m}.$$

Therefore

$$w(x) \ge \inf_{\alpha\in\Gamma} \sup_{u_2(\cdot)\in\mathcal{A}_2} \left(\int_0^T r(x(s), \alpha[u_2](s), u_2(s))\, ds + w(x(T)) \right) - \rho T.$$

We can prove the other inequality similarly. □

We now turn our attention to the uniqueness of w. Define a set Z as follows: $z \in Z$ if $z = \lim_{t_n\to\infty} x(t_n)$, where $t_n \to \infty$ and $x(\cdot)$ is a solution of (1.3) with an initial condition $x(0) = x_0$ for some $x_0 \in \mathbf{R}^d$ under some controls $(u_1(\cdot), u_2(\cdot)) \in \mathcal{A}_1 \times \mathcal{A}_2$. Then Z is nonempty under assumption (A3). We now show that if (ρ, w_1) and (ρ, w_2) are two viscosity solutions of (2.1) such that $w_1 \equiv w_2$ on Z, then $w_1 \equiv w_2$.

THEOREM 2.12. *Let (ρ, w_1) and (ρ, w_2) be two viscosity solutions of* (2.1) *such that $w_1 \equiv w_2$ on Z. Then $w_1 \equiv w_2$. An analogous result holds for* (2.2).

Proof. We prove this for the case when w_1, w_2 are $C^{1,1}$. The general case follows similarly as in the proof of Theorem 2.3. Let m be a positive integer. Let α^m be as in Claim C when we take $w = w_2$, and let β^m be as in Claim A when we take $w = w_1$. Taking α^m as α in (2.9), we obtain $\tilde{u}_1(\cdot) \in \mathcal{A}_1$ and $\tilde{u}_2(\cdot) \in \mathcal{A}_2$ such that

$$\beta^m[\tilde{u}_1](\cdot) = \tilde{u}_2(\cdot) \text{ and } \alpha^m[\tilde{u}_2](\cdot) = \tilde{u}_1(\cdot).$$

Using this, we obtain

$$w_1(x(N)) - w_1(x) \ge -\int_0^N r(x(s), \alpha^m[\tilde{u}_2](s), \tilde{u}_2(s))\, ds + (\rho - \epsilon)N - C\frac{N}{m}$$

and

$$w_2(x(N)) - w_2(x) \leq -\int_0^N r(x(s), \alpha^m[\tilde{u}_2](s), \tilde{u}_2(s))\ ds + (\rho - \epsilon)N + C\frac{N}{m}.$$

From these two inequalities, we obtain

$$w_1(x) - w_2(x) \leq w_1(x(N)) - w_2(x(N)) + 2C\frac{N}{m}. \tag{2.11}$$

Using the compactness and equi-Lipschitz continuity of trajectories, we get a trajectory $\bar{x}(\cdot)$ such that $x(\cdot) \to \bar{x}(\cdot)$ as $m \to \infty$. (Note that $x(\cdot)$ above depends on m.) Now from (2.11) we obtain by letting $m \to \infty$

$$w_1(x) - w_2(x) \leq w_1(\bar{x}(N)) - w_2(\bar{x}(N)).$$

Now letting $N \to \infty$, we see that

$$w_1(x) - w_2(x) \leq 0.$$

Similarly, we can prove

$$w_2(x) - w_1(x) \leq 0.$$

Thus $w_1 \equiv w_2$. □

Remark 2.13. The uniqueness result in [13] is established under certain growth conditions on the solutions. Here we have obtained similar results without any such condition. Our uniqueness result, however, is not complete. We have shown that if two solutions coincide on the set Z, then they are identical. In view of this, it would be interesting to investigate the structure of Z.

3. Existence results. In the previous section, we studied some representation formulas related to the viscosity solutions of (2.1) and (2.2). We now study the existence of viscosity solutions to (2.1) and (2.2). We refer to [9] for analogoues results. Here we present two simple proofs of the existence result.

To this end we make the following assumption.

(A4) There exists a constant $C_3 > 0$ such that for all $x, y \in \mathbf{R}^d$ and $(u_1, u_2) \in U_1 \times U_2$,

$$\langle b(x, u_1, u_2) - b(y, u_1, u_2),\ x - y\rangle \leq -C_3|x - y|^2.$$

Remark 3.1. (i) Let $(u_1(\cdot), u_2(\cdot)) \in \mathcal{A}_1 \times \mathcal{A}_2$. Let $x(\cdot)$ and $y(\cdot)$ denote the solutions of (1.3) with the initial conditions $x(0) = x$ and $y(0) = y$, respectively, under these controls. Then using (A4), we get

$$\frac{d}{dt}|x(t) - y(t)|^2 \leq -C_3|x(t) - y(t)|^2.$$

Now using Gronwall's inequality, we obtain

$$|x(t) - y(t)| \leq |x - y|e^{-C_4 t}$$

for a constant $C_4 > 0$.

(ii) Using Gronwall's inequality, it is easy to see that (A1) and (A4) together imply (A3).

We now give some examples where (A4) holds.

Example 3.2. (i) *Let U_1, U_2 be subsets of $\mathbf{R}^m$ and $\mathbf{R}^q$, respectively, for some m and q. Let b be given by*

$$b(x, u_1, u_2) = Bx + C_1u_1 + C_2u_2 + b_1(x, u_1, u_2),$$

where B is a $d\times d$ matrix, C_1 a $d\times m$ matrix, C_2 a $d\times q$ matrix, and $b_1 : \mathbf{R}^d\times U_1\times U_2 \to \mathbf{R}^d$. We assume the following:

$$\exists \alpha > 0 \text{ such that } \langle Bx, x\rangle \le -\alpha|x|^2$$

and

$$|b_1(x, u_1, u_2) - b(y, u_1, u_2)| \le \alpha_1|x - y| \text{ for some } \alpha_1 < \alpha.$$

Under these assumptions, it is easy to verify that (A4) *is satisfied.*

(ii) *Let U_1, U_2 be as above and let b be given by*

$$b(x, u_1, u_2) = A + B_1u_1 + B_2u_2 + \bar{b}(x),$$

where A is a $d\times d$ matrix, B_1 a $d\times m$ matrix, and B_2 a $d\times q$ matrix. Assume that there are matrices C_1, C_2 of orders $d\times m$ and $d\times q$, respectively, such that $A+B_1C_1+B_2C_2$ is stable. Further assume that $\bar{b}$ is bounded and Lipschitz continuous. Then (A4) *is satisfied.*

We now prove the existence via the vanishing limit in the discounted payoff criterion.

THEOREM 3.3. *Assume* (A1), (A2), *and* (A4). *Let w_λ be the unique viscosity solution in the class of linear growth functions of*

$$\lambda w_\lambda(x) = \inf_{u_1(\cdot)\in\mathcal{A}_1} \sup_{u_2(\cdot)\in\mathcal{A}_2} \left(b(x, u_1, u_2)\cdot Dw_\lambda(x) + r(x, u_1, u_2)\right). \tag{3.1}$$

Then $\lambda w_\lambda(x) \to \rho$, a constant as $\lambda \to 0$. Also for any $\bar{x} \in \mathbf{R}^d$, $w_\lambda(\cdot) - w_\lambda(\bar{x})$ converges uniformly on compact sets to a continuous function $w(\cdot)$. Thus (ρ, w) is a viscosity solution of (2.1) *for any $\bar{x} \in \mathbf{R}^d$. Moreover, $\rho = \rho^+(x)$ for all $x \in \mathbf{R}^d$. An analogous result holds for the existence of a viscosity solution to* (2.2).

Proof. Using standard results in differential games and viscosity solutions [1], we have

$$w_\lambda(x) = \sup_{\beta\in\Delta} \inf_{u_1(\cdot)\in\mathcal{A}_1} \int_0^\infty e^{-\lambda t} r(x(t), u_1(t), \beta[u_1](t))\, dt.$$

Let $u_1(\cdot) \in \mathcal{A}_1$ and $u_2(\cdot) \in \mathcal{A}_2$. Then using Remark 3.1(i), we see that

$$\left| \int_0^\infty e^{-\lambda s} r(x(s), u_1(s), u_2(s))\, ds - \int_0^\infty e^{-\lambda s} r(y(s), u_1(s), u_2(s))\, ds \right| \le \frac{1}{C_4 + \lambda}|x - y|,$$

where $x(\cdot), y(\cdot)$ are solutions of (1.3) with initial conditions $x(0) = x$ and $y(0) = y$, respectively, under the controls $(u_1(\cdot), u_2(\cdot))$. Using this fact, it is easy to note that w_λ is Lipschitz continuous where the Lipschitz constant is independent of λ. Therefore by Ascoli-Arzela's theorem for a fixed $\bar{x}$, $w_\lambda(x) - w_\lambda(\bar{x})$ converges locally uniformly to a continuous function $w(x)$ and $\lambda w_\lambda(x)$ converges to a constant ρ. By the stability of viscosity solutions, we note that (ρ, w) is a viscosity solution of (2.1). Now by Theorem 2.6, $\rho = \rho^+(x)$ for all $x \in \mathbf{R}^d$. ∎

We now turn our attention to the increasing horizon limit case. Let $T > 0$ and w_0 be any Lipschitz continuous function. Now consider the HJI equation in $(0,T) \times \mathbf{R}^d$,

$$\left\{ \begin{array}{rcl} w_t(t,x) & = & \inf\limits_{u_1 \in U_1} \sup\limits_{u_2 \in U_2} \{b(x,u_1,u_2) \cdot Dw(t,x) + r(x,u_1,u_2)\}, \\ w(0,x) & = & w_0(x). \end{array} \right. \tag{3.2}$$

Then we have the following theorem.

THEOREM 3.4. *Assume* (A1), (A2), *and* (A4). *Let* $w(t,x)$ *be the unique viscosity solution of* (3.2) *in the class of linear growth functions. Then* $\frac{w(T,x)}{T} \to \rho$, *a constant, and* $w(T,x) - \rho T$ *converges locally uniformly to a continuous function* $w_\infty(x)$ *such that* (ρ, w_∞) *is a viscosity solution of* (2.1). *Moreover,* $\rho = \rho^+(x)$ *for all* $x \in \mathbf{R}^d$. *An analogous results holds for* (2.2).

Proof. Using standard results in differential games and viscosity solutions [7], we have the following representation formula for $w(t,x)$:

$$w(T,x) = \sup_{\beta \in \Delta} \inf_{u_1(\cdot) \in \mathcal{A}_1} \left[\int_0^T r(x(s), u_1(s), \beta[u_1](s))\, ds + w_0(x(T)) \right].$$

As in above theorem, using Remark 3.1(i), we can show that

$$|w(T,x) - w(T,y)| \leq \frac{1 - e^{-C_{10}T}}{C_{10}} |x - y|.$$

Using Ascoli-Arzela's theorem, it is easy to see that $\frac{w(T,x)}{T} \to \rho$, a constant $w(T,x) - \rho T \to w_\infty(x)$ locally uniformly to a continuous function $w_\infty(x)$. We now need to show that (ρ, w_∞) is a viscosity solution of (2.1). Let

$$w^\epsilon(t,x) = w(t/\epsilon, x) \text{ for } t \in [0,1].$$

Then $w^\epsilon(t,x) - \rho \frac{t}{\epsilon} \to w_\infty(t,x)$ locally uniformly as $\epsilon \to 0$. Now it is easy to see that w^ϵ is viscosity solution of

$$\left\{ \begin{array}{rcl} \epsilon w_t^\epsilon(t,x) & = & \inf_{u_1 \in U_1} \sup_{u_2 \in U_2} \{b(x,u_1,u_2) \cdot Dw^\epsilon(t,x) + r(x,u_1,u_2)\}, \\ w(0,x) & = & w_0(x) \end{array} \right.$$

in $(0,1) \times \mathbf{R}^d$. Using the stability of viscosity solutions [5], we get that (ρ, w_∞) is a viscosity solution of (2.1). This completes the proof. □

4. Conclusions. In this paper, we have studied a zero sum differential game with ergodic payoff. We have identified the scalar appearing in the HJI equation as the ergodic value. Under a dissipativity-type condition, we have also established the existence of a viscosity solution to HJI equations. We have carried out two asymptotics, namely, we have shown that the ergodic value is the vanishing limit of the discounted value. At the same time, the ergodic value is also the time averaged limit of the finite horizon value. Finally we wish to mention that although we have identified the scalar appearing in the HJI equation as the ergodic value, we have not been able to establish the uniqueness (in some sense) of the solution of the HJI equation. We have obtained only a partial uniqueness result. Thus the uniqueness issue and the existence of viscosity solution to HJI equations under (A3) alone still remain problems that need further investigation.

Acknowledgments. The authors thank two anonymous referees for pointing out some errors and for drawing attention to related literature in an earlier version of this paper. They also thank M. Rajesh for pointing out an error in a preliminary draft of this paper.

REFERENCES

[1] M. Bardi and I. C. Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[2] G. Barles and P. E. Souganidis, *On the large time behavior of solutions of Hamilton–Jacobi equations*, SIAM J. Math. Anal., 31 (2000), pp. 925–939.

[3] T. Basar and P. Bernhard, $\mathcal{H}_\infty$ *Optimal Control and Related Minimax Design Problems*, Birkhäuser, Boston, 1990.

[4] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[5] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[6] P. M. Dower and R. D. James, *Worst case power generating capabilities of nonlinear systems*, Math. Control Signal Systems, 15 (2002), pp. 13–41.

[7] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J, 33 (1984), pp. 773–797.

[8] K. Fan, *Fixed point and minimax theorems in locally convex topological linear spaces*, Proc. Nat. Acad. Sci. U.S.A., 38 (1952), pp. 121–126.

[9] W. H. Fleming and W. M. McEnneaney, *Risk-sensitive control on an infinite horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.

[10] R. Jensen, P. L. Lions, and P. E. Souganidis, *A uniqueness result for viscosity solutions of second order fully nonlinear partial differential equations*, Proc. Amer. Math. Soc., 102 (1988), pp. 975–978.

[11] J. M. Lasry and P. L. Lions, *A remark on regularization in Hilbert spaces*, Israel J. Math. 55 (1986), pp. 257–266.

[12] P. L. Lions, G. Papanicolaou, and S. R. S. Varadhan, *Homogenization of HJB Equations*, manuscript.

[13] W. M. McEneaney, *A Uniqueness result for the Isaacs equation corresponding to nonlinear* $\mathcal{H}_\infty$ *control*, Math. Control Signals Systems, 11 (1998), pp. 303–334.

[14] G. Namah and J. M. Roquejoffre, *Remarks on the long time behaviour of the solutions of Hamilton-Jacobi equations*, Comm. Partial Differential Equations, 24 (1999), pp. 883–893.

[15] P. Soravia, *Stability of dynamical systems with competitive controls: The degenerate case*, J. Math. Anal. Appl., 191 (1995), 428–449.

[16] P. Soravia, $\mathcal{H}_\infty$ *Control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.

[17] A. Świech, *Sub- and superoptimality principles of dynamic programming revisited*, Nonlinear Anal., 26 (1996), pp. 1429–1436.

[18] A. Świech, *Another approach to the existence of value functions of stochastic differential games*, J. Math. Anal. Appl., 204 (1996), pp. 884–897.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2036–2048

NONCONVEX DUALITY IN OPTIMAL CONTROL*

F. H. CLARKE† AND C. NOUR†

Abstract. We derive a representation of the minimum cost of an optimal control problem in terms of the upper envelope of generalized semisolutions of the Hamilton–Jacobi equation.

Key words. optimal control, duality, Hamilton–Jacobi equations, viscosity solutions, proximal analysis, nonsmooth analysis

AMS subject classifications. 49J24, 49L25, 49J52

DOI. 10.1137/S0363012903429554

1. Introduction.

We consider in this article the following optimal control problem (P) in Mayer form: to minimize the cost functional

$$\ell(T, x(T))$$

over the (absolutely continuous) trajectories x on some interval $[0, T]$ of a differential inclusion

$$\dot{x}(t) \in F(x(t)) \text{ a.e.}$$

subject to the boundary conditions and state constraint

$$x(0) = x_0,\ x(T) \in C,\ x(t) \in A\ \forall\, t \in [0, T].$$

The horizon $T \geq 0$ is free in this problem. We assume that the set A is closed, that C is compact, and that the extended-valued function $\ell : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous and satisfies the following growth condition:

$$\text{(GC)} \qquad \lim_{t \to +\infty} \inf_x \ell(t, x) = +\infty.$$

As for the multivalued function F, we assume that it takes nonempty compact convex values, has closed graph, and satisfies a linear growth condition: for some positive constants γ and c, and for all $x \in \mathbb{R}^n$,

$$v \in F(x) \Longrightarrow \|v\| \leq \gamma \|x\| + c.$$

Finally, we assume that (P) is nontrivial in the sense that there is at least one admissible trajectory for which the cost is finite.

We associate with F the following function h, the (lower) Hamiltonian:

$$h(x, p) := \min\{\langle p, v\rangle : v \in F(x)\}.$$

*Received by the editors June 5, 2003; accepted for publication (in revised form) July 30, 2004; published electronically April 14, 2005.
http://www.siam.org/journals/sicon/43-6/42955.html

†Institut Girard Desargues, Université Lyon I, 21 avenue Claude Bernard, 69622 Villeurbanne Cedex, France (clarke@igd.univ-lyon1.fr, chadi@igd.univ-lyon1.fr). The first author is a member of IUF (Institut Universitaire de France), and the research of the second author was supported by LNCSR (Lebanese National Council of Scientific Research).

There is a well-known relationship between the value of the problem (P) on the one hand and certain solutions of a Hamilton–Jacobi inequality in terms of h on the other; let us recall it now. Let $\psi(t,x)$ be a smooth function satisfying

$$\psi_t(t,x) + h(x,\psi_x(t,x)) \geq 0 \ \ \forall (t,x), \ \ \psi(t,x) \leq \ell(t,x) \text{ if } x \in C.$$

Then it follows from the Hamiltonian inequality that for any trajectory $x(\cdot)$ of the differential inclusion, the function $t \mapsto \psi(t,x(t))$ is nondecreasing. Accordingly, for any $T \geq 0$ we have $\psi(T,x(T)) \geq \psi(0,x(0))$. If we now restrict the trajectories to those that are admissible for (P), we deduce

$$\ell(T,x(T)) \geq \psi(0,x_0).$$

There results a lower bound for the value of (P):

$$\inf(P) \geq \sup \psi(0,x_0),$$

where the supremum is taken over all functions ψ as described above.

The term *nonconvex duality* has been applied in optimal control to situations in which equality holds in this last relation. The basic idea is at the heart of Carathéodory's method in the calculus of variations, which is also known as that of *verification functions*. It is also related to the *generalized flows* of Young, as observed by Vinter [27], and Vinter and Lewis [29], [30], who have extended this duality to the setting of optimal control; see also [10], [12], [16], [21], [25], [26], [33], and [34].

In addition to the above equality, it is natural to ask under what conditions the supremum on the right is attained (see Remark 2.5). Since this fails in general for smooth functions ψ, this question led Clarke and others to introduce a generalized solution concept for nonsmooth solutions of the Hamilton–Jacobi equation [4], [11], [19], [22]; these have turned out to be what is now known as viscosity semisolutions. The method of verification functions extended in this way has been used to solve explicitly a number of problems in optimal control; see Clarke [5] for a thorough discussion of the method and for examples.

The issue of generalized solutions inevitably involves nonsmooth analysis, and here we employ the tools of proximal analysis [8]. Let us recall a basic definition in the subject. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function with dom $f := \{x : f(x) < +\infty\} \neq \emptyset$. A vector $\xi \in \mathbb{R}^n$ is a proximal subgradient of f at $x \in \text{dom} f$ if and only if there exist positive numbers σ and ν such that

$$f(y) - f(x) + \sigma \|y - x\|^2 \geq \langle \xi, y - x \rangle \ \ \forall y \in B(x;\nu).^1$$

The set (which could be empty) of all proximal subgradients of $f(\cdot)$ at x is denoted by $\partial_P f(x)$ and is referred to as the proximal subdifferential. The proximal density theorem asserts that $\partial_P f(x) \neq \emptyset$ for all x in a dense subset of dom$f$. The limiting subdifferential of f at $x \in \text{dom} f$ is defined by

$$\partial_L f(x) := \{\lim \xi_i : \xi_i \in \partial_P f(x_i), x_i \longrightarrow x \text{ and } f(x_i) \longrightarrow f(x)\}.$$

We remark that one can use the proximal subdifferential to define generalized solutions of the Hamilton–Jacobi equation, and the resulting solutions coincide (in the

[1] $B(x;\nu) := \{y \in \mathbb{R}^n : \|y - x\| < \nu\}$ and its closure is denoted by $\bar{B}(x;\nu)$. The open unit ball is denoted by B, and its closure by $\bar{B}$.

present context) with the viscosity solutions of Crandall and Lions and the minimax solutions of Subbotin (see, for example, [7], [8], [9], [13], [15], [20], and [31]). Let the extended Hamiltonian $\bar{h}$ be the function given by

$$\bar{h}(x,\theta,\zeta) := \theta + h(x,\zeta).$$

We define Ψ to be the set of all locally Lipschitz functions ψ on $\mathbb{R}\times\mathbb{R}^n$ that satisfy the proximal Hamilton–Jacobi inequality

$$\bar{h}(x,\partial_L\psi(t,x)) \geq 0 \ \ \forall (t,x)\in\mathbb{R}\times A$$

as well as the boundary condition

$$\psi(t,x)\leq \ell(t,x) \ \ \forall (t,x)\in\mathbb{R}\times C.$$

The following is the main result.

THEOREM 1.1.

$$\min(P) = \sup_{\psi\in\Psi}\psi(0,x_0).$$

This result unifies, and extends in a number of ways, the ones in the literature, which treat for the most part the fixed-horizon case (see, for example, [2], [3], [8, Chapter 4], [17], [18], [23], [24], and [32]). We remark that the fixed-horizon case is obtained by taking $\ell(T,x)$ equal to $+\infty$ whenever T differs from the given horizon T_0; see section 3 for a discussion of this and other special cases. Our theorem, whose proof is self-contained modulo some basic facts from proximal analysis, is also new with respect to its very mild regularity hypotheses on F (which need not even be continuous), as well as the presence of a unilateral state constraint. The fact that locally Lipschitz functions figure in our duality also gives easy access to smooth duality of the type found by Vinter [27]. Furthermore, we extend his result by obtaining in the case of the minimal time problem a duality which features only smooth solutions of an *autonomous* Hamilton–Jacobi inequality. Section 3 is devoted to specializations such as these, while the next section gives the proof of the theorem.

2. Proof of Theorem 1.1. First we note that under our hypotheses on F, any trajectory can be extend indefinitely both forward and backward, so all trajectories can be considered as being defined on $]-\infty,+\infty[$. Using (GC) and since C is compact, we have that ℓ is bounded below over $[0,+\infty[\times C$. Then we can assume that ℓ is bounded below over $\mathbb{R}\times\mathbb{R}^n$ by a constant ω. By the compactness property of trajectories and by (GC), it is easy to prove that the problem (P) admits a solution. For all $k\in\mathbb{N}^*$, we consider the function ℓ_k defined by

$$\ell_k(t,x) := \inf_{(\tau,y)\in\mathbb{R}\times\mathbb{R}^n}\{\ell(\tau,y)+k\|(t-\tau,x-y)\|^2\}. \tag{2.1}$$

The sequence $(\ell_k)_k$ is the quadratic inf-convolution sequence of ℓ. The following lemma, whose standard proof we omit (see [8, Theorem 1.5.1]), gives some properties of ℓ_k.

LEMMA 2.1. *For all $k\in\mathbb{N}^*$, we have*

(i) *$\ell_k(\cdot)\leq\ell(\cdot)$ and the set of minimizing points (τ,y) in (2.1) is nonempty;*

(ii) *ℓ_k is locally Lipschitz, bounded below by ω and satisfies (GC);*

(iii) *For all* $(t,x) \in \mathbb{R} \times \mathbb{R}^n$,

$$\lim_{k \longrightarrow +\infty} \ell_k(t,x) = \ell(t,x).$$

We also consider a locally Lipschitz approximation for the multifunction F. By [8, Proposition 4.4.4] there exists a sequence of locally Lipschitz multifunctions $\{F_k\}$ also satisfying the hypotheses of F such that

- for each $k \in \mathbb{N}$, for every $x \in \mathbb{R}^n$,

$$F(x) \subseteq F_{k+1}(x) \subseteq F_k(x) \subseteq \overline{co}\, F(x + 3^{-k+1}B);$$

- $\bigcap_{k \geq 1} F_k(x) = F(x)\ \forall x \in \mathbb{R}^n$.

A well-known method of approximating the terminally constrained problem (P) by a problem free of such constraints involves the imposition of a penalty term. To use this technique, the inf-convolution technique and the preceding approximation we consider for all $k \geq 1$ the following optimal control problem:

(P_k): Minimize $\ell_k(T, x(T)) + k d_C(x(T)) + k \int_0^T d_A(x(t))\, dt$ over points $T \geq 0$ and absolutely continuous functions $x : [0, +\infty[\longrightarrow \mathbb{R}^n$ that satisfy

$$\dot{x}(t) \in F_k(x(t)) \quad \text{a.e. } t \in [0, +\infty[, \quad x(0) = x_0.$$

LEMMA 2.2. *There exists a sequence* λ_n *strictly increasing in* $\mathbb{N}^*$ *such that*

$$\lim_{n \longrightarrow +\infty} \min(P_{\lambda_n}) = \min(P).$$

Proof. We denote by $(\bar{T}, \bar{x}(\cdot))$ a solution of (P) and by $(T_k, \bar{y}_k(\cdot))$ a solution of (P_k), (the existence of the solutions is easy to check). For all $k \in \mathbb{N}^*$ we have

$$\min(P_k) = \ell_k(T_k, \bar{y}_k(T_k)) + k d_C(\bar{y}_k(T_k)) + k \int_0^{T_k} d_A(\bar{y}_k(t))\, dt$$

and

$$\min(P) = \ell(\bar{T}, \bar{x}(\bar{T})).$$

Since $F \subset F_k$, $(\bar{T}, \bar{x}(\cdot))$ is admissible for (P_k) we have

$$\begin{aligned} \min(P_k) &= \ell_k(T_k, \bar{y}_k(T_k)) + k d_C(\bar{y}_k(T_k)) + k \int_0^{T_k} d_A(\bar{y}_k(t))\, dt \\ &\leq \ell_k(\bar{T}, \bar{x}(\bar{T})) \\ &\leq \ell(\bar{T}, \bar{x}(\bar{T})) \\ &= \min(P), \end{aligned}$$

hence

$$\omega \leq \ell_k(T_k, \bar{y}_k(T_k)) \leq \ell(\bar{T}, \bar{x}(\bar{T})) \tag{2.2}$$

and

$$k d_C(\bar{y}_k(T_k)) + k \int_0^{T_k} d_A(\bar{y}_k(t))\, dt \leq \ell(\bar{T}, \bar{x}(\bar{T})) - \omega. \tag{2.3}$$

By Lemma 2.1 there exists $(s_k, z_k) \in \mathbb{R} \times \mathbb{R}^n$ such that

$$\ell_k(T_k, \bar{y}_k(T_k)) = \ell(s_k, z_k) + k\|(s_k - T_k, z_k - \bar{y}_k(T_k))\|^2, \tag{2.4}$$

and hence using (2.2) we get that

$$k\|(s_k - T_k, z_k - \bar{y}_k(T_k))\|^2 \leq \ell_k(T_k, \bar{y}_k(T_k)) - \omega \leq \ell(\bar{T}, \bar{x}(\bar{T})) - \omega. \tag{2.5}$$

We claim that $\lim_{k \longrightarrow +\infty} T_k \neq +\infty$. Indeed, by (2.5) we have that there exists a sequence a_n strictly increasing in $\mathbb{N}^*$ such that $\lim_{n \longrightarrow +\infty}(s_{a_n} - T_{a_n}) = 0$. Then if $\lim_{n \longrightarrow +\infty} T_{a_n} = +\infty$, we get that $\lim_{n \longrightarrow +\infty} s_{a_n} = +\infty$. Since ℓ satisfies (GC) we have

$$\lim_{n \longrightarrow +\infty} \ell(s_{a_n}, z_{a_n}) = +\infty,$$

and hence by (2.4)

$$\lim_{n \longrightarrow +\infty} \ell_{a_n}(T_{a_n}, \bar{y}_{a_n}(T_{a_n})) = +\infty,$$

but this contradicts (2.2). Then we can assume that there exists $b \geq 0$ such that $T_k \in [0, b]$.

Since $F_k(\cdot) \subseteq \overline{co}\, F(\cdot + 3^{-k+1}B)$ and using the compactness of trajectories (see [1] and [14]), there exists a trajectory $\bar{y}(\cdot)$ of F on $[0, +\infty[$ which satisfies $\bar{y}(0) = x_0$, and there is a subsequence $\bar{y}_{k_i}$ of $\bar{y}_k$ such that $\bar{y}_{k_i}$ converges uniformly to $\bar{y}(\cdot)$ on $[0, b]$. Moreover, since $T_k \in [0, b]$, there exists a subsequence of T_k, which converges to a point in $[0, b]$. These considerations with (2.2), (2.3), and (2.5) give that there exists a sequence λ_n strictly increasing in $\mathbb{N}^*$ such that

- the sequence T_{λ_n} converges to a $\bar{T}_0 \in [0, b]$,
- the sequence $\bar{y}_{\lambda_n}$ converges uniformly on $[0, b]$ to the trajectory $\bar{y}(\cdot)$,
- the sequence $\lambda_n\|(s_{\lambda_n} - T_{\lambda_n}, z_{\lambda_n} - \bar{y}_{\lambda_n}(T_{\lambda_n}))\|^2$ is convergent,
- the sequence $\lambda_n d_C(\bar{y}_{\lambda_n}(T_{\lambda_n})) + \lambda_n \int_0^{T_{\lambda_n}} d_A(y_{\lambda_n}(t))\, dt$ is convergent, and
- the sequence $\ell_{\lambda_n}(T_{\lambda_n}, \bar{y}_{\lambda_n}(T_{\lambda_n}))$ is convergent.

Since the sequence λ_n is strictly increasing in $\mathbb{N}^*$ we get

$$\lim_{n \longrightarrow +\infty} z_{\lambda_n} = \bar{y}(\bar{T}_0) \quad \text{and} \quad \lim_{n \longrightarrow +\infty} s_{\lambda_n} = \bar{T}_0.$$

Hence by the convergence of the sequence $\lambda_n d_C(\bar{y}_{\lambda_n}(T_{\lambda_n})) + \lambda_n \int_0^{T_{\lambda_n}} d_A(y_{\lambda_n}(t))\, dt$ and using Lebesgue's theorem we have

$$\bar{y}(\bar{T}_0) \in C \ \text{ and } \ \bar{y}(t) \in A \ \ \forall\, t \in [0, \bar{T}_0].$$

Then $(\bar{T}_0, \bar{y}(\cdot))$ is admissible for the problem (P). Hence

$$\begin{aligned}
\min(P) &= \ell(\bar{T}, \bar{x}(\bar{T})) \\
&\geq \lim_{n \longrightarrow +\infty} \min(P_{\lambda_n}) \\
&= \lim_{n \longrightarrow +\infty} [\ell_{\lambda_n}(T_{\lambda_n}, \bar{y}_{\lambda_n}(T_{\lambda_n})) + \lambda_n d_C(\bar{y}_{\lambda_n}(T_{\lambda_n})) + \lambda_n \int_0^{T_{\lambda_n}} d_A(\bar{y}_{\lambda_n}(t))\, dt] \\
&\geq \lim_{n \longrightarrow +\infty} \ell_{\lambda_n}(T_{\lambda_n}, \bar{y}_{\lambda_n}(T_{\lambda_n})) \\
&= \lim_{n \longrightarrow +\infty} [\ell(s_{\lambda_n}, z_{\lambda_n}) + \lambda_n\|(s_{\lambda_n} - T_{\lambda_n}, z_{\lambda_n} - \bar{y}_{\lambda_n}(T_{\lambda_n}))\|^2] \\
&\geq \liminf_{(s', z') \longrightarrow (\bar{T}_0, \bar{y}(\bar{T}_0))} \ell(s', z') \\
&\geq \ell(\bar{T}_0, \bar{y}(\bar{T}_0)) \\
&\geq \min(P).
\end{aligned}$$

Then

$$\min(P) = \lim_{n \longrightarrow +\infty} \min(P_{\lambda_n}) = \ell(\bar{T}_0, \bar{y}(\bar{T}_0)),$$

which completes the proof of the lemma. □

We continue the proof of Theorem 1.1 and we remark that problem (P_{λ_n}) is exactly the following problem:

Minimize $\hat{\ell}_{\lambda_n}(T, z(T))$ over points $T \geq 0$ and absolutely continuous functions $z : [0, +\infty[\longrightarrow \mathbb{R}^{n+1}$ that satisfy

$$\dot{z}(t) \in \hat{F}_{\lambda_n}(z(t)) \quad \text{a.e. } t \in [0, +\infty[, \quad z(0) = (0, x_0), \tag{2.6}$$

where $\hat{F}_{\lambda_n}$ is the augmented locally Lipschitz multifunction defined by $\hat{F}_{\lambda_n}(y, x) := \{\lambda_n d_A(x)\} \times F_{\lambda_n}(x)\ \forall (y, x) \in \mathbb{R} \times \mathbb{R}^n$ and $\hat{\ell}_{\lambda_n}$ is the locally Lipschitz and bounded below function defined by $\hat{\ell}_{\lambda_n}(t, y, x) = \ell_{\lambda_n}(t, x) + \lambda_n d_C(x) + |y|\ \forall (t, y, x) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$. Let $\hat{V}_{\lambda_n} : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}$ be the value function of the problem (P_{λ_n}); that is, for every $(\tau, \beta, \alpha) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$, $\hat{V}_{\lambda_n}(\tau, \beta, \alpha)$ is the minimum of the following problem:

Minimize $\hat{\ell}_{\lambda_n}(T, z(T))$ over points $T \geq \tau$ and absolutely continuous functions $z : [\tau, +\infty[\longrightarrow \mathbb{R}^{n+1}$ that satisfy

$$\dot{z}(t) \in \hat{F}_{\lambda_n}(z(t)) \quad \text{a.e. } t \in [\tau, +\infty[, \quad z(\tau) = (\beta, \alpha). \tag{2.7}$$

It is easy to verify that $\hat{V}_{\lambda_n}$ is locally Lipschitz and satisfies

$$\hat{V}_{\lambda_n}(\cdot) \leq \hat{\ell}_{\lambda_n}(\cdot) \tag{2.8}$$

and

$$\hat{V}_{\lambda_n}(\tau, \beta, \alpha) = \hat{V}_{\lambda_n}(\tau, 0, \alpha) + \beta \quad \forall (\tau, \beta, \alpha) \in \mathbb{R} \times [0, +\infty[\times \mathbb{R}^n. \tag{2.9}$$

Using the logic known as the *principle of optimality* and since the minimum which defines $\hat{V}_{\lambda_n}(\tau, \beta, \alpha)$ is attained for all $(\tau, \beta, \alpha) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$, we can easily prove that the system $(\hat{V}_{\lambda_n}, \hat{F}_{\lambda_n})$ is *strongly increasing* on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$ (that is, the function $\hat{V}_{\lambda_n}(\cdot, z(\cdot))$ is increasing on $[a, b]$ whenever z is a trajectory of $\hat{F}_{\lambda_n}$ on some interval $[a, b]$). Then by [8, Proposition 4.6.5][2] we have

$$\theta + \lambda_n d_A(x)\xi + h_{\lambda_n}(x, \zeta) \geq 0^3 \tag{2.10}$$

$\forall (\theta, \xi, \zeta) \in \partial_P \hat{V}_{\lambda_n}(t, y, x)$, $\forall (t, y, x) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$.

If we consider $\psi_{\lambda_n} : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}$ defined by $\psi_{\lambda_n}(t, x) = \hat{V}_{\lambda_n}(t, 0, x)\ \forall (t, x) \in \mathbb{R} \times \mathbb{R}^n$, then since $\hat{V}_{\lambda_n}$ is locally Lipschitz and by (2.8), (2.9), and (2.10) we have

1. ψ_{λ_n} is locally Lipschitz on $\mathbb{R} \times \mathbb{R}^n$,
2. $\psi_{\lambda_n}(0, x_0) = \hat{V}_{\lambda_n}(0, 0, x_0) = \min(P_{\lambda_n})$,
3. $\lambda_n d_A(x) + \bar{h}_{\lambda_n}(x, \partial_P \psi(t, x)) \geq 0\ \forall (t, x) \in \mathbb{R} \times \mathbb{R}^n$, and
4. $\psi_{\lambda_n}(t, x) \leq \ell_{\lambda_n}(t, x) \leq \ell(t, x)\ \forall (t, x) \in \mathbb{R} \times C$.

By point 3 and by a simple argument we have that ψ_{λ_n} satisfies the following limiting proximal Hamilton–Jacobi equation: $\bar{h}_{\lambda_n}(x, \partial_L \psi_{\lambda_n}(t, x)) \geq 0\ \forall (t, x) \in \mathbb{R} \times A$. Then since $\bar{h}_{\lambda_n}(\cdot) \leq \bar{h}(\cdot)$ we get $\psi_{\lambda_n} \in \Psi$. Therefore

$$\sup_{\psi \in \Psi} \psi(0, x_0) \geq \psi_{\lambda_n}(0, x_0) = \min(P_{\lambda_n});$$

[2] This proposition gives a proximal characterization for the strong increase property.

[3] h_{λ_n} is the lower Hamiltonian corresponding to F_{λ_n}.

then

$$\min(P) = \lim_{n \longrightarrow +\infty} \min(P_{\lambda_n}) \leq \sup_{\psi \in \Psi} \psi(0, x_0).$$

Now we show the reverse inequality.[4] Let $\psi \in \Psi$. We make the *temporary hypothesis* that F is locally Lipschitz. Then we have the following lemma.

LEMMA 2.3. *For all open and bounded subset $S \subset \mathbb{R}^{n+1}$, for all $\varepsilon > 0$, there exists a neighborhood U of A such that*

$$\bar{h}(x, \partial_P \psi(t, x)) \geq -\varepsilon \quad \forall (t, x) \in S \cap \{\mathbb{R} \times U\}.$$

Proof. We reason by the absurd. We assume that there exist an open and bounded subset $S \subset \mathbb{R}^{n+1}$ and a $\varepsilon > 0$ such that for all neighborhood U of A, there exist $(t, x) \in S \cap \{\mathbb{R} \times U\}$ and $(\theta, \zeta) \in \partial_P \psi(t, x)$ such that

$$\theta + h(x, \zeta) < -\varepsilon.$$

Then there exist two sequences $(t_n, x_n) \in S$ and (θ_n, ζ_n) such that

$$(t_n, x_n) \longrightarrow (t_0, x_0) \in \mathbb{R} \times A,$$

$$(\theta_n, \zeta_n) \in \partial_P \psi(t_n, x_n),$$

and

$$\theta_n + h(x_n, \zeta_n) < -\varepsilon. \tag{2.11}$$

Since ψ is locally Lipschitz and S is bounded, the sequence (θ_n, ζ_n) is bounded and then we can assume that it converges to a point (θ_0, ζ_0). By the definition of ∂_L we get that $(\theta_0, \zeta_0) \in \partial_L \psi(t_0, x_0)$. By (2.11) and since F is locally Lipschitz we find that

$$\theta_0 + h(x_0, \zeta_0) \leq -\varepsilon$$

and this gives a contradiction since $\psi \in \Psi$. □

Now let $(\bar{T}, \bar{x}(\cdot))$ be a solution of the problem (P); then by [8, Proposition 4.1.4] there exists $\rho > 0$ such that $\bar{x}(t) \in B(0; \rho)\ \forall\, t \in [0, \bar{T}]$. We apply the preceding lemma for $S =]-1, \bar{T}+1[\times B(0; \rho)$, and for $\varepsilon > 0$, we get that there exists a neighborhood U_ε of A such that

$$\bar{h}(x, \partial_P \psi(t, x)) \geq -\varepsilon \quad \forall (t, x) \in S \cap \{\mathbb{R} \times U_\varepsilon\}.$$

But $S \cap \{\mathbb{R} \times U_\varepsilon\} =]-1, \bar{T}+1[\times \{B(0; \rho) \cap U_\varepsilon\}$; then since $[0, \bar{T}] \subset]-1, \bar{T}+1[$, $\bar{x}(t) \in B(0; \rho) \cap U_\varepsilon\ \forall\, t \in [0, \bar{T}]$ and by [8, Proposition 4.6.5] we get that

$$\psi(0, x_0) \leq \varepsilon \bar{T} + \psi(\bar{T}, \bar{x}(\bar{T}))$$

but $\psi(\bar{T}, \bar{x}(\bar{T})) \leq \ell(\bar{T}, \bar{x}(\bar{T})) = \min(P)$; then

$$\psi(0, x_0) \leq \min(P) + \varepsilon \bar{T},$$

[4] We cannot directly deduce this inequality like in the smooth case since the multifunction F is not locally Lipschitz and the Hamilton–Jacobi inequality is satisfied only on a closed subset and then we cannot apply the monotonicity of trajectories [8, Proposition 4.6.5].

and hence by taking $\varepsilon \longrightarrow 0$ we get

$$\psi(0, x_0) \leq \min(P),$$

and therefore

$$\min(P) \geq \sup_{\psi \in \Psi} \psi(0, x_0).$$

To remove the need for the locally Lipschitz hypothesis on F we use the sequence F_k. First we have the following lemma.

LEMMA 2.4. *For all $n \in \mathbb{N}$ there exists $k_n \geq n$ such that*

$$\bar{h}_{k_n}(x, \partial_L \psi(t, x)) \geq \frac{-1}{n} \quad \forall (t, x) \in]-1, \bar{T}+1[\times\{A \cap \bar{B}(0; \rho)\}.$$

Proof. We reason again by the absurd. Assume that there exists $n_0 \in \mathbb{N}$ such that for all $k > n_0$ there exists $(t_k, x_k) \in]-1, \bar{T}+1[\times\{A \cap \bar{B}(0; \rho)\}$ and $(\theta_k, \zeta_k) \in \partial_L \psi(t_k, x_k)$ such that

$$\theta_k + h_k(x_k, \zeta_k) < -\frac{1}{n_0}. \tag{2.12}$$

Since the sequence (t_k, x_k) is bounded, we assume that it converges to a point $(t_0, x_0) \in \mathbb{R} \times A$. On the other hand, the sequence (θ_k, ζ_k) is also bounded since ψ is locally Lipschitz, and then we can assume that it converges to a point (θ_0, ζ_0). Using [8, Exercise 1.11.21] we have $(\theta_0, \zeta_0) \in \partial_L \psi(t_0, x_0)$. Now let $\varepsilon > 0$; then since F is upper semicontinuous we have that for k sufficiently large $F(x_k + 3^{-k+1}B) \subset F(x_0) + \varepsilon \bar{B}$. Hence $F_k(x_k) \subset F(x_0) + \varepsilon \bar{B}$. Using the definition of the lower Hamiltonian, we get that $h_k(x_k, \zeta_k) \geq h(x_0, \zeta_k) - \varepsilon \|\zeta_k\|$. Then by (2.12), we have

$$\theta_k + h(x_0, \zeta_k) - \varepsilon \|\zeta_k\| < -\frac{1}{n_0}.$$

But h is continuous in the second variable. Then if we take $k \longrightarrow +\infty$ in the preceding inequality we get

$$\theta_0 + h(x, \zeta_0) - \varepsilon \|\zeta_0\| \leq -\frac{1}{n_0}.$$

Since ε is arbitrary we find that

$$\theta_0 + h(x, \zeta_0) \leq -\frac{1}{n_0},$$

and this gives a contradiction with $\psi \in \Psi$. □

By the preceding lemma, there exists a subsequence F_{k_i} of F_k and a sequence $\varepsilon_i > 0$ such that $\varepsilon_i \longrightarrow 0$ and

$$\bar{h}_{k_i}(x, \partial_L \psi(t, x)) \geq -\varepsilon_i \quad \forall (t, x) \in]-1, \bar{T}+1[\times\{A \cap \bar{B}(0; \rho)\}.$$

We continue as in the Lipschitz case and we find the result. □

Remark 2.5. If we suppose in the definition of Ψ that the functions ψ are continuous and satisfy the proximal Hamilton–Jacobi equation (∂_P instead of ∂_L), then the first inequality

$$\min(P) \leq \sup_{\psi \in \Psi} \psi(0, x_0)$$

remains true, and the equality holds if we assume in addition that $A = \mathbb{R}^n$ and F is locally Lipschitz. Moreover, in this case if we eliminate the terminal constraint $x(T) \in C$, then the supremum in $\sup \psi(0, x_0)$ is attained since the value function of (P) becomes an element of Ψ. Finally, an earlier result of Clarke [6, Theorem 3.7.6] shows that even in the presence of terminal constraints, but still under the extra hypotheses, the supremum is attained if solutions are *normal*. It is an open question whether the attainment of the supremum can be asserted in the fully general context.

3. Consequences and extensions. In this section we give some applications of our main result. In the next subsection we give necessary and sufficient optimality conditions. In subsection 3.2, we treat several types of optimal control problems (fixed time problem and free time problem with finite and infinite horizon). Vinter's smooth duality is studied in subsection 3.3.

3.1. Characterization of optimality. Theorem 1.1 leads directly to the following optimality conditions.

COROLLARY 3.1. *Let $(\bar{T}, \bar{x}(\cdot))$ be an admissible trajectory for (P). Then $(\bar{T}, \bar{x}(\cdot))$ is a minimizer for (P) iff there exists a sequence of functions $\{\psi_i\}$ in Ψ such that*

$$\lim_{i \longrightarrow +\infty} \psi_i(0, x_0) = \ell(\bar{T}, \bar{x}(\bar{T})).$$

Proof. Let $(\bar{T}, \bar{x}(\cdot))$ be an admissible trajectory for (P) and assume that $(\bar{T}, \bar{x}(\cdot))$ is a minimizer for (P). Then by Theorem 1.1 we have

$$\ell(\bar{T}, \bar{x}(\bar{T})) = \sup_{\psi \in \Psi} \psi(0, x_0).$$

Hence the necessary condition follows by taking a maximizing sequence $\{\psi_i\}$ for the supremum $\sup_{\psi \in \Psi} \psi(0, x_0)$.

For the sufficient condition, let $(\bar{T}, \bar{x}(\cdot))$ be an admissible trajectory for (P) and suppose that there exists a sequence of functions $\{\psi_i\}$ in Ψ such that

$$\lim_{i \longrightarrow +\infty} \psi_i(0, x_0) = \ell(\bar{T}, \bar{x}(\bar{T})).$$

Since $\psi_i \in \Psi$ and by Theorem 1.1 we have

$$\lim_{i \longrightarrow +\infty} \psi_i(0, x_0) \leq \min(P).$$

But $(\bar{T}, \bar{x}(\cdot))$ is an admissible trajectory for (P); then

$$\min(P) \geq \lim_{i \longrightarrow +\infty} \psi_i(0, x_0) = \ell(\bar{T}, \bar{x}(\bar{T})) \geq \min(P),$$

and then

$$\ell(\bar{T}, \bar{x}(\bar{T})) = \min(P),$$

which completes the proof. □

3.2. Special cases.

1. *Fixed time problem.* If we take $\ell(t,x) = I_{\{T_0\}\times C}(t,x) + \ell_0(x)$, where $T_0 \geq 0$ is fixed, $\ell_0 : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a lower semicontinuous function, and $I_{\{T_0\}\times C}$ denotes the indicator function of $\{T_0\} \times C$, then we can treat the fixed time case. In this case we obtain the same duality as in Theorem 1.1 but the boundary condition becomes

$$\psi(T_0, x) \leq \ell_0(x) \ \ \forall x \in C.$$

   If we assume that $A = \mathbb{R}^n$ and F is locally Lipschitz, then we can consider in the preceding duality continuous functions ψ and we can replace ∂_L by ∂_P, and this gives a slightly strengthened version of [8, Corollary 4.7.7].[5]
2. *Free time problem with finite horizon.* Problems in which T varies in a compact interval (as is the case in Vinter's work) can be treated by taking (for example) $\ell(t,x) = I_{[0,1]\times C}(t,x) + \ell_0(x)$, where $\ell_0 : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a lower semicontinuous function. In this case the boundary condition is

$$\psi(t,x) \leq \ell_0(x) \ \ \forall (t,x) \in [0,1] \times C.$$

3. *Free time problem with infinite horizon.* We now consider the free time problem in its most familiar form, where T is completely unrestricted. We suppose that in (P) we have $\ell(t,x) = t + \ell_0(x)$ $\forall (t,x) \in \mathbb{R} \times \mathbb{R}^n$, where $\ell_0 : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a lower semicontinuous function bounded below. In this case and by Theorem 1.1 we have that

$$\min(P) = \sup_{\psi \in \Psi} \psi(x_0),$$

   where Ψ is the set of all functions $\psi : \mathbb{R}^n \longrightarrow \mathbb{R}$ which satisfy
   - ψ is locally Lipschitz on $\mathbb{R}^n$,
   - $1 + h(x, \partial_L \psi(x)) \geq 0$ $\forall x \in A$, and
   - $\psi(x) \leq \ell_0(x)$ $\forall x \in C$.

   We remark that in this result, only autonomous functions ψ contribute to the upper envelope. This follows since in this case we can take in the proof of Theorem 1.1, $\ell_k := t + \ell_0^k$, where ℓ_0^k is the quadratic inf-convolution of ℓ_0, and then we get that the function $\hat{V}_{\lambda_n}(\cdot,\cdot,\cdot)$ satisfies

$$\hat{V}_{\lambda_n}(\tau,\beta,\alpha) = \tau + \hat{V}_{\lambda_n}(0,0,\alpha) + \beta \quad \forall (\tau,\beta,\alpha) \in \mathbb{R} \times [0,+\infty[\times\mathbb{R}^n.$$

3.3. Smooth duality. Another important application of our main result extends the smooth duality theory of Vinter [27], whose methods do not apply to the cases in which T is either fixed or totally unrestricted. For this, we need the following technical extension of Theorem 1.1.

COROLLARY 3.2. *There exists $\delta_{x_0} > 0$ such that*

$$\min(P) = \sup_{\varphi \in \Phi} \varphi(0, x_0),$$

where Φ is the set of all functions $\varphi : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}$ which satisfy
- $\varphi \in C^1(\mathbb{R}^{n+1}, \mathbb{R})$,
- $\varphi_t(t,x) + \langle \varphi_x(t,x), v \rangle \geq 0$ $\forall (t,x) \in \mathbb{R} \times A$ $\forall v \in F(x)$, *and*

[5] In [8, Corollary 4.7.7] ℓ is continuous and the functions ψ are defined on $]-\infty, T_0] \times \mathbb{R}^n$.

• $\varphi(t,x) \le \ell(t,x)\ \forall (t,x) \in [0,\delta_{x_0}]\times C$.

Proof. We fix a constant δ_{x_0} such that for any trajectory $(\bar{T}, \bar{x}(\cdot))$ solving the problem (P) we have $\bar{T} \le \delta_{x_0}$ (the existence of δ_{x_0} follows from (GC)). Now let $\varphi \in \Phi$. Since φ is a C^1 function on $\mathbb{R}^{n+1}$, we have

$$\partial_L \varphi(t,x) = \{\varphi'(t,x)\}$$

for all $(t,x) \in \mathbb{R}\times\mathbb{R}^n$. Then we have $\varphi \in \Psi$.[6] By Theorem 1.1 we get

$$\min(P) = \sup_{\psi\in\Psi} \psi(0,x_0) \ge \sup_{\varphi\in\Phi} \varphi(0,x_0).$$

For the reverse inequality, let $\psi \in \Psi$. Using the fact that if ψ is differentiable at $(\tau,\alpha) \in \mathbb{R}\times\mathbb{R}^n$, then $\psi'(\tau,\alpha) \in \partial_L\psi(\tau,\alpha)$, we have that for all $(\tau,\alpha) \in \mathbb{R}\times A$ such that ψ is differentiable at (τ,α),

$$\psi_t(\tau,\alpha) + \langle \psi_x(\tau,\alpha), v\rangle \ge 0\ \forall v \in F(\alpha).$$

Since ψ is locally Lipschitz and by Rademacher's theorem, we have that ψ is differentiable a.e. $(\tau,\alpha) \in \mathbb{R}\times\mathbb{R}^n$. Now we make the *temporary hypothesis* that F is Lipschitz and $A = \mathbb{R}^n$. Then using a standard mollification technique (convolution with mollifier sequence) we have the following lemma.

LEMMA 3.3. *For all $\varepsilon > 0$ there exists $\varphi_\varepsilon \in \Phi$ such that*

$$\varphi_\varepsilon(0,x_0) \ge \psi(0,x_0) - \varepsilon.$$[7]

Clearly this lemma gives the desired inequality. To remove the need for the Lipschitz hypothesis on F and the assumption $A = \mathbb{R}^n$, it is sufficient to use the sequence F_k and the penalization term $k\int_0^T d_A(x(t))\,dt$ as in the proof of Theorem 1.1. □

It is clear that Corollary 3.2 leads to a version of the necessary and sufficient conditions of Corollary 3.1 in which only smooth semisolutions are used. Let us now examine more closely this type of smooth duality in the three special cases of subsection 3.2.

1. *Fixed time problem.* We take $\ell(t,x) = I_{\{T_0\}\times C}(t,x) + \ell_0(x)$, where $T_0 \ge 0$ is fixed and $\ell_0 : \mathbb{R}^n \longrightarrow \mathbb{R}\cup\{+\infty\}$ is a lower semicontinuous function, then by Corollary 3.2 (we take $\delta_{x_0} = T_0$) we obtain a similar duality but as in the nonsmooth case the boundary condition becomes

$$\varphi(T_0,x) \le \ell_0(x)\ \ \forall x \in C.$$

2. *Free time problem with finite horizon.* We take $\ell(t,x) = I_{[0,1]\times C}(t,x) + \ell_0(x)$, where $\ell_0 : \mathbb{R}^n \longrightarrow \mathbb{R}\cup\{+\infty\}$ is a lower semicontinuous function. By Corollary 3.2 (we take $\delta_{x_0} = 1$) we obtain

$$\min(P) = \sup_{\varphi\in\Phi} \varphi(0,x_0),$$

where Φ is the set of all functions $\varphi : \mathbb{R}\times\mathbb{R}^n \longrightarrow \mathbb{R}$ which satisfy

[6] We can replace in Theorem 1.1 the condition $\psi(t,x) \le \ell(t,x)\ \forall (t,x) \in \mathbb{R}\times C$ by $\psi(t,x) \le \ell(t,x)$ $\forall (t,x) \in [0,\delta_{x_0}]\times C$.

[7] As mentioned above, this lemma follows using a standard mollification technique. We also use the following remark: if $\varphi(t,x)$ is a smooth function which satisfies $\varphi_t(t,x) + \langle \varphi_x(t,x), v\rangle \ge \varepsilon$ $\forall (t,x) \in \mathbb{R}\times A$, $\forall v \in F(x)$, then the function $\varphi^\varepsilon(t,x) := \varphi(t,x) - t\varepsilon$ satisfies $\varphi^\varepsilon_t(t,x) + \langle \varphi^\varepsilon_x(t,x), v\rangle \ge 0$ $\forall (t,x) \in \mathbb{R}\times A$, $\forall v \in F(x)$.

- $\varphi \in C^1(\mathbb{R}^{n+1}, \mathbb{R})$,
- $\varphi_t(t,x) + \langle \varphi_x(t,x), v\rangle \geq 0 \ \forall (t,x) \in \mathbb{R}\times A$, $\forall v \in F(x)$, and
- $\varphi(t,x) \leq \ell_0(x) \ \forall (t,x) \in [0,1]\times C$.

This is essentially the case treated by Vinter in [27]. We remark that the nonautomous case (when F and ℓ_0 depend on t) can be obtained by the well-known device of state augmentation (see [8, Chapter 4]).

3. *Free time problem with infinite horizon.* We take $\ell(t,x) = t + \ell_0(x)\ \forall (t,x) \in \mathbb{R}\times\mathbb{R}^n$, where $\ell_0 : \mathbb{R}^n \longrightarrow \mathbb{R}\cup\{+\infty\}$ is a lower semicontinuous function bounded below. This is the most familiar case of the minimal time problem, and Vinter has remarked [28] that his generalized flows approach does not appear to yield a duality involving solutions of the autonomous Hamilton–Jacobi inequality (as one would hope). However, we obtain the following:[8]

$$\min(P) = \sup_{\varphi\in\Phi} \varphi(x_0),$$

where Φ is the set of all functions $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}$ which satisfy

- $\varphi \in C^1(\mathbb{R}^n, \mathbb{R})$,
- $1 + \langle \varphi'(x), v\rangle \geq 0 \ \forall x \in A$, $\forall v \in F(x)$, and
- $\varphi(x) \leq \ell_0(x) \ \forall x \in C$.

A well-known and more special case of the present framework involves the minimal time function associated to the target C and under the state constraint A:

$T_C^A(\alpha) := \inf T \geq 0$ over trajectories of $\dot{x}(t) \in F(x(t))$ which satisfy

$$x(0) = \alpha, \quad x(T) \in .C, \quad x(t) \in A \ \ \forall\, t \in [0,T].$$

We then obtain the following new characterization of T_C^A:

$$T_C^A(\alpha) = \sup_{\varphi\in\Phi} \varphi(\alpha),$$

where Φ is the set of all functions $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}$ which satisfy

- $\varphi \in C^1(\mathbb{R}^n, \mathbb{R})$,
- $1 + \langle \varphi'(x), v\rangle \geq 0 \ \forall x \in A$, $\forall v \in F(x)$, and
- $\varphi(x) \leq 0 \ \forall x \in C$.

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, New York, 1984.

[2] M. Bardi and I. Capuzzo-Dolcetta, *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*, Birkhäuser, Boston, 1997.

[3] G. Barles, *Solutions de viscosité des équations de Hamilton-Jacobi*, Math. Appl. Berlin, 17, Springer-Verlag, Paris, 1994.

[4] F. H. Clarke, *The applicability of the Hamilton-Jacobi verification technique*, in Proceedings of the 10th IFIP Conference (New York, 1981), System Model. Optim. Ser. 38, R. F. Drenick and F. Kizin, eds., Springer-Verlag, New York, 1982, pp. 88–94.

[5] F. H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conf. Ser. Appl. Math., 57, SIAM, Philadelphia, 1989.

[6] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983. Reprinted as vol. 5 of Classics in Applied Mathematics, SIAM, Philadelphia, 1990.

[7] F. H. Clarke and Yu. Ledyaev, *Mean value inequalities in Hilbert space*, Trans. Amer. Math. Soc., 344 (1994), pp. 307–324.

[8] This smooth duality follows using our nonsmooth duality presented in subsection 3.2 (for free time problem with infinite horizon) and using the same techniques as in Corollary 3.2.

[8] F. H. Clarke, Yu. Ledyaev, R. Stern, and P. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer-Verlag, New York, 1998.

[9] F. H. Clarke, Yu. Ledyaev, R. Stern, and P. Wolenski, *Qualitative properties of trajectories of control systems: A survey*, J. Dynam. Control Systems, 1 (1995), pp. 1–48.

[10] F. H. Clarke and P. D. Loewen, *The value function in optimal control: Sensitivity, controllability, and time-optimality*, SIAM J. Control Optim., 24 (1994), pp. 243–263.

[11] F. H. Clarke and R. B. Vinter, *Local optimality conditions and Lipschitizian solutions to the Hamilton–Jacobi equation*, SIAM J. Control Optim., 21 (1983), pp. 856–870.

[12] F. H. Clarke and R. B. Vinter, *The relationship between the maximum principle and dynamic programming*, SIAM J. Control Optim., 25 (1987), pp. 1291–1311.

[13] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[14] K. Deimling, *Multivalued Differential Equations*, de Gruyter Ser. Nonlinear Anal. Appl. 1, Walter de Gruyter, Berlin, 1992.

[15] W. H. Fleming, *Generalized solutions and convex duality in optimal control*, in Partial Differential Equations and the Calculus of Variations: Essays in Honor of Ennio De Giorgi, Progr. Nonlinear Differential Equations Appl. 1, Birkhäuser, Boston, 1989, pp. 461–471.

[16] W. H. Fleming and D. Vermes, *Convex duality approach to the optimal control of diffusions*, SIAM J. Control Optim., 27 (1989), pp. 1136–1155.

[17] H. Frankowska, *Optimal trajectories associated with solution of the Hamilton-Jacobi equation*, Appl. Math. Optim., 19 (1989), pp. 291–311.

[18] H. Frankowska, *Lower semicontinuous solutions of the Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.

[19] D. Havelock, *A Generalization of the Hamilton-Jacobi Equation*, M.Sc. thesis, University of British Columbia, Canada, 1977.

[20] P. L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations* I: *The dynamic programming principle and applications*, Comm. Partial Differential Equations, 8 (1983), pp. 1101–1174.

[21] P. D. Loewen, *Optimal Control Via Nonsmooth Analysis*, CRM Proc. Lecture Notes 2, AMS, Providence, RI, 1993.

[22] D. Offin, *A Hamilton-Jacobi Approach to the Differential Inclusion Problem*, M.Sc. thesis, University of British Columbia, Canada, 1979.

[23] P. Soravia, *Discontinuous viscosity solutions to Dirichlet problems for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 18 (1993), pp. 1493–1514.

[24] P. Soravia, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations.* II. *Equations of control problems with state constraints*, Differential Integral Equations, 12 (1999), pp. 275–293.

[25] V. M. Veliov, *Lipschitz continuity of the value function in optimal control*, J. Optim. Theory Appl., 94 (1997), pp. 335–363.

[26] R. B. Vinter, *Optimal Control*, Birkhäuser, Boston, 2000.

[27] R. B. Vinter, *Convex duality and nonlinear optimal control*, SIAM J. Control Optim., 31 (1993), pp. 518–538.

[28] R. B. Vinter, *private communication.*

[29] R. B. Vinter and R. M. Lewis, *The equivalence of strong and weak formulations for certain problems in optimal control*, SIAM J. Control Optim., 16 (1978), pp. 546–570

[30] R. B. Vinter and R. M. Lewis, *A necessary and sufficient condition for optimality of dynamic programming type, making no a priori assumptions on the controls*, SIAM J. Control Optim., 16 (1978), pp. 571–583.

[31] R. B. Vinter and P. Wolenski, *Hamilton–Jacobi theory for optimal control problems with data measurable in time*, SIAM J. Control Optim., 28 (1990), pp. 1404–1419.

[32] P. R. Wolenski and Y. Zhuang, *Proximal analysis and the minimal time function*, SIAM J. Control Optim., 36 (1998), pp. 1048–1072.

[33] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

[34] H. Zhu, *Convex duality and generalized solutions in optimal control problem for stopped processes: The deterministic model*, SIAM J. Control Optim., 30 (1992), pp. 465–476.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2049–2070

FEEDBACK CLASSIFICATION OF MULTI-INPUT NONLINEAR CONTROL SYSTEMS*

ISSA AMADOU TALL†

Abstract. We study the feedback group action on multi-input nonlinear control systems with uncontrollable mode. We follow slightly an approach proposed in Kang and Krener [W. Kang and A. J. Krener, *SIAM J. Control. Optim.,* 30 (1992), pp. 1319–1337] which consists of analyzing the system and the feedback group step by step. We construct a normal form which generalizes, on one hand, the results obtained in the single-input case and, on the other hand, those recently obtained by the same author in the controllable case. We illustrate our results by studying the Caltech Multi-Vehicle Wireless Testbed (MVWT) and the prototype of Planar Vertical TakeOff and Landing aircraft (PVTOL). We also study the notion of *bifurcation of controllability* for systems with one nonzero uncontrollable mode. We first show that the equilibria for those systems is a p-dimensional submanifold (p equals number of inputs). Provided that one term in their normal form is nonzero, we show that these systems are linearly controllable, hence stabilizable, at any nearby equilibrium point of the origin.

Key words. feedback, multi-input, homogeneous, normal form, uncontrollable, control system

AMS subject classifications. 93B11, 93B17, 93B27

DOI. 10.1137/S0363012902410502

1. Introduction.

During the last twenty years the problem of transforming a nonlinear control system

$$\Pi \ : \ \dot{\zeta} = f(\zeta, u), \quad \zeta(\cdot) \in \mathbb{R}^n \quad u(\cdot) = (u_1(\cdot), \ldots, u_p(\cdot))^T \in \mathbb{R}^p$$

by a feedback transformation of the form

$$\Upsilon \ : \ \begin{array}{rcl} \bar{\zeta} & = & \varphi(\zeta), \\ u & = & \gamma(\zeta, \bar{u}) \end{array}$$

to a simpler form has been extensively studied by several authors. Necessary and sufficient geometric conditions for smooth linearizability, that is, smooth feedback equivalence to a linear system, have been obtained independently by Hunt and Su [16], Hunt, Su, and Meyer [17], and Jakubczyk and Respondek [21] among others. Those conditions turn out to be restrictive, except for the planar case, and a natural problem that arises is to find normal forms for nonlinearizable systems. Four basic methods have been proposed to study feedback equivalence problems. The first method is based on the theory of singularities of vector fields and distributions and their invariants, and using that method on a large variety of feedback classification problems have been solved; see, e.g., [4, 6, 10, 17, 18, 21, 22, 23, 31, 36, 39, 51]. The second approach, proposed by Gardner [10], uses Cartan's method of equivalence [5] and describes the geometry of feedback equivalence, [11, 12, 13, 35]. The third method, inspired by the Hamiltonian formalism for optimal control problems, has been developed by Bonnard [3, 4] and Jakubczyk [19, 20] and has led to a very nice

*Received by the editors July 1, 2002; accepted for publication (in revised form) June 3, 2004; published electronically April 14, 2005.
http://www.siam.org/journals/sicon/43-6/41050.html
†Department of Mathematics, University of California, One Shields Avenue, Davis, CA 95616 (tall@math.ucdavis.edu). This research was conducted while the author was on leave from the Department of Mathematics, Université Cheikh Anta Diop, Dakar, Sénégal.

description of feedback invariants in terms of singular extremals. Most recently, Kang and Krener [29] adapted Poincaré's technique for linearization of dynamical systems (see, e.g., [1]) to control systems. Their idea consists of analyzing the system Π and the feedback transformation Υ step by step in order to produce a simpler equivalent system, also step by step. They first obtained quadratic normal forms under quadratic changes of coordinates and feedback for single-input control systems with controllable linearization. Later, Kang [24] generalizes this result to all degrees for the same class of control systems. He also obtained [25] quadratic normal forms for systems with uncontrollable linearization. The method introduced by Kang and Krener finds its importance in replacing the solving of partial differential equations by that of algebraic equations.

Since then many results have followed. Tall [41, 43] and Tall and Respondek [49] obtained canonical forms and dual canonical forms for single-input nonlinear control systems with controllable linearization, then normal forms for single-input nonlinear control systems with uncontrollable linearization [44] (see also Krener, Kang, and [32]), as well as the corresponding homogeneous invariants. Hence, the feedback classification of single-input nonlinear control systems is almost complete, and the aim of this paper is to deal with the multi-input nonlinear control systems. Preliminary results for two-input control systems, with controllable mode, have been recently obtained by Tall and Respondek [47] and completed by Tall [42] for multi-input systems with controllable mode. This paper gives a generalization of those results to multi-input systems with uncontrollable mode.

Motivations for studying normal forms for multi-input systems are underlined by the huge varieties of applications derived for single-input systems. Indeed, in the single-input case, the theory of normal forms has proved to be very useful in analyzing structural properties of nonlinear control systems. It has been used to study bifurcations and stabilizations of nonlinear systems [7, 14, 26, 27, 28], has led to a complete description of symmetries around equilibrium [37, 38, 48], and allowed the characterization of systems equivalent to feedforward forms [45, 46, 50]. The same approach has been introduced to study observability of control systems [33, 34, 2], the problem of output regulation, and the model matching problem.

The study of linearly uncontrollable systems is also motivated by the numerous engineering applications and the fact that the qualitative properties like controllability and stabilizability are generic, that is, invariant under a small variation of parameters at a point where the system is linearly controllable. Furthermore, it is known that local bifurcations at a point where the system is linearly controllable can be removed or delayed by pole placement. For those systems that are not linearly controllable, nonlinear phenoma like bifurcations are expected around the critical points.

In this paper we construct a normal form for multi-input nonlinear control systems with uncontrollable linearization which generalizes the results obtained in the single-input case [24, 41, 43, 44, 49] and the two-input case [47].

The organization of the paper is as follows. Section 2 deals with basic notations. In section 3, we construct a normal form for multi-input nonlinear control systems with uncontrollable linearization. We illustrate our results by two physical examples. We also discuss the notion of *bifurcation of controllability* for systems with one nonzero uncontrollable mode. We first show that the set of equilibria of these systems is a p-dimensional surface, and at any nearby equilibrium point of the origin, these systems became linearly controllable. Section 4 deals with the proofs of our results.

2. Notations and preliminaries. All objects, that is, functions, maps, vector fields, control systems, etc., are considered in a neighborhood of $0 \in \mathbb{R}^n$ and assumed to be C^∞-smooth. Consider the system

$$\Pi \ : \ \dot{\zeta} = f(\zeta, u), \quad \zeta(\cdot) \in \mathbb{R}^n, \ u(\cdot) = (u_1(\cdot), \dots, u_p(\cdot))^T \in \mathbb{R}^p.$$

We will assume throughout the paper that the point $(0,0) \in \mathbb{R}^n \times \mathbb{R}^p$ is an equilibrium point, that is, $f(0,0) = 0$, and let

$$\Pi^{[1]} \ : \ \dot{\zeta} = F\zeta + Gu = F\zeta + G_1 u_1 + \dots + G_p u_p$$

be its linear approximation around the equilibrium point $(0,0) \in \mathbb{R}^n \times \mathbb{R}^p$, where

$$F = \frac{\partial f}{\partial \zeta}(0,0), \qquad G_1 = \frac{\partial f}{\partial u_1}(0,0), \dots, G_p = \frac{\partial f}{\partial u_p}(0,0).$$

We assume that the linear approximation is uncontrollable which means that there exists a nonnegative integer $q \in \mathbb{N}^*$ such that

$$\operatorname{span}\left\{ F^i G_s : \ 0 \le i \le n-1, \ 1 \le s \le p \right\} = \mathbb{R}^{n-q}.$$

We will also assume that $G_1 \wedge \dots \wedge G_p \neq 0$, that is, the $n \times p$ matrix whose columns are $G_1, \dots, G_p$ to be of constant rank p.

Let $(r_1, \dots, r_p)$, $1 \le r_1 \le \dots \le r_p = r$, be the largest, in the lexicographic ordering, p-tuple of nonnegative integers, with $r_1 + \dots + r_p = n$, such that

$$\operatorname{span}\left\{ F^i G_s : \ 0 \le i \le r_s - 1, \ 1 \le s \le p \right\} = \mathbb{R}^{n-q}.$$

For the simplicity of the presentation we will suppose that $r_1 = \dots = r_p = r$, and we will show how the general case deduces.

By a smooth linear feedback transformation it is always possible to bring the pair (F, G) into the Brunovský–Jordan canonical pair $(\tilde{A}, \tilde{B})$, where

$$\tilde{A} = \operatorname{diag}(J, A_1, \dots, A_p), \ \tilde{B} = (0, B_1, \dots, B_p) = \operatorname{diag}(0, b_1, \dots, b_p)$$

with J the Jordan canonical form of dimension q, (A_s, b_s) the Brunovský single-input canonical form of dimension $r_s = r$ for any $1 \le s \le p$.

For simplicity we will set

$$A = \operatorname{diag}(A_1, \dots, A_p), \ B = (B_1, \dots, B_p) = \operatorname{diag}(b_1, \dots, b_p).$$

We will denote coordinates of $\mathbb{R}^q \times \mathbb{R}^{n-q}$ by $(z^T, x^T)^T$, where $z = (z_1, \dots, z_q)^T$ and $x = (x_1^T, \dots, x_p^T)^T$ with $x_s = (x_{s,1}, \dots, x_{s,r})^T$ for any $1 \le s \le p$. For the fixed value q, we will denote by $\mathcal{S}_q(\mathbb{R}, 0)$, the set of all functions, either smooth or formal, depending on the variables $z = (z_1, \dots, z_q)^T \in \mathbb{R}^q$.

Let $h = h(z, x, u)$ be a smooth $\mathbb{R}$-valued function defined in a neighborhood of the point $(0,0,0) \in \mathbb{R}^q \times \mathbb{R}^{n-q} \times \mathbb{R}^p$. By

$$h(z,x,u) = h^{[0]}(z,x,u) + h^{[1]}(z,x,u) + h^{[2]}(z,x,u) + \dots = \sum_{m=0}^{\infty} h^{[m]}(z,x,u)$$

we denote its Taylor series expansion at $(0,0,0) \in \mathbb{R}^q \times \mathbb{R}^{n-q} \times \mathbb{R}^p$ with respect to the variables x and u, where $h^{[m]}(z,x,u)$ stands for a homogeneous polynomial of degree m

of the variables x and u whose coefficients are functions of the variable $z \in \mathbb{R}^q$, that is, in $\mathcal{S}_q(\mathbb{R},0)$.

To fix the ideas, the functions $x_{1,1}^3$, $z^2x_{1,1}^3$, $\cos z\ x_{1,1}^3$, $(1-e^z)x_{1,1}x_{2,1}^2$, $\sin z\ x_{1,2}u_2^2$, and $z_1z_2u_1^2u_2 + x_{1,1}x_{2,1}u_1$ are all polynomials of degree 3.

Choose $d \in \mathbb{N} \cup \{\infty\}$ large enough and consider the Taylor series expansion of order d of the system Π

$$\Pi^{\leq d} : \begin{cases} \dot{z} = Jz + \sum\limits_{m=1}^{d} g^{[m-1]}(z,x,u) + O(z,x,u)^d, \\ \dot{x} = Ax + Bu + \sum\limits_{m=0}^{d} f^{[m]}(z,x,u) + O(z,x,u)^{d+1}, \end{cases} \tag{2.1}$$

where we already assumed that the linear part is in Brunovský–Jordan form, and the Taylor series expansion of order d of the transformation Υ,

$$\Upsilon^{\leq d} : \begin{cases} \bar{z} &= \psi(z,x) = z + \sum\limits_{m=1}^{d} \psi^{[m-1]}(z,x) + O(z,x)^d, \\ \bar{x} &= \phi(z,x) = x + \sum\limits_{m=0}^{d} \phi^{[m]}(z,x) + O(z,x)^{d+1}, \\ u &= \gamma(z,x,\bar{u}) = \bar{u} + \sum\limits_{m=0}^{d} \gamma^{[m]}(z,x,\bar{u}) + O(z,x,\bar{u})^{d+1}. \end{cases} \tag{2.2}$$

The variables z and $\bar{z}$ (resp., (x,u) and $(\bar{x},\bar{u})$) will be called the *uncontrollable variables* associated with the uncontrollable part (resp., *controllable variables* associated with the controllable part) of the system.

Above, and throughout the paper, we mean by $g^{[m-1]}(z,x,u)$ and $\psi^{[m-1]}(z,x)$ (resp., $f^{[m]}(z,x,u)$, $\phi^{[m]}(z,x)$ and $\gamma^{[m]}(z,x,\bar{u})$) that each of their components is a homogeneous polynomial of degree $m-1$ (resp., of degree m) of the *controllable variables*. Moreover, $O(\cdot)^k$ denotes terms of degree k and higher of the controllable variables.

Notice that although we bring the linear approximation (F,G) of the system into Brunovský–Jordan canonical form, that is, the uncontrollable part in Jordan form of dimension q, and the controllable part in Brunovský form of dimension $n-q$, we still have terms of degree 0 and degree 1. However, the first jets of these terms is zero at the origin.

3. Main results. In this section we will establish our main results. We will give, in subsection 3.1 below, the normal forms we obtain for general control systems. The results are given in the simplest case where the controllability indices are equal. We will show that the general case deduces by extended the system. In subsection 3.2, we will study two physical examples: The Caltech Multi-Vehicle Wireless Testbed (MVWT) and the prototype of a Planar Vertical TakeOff and Landing (PVTOL) aircraft. In subsection 3.3, we discuss the notion of *bifurcation of controllability* for systems with one nonzero uncontrollable mode. We first show that the set of equilibria of these systems is a p-dimensional surface, and at any nearby equilibrium point of the origin, these systems become linearly controllable.

3.1. Normal forms. Let $1 \leq s \leq p$. We denote

$$\bar{z} = (\bar{z}_1, \ldots, \bar{z}_q)^T, \qquad \bar{x}_s = (\bar{x}_{s,1}, \ldots, \bar{x}_{s,r})^T, \qquad \text{and} \qquad \bar{x}_{s,r+1} = \bar{u}_s$$

and we set $\hat{x}_{s,i} = (\bar{x}_{s,1}, \ldots, \bar{x}_{s,i})$ for any $1 \leq i \leq r+1$.

For any $1 \leq s \leq t \leq p$ and any $1 \leq i \leq r+1$, we also denote

$$\pi^s_{t,i}(\bar{x}) = (\hat{x}_{1,i}, \dots, \hat{x}_{s,i}, \hat{x}_{s+1,i-1}, \dots, \hat{x}_{t-1,i-1}, \hat{x}_{t,i}, \hat{x}_{t+1,i-1}, \dots, \hat{x}_{p,i-1})^T.$$

For $i = 1$ the expressions $\hat{x}_{t,i-1}$ will be taken to be empty. Few examples are given right after the theorem to make these notations comprehensible.

Our main result for general control systems, that is, for control systems with *uncontrollable linearization* is as follows.

THEOREM 3.1. *For any $d \in \mathbb{N} \cup \{\infty\}$, the system $\Pi^{\leq d}$, defined by* (2.1)*, with* uncontrollable linearization *is feedback equivalent, by a feedback transformation $\Upsilon^{\leq d}$ of the form* (2.2)*, to the following normal form:*

$$\Pi^{\leq d}_{NF} : \begin{cases} \dot{\bar{z}} = J\bar{z} + \bar{g}^{[0]}(\bar{z}) + \sum\limits_{m=2}^{d} \bar{g}^{[m-1]}(\bar{z}, \bar{x}, \bar{u}) + O(\bar{z}, \bar{x}, \bar{u})^d, \\ \dot{\bar{x}} = A\bar{x} + B\bar{u} + \sum\limits_{m=2}^{d} \bar{f}^{[m]}(\bar{z}, \bar{x}, \bar{u}) + O(\bar{z}, \bar{x}, \bar{u})^{d+1}, \end{cases}$$

where for any m,

$$\begin{aligned} \bar{g}^{[m-1]}(\bar{z}, \bar{x}, \bar{u}) &= \sum_{j=1}^{q} \bar{g}^{[m-1]}_j(\bar{z}, \bar{x}, \bar{u}) \tfrac{\partial}{\partial \bar{z}_j} \\ \bar{f}^{[m]}(\bar{z}, \bar{x}, \bar{u}) &= \sum_{k=1}^{p} \sum_{j=1}^{r-1} \bar{f}^{k[m]}_j(\bar{z}, \bar{x}, \bar{u}) \tfrac{\partial}{\partial \bar{x}_{k,j}}, \end{aligned} \tag{3.1}$$

with

$$\begin{aligned} \bar{g}^{[1]}_j(\bar{z}, \bar{x}, \bar{u}) &= \sum_{1 \leq s \leq p} \bar{x}_{s,1} R_{j,s}(\bar{z}) \\ \bar{g}^{[m-1]}_j(\bar{z}, \bar{x}, \bar{u}) &= \sum_{1 \leq s \leq t \leq p} \sum_{i=1}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i} P^{[m-3]}_{j,i,s,t}(\bar{z}, \pi^s_{t,i}(\bar{x})) \\ &+ \sum_{1 \leq s < t \leq p} \sum_{i=2}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i-1} Q^{[m-3]}_{j,i,s,t}(\bar{z}, \pi^t_{t,i-1}(\bar{x})) \end{aligned} \tag{3.2}$$

for any $1 \leq j \leq q$ and

$$\begin{aligned} \bar{f}^{k[m]}_j(\bar{z}, \bar{x}, \bar{u}) &= \sum_{1 \leq s \leq t \leq p} \sum_{i=j+2}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i} P^{k[m-2]}_{j,i,s,t}(\bar{z}, \pi^s_{t,i}(\bar{x})) \\ &+ \sum_{1 \leq s < t \leq p} \sum_{i=j+2}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i-1} Q^{k[m-2]}_{j,i,s,t}(\bar{z}, \pi^t_{t,i-1}(\bar{x})) \end{aligned} \tag{3.3}$$

for any $1 \leq k \leq p$ and any $1 \leq j \leq r-1$.

Above, the functions $P^{[m-3]}_{j,i,s,t}$, $Q^{[m-3]}_{j,i,s,t}$, $P^{k[m-2]}_{j,i,s,t}$, and $Q^{k[m-2]}_{j,i,s,t}$ stand for homogeneous polynomials of the indicated controllable variables $\bar{x}$ and $\bar{u}$ whose coefficients depend on the uncontrollable variable $\bar{z}$.

To make the notations $\pi^s_{t,i}(\bar{x})$ somewhat understandable, suppose $p = 3$. Then we have

$$\pi^1_{2,2}(\bar{x}) = (\bar{x}_{1,1}, \bar{x}_{1,2}, \bar{x}_{2,1}, \bar{x}_{2,2}, \bar{x}_{3,1}) \text{ and } \pi^2_{2,2}(\bar{x}) = (\bar{x}_{1,1}, \bar{x}_{1,2}, \bar{x}_{2,1}, \bar{x}_{2,2}, \bar{x}_{3,1}).$$

We also have

$$\pi^1_{2,1}(\bar{x}) = (\bar{x}_{1,1}, \bar{x}_{2,1}), \quad \pi^2_{2,1}(\bar{x}) = (\bar{x}_{1,1}, \bar{x}_{2,1}), \quad \text{and} \quad \pi^1_{3,1}(\bar{x}) = (\bar{x}_{1,1}, \bar{x}_{3,1}).$$

Notice that the above normal form is a natural combination of the two extreme cases: that of dynamical systems and that of systems with controllable linearization.

Indeed, if $q = n$, that is, we deal with a dynamical system, then the coordinates $(\bar{x}_1^T, \dots, \bar{x}_p^T)^T$ are not present and the normal form $\Pi_{NF}^{\leq d}$ reduces to a dynamical system $\dot{\bar{z}} = J\bar{z} + \bar{g}^{[0]}(\bar{z})$ containing resonant terms only. This is, of course, Poincaré normal form of a vector field under a formal diffeomorphism. On the other hand, if $q = 0$, that is, if the linearization of the system is controllable, the coordinates $\bar{z} = (\bar{z}_1, \dots, \bar{z}_q)^T$ are not present and our normal form reduces to

$$\begin{aligned}\bar{f}_j^{k[m]}(\bar{x}, \bar{u}) = &\sum_{1\leq s\leq t\leq p} \sum_{i=j+2}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i} P_{j,i,s,t}^{k[m]}(\pi_{t,i}^s(\bar{x})) \\ &+ \sum_{1\leq s<t\leq p} \sum_{i=j+2}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i-1} Q_{j,i,s,t}^{k[m]}(\pi_{t,i-1}^t(\bar{x}))\end{aligned}$$

for any $1 \leq k \leq p$ and any $1 \leq j \leq r-1$, and $\bar{f}_j^{k[m]}(\bar{x}, \bar{u}) = 0$ otherwise. This latter case will be summarized in the following corollary. It gives the normal form for multi-input control systems with controllable linearization (see [42]).

COROLLARY 3.2. *The system* $\Pi^{\leq d}$, *defined by* (2.1), *with* controllable linearization, *is feedback equivalent by a polynomial feedback transformation* $\Upsilon^{\leq d}$ *of the form* (2.2), *to the following normal form:*

$$\Pi_{NF}^{\leq d} \ : \ \dot{\bar{x}} = A\bar{x} + B\bar{u} + \sum_{m=2}^{d} \bar{f}^{[m]}(\bar{x}, \bar{u}) + O(\bar{x}, \bar{u})^{d+1},$$

where

$$\bar{f}^{[m]}(\bar{x}, \bar{u}) = \sum_{k=1}^{p} \sum_{j=1}^{r-1} \bar{f}_j^{k[m]}(\bar{x}, \bar{u}) \frac{\partial}{\partial \bar{x}_{k,j}},$$

with

$$\begin{aligned}\bar{f}_j^{k[m]}(\bar{x}, \bar{u}) = &\sum_{1\leq s\leq t\leq p} \sum_{i=j+2}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i} P_{j,i,s,t}^{k[m-2]}(\pi_{t,i}^s(\bar{x})) \\ &+ \sum_{1\leq s<t\leq p} \sum_{i=j+2}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i-1} Q_{j,i,s,t}^{k[m-2]}(\pi_{t,i-1}^t(\bar{x}))\end{aligned}$$

for any $1 \leq k \leq p$ *and any* $1 \leq j \leq r-1$.

Particular case $p = 2$ and $r = 3$. In this particular case the normal form will be given by

$$\Pi^{\leq d} \ : \ \begin{cases} \dot{\bar{x}}_{1,1} &= \bar{x}_{1,2} + \bar{f}_1^{1[2]}(\bar{x}, \bar{u}) + \cdots + \bar{f}_1^{1[d]}(\bar{x}, \bar{u}) + O(\bar{x}, \bar{u})^{d+1}, \\ \dot{\bar{x}}_{1,2} &= \bar{x}_{1,3} + \bar{f}_2^{1[2]}(\bar{x}, \bar{u}) + \cdots + \bar{f}_2^{1[d]}(\bar{x}, \bar{u}) + O(\bar{x}, \bar{u})^{d+1}, \\ \dot{\bar{x}}_{1,3} &= \bar{u}_1, \\ \dot{\bar{x}}_{2,1} &= \bar{x}_{2,2} + \bar{f}_1^{2[2]}(\bar{x}, \bar{u}) + \cdots + \bar{f}_1^{2[d]}(\bar{x}, \bar{u}) + O(\bar{x}, \bar{u})^{d+1}, \\ \dot{\bar{x}}_{2,2} &= \bar{x}_{2,3} + \bar{f}_2^{2[2]}(\bar{x}, \bar{u}) + \cdots + \bar{f}_2^{2[d]}(\bar{x}, \bar{u}) + O(\bar{x}, \bar{u})^{d+1}, \\ \dot{\bar{x}}_{2,3} &= \bar{u}_2, \end{cases}$$

where for any $2 \leq m \leq d$ and any $k = 1, 2$ we have

$$\begin{aligned}\bar{f}_1^{k[m]}(\bar{x},\bar{u}) &= \bar{x}_{1,3}^2 P_{1,3,1,1}^{k[m-2]}(\hat{x}_{1,3},\hat{x}_{2,2}) + \bar{x}_{2,3}^2 P_{1,3,2,2}^{k[m-2]}(\hat{x}_{1,2},\hat{x}_{2,3}) \\ &+ \bar{x}_{1,3}\bar{x}_{2,3} P_{1,3,1,2}^{k[m-2]}(\hat{x}_{1,3},\hat{x}_{2,3}) + \bar{x}_{1,3}\bar{x}_{2,2} Q_{1,3,1,2}^{k[m-2]}(\hat{x}_{1,2},\hat{x}_{2,2}) \\ &+ \bar{u}_1^2 P_{1,4,1,1}^{k[m-2]}(\hat{u}_1,\hat{x}_{2,3}) + \bar{u}_2^2 P_{1,4,2,2}^{k[m-2]}(\hat{x}_{1,3},\hat{u}_2) + \bar{u}_1\bar{u}_2 P_{1,4,1,2}^{k[m-2]}(\hat{u}_1,\hat{u}_2) \\ &+ \bar{u}_1\bar{x}_{2,3} Q_{1,4,1,2}^{k[m-2]}(\hat{x}_{1,3},\hat{x}_{2,3})\end{aligned}$$

and

$$\begin{aligned}\bar{f}_2^{k[m]}(\bar{x},\bar{u}) &= \bar{u}_1^2 P_{2,4,1,1}^{k[m-2]}(\hat{u}_1,\hat{x}_{2,3}) + \bar{u}_2^2 P_{2,4,2,2}^{k[m-2]}(\hat{x}_{1,3},\hat{u}_2) + \bar{u}_1\bar{u}_2 P_{2,4,1,2}^{k[m-2]}(\hat{u}_1,\hat{u}_2) \\ &+ \bar{u}_1\bar{x}_{2,3} Q_{2,4,1,2}^{k[m-2]}(\hat{x}_{1,3},\hat{x}_{2,3}).\end{aligned}$$

We recall that $\hat{x}_{1,i} = (\bar{x}_{1,1},\ldots,\bar{x}_{1,i})$ and $\hat{x}_{2,i} = (\bar{x}_{2,1},\ldots,\bar{x}_{2,i})$. Moreover,

$$\hat{u}_1 = (\bar{x}_{1,1},\bar{x}_{1,2},\bar{x}_{1,3},\bar{u}_1) \text{ and } \hat{u}_2 = (\bar{x}_{2,1}\bar{x}_{2,2},\bar{x}_{2,3},\bar{u}_2).$$

When the initial system is affine in the control, then in the normal form, the homogeneous polynomials $P_{j,4,1,1}^{k[m-2]}(\hat{u}_1,\hat{x}_{2,3})$, $P_{j,4,2,2}^{k[m-2]}(\hat{x}_{1,3},\hat{u}_2)$, and $P_{j,4,1,2}^{k[m-2]}(\hat{u}_1,\hat{u}_2)$ are all zero.

Generalization. Now let us assume that the controllability indices are not equal. Without loss of generality, we suppose that $1 \leq r_1 \leq \cdots \leq r_p = r$. We then define the sequence of indices $d_1 \geq \cdots \geq d_p = 0$ so that $r_1 + d_1 = \cdots = r_p + d_p = r$.

It thus suffices to extend each subsystem, say the kth subsystem of (2.1) given by

$$\Pi_k^{\leq d} : \begin{cases} \dot{x}_{k,1} &= x_{k,2} + \sum_{m=0}^{d} f_1^{k[m]}(z,x,u) + O(z,x,u)^{d+1}, \\ &\vdots \\ \dot{x}_{k,r-1} &= x_{k,r} + \sum_{m=0}^{d} f_{r_k-1}^{k[m]}(z,x,u) + O(z,x,u)^{d+1}, \\ \dot{x}_{k,r} &= u_k, \end{cases}$$

as follows. We define $\tilde{x}_k = (\tilde{x}_{k,1},\ldots,\tilde{x}_{k,r})$ so that

$$\tilde{x}_{k,d_k+1} = x_{k,1},\ldots,\tilde{x}_{k,r} = x_{k,r_k} \text{ and } \dot{\tilde{x}}_{k,1} = \tilde{x}_{k,2},\ldots,\dot{\tilde{x}}_{k,d_k} = x_{k,1}.$$

This means that the kth subsystem is extended as

$$\tilde{\Pi}_k^{\leq d} : \begin{cases} \dot{\tilde{x}}_{k,1} &= \tilde{x}_{k,2}, \\ &\vdots \\ \dot{\tilde{x}}_{k,d_k} &= \tilde{x}_{k,d_k+1}, \\ \dot{\tilde{x}}_{k,d_k+1} &= \tilde{x}_{k,d_k+2} + \sum_{m=0}^{d} \tilde{f}_{d_k+1}^{k[m]}(z,\tilde{x},u) + O(z,\tilde{x},u)^{d+1}, \\ &\vdots \\ \dot{\tilde{x}}_{k,r-1} &= \tilde{x}_{k,r} + \sum_{m=0}^{d} \tilde{f}_{r-1}^{k[m]}(z,\tilde{x},u) + O(z,\tilde{x},u)^{d+1}, \\ \dot{\tilde{x}}_{k,r} &= u_k, \end{cases}$$

where

$$\tilde{f}_{d_k+1}^{k[m]}(z,\tilde{x},u) = f_1^{k[m]}(z,x,u),\ldots,\tilde{f}_{r-1}^{k[m]}(z,\tilde{x},u) = f_{r_k-1}^{k[m]}(z,x,u).$$

In this case all extended subsystems will have the same controllability index r, and by Theorem 3.1 the extended system will be equivalent to the normal form $\Pi_{NF}^{\leq d}$ given by (3.2)–(3.3) with $\bar{g}_j^{[m-1]}(\bar{z},\bar{x},\bar{u})$ and $\bar{f}_j^{k[m]}(\bar{z},\bar{x},\bar{u})$ *depending exclusively on the variables* $\bar{z},\bar{u}$ and $\bar{x}_{s,d_s+1},\ldots,\bar{x}_{s,r}$ (not on the added variables $\bar{x}_{s,1},\ldots,\bar{x}_{s,d_s}$) for all $1 \leq s \leq p$. Moreover, the first d_k components $\bar{f}_1^{k[m]}(\bar{z},\bar{x},\bar{u}),\ldots,\bar{f}_{d_k}^{k[m]}(\bar{z},\bar{x},\bar{u})$ remain zero (see Example 1 for illustration).

3.2. Examples. In this subsection we will illustrate our results by considering two physical examples: The MVWT and the prototype of a PVTOL.

Example 1. Multi-Vehicle Wireless Testbed. We consider the MVWT, presented in [8, 9] and we study the normal form of one vehicle. The equations of motion of an MVWT vehicle (see [8, 9]) are given by

$$\begin{cases} m\ddot{x} &= -\eta\dot{x} + (F_s + F_p)\cos\theta, \\ m\ddot{y} &= -\eta\dot{y} + (F_s + F_p)\sin\theta, \\ J\ddot{\theta} &= -\psi\dot{\theta} + (F_s - F_p)l, \end{cases}$$

where (x,y) denotes the position of the center mass of the vehicle, θ the angle of the axis of the vehicle with the horizontal (x-axis), m the mass of the vehicle, J the rotational inertia, F_s and F_p denote the starboard and port fan forces, respectively, and l (r in [8, 9]) the common moment arm of the forces. The center mass of the vehicle and the center of geometry are assumed to coincide. The constants η and ψ stand, respectively, for the coefficients of viscous friction and rotational friction.

Let us introduce the variables

$$\begin{array}{lllllllllll} z_1 &=& y, & x_{1,1} &=& x, & x_{2,1} &=& \theta, & u_1 &=& F_s + F_p, \\ z_2 &=& \dot{z_1}, & x_{1,2} &=& \dot{x}_{1,1}, & x_{2,2} &=& \dot{x}_{2,1}, & u_2 &=& F_s - F_p. \end{array}$$

The equations of motion of an MVWT vehicle are rewritten as

$$\begin{cases} \dot{z}_1 &= z_2, \\ \dot{z}_2 &= -\eta z_2 + u_1 \sin x_{2,1}, \\ \dot{x}_{1,1} &= x_{1,2}, \\ \dot{x}_{1,2} &= -\eta x_{1,2} + u_1 \cos x_{2,1}, \\ \dot{x}_{2,1} &= x_{2,2}, \\ \dot{x}_{2,2} &= -\phi x_{2,2} + u_2 l. \end{cases} \tag{3.4}$$

We can notice that the system is affine and its distribution $\mathcal{G} = \operatorname{span}\{g_1, g_2\}$, where

$$g_1 = \begin{pmatrix} 0 \\ \sin x_{2,1} \\ 0 \\ \cos x_{2,1} \\ 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad g_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

is involutive and of constant rank 2. An equilibrium point for the system (3.4) is given by any constant position and orientation

$$(z_1^e, z_2^e, x_{1,1}^e, x_{1,2}^e, x_{2,1}^e, x_{2,2}^e)^T = (z_1, 0, x_{1,1}, 0, x_{2,1}, 0)^T.$$

The linearization of the system (3.4) around an equilibrium (we assume $x_{2,1} = 0$) is given by

$$\left\{\begin{array}{rcl} \dot{z}_1 &=& z_2, \\ \dot{z}_2 &=& -\eta z_2, \\ \dot{x}_{1,1} &=& x_{1,2}, \\ \dot{x}_{1,2} &=& -\eta x_{1,2} + u_1, \\ \dot{x}_{2,1} &=& x_{2,2}, \\ \dot{x}_{2,2} &=& -\phi x_{2,2} + u_2 l. \end{array}\right.$$

It is easy to see that this linear system is not controllable because

$$\operatorname{span}\left\{ F^i G_k,\ 0 \le i \le 5,\ 1 \le k \le 2 \right\} = \mathbb{R}^4,$$

where

$$F = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -\eta & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\eta & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -\phi \end{pmatrix}, \qquad G_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \text{and} \qquad G_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

It thus follows that $q = 2$, and the computation of the controllability matrix shows that $r_1 = r_2 = 2$.

The feedback transformation defined by

$$u_1 = \frac{1}{\cos x_{2,1}} \bar{u}_1 + \eta \frac{x_{1,2}}{\cos x_{2,1}} \qquad \text{and} \qquad \bar{u}_2 = \frac{u_2}{l} + \frac{\phi}{l} x_{2,2}$$

takes the system into the following form:

$$\left\{\begin{array}{rcl} \dot{z}_1 &=& z_2 \\ \dot{z}_2 &=& -\eta z_2 + \eta x_{1,2} \tan x_{2,1} + \bar{u}_1 \tan x_{2,1}, \\ \dot{x}_{1,1} &=& x_{1,2}, \\ \dot{x}_{1,2} &=& \bar{u}_1, \\ \dot{x}_{2,1} &=& x_{2,2}, \\ \dot{x}_{2,2} &=& \bar{u}_2. \end{array}\right.$$

The change of coordinates given by

$$\left\{\begin{array}{rcl} \bar{z}_1 &=& z_1, \\ \bar{z}_2 &=& z_2 - x_{1,2} \tan x_{2,1}, \\ \bar{x}_{1,1} &=& x_{1,1}, \\ \bar{x}_{1,2} &=& x_{1,2}, \\ \bar{x}_{2,1} &=& x_{2,1}, \\ \bar{x}_{2,2} &=& x_{2,2} \end{array}\right.$$

brings the system into the form

$$\left\{\begin{array}{rcl} \dot{\bar{z}}_1 &=& \bar{z}_2 + \bar{x}_{1,2} \tan \bar{x}_{2,1}, \\ \dot{\bar{z}}_2 &=& -\eta \bar{z}_2 - \bar{x}_{1,2} \bar{x}_{2,2} (1 + \tan^2 \bar{x}_{2,1}), \\ \dot{\bar{x}}_{1,1} &=& \bar{x}_{1,2}, \\ \dot{\bar{x}}_{1,2} &=& \bar{u}_1, \\ \dot{\bar{x}}_{2,1} &=& \bar{x}_{2,2}, \\ \dot{\bar{x}}_{2,2} &=& \bar{u}_2. \end{array}\right.$$

Since

$$\bar{x}_{1,2}\tan\bar{x}_{2,1}=\bar{x}_{1,2}\sum_{\nu=0}^{\infty}(-1)^{\nu}\frac{\bar{x}_{2,1}^{2\nu+1}}{(2\nu+1)!}=\bar{x}_{1,2}\bar{x}_{2,1}\sum_{\nu=0}^{\infty}(-1)^{\nu}\frac{\bar{x}_{2,1}^{2\nu}}{(2\nu+1)!},$$

$$\bar{x}_{1,2}\bar{x}_{2,2}(1+\tan^2\bar{x}_{2,1})=\bar{x}_{1,2}\bar{x}_{2,2}\left(1+\sum_{\nu=0}^{\infty}(-1)^{\nu}\frac{\bar{x}_{2,1}^{2\nu+1}}{(2\nu+1)!}\right),$$

we conclude that the system is in normal form (compare with Theorem 3.1), with

$$\bar{g}_1^{[m-1]}(\bar{x})=\begin{cases}\bar{x}_{1,2}\bar{x}_{2,1}Q_{1,2}^{[m-3]}(\bar{x})=\pm\bar{x}_{1,2}\bar{x}_{2,1}\frac{\bar{x}_{2,1}^{m-3}}{(m-2)!} & \text{if } m \text{ is even,}\\ 0 & \text{if } m \text{ is odd}\end{cases}$$

and

$$\bar{g}_2^{[m-1]}(\bar{x})=\begin{cases}-\bar{x}_{1,2}\bar{x}_{2,2} & \text{if } m=3,\\ \bar{x}_{1,2}\bar{x}_{2,2}P_{2,2}^{[m-3]}(\bar{x})=\pm\bar{x}_{1,2}\bar{x}_{2,2}\frac{\bar{x}_{2,1}^{m-3}}{(m-3)!} & \text{if } m\geq 4 \text{ is odd,}\\ 0 & \text{if } m \text{ is even.}\end{cases}$$

Example 2. Planar Vertical TakeOff and Landing. In this example we study a simple toy aircraft of prototype PVTOL presented in [15, 40]. The equations of motion of the PVTOL (see [15, 40]) are given by

$$\begin{cases}\ddot{x} &= -\sin\theta u_1+\epsilon^2\cos\theta u_2,\\ \ddot{y} &= \cos\theta u_1+\epsilon^2\sin\theta u_2-1,\\ \ddot{\theta} &= u_2,\end{cases}$$

where (x,y) denotes the position of the center mass of the aircraft, θ the angle of the aircraft relative to the x-axis, " -1" the gravitational acceleration, and $\epsilon\neq 0$ the (small) coefficient giving the coupling between the rolling moment and the lateral acceleration of the aircraft. The control inputs u_1 and u_2 are the thrust (directed out the bottom of the aircraft) and the rolling moment.

We introduce the variables

$$\begin{array}{ccccccccccc} x_{1,1} & = & y, & x_{2,1} & = & x, & x_{2,3} & = & \theta, & w_1 & = & u_1-1,\\ x_{1,2} & = & \dot{x}_{1,1}, & x_{2,2} & = & \dot{x}_{2,1}, & x_{2,4} & = & \dot{x}_{2,3}, & w_2 & = & u_2.\end{array}$$

The equations of motion of the PVTOL are rewritten as

$$\begin{cases}\dot{x}_{1,1} &= x_{1,2},\\ \dot{x}_{1,2} &= \cos x_{2,3}w_1+\epsilon^2\sin x_{2,3}w_2+\cos x_{2,3}-1,\\ \dot{x}_{2,1} &= x_{2,2},\\ \dot{x}_{2,2} &= -\sin x_{2,3}w_1+\epsilon^2\cos x_{2,3}w_2-\sin x_{2,3},\\ \dot{x}_{2,3} &= x_{2,4},\\ \dot{x}_{2,4} &= w_2.\end{cases} \tag{3.5}$$

The system is affine and its distribution $\mathcal{G}=\operatorname{span}\{g_1,g_2\}$, given by

$$g_1=\begin{pmatrix}0\\ \cos x_{2,3}\\ 0\\ -\sin x_{2,3}\\ 0\\ 0\end{pmatrix}\qquad\text{and}\qquad g_2=\begin{pmatrix}0\\ \epsilon^2\sin x_{2,3}\\ 0\\ \epsilon^2\cos x_{2,3}\\ 0\\ 1\end{pmatrix},$$

is involutive and of constant rank 2. The equilibria is defined by

$$(x^e_{1,1}, x^e_{1,2}, x^e_{2,1}, x^e_{2,2}, x^e_{2,3}, x^e_{2,4}, w^e_1, w^e_2)^T = (c, 0, 0, 0, 0, 0, 0, 0)^T,$$

where c is any constant. The linearization of the system (3.5) around the equilibria is given by

$$\begin{cases} \dot{x}_{1,1} &= x_{1,2}, \\ \dot{x}_{1,2} &= w_1, \\ \dot{x}_{2,1} &= x_{2,2}, \\ \dot{x}_{2,2} &= -x_{2,3} + \epsilon^2 w_2, \\ \dot{x}_{2,3} &= x_{2,4}, \\ \dot{x}_{2,4} &= w_2. \end{cases}$$

It is easy to see that the linear system is controllable with controllability indices $r_1 = 2$ and $r_2 = 4$. Indeed,

$$\text{span}\left\{ G_1, FG_1, G_2, FG_2, F^2G_2, F^3G_2 \right\} = \mathbb{R}^6,$$

where

$$F = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad G_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \text{and} \qquad G_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \epsilon^2 \\ 0 \\ 1 \end{pmatrix}.$$

Since $r_1 = 2 < r_2 = 4$ we have $d_1 = 2$ and $d_2 = 0$. Thus we extend the system as

$$\begin{cases} \dot{\tilde{x}}_{1,1} &= \tilde{x}_{1,2}, \\ \dot{\tilde{x}}_{1,2} &= \tilde{x}_{1,3}, \\ \dot{\tilde{x}}_{1,3} &= \tilde{x}_{1,4}, \\ \dot{\tilde{x}}_{1,4} &= \cos\tilde{x}_{2,3} w_1 + \epsilon^2 \sin\tilde{x}_{2,3} w_2 + \cos\tilde{x}_{2,3} - 1, \\ \dot{\tilde{x}}_{2,1} &= \tilde{x}_{2,2}, \\ \dot{\tilde{x}}_{2,2} &= -\sin\tilde{x}_{2,3} w_1 + \epsilon^2 \cos\tilde{x}_{2,3} w_2 - \sin\tilde{x}_{2,3}, \\ \dot{\tilde{x}}_{2,3} &= \tilde{x}_{2,4}, \\ \dot{\tilde{x}}_{2,4} &= w_2, \end{cases}$$

where

$$\tilde{x}_{1,3} = x_{1,1}, \quad \tilde{x}_{1,4} = x_{1,2}, \quad \tilde{x}_{2,1} = x_{2,1}, \quad \tilde{x}_{2,2} = x_{2,2}, \quad \tilde{x}_{2,3} = x_{2,3}, \quad \tilde{x}_{2,4} = x_{2,4}.$$

The feedback transformation defined by

$$w_1 = \frac{1}{\cos\tilde{x}_{2,1}} v_1 - \epsilon^2 \tan\tilde{x}_{2,1} v_2 + \frac{1}{\cos\tilde{x}_{2,1}} - 1 \qquad \text{and} \qquad w_2 = v_2$$

takes the system into the following form:

$$\begin{cases} \dot{\tilde{x}}_{1,1} &= \tilde{x}_{1,2}, \\ \dot{\tilde{x}}_{1,2} &= \tilde{x}_{1,3}, \\ \dot{\tilde{x}}_{1,3} &= \tilde{x}_{1,4}, \\ \dot{x}_{1,4} &= v_1, \\ \dot{\tilde{x}}_{2,1} &= \tilde{x}_{2,2}, \\ \dot{\tilde{x}}_{2,2} &= -\tan\tilde{x}_{2,3}v_1 + \frac{\epsilon^2}{\cos\tilde{x}_{2,3}}v_2 - \tan\tilde{x}_{2,3}, \\ \dot{\tilde{x}}_{2,3} &= \tilde{x}_{2,4}, \\ \dot{\tilde{x}}_{2,4} &= v_2. \end{cases}$$

The change of coordinates given by

$$\begin{cases} \bar{x}_{1,1} &= \tilde{x}_{1,1}, \\ \bar{x}_{1,2} &= \tilde{x}_{1,2}, \\ \bar{x}_{1,3} &= \tilde{x}_{1,3} - \epsilon^2 \int_0^{\tilde{x}_{2,3}} \frac{dt}{\cos t}, \\ \bar{x}_{1,4} &= \tilde{x}_{1,4} + \tilde{x}_{1,4}\tan\tilde{x}_{2,3} - \frac{\epsilon^2}{\cos\tilde{x}_{2,3}}\tilde{x}_{2,4}, \\ \bar{x}_{2,1} &= \tilde{x}_{2,1}, \\ \bar{x}_{2,2} &= \tilde{x}_{2,2}, \\ \bar{x}_{2,3} &= -\tan\tilde{x}_{2,3}, \\ \bar{x}_{2,4} &= -\tilde{x}_{2,4}(1+\tan^2\tilde{x}_{2,3}) = \dot{\tilde{x}}_{2,3} \end{cases}$$

followed by the feedback

$$\bar{u}_1 = v_1, \text{ and } \bar{u}_2 = \dot{\tilde{x}}_{2,4} = -v_2(1+\tan^2\tilde{x}_{2,3}) - 2\tilde{x}_{2,4}^2\tan\tilde{x}_{2,3}(1+\tan^2\tilde{x}_{2,3})$$

brings the system into

$$\begin{cases} \dot{\bar{x}}_{1,1} &= \bar{x}_{1,2}, \\ \dot{\bar{x}}_{1,2} &= \bar{x}_{1,3}, \\ \dot{\bar{x}}_{1,3} &= \bar{x}_{1,4}, \\ \dot{\bar{x}}_{1,4} &= \bar{u}_1, \\ \dot{\bar{x}}_{2,1} &= x_{2,2} + \bar{x}_{1,4}\bar{x}_{2,3}, \\ \dot{\bar{x}}_{2,2} &= \bar{x}_{2,3} - \bar{x}_{1,4}\bar{x}_{2,4} + \epsilon^2(1-\bar{x}_{2,3}^2)\bar{x}_{2,4}^2, \\ \dot{\bar{x}}_{2,3} &= \bar{x}_{2,4}, \\ \dot{\bar{x}}_{2,4} &= \bar{u}_2. \end{cases}$$

Comparing with Corollary 3.2 we get

$$\begin{aligned} \bar{f}_1^{2[2]}(\bar{x}) &= \bar{x}_{1,4}\bar{x}_{2,3}Q_{1,4,1,2}^{2[0]}(\bar{x}), \\ \bar{f}_2^{2[2]}(\bar{x}) &= \bar{x}_{1,4}\bar{x}_{2,4}P_{2,4,1,2}^{2[0]}(\bar{x}) + \bar{x}_{2,4}\bar{x}_{2,4}P_{2,4,2,2}^{2[0]}(\bar{x}), \\ \bar{f}_2^{2[4]}(\bar{x}) &= \bar{x}_{2,4}\bar{x}_{2,4}P_{2,4,2,2}^{2[2]}(\pi_{2,4}^2(\bar{x})), \end{aligned}$$

where

$$Q_{1,4,1,2}^{2[0]}(\bar{x}) \equiv 1, \quad P_{2,4,1,2}^{2[0]}(\bar{x}) \equiv -1, \quad P_{2,4,2,2}^{2[0]}(\bar{x}) = \epsilon^2, \quad P_{2,4,2,2}^{2[2]}(\bar{x}) = -\epsilon^2\bar{x}_{2,3}^2.$$

This means that the system (3.5) is equivalent to the normal form

$$\begin{cases} \dot{\bar{x}}_{1,3} &= \bar{x}_{1,4}, \\ \dot{\bar{x}}_{1,4} &= \bar{u}_1, \\ \dot{\bar{x}}_{2,1} &= x_{2,2} + \bar{x}_{1,4}\bar{x}_{2,3}, \\ \dot{\bar{x}}_{2,2} &= \bar{x}_{2,3} - \bar{x}_{1,4}\bar{x}_{2,4} + \epsilon^2(1-\bar{x}_{2,3}^2)\bar{x}_{2,4}^2, \\ \dot{\bar{x}}_{2,3} &= \bar{x}_{2,4}, \\ \dot{\bar{x}}_{2,4} &= \bar{u}_2. \end{cases}$$

Notice that the *added variables* $\bar{x}_{1,1}$ and $\bar{x}_{1,2}$ are not present in the normal form. We may also notice that in both *Examples* 1 and 2 the transformations taking the corresponding systems into their normal forms are smooth, actually they are analytic. Though little is known about the convergence of the formal transformations taking a system into its normal form, this gives hope.

3.3. Nearby controllability. In this subsection we generalize a result obtained earlier in collaboration with Kang et al. [30]. We proved that for systems with one nonzero uncontrollable mode the set of equilibria is a smooth curve passing by the origin. Moreover, provided that some term in the normal form is nonzero, the system becomes linearly controllable at these equilibria (except at the origin). This is called a *bifurcation of controllability* and the conclusion drawn from this study is that we can stabilize the system at any nearby point of the origin.

A parallel analysis could be made for multi-input systems with one nonzero uncontrollable mode. Indeed, consider the system $\Pi^{\leq d}$ defined by (2.1) and assume that $q = 1$, i.e., $J = \lambda$. The equilibria set of this system is

$$E = \left\{ (z, x) \in \mathbb{R} \times \mathbb{R}^{n-1} \text{ such that } \exists u \in \mathbb{R}^p : H(z, x, u) = 0 \right\},$$

where $H(z, x, u) = (g(z, x, u), f(z, x, u))$ with

$$\begin{aligned} g(z, x, u) &= \lambda z + \sum_{m=1}^{d} g^{[m-1]}(z, x, u) + O(z, x, u)^d, \\ f(z, x, u) &= Ax + Bu + \sum_{m=0}^{d} f^{[m]}(z, x, u) + O(z, x, u)^{d+1}. \end{aligned}$$

We will show that E is a surface parameterized by $x^1 = (x_{1,1}, x_{2,1}, \ldots, x_{p,1})^T$. If we denote by $\mathbf{x}_s = (x_{s,2}, \ldots, x_{s,r})^T$ for all $1 \leq s \leq p$ and $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_p^T)$, then we have

$$\frac{\partial H(z, x, u)}{\partial (z, \mathbf{x}, u)}\Big|_{(z,x,u)=0} = \text{diag } (\lambda, \text{Id}_{\mathbb{R}^{n-1}}).$$

Since $\lambda \neq 0$, the matrix diag $(\lambda, \text{Id}_{\mathbb{R}^{n-1}})$ is invertible. The *implicit function theorem* implies that the equation

$$H(z, x, u) = H(z, x^1, \mathbf{x}, u) = 0$$

has a solution in a neighborhood of the origin parameterized by the variables x^1, that is, there exist functions

$$z = z^e(x^1), \quad \mathbf{x} = \mathbf{x}^e(x^1), \quad u = u^e(x^1)$$

so that for some open neighborhood $\mathcal{V}$ of the origin in $\mathbb{R}^p$, we have

$$H(z^e(x^1), x^1, \mathbf{x}^e(x^1), u^e(x^1)) = 0 \text{ for all } x^1 \in \mathcal{V} \subset \mathbb{R}^p.$$

We thus deduce that the equilibria set is a surface parameterized by the variables x^1.

Let us denote by E_0 the subset of E defined by

$$E_0 = \left\{ (z, x) \in E \text{ such that } x_{s,1} \neq 0 \text{ for all } 1 \leq s \leq p \right\}.$$

Consider the normal form of $\Pi^{\leq d}$ given by (3.1)–(3.3), where (since $q = 1$) we have

$$\begin{aligned}\bar{g}_1^{[m-1]}(\bar{z},\bar{x},\bar{u}) &= \sum_{1\leq s\leq t\leq p}\sum_{i=1}^{r+1}\bar{x}_{s,i}\bar{x}_{t,i}P_{1,i,s,t}^{[m-3]}(\bar{z},\pi_{t,i}^{s}(\bar{x}))\\ &+ \sum_{1\leq s<t\leq p}\sum_{i=2}^{r+1}\bar{x}_{s,i}\bar{x}_{t,i-1}Q_{1,i,s,t}^{[m-3]}(\bar{z},\pi_{t,i-1}^{t}(\bar{x}))\end{aligned}$$

for any $m \geq 3$.

An analogous result to that given in [30] could be formulated as follows.

THEOREM 3.3. *Consider the system $\Pi^{\leq d}$ defined by (2.1) for $d \in \mathbb{N}\cup\{\infty\}$ sufficiently large enough. If there are integers $3 \leq m \leq d$ and $1 \leq s \leq t \leq p$ so that*

$$P_{1,1,s,t}^{[m-3]}(\bar{z},\pi_{t,1}^{s}(\bar{x}))\Big|_{\bar{z}=0} = P_{1,1,s,t}^{[m-3]}(\bar{x}_{1,1},\ldots,\bar{x}_{s,1},\bar{x}_{t,1}) \neq 0,$$

then the system $\Pi^{\leq d}$ is linearly controllable at any point of E_0.

The proof of this result is straightforward and follows the same steps as in [30]. The domain where the system is linearly controllable could be, of course, larger than the subset E_0, but this will depend on the normal form.

Example 3. Consider the system

$$\left\{\begin{array}{rcl}\dot{z} &=& z + x_{1,1}x_{2,1},\\ \dot{x}_{1,1} &=& x_{1,2},\\ \dot{x}_{1,2} &=& x_{1,3},\\ \dot{x}_{1,3} &=& u_1,\\ \dot{x}_{2,1} &=& x_{2,2},\\ \dot{x}_{2,2} &=& u_2,\end{array}\right.$$

which is already in normal form. Its equilibria set is given by

$$E = \left\{\, (-x_{1,1}x_{2,1}, x_{1,1}, 0, 0, x_{2,1}, 0)^T \in \mathbb{R}^6 :\ (x_{1,1}, x_{2,1})^T \in \mathbb{R}^2 \,\right\}.$$

The system is not linearly controllable at the origin but it is at any other point of E. Indeed, put

$$f(z,x,u) = (z + x_{1,1}x_{2,1}, x_{1,2}, x_{1,3}, 0, x_{2,2}, 0)^T,$$

$$g_1(z,x,u) = (0,0,0,1,0,0)^T, \text{ and } g_2(z,x,u) = (0,0,0,0,0,1)^T.$$

(i) If $x_{1,1} \neq 0$, we have

$$\operatorname{span}\left\{\, g_1, ad_f g_1, ad_f^2 g_1, g_2, ad_f g_2, ad_f^2 g_2 \,\right\}(z,x) = \mathbb{R}^6$$

for all $(z,x) \neq (0,0)$ in E.

(ii) If $x_{2,1} \neq 0$, we have

$$\operatorname{span}\left\{\, g_1, ad_f g_1, ad_f^2 g_1, ad_f^3 g_1, g_2, ad_f g_2, \,\right\}(z,x) = \mathbb{R}^6$$

for all $(z,x) \neq (0,0)$ in E.

In this example the linear controllability occurs outside the subset E_0 for which we have $\bar{x}_{1,1} \neq 0$ and $\bar{x}_{2,1} \neq 0$.

However, if we replace $\dot{z} = z + x_{1,1}x_{2,1}$ by $\dot{z} = z + x_{1,1}^2 x_{2,1}^2$, then the linear controllability will occur only inside E_0.

4. Proofs of main results. The aim of this section is to prove Theorem 3.1. The proof of Corollary 3.2 is given in [42].

Consider the system $\Pi^{\leq d}$ defined by (2.1). Following [44] it is possible to show that the terms $f^{[0]}(z)$ and $f^{[1]}(z,x)$ can be removed. Thus, without loss of generality we will assume that the system $\Pi^{\leq d}$ is of the form

$$\Pi^{\leq d}: \begin{cases} \dot{z} = Jz + g^{[0]}(z) + \sum_{m=2}^{d} g^{[m-1]}(z,x,u) + O(z,x,u)^d, \\ \dot{x} = Ax + Bu + \sum_{m=2}^{d} f^{[m]}(z,x,u) + O(z,x,u)^{d+1}. \end{cases} \tag{4.1}$$

We like to study the action of the feedback transformation

$$\Upsilon^m: \begin{cases} \bar{z} &= z + \psi^{[m-1]}(z,x), \\ \bar{x} &= x + \phi^{[m]}(z,x), \\ u &= \bar{u} + \gamma^{[m]}(z,x,\bar{u}) \end{cases}$$

on the system $\Pi^{\leq d}$ up to some degree. First, remark that the inverse of this transformation is such that

$$\begin{cases} z &= \bar{z} - \psi^{[m-1]}(\bar{z},\bar{x}) + O(\bar{z},\bar{x})^m, \\ x &= \bar{x} - \phi^{[m]}(\bar{z},\bar{x}) + O(\bar{z},\bar{x})^{m+1}, \\ u &= \bar{u} + \gamma^{[m]}(\bar{z},\bar{x},\bar{u}) + O(\bar{z},\bar{x},\bar{u})^{m+1}. \end{cases}$$

Then the uncontrollable part is transformed as

$$\begin{aligned} \dot{\bar{z}} &= \dot{z} + \frac{\partial \psi^{[m-1]}}{\partial z}(z,x)\dot{z} + \frac{\partial \psi^{[m-1]}}{\partial x}(z,x)\dot{x} \\ &= Jz + g^{[0]}(z) + \cdots + g^{[m-1]}(z,x,u) + O(z,x,u)^m \\ &+ \frac{\partial \psi^{[m-1]}}{\partial z}(z,x)(Jz + g^{[0]}(z)) + \frac{\partial \psi^{[m-1]}}{\partial x}(z,x)(Ax + Bu) + O(z,x,u)^m \\ &= J\bar{z} + g^{[0]}(\bar{z}) + \cdots + g^{[m-1]}(\bar{z},\bar{x},\bar{u}) - \left(J + \frac{\partial g^{[0]}}{\partial z}(\bar{z})\right)\psi^{[m-1]}(\bar{z},\bar{x}) \\ &+ \frac{\partial \psi^{[m-1]}}{\partial z}(\bar{z},\bar{x})(J\bar{z} + g^{[0]}(\bar{z})) + \frac{\partial \psi^{[m-1]}}{\partial x}(\bar{z},\bar{x})(A\bar{x} + B\bar{u}) + O(\bar{z},\bar{x},\bar{u})^m. \end{aligned}$$

It clearly appears that the terms of degree $m-2$ or less of the uncontrollable part remain unmodified while the terms of degree $m-1$ or higher are modified.

Similarly, we can show that

$$\begin{aligned} \dot{\bar{x}} &= A\bar{x} + B\bar{u} + f^{[2]}(\bar{z},\bar{x},\bar{u}) + \cdots + f^{[m]}(\bar{z},\bar{x},\bar{u}) - A\phi^{[m]}(\bar{z},\bar{x}) - B\gamma^{[m]}(\bar{z},\bar{x},\bar{u}) \\ &+ \frac{\partial \phi^{[m]}}{\partial z}(\bar{z},\bar{x})(J\bar{z} + g^{[0]}(\bar{z})) + \frac{\partial \phi^{[m]}}{\partial x}(\bar{z},\bar{x})(A\bar{x} + B\bar{u}) + O(\bar{z},\bar{x},\bar{u})^{m+1}, \end{aligned}$$

which means that the terms of degree $m-1$ or less of the controllable part are preserved while the terms of degree m or higher are modified.

To study the action of the feedback transformation Υ^m on the terms of degree m (terms of degree $m-1$ of the uncontrollable part and of degree m for the controllable part) of the system $\Pi^{\leq d}$, it is enough to study their action on a *homogeneous system* of the form

$$\Pi^m: \begin{cases} \dot{z} = Jz + g^{[0]}(z) + g^{[m-1]}(z,x,u), \\ \dot{x} = Ax + Bu + f^{[m]}(z,x,u). \end{cases} \tag{4.2}$$

The proof of Theorem 3.1 will follow if we show that, by a feedback transformation Υ^m, we can take the system (4.2) into the normal form

$$\Pi^m_{NF}: \begin{cases} \dot{\bar{z}} = J\bar{z} + g^{[0]}(\bar{z}) + \bar{g}^{[m-1]}(\bar{z},\bar{x},\bar{u}), \\ \dot{\bar{x}} = A\bar{x} + B\bar{u} + \bar{f}^{[m]}(\bar{z},\bar{x},\bar{u}), \end{cases}$$

where the components of $\bar{g}^{[m-1]}(\bar{z},\bar{x},\bar{u})$ and $\bar{f}^{[m]}(\bar{z},\bar{x},\bar{u})$ are given by (3.1)–(3.3).

Indeed, if this is true we then consider the system $\Pi^{\leq d}$ of the form (4.1) and we first apply a quadratic feedback transformation Υ^2 to take it to the form

$$\Pi^{\leq d}: \begin{cases} \dot{\bar{z}} = J\bar{z} + g^{[0]}(\bar{z}) + \bar{g}^{[1]}(\bar{z},\bar{x},\bar{u}) + \sum\limits_{m=3}^{d} g^{[m-1]}(\bar{z},\bar{x},\bar{u}) + O(\bar{z},\bar{x},\bar{u})^d, \\ \dot{\bar{x}} = A\bar{x} + B\bar{u} + \bar{f}^{[2]}(\bar{z},\bar{x},\bar{u}) + \sum\limits_{m=3}^{d} f^{[m]}(\bar{z},\bar{x},\bar{u}) + O(\bar{z},\bar{x},\bar{u})^{d+1}, \end{cases}$$

where the vector fields $\bar{g}^{[1]}(\bar{z},\bar{x},\bar{u})$, and $\bar{f}^{[2]}(\bar{z},\bar{x},\bar{u})$ are in their normal forms, and the vector fields $g^{[m-1]}(\bar{z},\bar{x},\bar{u})$ and $f^{[m]}(\bar{z},\bar{x},\bar{u})$ stand for the new transformed vector fields. We thus apply a cubic transformation Υ^3 to take the system above into the form

$$\Pi^{\leq d}: \begin{cases} \dot{\bar{z}} = J\bar{z} + g^{[0]}(\bar{z}) + \bar{g}^{[1]}(\bar{z},\bar{x},\bar{u}) + \bar{g}^{[2]}(\bar{z},\bar{x},\bar{u}) + \sum\limits_{m=4}^{d} g^{[m-1]}(\bar{z},\bar{x},\bar{u}) + O(\bar{z},\bar{x},\bar{u})^d, \\ \dot{\bar{x}} = A\bar{x} + B\bar{u} + \bar{f}^{[2]}(\bar{z},\bar{x},\bar{u}) + \bar{f}^{[3]}(\bar{z},\bar{x},\bar{u}) + \sum\limits_{m=4}^{d} f^{[m]}(\bar{z},\bar{x},\bar{u}) + O(\bar{z},\bar{x},\bar{u})^{d+1}, \end{cases}$$

where $\bar{g}^{[1]}(\bar{z},\bar{x},\bar{u})$, $\bar{g}^{[2]}(\bar{z},\bar{x},\bar{u})$, $\bar{f}^{[2]}(\bar{z},\bar{x},\bar{u})$, and $\bar{f}^{[3]}(\bar{z},\bar{x},\bar{u})$ are in their normal forms. The vectors $g^{[m-1]}(\bar{z},\bar{x},\bar{u})$ and $f^{[m]}(\bar{z},\bar{x},\bar{u})$, for $m \geq 4$ are the new transformed vector fields. The process continues until the original system is in the desired normal form.

Proof of Theorem 3.1. As stated above, we need to prove that the homogeneous system $\Pi^{[m]}$ could be transformed into the normal form $\Pi^{[m]}_{NF}$ by a homogeneous transformation Υ^m. The proof will be divided into two parts. In the first part we will deal with the controllable mode and in the second part we will consider the uncontrollable mode.

(i) Consider the kth subsystem

$$\Pi^{k[m]}: \begin{cases} \dot{x}_{k,1} &= x_{k,2} + f_1^{k[m]}(z,x,u), \\ &\vdots \\ \dot{x}_{k,r-1} &= x_{k,r} + f_{r-1}^{k[m]}(z,x,u), \\ \dot{x}_{k,r} &= u_k. \end{cases}$$

Let us denote by $\mathfrak{P}^m(\mathbb{R}^q \times \mathbb{R}^{n-q} \times \mathbb{R}^p)$ the set of all homogeneous polynomials of degree m in the variables $(x,u) \in \mathbb{R}^{n-q} \times \mathbb{R}^p$ whose coefficients are functions of the variable $z \in \mathbb{R}^q$.

For a fixed j, $1 \leq j \leq r-1$ we define the set $\mathfrak{F}^m_j(\mathbb{R}^q \times \mathbb{R}^{n-q} \times \mathbb{R}^p)$ of all

homogeneous polynomials $h^{[m]}(z,x,v) \in \mathfrak{P}^m(\mathbb{R}^q \times \mathbb{R}^{n-q} \times \mathbb{R}^p)$ such that

$$h^{[m]}(z,x,v) = \sum_{1\le s\le t\le p} \sum_{i=j+2}^{r+1} x_{s,i}x_{t,i} P^{k[m-2]}_{j,i,s,t}(z, \pi^s_{t,i}(x)) + \sum_{1\le s<t\le p} \sum_{i=j+2}^{r+1} x_{s,i}x_{t,i-1} Q^{k[m-2]}_{j,i,s,t}(z, \pi^t_{t,i-1}(x)).$$

For simplicity we will just refer to $\mathfrak{P}^m(\mathbb{R}^q \times \mathbb{R}^{n-q} \times \mathbb{R}^p)$ and $\mathfrak{F}^m_j(\mathbb{R}^q \times \mathbb{R}^{n-q} \times \mathbb{R}^p)$ as $\mathfrak{P}^m$ and $\mathfrak{F}^m_j$, respectively. Denote by $\mathfrak{E}^m_j$ the subspace of $\mathfrak{P}^m$ so that

$$\mathfrak{P}^m = \mathfrak{F}^m_j \oplus \mathfrak{E}^m_j.$$

We want to prove that the subsystem $\Pi^{k[m]}$ could be transformed into the normal form

$$\Pi^{k[m]}_{NF} : \begin{cases} \dot{\bar{x}}_{k,1} &= \bar{x}_{k,2} + \bar{f}^{k[m]}_1(\bar{z},\bar{x},\bar{u}), \\ &\vdots \\ \dot{\bar{x}}_{k,r-1} &= \bar{x}_{k,r} + \bar{f}^{k[m]}_{r-1}(\bar{z},\bar{x},\bar{u}), \\ \dot{\bar{x}}_{k,r} &= \bar{u}_k, \end{cases}$$

where for any $1 \le j \le r-1$, the homogeneous polynomial $\bar{f}^{k[m]}_j(\bar{z},\bar{x},\bar{u})$ is of the form (3.3). Assume that the first $j-1$ components $f^{k[m]}_1(z,x,u), \dots, f^{k[m]}_{j-1}(z,x,u)$ of $\Pi^{k[m]}$ are already in their normal forms and let us focus exclusively on the jth component $f^{k[m]}_j(z,x,u)$.

Since $f^{k[m]}_j(z,x,u) \in \mathfrak{P}^m$, it decomposes uniquely as

$$f^{k[m]}_j(z,x,u) = \bar{f}^{k[m]}_j(z,x,u) + \tilde{f}^{k[m]}_j(z,x,u),$$

where $\bar{f}^{k[m]}_j(z,x,u) \in \mathfrak{F}^m_j$ and $\tilde{f}^{k[m]}_j(z,x,u) \in \mathfrak{E}^m_j$. We may remark that $\tilde{f}^{k[m]}_j(z,x,u)$ is necessarily affine in u; otherwise its projection on $\mathfrak{F}^m_j$ will not be zero.

We may also suppose that $\tilde{f}^{k[m]}_j(z,x,u)$ doesn't depend on u. Indeed, if

$$\tilde{f}^{k[m]}_j(z,x,u) = \sum_{t=1}^{p} u_t R^{k[m-1]}_{j,t}(z,x) + \hat{f}^{k[m]}_j(z,x)$$

it suffices to take the change of variable

$$\bar{x}_{k,j} = x_{k,j} - \sum_{1\le t\le p} \int_0^{x_{t,r}} R^{k[m-1]}_{j,t}(z,x)d\epsilon,$$

to get rid of those terms that depend on u. Of course, in order for the integral to make sense, the variable $x_{t,r}$ of the polynomial $R^{k[m-1]}_{j,t}(z,x)$ will be replaced by the parameter of integration ϵ.

Now, if we assume that $\tilde{f}^{k[m]}_j(z,x,u)$ doesn't depend on u, it suffices to take the change of coordinates

$$\bar{x}_{k,j+1} = x_{k,j+1} + \tilde{f}^{k[m]}_j(z,x)$$

to get the jth component $f_j^{k[m]}(z,x,u)$ into its normal form. As we now see, the procedure didn't modify the previous $j-1$ components because the change of coordinates involves only the variables $x_{k,j}$ and $x_{k,j+1}$, which didn't appear *linearly* in these components. For the same reason, it doesn't modify the other subsystems. This ends the proof of this part.

(ii) (a) The proof of this part will be done by induction. Consider the subsystem

$$\Pi^{[m-1]} \ : \ \dot{z} = Jz + g^{[0]}(z) + g^{[m-1]}(z,x,u) \tag{4.3}$$

with $m \geq 3$ and assume that for some $2 \leq l \leq r$, this system has been transformed so that

$$g^{[m-1]}(z,x,u) = \tilde{g}^{[m-1]}(z,x,u) + \hat{g}^{[m-1]}(z,x,u),$$

where for any $1 \leq j \leq q$ we have

$$\begin{aligned}\tilde{g}_j^{[m-1]}(z,x,u) = &\sum_{1\leq s\leq t\leq p} \sum_{i=l+1}^{r+1} x_{s,i}x_{t,i}P_{j,i,s,t}^{[m-3]}(z,\pi_{t,i}^s(x)) \\ &+ \sum_{1\leq s<t\leq p} \sum_{i=l+1}^{r+1} x_{s,i}x_{t,i-1}Q_{j,i,s,t}^{[m-3]}(z,\pi_{t,i-1}^t(x))\end{aligned}$$

and $\hat{g}^{[m-1]}(z,x,u)$ depends only on the variables z and $x_{s,1},\ldots,x_{s,l}$ for $1 \leq s \leq p$. We emphasize here that $\hat{g}^{[m-1]}(z,x,u)$ doesn't depend on any variable $x_{s,k}$ for $k \geq l+1$.

Each component of $\hat{g}^{[m-1]}(z,x,u)$ could be uniquely decomposed as follows:

$$\begin{aligned}\hat{g}_j^{[m-1]}(z,x,u) = &\sum_{1\leq s\leq t\leq p} x_{s,l}x_{t,l}P_{j,l,s,t}^{[m-3]}(z,\pi_{t,l}^s(x)) \\ &+ \sum_{1\leq s<t\leq p} x_{s,l}x_{t,l-1}Q_{j,l,s,t}^{[m-3]}(z,\pi_{t,l-1}^t(x)) \\ &+ \sum_{1\leq t\leq p} x_{t,l}R_{j,l,t}^{[m-2]}(z,x) + S_{j,l}^{[m-1]}(z,x),\end{aligned}$$

where the polynomials $R_{j,l,t}^{[m-2]}(z,x)$ and $S_{j,l}^{[m-1]}(z,x)$ depend only on the variables z and $x_{s,1},\ldots,x_{s,l-1}$ for $1 \leq s \leq p$.

It then suffices to apply the change of variables given, for any $1 \leq j \leq q$, by

$$\bar{z}_j = z_j - \sum_{1\leq t\leq p} \int_0^{x_{t,l-1}} R_{j,l,t}^{[m-2]}(z,x)d\epsilon,$$

to get rid of the terms $\sum 1 \leq t \leq p x_{t,l}R_{j,l,t}^{[m-2]}(z,x)$. For the need of the integral we replace the variable $x_{t,l-1}$ in $R_{j,l,t}^{[m-2]}(z,x)$ by the parameter of integration ϵ.

This means that we transform the subsystem (4.3) into the form

$$\Pi^{[m-1]} \ : \ \dot{\bar{z}} = J\bar{z} + g^{[0]}(\bar{z}) + g^{[m-1]}(\bar{z},\bar{x},\bar{u})$$

with

$$g^{[m-1]}(\bar{z},\bar{x},\bar{u}) = \tilde{g}^{[m-1]}(\bar{z},\bar{x},\bar{u}) + \hat{g}^{[m-1]}(\bar{z},\bar{x},\bar{u}),$$

where for any $1 \leq j \leq q$ we have

$$\tilde{g}_j^{[m-1]}(\bar{z},\bar{x},\bar{u}) = \sum_{1\leq s\leq t\leq p}\sum_{i=l}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i}P_{j,i,s,t}^{[m-3]}(\bar{z},\pi_{t,i}^{s}(\bar{x})) + \sum_{1\leq s<t\leq p}\sum_{i=l}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i-1}Q_{j,i,s,t}^{[m-3]}(\bar{z},\pi_{t,i-1}^{t}(\bar{x}))$$

and $\hat{g}^{[m-1]}(\bar{z},\bar{x},\bar{u})$ depends only on the variables $\bar{z}$ and $\bar{x}_{s,1},\ldots,\bar{x}_{s,l-1}$ for $1 \leq s \leq p$. This proves the induction argument. If we take $l = 2$, then $\hat{g}^{[m-1]}(\bar{z},\bar{x},\bar{u})$ will depend only on the variables $\bar{z}$ and $\bar{x}_{s,1}$ for $1 \leq s \leq p$, which means that

$$\hat{g}^{[m-1]}(\bar{z},\bar{x},\bar{u}) = \sum_{1\leq s\leq t\leq p} \bar{x}_{s,1}\bar{x}_{t,1}P_{j,1,s,t}^{[m-3]}(z,\pi_{t,1}^{s}(x)).$$

We thus deduce that

$$g_j^{[m-1]}(\bar{z},\bar{x},\bar{u}) = \sum_{1\leq s\leq t\leq p}\sum_{i=1}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i}P_{j,i,s,t}^{[m-3]}(\bar{z},\pi_{t,i}^{s}(\bar{x})) + \sum_{1\leq s<t\leq p}\sum_{i=2}^{r+1} \bar{x}_{s,i}\bar{x}_{t,i-1}Q_{j,i,s,t}^{[m-3]}(\bar{z},\pi_{t,i-1}^{t}(\bar{x}))$$

and this achieves the proof of this part.

(ii)(b) When $m = 2$, the homogeneous vector field $g^{[m-1]}(z,x,u)$ is linear with respect to the variables x and u, that is,

$$g^{[m-1]}(z,x,u) = \sum_{1\leq t\leq p}\sum_{i=1}^{r+1} x_{t,i}P_{i,t}^{[0]}(z),$$

where $P_{i,t}^{[0]}(z) = \big(P_{1,i,t}^{[0]}(z),\ldots,P_{q,i,t}^{[0]}(z)\big)^T$ is a vector field that depends exclusively on the variable z.

The method is to apply first a change of coordinates of the form

$$\tilde{z} = z - \sum_{1\leq t\leq p} x_{t,r}P_{r+1,t}^{[0]}(z)$$

to annihilate the terms

$$\sum_{1\leq t\leq p} x_{t,r+1}P_{r+1,t}^{[0]}(z) = \sum_{1\leq t\leq p} u_t P_{r+1,t}^{[0]}(z).$$

Then, apply a change of coordinates of the form

$$\bar{z} = \tilde{z} - \sum_{1\leq t\leq p} x_{t,r-1}\tilde{P}_{r,t}^{[0]}(\tilde{z})$$

to annihilate the terms

$$\sum_{1\leq t\leq p} x_{t,r}\tilde{P}_{j,r,t}^{[0]}(\tilde{z}),$$

where $\tilde{P}^{[0]}_{j,r,t}(\tilde{z})$ denotes the new terms obtained after the first change of coordinates.

We keep applying this method until we get

$$g^{[1]}(z,x,u) = \sum_{1\le t\le p} x_{t,1} P^{[0]}_{1,t}(z).$$

This completes the proof of item (ii) and that of the theorem. ∎

REFERENCES

[1] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, 2nd ed., Springer-Verlag, New York, 1988.

[2] J.-P. Barbot, I. Belmouhoub, L. Boutat-Baddas, *Observability normal forms*, in New Trends in Nonlinear Dynamics and Control, and Their Applications, W. Kang, M. Xiao, and C. Borges, eds., Springer-Verlag, Berlin, 2003, pp. 3–17.

[3] B. Bonnard, *Feedback equivalence for nonlinear systems and the time optimal control problem*, SIAM J. Control Optim., 29 (1991), pp. 1300–1321.

[4] B. Bonnard, *Quadratic control systems*, Math. Control Signals Systems, 4 (1991), pp. 139–160.

[5] E. Cartan, *La Théorie des Groupes Finis et Continus et la Géométrie Différentielle traitées par la Méthode du Repère Mobile*, Gauthier-Villars, Paris, 1937.

[6] S. Celikovsky and H. Nijmeijer, *Equivalence of nonlinear systems to triangular form: The singular case*, Systems Control Lett., 27 (1996), pp. 135–144.

[7] D. E. Chang, W. Kang, and A. J. Krener, *Normal forms and bifurcations of control systems*, in Proceedings of the 39th IEEE Conference on Decision and Control, Vol. 2, 2000, pp. 1602–1607.

[8] L. Cremean, W. B. Dunbar, D. V. Goh, J. Hickey, E. Kavins, J. Meltzer, and R. Murray, *The Caltech multi-vehicle wireless testbed*, in Proceedings of the 40th IEEE Conference on Decision and Control, Las Vegas, 2002, pp. 86–88.

[9] W. B. Dunbar and R. Murray, *Model predictive control of coordinated multi-vehicle formations*, in Proceedings of the 40th IEEE Conference on Decision and Control, Las Vegas, 2002, pp. 4631–4636.

[10] R. B. Gardner, *The Method of Equivalence and Its Applications*, CBMS-NSF Regional Conf. Ser. Appl. Math., 58, SIAM, Philadelphia, 1989.

[11] R. B. Gardner and W. Shadwick, *The GS algorithm for exact linearization to Brunovský normal form*, IEEE Trans. Automat. Control, 37 (1992), pp. 224–230.

[12] R. B. Gardner, W. F. Shadwick, and G. R. Wilkens, *A geometric isomorphism with applications to closed loop controls*, SIAM J. Control Optim., 27 (1989), pp. 1361–1368.

[13] R. B. Gardner, W. F. Shadwick, and G. R. Wilkens, *Feedback equivalence and symmetries of Brunovský normal forms*, in Dynamics and Control of Multibody Systems, Contemp. Math., 97, J. E. Marsden, P. S. Krishnaprasad, and J. C. Simo, eds., AMS, Providence, RI, 1989, pp. 115–130.

[14] B. Hamzi, *Analysis and stabilization of nonlinear systems with a zero-Hopf control bifurcation*, in Proceedings of the 41st IEEE Conference on Decision and Control, Vol. 4, 2002, pp. 3912–3917.

[15] J. Hauser, S. Sastry, and G. Meyer, *Nonlinear control design for slightly nonminimum phase systems: Application to v/stol aircraft*, Automatica J. IFAC, 28 (1992), pp. 665–679.

[16] L. R. Hunt and R. Su, *Linear equivalents of nonlinear time varying systems*, in Proceedings of the MTNS, Santa Monica, CA, 1981, pp. 119–123.

[17] L. R. Hunt, R. Su, and G. Meyer, *Design for multi-input nonlinear systems*, in Differential Geometric Control Theory, R. Brockett, R. Millman, and H. J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 268–298.

[18] B. Jakubczyk, *Equivalence and invariants of nonlinear control systems*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, New York, Basel, 1990, pp. 177–218.

[19] B. Jakubczyk, *Feedback Invariants, Critical Trajectories, and Hamiltonian Formalism*, in Nonlinear Control in the Year 2000, Vol. 1, A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, eds., Springer, London, 2001, pp. 545–568.

[20] B. Jakubczyk, *Critical Hamiltonians and feedback invariants*, in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Marcel Dekker, New York, Basel, 1998, pp. 219–256.

[21] B. Jakubczyk and W. Respondek, *On linearization of control systems*, Bull. Acad. Polon. Sci. Ser. Sci. Math., 28 (1980), pp. 517–522.
[22] B. Jakubczyk and W. Respondek, *Feedback classification of analytic control systems in the plane*, in Analysis of Controlled Dynamical Systems, B. Bonnard, B. Bride, J. P. Gauthier, and I. Kupka, eds., Birkhäuser, Boston, 1991, pp. 263–273.
[23] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[24] W. Kang, *Extended controller form and invariants of nonlinear control systems with a single input*, J. Math. System. Estim. Control, 4 (1994), pp. 253–256.
[25] W. Kang *Quadratic normal forms of nonlinear control systems with uncontrollable linearization*, in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, Vol. 1, 1995, pp. 608–612.
[26] W. Kang, *Bifurcation and normal form of nonlinear control systems, Part* I, SIAM J. Control Optim., 36 (1998), pp. 193–212.
[27] W. Kang, *Bifurcation and normal form of nonlinear control systems, Part* II, SIAM J. Control Optim., 36 (1998), pp. 213–232.
[28] W. Kang, *Bifurcation control via state feedback for systems with a single uncontrollable mode*, SIAM J. Control Optim., 38 (2000), pp. 1428–1452.
[29] W. Kang and A. J. Krener, *Extended quadratic controller normal form and dynamic state feedback linearization of nonlinear systems*, SIAM J. Control Optim., 30 (1992), pp. 1319–1337.
[30] W. Kang, M. Xiao, J. Mao, and I. A. Tall, *Controllability and local accessibility—A normal form approach*, IEEE Trans. Automat. Control, 48 (2003), pp. 1724–1736.
[31] A. J. Krener, *Approximate linearization by state feedback and coordinate change*, Systems Control Lett., 5 (1984), pp. 181–185.
[32] A. J. Krener, W. Kang, and D. E. Chang, *Normal forms of linearly uncontrollable nonlinear control systems with single-input*, in Proceedings NOLCOS'01, St. Petersburg, Russia, 2001, pp. 134–139.
[33] A. J. Krener and M. Xiao, *Nonlinear observer design in the Siegel domain*, SIAM J. Control Optim., 41 (2002), pp. 932–953.
[34] A. J. Krener and M. Xiao, *Erratum: Nonlinear observer design*, SIAM J. Control Optim., 43 (2004), pp. 377–378.
[35] I. Kupka, *On feedback equivalence*, CMS Conf. Proc., 12 (1990), pp. 105–117.
[36] W. Respondek, *Feedback classification of nonlinear control systems in* $\mathbb{R}^2$ *and* $\mathbb{R}^3$, in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Marcel Dekker, New York, 1998, pp. 347–381.
[37] W. Respondek, *Symmetries and minimal flat outputs of nonlinear control systems*, in New Trends in Nonlinear Dynamics and Control, and Their Applications, W. Kang, M. Xiao, and C. Borges, eds., Springer-Verlag, Berlin, 2003.
[38] W. Respondek and I. A. Tall, *How many symmetries does admit a nonlinear single-input control system around equilibrium*, in Proceedings of the 40th IEEE Conference on Decision and Control, FL, 2000, pp. 1795–1800.
[39] W. Respondek and M. Zhitomirskii, *Feedback classification of nonlinear control systems on 3-manifolds*, Math. Control Signals Systems, 8 (1995), pp. 299–333.
[40] S. Sastry, Mathematics of Control, Signals, and Systems, 8 (1995), pp. 299–333.
[41] I. A. Tall, *Classification par bouclage des systèmes de contrôles non linéaires mono-entrée: Formes normales, formes canoniques, invariants et symétries*, Ph.D. thesis, INSA de Rouen, Rouen, France, 2000.
[42] I. A. Tall, *Normal forms of multi-input nonlinear control systems with controllable linearization*, in New Trends in Nonlinear Dynamics and Control, and Their Applications, W. Kang, M. Xiao, and C. Borges, eds., Springer-Verlag, Berlin, 2003, pp. 87–100.
[43] I. A. Tall and W. Respondek, *Normal forms, canonical forms, and invariants of single-input control systems under feedback*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, 2000, pp. 1625–1630.
[44] I. A. Tall and W. Respondek, *Normal forms and invariants of nonlinear single-input systems with noncontrollable linearization*, in Proceedings NOLCOS'01, St. Petersburg, Russia, 2001, pp. 122–127.
[45] I. A. Tall and W. Respondek, *Transforming a single-input nonlinear system to a strict feedforward form via feedback*, in Nonlinear Control in the Year 2000, Vol. 2, A. Isidori, F. Lamnabhi, and W. Respondek, eds., Springer-Verlag, London, 2001, pp. 527–542.
[46] I. A. Tall and W. Respondek, *Feedback equivalence to feedforward forms for nonlinear single-input systems*, in Dynamics, Bifurcations and Control, F. Colonius and L. Grune, eds., Springer-Verlag, Berlin, 2002, pp. 269–286.

[47] I. A. Tall and W. Respondek, *Normal forms of two-inputs nonlinear control systems*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, 2002, pp. 2732–2737.

[48] I. A. Tall and W. Respondek, *Nonlinearizable single-input control systems do not admit stationary symmetries*, Systems Control Lett., 46 (2002), pp. 1–16.

[49] I. A. Tall and W. Respondek, *Feedback classification of nonlinear single-input control systems with controllable linearization: Normal forms, canonical forms, and invariants*, SIAM J. Control Optim., 41 (2002), pp. 1498–1531.

[50] I. A. Tall and W. Respondek, *Feedback equivalence to a strict feedforward form for nonlinear single-input systems*, Internat. J. Control, to appear.

[51] M. Zhitomirskii and W. Respondek, *Simple germs of corank one affine distributions*, in Singularities Symposium—Lojasiewicz 70, B. Jakubczyk, W. Pawlucki, and J. Stasica, eds., Banach Center Publ. 44, Polish Academy of Science, Warsaw, 1998, pp. 269–276.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2071–2088

AN OUTER APPROXIMATION METHOD FOR THE VARIATIONAL INEQUALITY PROBLEM*

R. S. BURACHIK†, J. O. LOPES†, AND B. F. SVAITER‡

Abstract. We study two outer approximation schemes, applied to the variational inequality problem in reflexive Banach spaces. First we propose a generic outer approximation scheme, and its convergence analysis unifies a wide class of outer approximation methods applied to the constrained optimization problem. As is standard in this setting, boundedness and optimality of weak limit points are proved to hold under two alternative conditions: (i) boundedness of the feasible set, or (ii) coerciveness of the operator. To develop a convergence analysis where (i) and (ii) do not hold, we consider a second scheme in which the approximated subproblems use a coercive approximation of the original operator. Under conditions alternative to both (i) and (ii), we obtain standard convergence results. Furthermore, when the space is uniformly convex, we establish full strong convergence of the second scheme to a solution.

Key words. maximal monotone operators, Banach spaces, outer approximation algorithm, semi-infinite programs

AMS subject classifications. 49M27, 65J05, 65K05, 90C25

DOI. 10.1137/S0363012902415487

1. Introduction.

We investigate a broad class of outer approximation methods for solving the classical monotone variational inequality problem in a reflexive Banach space. First we recall this problem and then we describe those methods. Let B be a real reflexive Banach space with dual B^*. The notation $\langle v, x\rangle$ stands for the duality product $v(x)$ of $v \in B^*$ and $x \in B$. Given $T\colon B \rightrightarrows B^*$ a maximal monotone operator and $\Omega \subset B$ a nonempty closed and convex set, the *variational inequality problem* for T and Ω, $VIP(T, \Omega)$ is as follows. Find x^* such that

$$x^* \in \Omega,\ \exists\ u^* \in T(x^*):\ \ \langle u^*, x - x^*\rangle \geq 0\ \forall\ x \in \Omega. \tag{1.1}$$

The set Ω will be called the *feasible set* for problem (1.1). In the particular case in which T is the subdifferential of a proper, convex, and lower semicontinuous function $f\colon B \longrightarrow \mathbb{R} \cup \{+\infty\}$, problem (1.1) reduces to the *convex optimization problem*:

$$\min_{x\in\Omega} f(x). \tag{1.2}$$

Outer approximation methods solve problem (1.1) by generating and solving a sequence of problems with feasible sets Ω^k which contain the original feasible set Ω but have a simpler structure. These methods were introduced for solving optimization problems four decades ago in the form of cutting plane methods [10, 32].

*Received by the editors October 1, 2002; accepted for publication (in revised form) August 6, 2004; published electronically April 14, 2005.

http://www.siam.org/journals/sicon/43-6/41548.html

†Engenharia de Sistemas e Computação, COPPE-UFRJ, CP 68511, Rio de Janeiro, RJ, CEP 21941-972, Brazil (regi@cos.ufrj.br, jurandir@cos.ufrj.br). The research of the first author was supported by CAPES grant BEX 0664-02/2. The research of the second author was partially supported by PICDT/UFPI-CAPES.

‡Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ, CEP 22460-320, Brazil (benar@impa.br). The research of this author was partially supported by CNPq grant 301200/93-9(RN) and PRONEX–Optimization.

Outer approximation schemes typically arise when the set Ω is of the form $\Omega = \cap_{y\in Y}\Omega_y$, where Y is infinite and $\Omega_y := \{x \in B \mid g(x,y) \leq 0\}$ with each $g(\cdot,y) : B \to \mathbb{R}$. Feasible sets of this kind appear in several areas of applications (see, e.g., [23, 22, 17, 13, 3, 19]). In this situation, a common choice is to replace the feasible set of the original problem by $\Omega^k := \cap_{y\in Y^k}\Omega_y = \{x \in B \mid g(x,y) \leq 0\, \forall\, y \in Y^k\}$, where $Y^k \subset Y$ is finite and conveniently chosen.

For this kind of Ω, outer approximation methods for problem (1.2) are classified according to the way in which the sets $\Omega^k \supset \Omega = \cap_{y\in Y}\Omega_y$ are defined. We mention here *cutoff* methods (e.g., those in [10, 28, 32, 45, 48] for $B = \mathbb{R}^n$), *filtered* cutoff methods (e.g., those in [2, 14, 16, 43, 44] for $B = \mathbb{R}^n$), and *disintegration schemes* [20], like the ones proposed in [12, 21, 36, 38, 39] for the minimization of quadratic functions in Hilbert spaces and extended in [33] to the minimization of a convex function in a Banach space. Recently, Combettes gave in [11] a unified convergence analysis which includes and extends all the above-mentioned outer approximation methods. A basic assumption for obtaining convergence of these methods is that either all sets Ω^k are contained in some bounded set or some coerciveness property of the objective function f. These *boundedness* assumptions are a standard requirement in the analysis of all the mentioned methods. (See the excellent surveys [40, 24] and references therein.) The goal for the present work is twofold. First, we develop a convergence analysis which can be applied to more general and flexible schemes for successive approximation of variational inequalities, under the standard boundedness assumptions. Our analysis covers as a particular case the outer approximation scheme studied in [11] for problem (1.2) (and hence all the above-mentioned algorithms). We prove that our generic scheme generates a bounded sequence and that all weak accumulation points are solutions of $VIP(T,\Omega)$. Second, we obtain the same convergence results in the absence of boundedness assumptions. For doing this, we consider subproblems (P_k), where the original operator is replaced by a suitable coercive regularization. Our work is built around the following generic outer approximation scheme for solving $VIP(T,\Omega)$.

ALGORITHM.

Initialization. Take $\Omega^1 \supseteq \Omega$,

Iterations. For $k = 1, 2, \ldots$, find $x^k \in \Omega^k$, a solution of the approximated problem (P_k), defined as

$$\exists\, u^k \in T(x^k) \quad \text{with} \tag{1.3}$$

$$\langle u^k, x - x^k\rangle \geq 0 \forall\, x \in \Omega^k. \tag{1.4}$$

In our first generic scheme we relax the inequality in (1.4). Namely, the iterate $x^k \in \Omega^k$ is taken such that

$$\exists\, u^k \in T(x^k) \;\text{ with } \langle u^k, x - x^k\rangle \geq -\varepsilon_k \;\forall\; x \in \Omega^k, \tag{1.5}$$

where $\varepsilon_k > 0$ and $\Omega^k \supset \Omega$ is convex and closed. In the second scheme, we relax (1.3). More precisely, the approximate solution $x^k \in \Omega^k$ is such that

$$\exists\, u^k \in T_{\lambda_k}(x^k) \;\text{ with } \langle u^k, x - x^k\rangle \geq 0 \;\forall\; x \in \Omega^k, \tag{1.6}$$

where $\lambda_k > 0$, $\Omega^k \supset \Omega$ is closed and convex and T_{λ_k} is a suitable coercive approximation of T. Schemes in which the approximated subproblems use a coercive regularization of T are a common approach for solving noncoercive variational inequalities. Two classical examples of these are *proximal-like* regularizations (see, e.g., [29, 30])

and Tikhonov regularizations [42, 18]. The latter kind of regularization has been extensively studied in the last two decades (see [34] and the references therein). Mosco [35] studied the convergence of what is now called the Mosco scheme, which combines the Tikhonov regularization with a perturbation of the feasible set. This approach is followed in [34], where, as in [35, section 5], the approximating feasible sets are assumed to converge in the sense of set-convergence to the original feasible set. In [34], the authors studied variational inequalities in Hilbert spaces and the operator T is assumed to be point-to-point.

Classical application of all these approximating schemes is in perturbation theory of variational boundary value problems for the operator T. (See, e.g., problems (p') and (p'_n) and Corollary 1 in [35, pp. 555–556].)

The paper is organized as follows. Section 2 contains some theoretical preliminaries which are necessary for our analysis. In section 3 we give a unified analysis for a broad family of outer approximation algorithms, in which the iterates solve problem (1.5). We prove existence of this sequence and establish optimality of all weak accumulation points under standard boundedness assumptions. In section 4 we relax the boundedness assumptions and consider a sequence $\{x^k\}$ as in (1.6). Under suitable assumptions, we prove that the iterates are bounded and all its accumulation points are optimal. Moreover, we establish strong convergence of the whole sequence to a solution when B is uniformly convex.

2. Theoretical preliminaries. From now on B is a real Banach space. Let $T\colon B \rightrightarrows B^*$ be an arbitrary point-to-set operator. We recall some basic definitions:

- domain of T, $D(T) := \{x \in B \,|\, T(x) \neq \emptyset\}$;
- graph of T, $G(T) := \{(x,u) \in B \times B^* \,|\, u \in T(x)\}$;
- range of T, $R(T) := \{u \in B^* \,|\, u \in T(x) \text{ for some } x \in B\}$;
- the operator T is *monotone* if for all $x, y \in B$, $u \in T(x)$, and $v \in T(y)$,

$$\langle u - v, x - y\rangle \geq 0;$$

  if this inequality holds strictly whenever $x, y \in B$, $u \in T(x)$, $v \in T(y)$, and $x \neq y$, then T is *strictly monotone*;
- the operator T is *maximal monotone* if it is monotone and for any monotone $\tilde{T}\colon B \rightrightarrows B^*$, $G(T) \subset G(\tilde{T}) \Rightarrow T = \tilde{T}$.

An example of maximal monotone operator is the *normality operator*. Let $\Omega \subset B$ be closed, convex, and nonempty.

- The normality operator of Ω is $N_\Omega\colon B \rightrightarrows B^*$,

$$N_\Omega(x) = \begin{cases} \{u \in B^* | \langle u, z - x\rangle \leq\ 0\ \forall z \in \Omega\} & \text{if } x \in \Omega, \\ \emptyset & \text{otherwise.} \end{cases}$$

Maximal monotonicity of N_Ω follows from Ω being closed, convex, and nonempty.

It is easy to verify that $VIP(T, \Omega)$, as defined in (1.1), is equivalent to the inclusion problem (or generalized equation): Find x^* such that

$$0 \in (T + N_\Omega)(x^*).$$

If $T + N_\Omega$ is onto, then this problem, and hence $VIP(T, \Omega)$, has a solution. In this paper, existence of solutions of variational inequality problems will be based on surjectivity of sums of maximal monotone operators. The following definitions will be needed.

DEFINITION 2.1 ((see [37])). *An operator $T: B \rightrightarrows B^*$ is*

1. coercive *if $D(T)$ is bounded or for any $x \in D(T)$,*

$$\lim_{\|z\| \to +\infty} \langle v, z - x\rangle / \|z\| = +\infty \text{ holds for each selection } v \in T(z);$$

2. regular *if for any $y \in D(T)$ and $u \in R(T)$*

$$\sup_{(z,v) \in G(T)} \langle v - u, y - z\rangle < \infty.$$

The next proposition will be used for establishing existence of solutions of the subproblems (P_k).

PROPOSITION 2.2 (see [8, Lemma 2.7]). *Suppose that B is reflexive. Let $T_1, T_2: B \rightrightarrows B^*$ be maximal monotone operators such that*

(a) *$T_1 + T_2$ is maximal monotone;*

(b) *T_1 is regular and onto. Then $T_1 + T_2$ is onto.*

To establish condition (a) of the above proposition we will use a classical theorem, due to Rockafellar [41]. Denote by $\text{int}(A)$ the topological interior of the set A.

PROPOSITION 2.3 (see [41, Theorem 1]). *Suppose that B is reflexive. Let $T_1, T_2: B \rightrightarrows B^*$ maximal monotone operators . If $D(T_1) \cap \text{int}(D(T_2)) \neq \emptyset$, then $T_1 + T_2$ is a maximal monotone operator.*

To check for condition (b) of Proposition 2.2, we will need two auxiliary results.

THEOREM 2.4 (see [4, p. 147]). *Suppose that B is reflexive. Let $T: B \rightrightarrows B^*$ be a maximal monotone operator. If T is coercive (in particular, if $D(T)$ is bounded), then T is surjective.*

The fact stated next relates the two concepts given in Definition 2.1.

THEOREM 2.5 (see [37, p. 122]). *Let $T: B \rightrightarrows B^*$ be a monotone operator. If T is coercive, then T is regular.*

The discussion of the maximality of monotone operators and their surjective properties requires the introduction of duality maps. Asplund [1] showed that when B is a reflexive Banach space, there exists an equivalent norm on B which is everywhere Gâteaux differentiable except at the origin and such that the corresponding dual norm on B^* is also everywhere Gâteaux differentiable except at the origin.

From now on, we assume that B is a *reflexive* real Banach space. For simplifying the notation, we also assume that the given norm on B already has these special properties. We use the same notation $\|\cdot\|$ for this norm on B and its associated norm on the dual B^*. Denote by J the Gâteaux gradient of the function $\varphi(x) := (1/2)\|x\|^2$. Thus, J is the *duality mapping*, which assigns to each $x \in B$ the unique $J(x) \in B^*$ such that

$$\langle x, J(x)\rangle = \|x\|^2 = \|J(x)\|^2. \tag{2.1}$$

PROPOSITION 2.6. *Let $J: B \to B^*$ be the duality mapping described above. The following assertions hold:*

(i) *$J(-x) = -J(x)$ and $J(\lambda x) = \lambda J(x)$ $\forall \lambda > 0$.*

(ii) *Let $w = J(x)$; then $\langle w, z - x\rangle \leq \frac{1}{2}(\|z\|^2 - \|x\|^2)$ $\forall z \in B$.*

(iii) *If $\widetilde{T}$ is maximal monotone, then $\forall \lambda > 0$, $\widetilde{T} + \lambda J$ is maximal monotone and onto, and $(\widetilde{T} + \lambda J)^{-1}$ is a single-valued and maximal monotone operator.*

Proof. Item (i) follows from (2.1), and (ii) uses the fact that $J(\cdot) = \partial(\frac{1}{2}\|\cdot\|^2)$. To prove (iii), note that Proposition 2.3 implies maximal monotonicity of $\widetilde{T} + \lambda J$. The

surjectivity of $\widetilde{T} + \lambda J$ and the assertion on $(\widetilde{T} + \lambda J)^{-1}$ follow from [41, Proposition 1]. □

Our convergence theorems require two conditions on the operator T, namely, para- and pseudomotonicity, which we discuss next. The notion of paramonotonicity was introduced in [7] and further studied in [9, 26]. It is defined as follows.

DEFINITION 2.7. *The operator T is* paramonotone *in Ω if it is monotone and $\langle v-u, y-z\rangle = 0$ with $y, z \in \Omega$, $v \in T(y), u \in T(z)$ implies that $u \in T(y), v \in T(z)$. The operator T is* paramonotone *if this property holds in the whole space.*

PROPOSITION 2.8 ((see [26, Proposition 4])). *Assume that T is paramonotone on Ω and $\overline{x}$ is a solution of $VIP(T, \Omega)$. Let $x^* \in \Omega$ be such that there exists an element $u^* \in T(x^*)$ with $\langle u^*, x^* - \overline{x}\rangle \leq 0$. Then x^* also solves $VIP(T, \Omega)$.* Paramonotonicity can be seen a condition which is weaker than strict monotonicity. The remark below contains some examples of operators which are paramonotone.

Remark 2.9. If T is the subdifferential of a convex function $f : B \to \mathbb{R} \cup \{\infty\}$, then T is paramonotone. When $B = \mathbb{R}^n$, a condition which guarantees paramonotonicity of $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is when T is differentiable and the symmetrization of its Jacobian matrix has the same rank as the Jacobian matrix itself. However, relevant operators fail to satisfy this condition. More precisely, the *saddle-point operator* $\mathbf{\Lambda}(x, y) := (\partial_x L(x, y), -\partial_y L(x, y))$, where L is the Lagrangian associated to a constrained convex optimization problem, is not paramonotone, except in trivial instances. For more details on paramonotone operators see [26].

Next we recall the definition of pseudomonotonicity, which was taken from [6] and should not be confused with other uses of the same word (see, e.g., [31]).

DEFINITION 2.10. *Let B be a reflexive Banach space and the operator T such that $D(T)$ is closed and convex. T is said to be* pseudomonotone *if it satisfies the following condition. If the sequence $\{(x^k, u^k)\} \subset G(T)$ satisfies that*

(a) *$\{x^k\}$ converges weakly to $x^* \in D(T)$,*

(b) *$\limsup_k \langle u^k, x^k - x^*\rangle \leq 0$,*

then for every $w \in D(T)$ there exists an element $u^* \in T(x^*)$ such that

$$\langle u^*, x^* - w\rangle \leq \liminf_k \langle u^k, x^k - w\rangle.$$

Remark 2.11. If T is the gradient of a Gâteaux differentiable convex function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, then T is pseudomonotone. Indeed, using a fact proved in [37, p. 94], it holds that $\nabla\varphi$ is *hemicontinuous* (i.e., for every fixed $x, y \in \mathbb{R}^n$, the real-valued mapping $t \mapsto \langle \nabla\varphi((1-t)x + ty), x - y\rangle$ is continuous). On the other hand, point-to-point hemicontinuous operators defined on $\mathbb{R}^n$ are always pseudomonotone (see e.g., [37, p. 107]), which yields $T = \nabla\varphi$ pseudomonotone, as claimed. Combining the latter statement with Remark 2.9, we conclude that every T of this kind is both para- and pseudomonotone. An example of a nonstrictly monotone operator which is both para- and pseudomonotone is the subdifferential of the function $\varphi : \mathbb{R} \to \mathbb{R}$ defined by $\varphi(t) = |t|$ for all t.

DEFINITION 2.12 (see [5, p. 51]). *We say that B is* uniformly convex *if $\forall\ \varepsilon > 0\ \exists\ \delta > 0$ such that*

$$\forall\ x, y \in B, \|x\| \leq 1, \|y\| \leq 1 \text{ and } \|x - y\| > \varepsilon \ \Rightarrow\ \left\|\frac{x+y}{2}\right\| < 1 - \delta.$$

PROPOSITION 2.13 (see [5, Proposition III.30]). *Assume that B is uniformly convex and that $\{y^k\}$ is a sequence converging weakly to y^*. Suppose further that $\limsup \|y^k\| \leq \|y^*\|$; then $\{y^k\}$ converges strongly to y^*.*

3. A general outer approximation scheme for $VIP(T,\Omega)$. Let $\Omega \subset B$ be a nonempty closed and convex set and let $T\colon B \rightrightarrows B^*$ be a maximal monotone operator. In this section we present a unified convergence analysis of a general and flexible scheme for successive approximation of variational inequalities. Recall that $VIP(T,\Omega)$ is defined by

$$\begin{aligned} &\text{find } x^* \in \Omega \text{ such that there exists } u^* \in T(x^*) \text{ with} \\ &\langle u^*, x - x^* \rangle \geq 0 \ \forall\, x \in \Omega. \end{aligned} \tag{3.1}$$

To present a convergence analysis which can be applied to a wide family of outer approximation schemes, we fix a sequence $\{\Omega^k\}$ of convex closed subsets of B and a sequence $\{\varepsilon_k\} \subset \mathbb{R}_+ := \{t \in \mathbb{R} \,:\, t \geq 0\}$ verifying

(i) $\Omega \subset \Omega^k$ for all k,

(ii) $\lim_k \varepsilon_k = 0$.

These sequences define our approximating problems. Namely, given Ω^k and ε_k, define the *kth approximating problem* as

$$(P_k) \qquad \begin{cases} \text{find } x^k \in \Omega^k \text{ such that there exists } u^k \in T(x^k) \text{ with} \\ \langle u^k, x - x^k \rangle \geq -\varepsilon_k \ \ \forall\, x \in \Omega^k. \end{cases}$$

Definition 3.1. *Fix $\{\Omega^k\}$ and $\{\varepsilon_k\}$ as in* (i) *and* (ii).

(a) *A sequence $\{x^k\}$ will be called an* orbit *when x^k solves (P_k) for all k.*

(b) *An orbit $\{x^k\}$ will be called* asymptotically feasible *(AF, for short) when all weak accumulation points of $\{x^k\}$ belong to Ω.*

Example 1. A broad family of outer approximation methods for the convex constrained optimization problem (1.2) generates AF orbits. This is shown in [11], where the author provides a unified convergence analysis for a wide class of outer approximation schemes devised for problem (1.2). The feasible set considered in [11] is explicitly defined in the form

$$\Omega := \{x \in B \mid g(x,y) \leq 0 \ \forall\, y \in Y\}, \tag{3.2}$$

where Y is an arbitrary set of indexes and the constraint functions $g(\cdot,y)\colon B \longrightarrow \mathbb{R} \cup \{+\infty\}$ satisfy the basic assumptions:

(G_1) $\{x \in B \mid g(x,y) \leq 0\}$ is nonempty and convex for all $y \in Y$.

(G_2) For all $y \in Y$ and for very sequence $\{z^k\} \subset B$ such that

$$\begin{cases} w - \lim_k z^k = z \\ \limsup_k g(z^k, y) \leq 0, \end{cases} \quad \Longrightarrow \quad g(z,y) \leq 0,$$

where $w-\lim$ stands for weak limit in B. In [11] $T = \partial f$, where f is lower semicontinuous, convex, and verifies the following assumptions:

(F_1) For some closed convex set $E \supset \Omega$, there exists a point $u \in \mathrm{dom} f \cap \Omega$ such that the set $C := \{x \in E \mid f(x) \leq f(u)\}$ is bounded.

(F_2) f is *uniformly convex with modulus of convexity c* on C, i.e., [46, 47]

$$\text{for all } x,y \in C, \quad f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2} - c\left(\|x-y\|\right), \tag{3.3}$$

where $c : \mathbb{R}_+ \to \mathbb{R}_+$ is nondecreasing and $(\forall\, \tau \in \mathbb{R})$ $c(\tau) = 0 \Leftrightarrow \tau = 0$. The outer approximation scheme studied in [11, Algorithm 1.1] has for kth approximating problem the minimization of f in a set $\Omega^k \supset \Omega$. Hence this (P_k) verifies conditions

(i) and (ii), where $T = \partial f$ and $\varepsilon_k = 0$ for all k. As a conclusion, [11, Algorithm 1.1] also generates an orbit in the sense of Definition 3.1(a). Therefore, our approach contains the methods extended by [11, Algorithm 1.1]. The analysis presented in [11] establishes conditions under which [11, Algorithm 1.1] generates an AF orbit. However, assumptions (F_1) and (F_2) force every AF orbit to converge strongly to the unique solution of (1.2). See also [40, Theorem 4.1], where other methods (from the family of cutting plane methods [10] for the convex semi-infinite programming problem) are proved to generate AF orbits. When the objective and constraint functions are smooth, a family of schemes which generates AF orbits is found in [24, Theorem 7.2]. Throughout our work we consider the following assumptions:

(H_1) $D(T) \cap \operatorname{int}(\Omega) \neq \emptyset$ or $\operatorname{int}(D(T)) \cap \Omega \neq \emptyset$.

(H_2) T paramonotone and pseudomonotone with closed domain.

(H_3) The solution set S^* of $VIP(T, \Omega)$ is nonempty.

For the case in which the feasible set Ω is given as in (3.2), we will assume that the set Y and the function $g\colon B \times Y \longrightarrow \mathbb{R} \cup \{+\infty\}$ satisfy the following assumptions:

(G_1) Y is a weakly compact set contained in a reflexive Banach space.

(G_2) $g(\cdot, y)$ is a proper, lower semicontinuous and convex function $\forall y \in Y$.

(G_3) g is weakly continuous on $B \times Y$.

A relevant question regarding AF orbits is which extra conditions guarantee optimality of all weak accumulation points. In our analysis, we use the assumption of para- and pseudomonotonicity (see Remarks 2.9 and 2.11).

LEMMA 3.2. *Let $\{x^k\}$ be an* AF *orbit for $VIP(T, \Omega)$. If (H_2) and (H_3) hold, then every weak accumulation point of $\{x^k\}$ is a solution of $VIP(T, \Omega)$.*

Proof. Assume that x^* is a weak accumulation point of $\{x^k\}$. Then, there exists a subsequence $\{x^{k_j}\}$ converging (weakly) to x^*. For each j, x^{k_j} solves (P_{k_j}), and therefore, there exists $u^{k_j} \in T(x^{k_j})$ such that

$$\langle u^{k_j}, x^{k_j} - x \rangle \leq \varepsilon_{k_j} \ \forall x \in \Omega^{k_j} \text{ and } \forall k_j.$$

Then by (i) we have

$$\langle u^{k_j}, x^{k_j} - x \rangle \leq \varepsilon_{k_j} \ \forall x \in \Omega \text{ and } \forall k_j. \tag{3.4}$$

Since $\{x^k\}$ is AF, $x^* \in \Omega$ and hence

$$\langle u^{k_j}, x^{k_j} - x^* \rangle \leq \varepsilon_{k_j} \quad \forall k_j.$$

Using also (ii) we have

$$\limsup_j \langle u^{k_j}, x^{k_j} - x^* \rangle \leq \limsup_j \varepsilon_{k_j} = 0. \tag{3.5}$$

Take $\overline{x} \in S^*$. By pseudomonotonicity of T, we conclude that there exists $u^* \in T(x^*)$ such that

$$\liminf_j \langle u^{k_j}, x^{k_j} - \overline{x} \rangle \geq \langle u^*, x^* - \overline{x} \rangle.$$

Since $\overline{x} \in \Omega$, (3.4) implies that

$$\liminf_j \langle u^{k_j}, x^{k_j} - \overline{x} \rangle \leq \liminf_j \varepsilon_{k_j} = 0.$$

Combining the last two inequalities we have that

$$\langle u^*, x^* - \overline{x} \rangle \leq 0.$$

Finally, by paramonotonicity of T and Proposition 2.8 we conclude that x^* is a solution of the $VIP(T,\Omega)$. ∎

We saw in Example 1 that many well-known outer approximation schemes for the convex constrained problem solve the subproblems (P_k) exactly, i.e., with $\varepsilon_k = 0$ for all k. Thus a relevant question is whether problem (P_k) admits an *exact* solution.

PROPOSITION 3.3. *Assume (H_1) holds and suppose that one of the following assumptions holds:*

(a) *There is a bounded set K such that $K \supset \Omega^k$ for all k.*

(b) *T is coercive.*

Then every (P_k) admits an exact solution.

Proof. Define the operator $T_k{=}T + N_{\Omega^k}$. For all k it holds that

(1) $D(T_k) = D(T) \cap D(N_{\Omega^k}) = D(T) \cap \Omega^k \neq \emptyset$ by (H_1) and (i).

(2) $T_k = T + N_{\Omega^k}$ is maximal monotone by (H_1) and Proposition 2.3. If (a) holds, then by (1) and (i), $D(T_k)$ is bounded; thus T_k is onto by Theorem 2.4. This implies that (P_k) admits an exact solution in this case. Now suppose that T is coercive. In this case, T is regular by Theorem 2.5. On the other hand, $R(T) = B^*$ by Theorem 2.4. Since (2) holds, it follows from Proposition 2.2 that T_k is onto. Thus (P_k) has a solution. ∎

Now we are in position to present our first convergence result. We point out that this result is an extension to $VIP(T,\Omega)$ of, e.g., [2, Theorem 2.1], [11, Proposition 3.1(i)], [24, Theorem 7.2], and [40, Theorem 4.1].

THEOREM 3.4. *Let the sequence $\{x^k\}$ be an* AF *orbit. Assume that (H_2) and (H_3) hold. If one of the conditions*

(a) *there exists $\overline{\Omega}$ a bounded set such that $\overline{\Omega} \supset \Omega^k$ for all k, or*

(b) *T is coercive*

holds, then $\{x^k\}$ is bounded and each accumulation point is a solution of $VIP(T,\Omega)$.

Proof. By Lemma 3.2, it is enough to establish boundedness of $\{x^k\}$. Assume (a) holds. By (i), we have that $\{x^k\} \subset \overline{\Omega}$, and hence the sequence is bounded. Assume now that (b) occurs, and suppose that $\{x^k\}$ is unbounded. By definition of (P_k), there exists $u^k \in Tx^k$ with

$$\lim_k \langle u^k, x^k - x\rangle / \|x^k\| \leq \lim_k \varepsilon_k / \|x^k\| = 0,$$

where we used unboundedness of $\{x^k\}$ and condition (ii). However, the above expression contradicts the coercivity of T. Hence $\{x^k\}$ is bounded. ∎

Example 2. Consider now problem $VIP(T,\Omega)$, where the feasible set Ω is given as in (3.2). We describe next an outer approximation scheme for this problem. Assume that the set Y and the function $g\colon B \times Y \longrightarrow \mathbb{R} \cup \{+\infty\}$ satisfy assumptions (G_1) and (G_3).

Before stating the algorithm, we need some notation:

- Y^k is a finite subset of Y.
- $\Omega^k := \{x \in B |\ g(x,y) \leq 0\ \forall\ y \in Y^k\}$.
- Given x^k solution of (P_k), define the *kth auxiliary problem* as

$$(A^k)\ \begin{cases} \text{find } y^{k+1} \in Y \text{ such that} \\ y^{k+1} \in \arg\max_{y \in Y}\ g(x^k, y). \end{cases}$$

ALGORITHM 1.
Step 0. Initialize: *Set $k = 1$, and choose any $Y^1 \subset Y$ finite and nonempty.*
Iteration: *For $k = 1, 2, \ldots,$*
Step 1. *Given Ω^k, find x^k solution of (P_k).*
Step 2. *For x^k obtained in Step 1, solve A^k.*
Step 3. (Check for solution and update if necessary.)
If $g(x^k, y^{k+1}) \leq 0$ stop. Otherwise, set

$$Y^{k+1} := Y^k \cup \{y^{k+1}\}. \tag{3.6}$$

Step 4. *Set $k := k + 1$ and return to Step 1.*

If Algorithm 1 stops at step 3, then $x^k \in S^*$. Indeed, if the solution y^{k+1} of the *kth auxiliary problem* A^k obtained in Step 2 satisfies

$$g(x^k, y^{k+1}) \leq 0,$$

then it holds that $g(x^k, y) \leq 0$ for all $y \in Y$, i.e., $x^k \in \Omega$. Thus,

$$\langle u^k, x - x^k \rangle \geq 0 \ \forall \ x \in \Omega,$$

so that x^k is a solution of problem (1.1). This justifies the stopping rule of Step 3. In the particular case in which B is finite dimensional, $T = \partial f$, and $\varepsilon_k = 0$ for all k in problem (P_k), the scheme above is the simplest exchange method [24, section 7.1] for the numerical solution of semi-infinite optimization problems. The analysis for this particular case was developed in [2, Theorem 2.1], where the authors proved, under the hypothesis of boundedness of Ω^1, that the orbit generated by Algorithm 1 is AF and every accumulation point of the orbit is a solution. The proof of the asymptotic feasibility of the orbit is omitted here, since it can be transferred, in a straightforward way, to our more general setting. Optimality of the weak accumulation points also holds in our setting, under the conditions of Theorem 3.4.

Remark 3.5. Under the assumptions $(G_1) - (G_3)$, define $h(x) := \max_{y \in Y} g(x, y)$. Then h is convex and weakly continuous. Therefore, if the conditions of Theorem 3.4 are met, then for every $\beta > 0$ and every orbit $\{x^k\}$ generated by Algorithm 1 in Example 2 there exists k_0 such that $h(x^k) \leq \beta$ for all $k \geq k_0$. Indeed, suppose that there exists $\beta_0 > 0$ and a subsequence $\{x^{k_j}\} \subset \{x^k\}$ such that $h(x^{k_j}) \geq \beta_0$ for all j. By Theorem 3.4, the subsequence $\{x^{k_j}\}$ is bounded and AF. Without loss of generality, assume the whole subsequence $\{x^{k_j}\}$ converges weakly to some $\bar{x} \in \Omega$. Using also the assumption on β_0, we have that

$$\beta_0 \leq \lim_j h(x^{k_j}) = h(\bar{x}) \leq 0,$$

a contradiction. This fact will be useful later on, to define suitable approximated solutions for this particular instance of $VIP(T, \Omega)$ in the case in which boundedness assumptions do not hold.

4. Convergence without boundedness assumptions. In the convergence analysis of the previous section, existence and boundedness of the iterates are proved under the boundedness assumptions of Theorem 3.4. Our aim in this section is to define outer approximating schemes with the same convergence properties, but under alternative assumptions. To achieve this goal, we define subproblems using a Tikhonov regularization of T. As is standard in this kind of regularization, we will force the

parameters to go to zero, to establish convergence of the method. We will consider outer approximations Ω^k for an arbitrary closed and convex set Ω as well as for the case in which the set Ω is given as in Example 2:

$$\Omega := \{x \in B \mid g(x,y) \leq 0 \ \forall\ y \in Y\}. \tag{4.1}$$

Throughout this section, we will always assume that the set Y of indexes and the constraint functions $g(\cdot, y)\colon B \longrightarrow \mathbb{R} \cup \{+\infty\}$ verify the assumptions $(G_1) - -(G_3)$.

4.1. Approximated solutions of $VIP(T, \Omega)$. Fix $x^0 \in B$ and $\lambda > 0$. Define $T_\lambda(x) := T(x) + \lambda J(x - x^0)$, where J is the duality mapping given in (2.1). It is well known that T_λ is coercive. We use T_λ for defining approximated solutions of $VIP(T, \Omega)$.

DEFINITION 4.1. *Fix $\lambda > 0$, $\beta > 0$, and $x^0 \in B$; an element $\widetilde{x} =: \widetilde{x}(\lambda, \beta, x^0)$ (i.e., depending on λ, β, x^0) is said to be an* approximated solution of $VIP(T, \Omega)$ *when*

(1) $\exists\ \widetilde{\Omega} \supseteq \Omega$ *and* $\widetilde{x} \in \widetilde{\Omega}$ *solves* $VIP(T_\lambda, \widetilde{\Omega})$;

(2) $\widetilde{x} \in x^0 + (1+\beta)(\Omega - x^0)$.

In the particular case in which Ω is given as in (4.1)*, an element $\widetilde{x} =: \widetilde{x}(\lambda, \beta, x^0)$ is an* approximated solution of $VIP(T, \Omega)$ *if it verifies condition* (1) *and*

(2′) $h(\widetilde{x}) \leq \beta$,

where $h(x) = \max_{y \in Y} g(x, y)$.

The parameter λ used in Definition 4.1(1) has a *regularizing* role, since it defines subproblems with coercive operator T_λ. The role of the parameter $\beta > 0$ in Definition 4.1(2) or (2′) is to control the infeasibility of the iterates $\widetilde{x}$. When Ω is as in (4.1), we can connect conditions (2) and (2′). To do this we need the following simple lemma.

LEMMA 4.2. *Fix $x^0 \in B$ and let $\hat{h} : B \to \mathbb{R} \cup \{\infty\}$ be convex and continuous, and suppose that $\emptyset \neq \Omega_0 \subset \{x \in B \mid \hat{h}(x) \leq 0\}$. Then for all $\gamma > 0$ there exists $\beta = \beta(\gamma)$ such that $\inf_{z \in \Omega_0} \hat{h}\left(x^0 + (1+\beta)(z - x^0)\right) < \gamma$.*

Proof. The proof is a consequence of the continuity of $\hat{h}$. Suppose that for some $\gamma_0 > 0$ the conclusion of the lemma does not hold and fix $z^0 \in \Omega_0$. Then for all k we must have $\hat{h}(x^0 + (1 + \frac{1}{k})(z^0 - x^0)) \geq \gamma_0$. Call $\tilde{x}^k := x^0 + (1 + \frac{1}{k})(z^0 - x^0)$. Then $\gamma_0 \leq \lim_k \hat{h}(\tilde{x}^k) = \hat{h}(z^0) \leq 0$, a contradiction. ∎

Take $\Omega_0 := \Omega$ and $\hat{h}(\cdot) := \max_{y \in Y} g(\cdot, y)$ in the lemma above and choose an arbitrary $\gamma > 0$. By the lemma, there exists β small enough and $z \in \Omega$ such that $\tilde{x} := x^0 + (1+\beta)(z - x^0)$ verifies $h(\tilde{x}) < \gamma$, i.e., $\tilde{x}$ verifies condition (2′) for γ. In this sense, we can say that condition (2) is stronger than (2′). On the other hand, condition (2) may never be met, no matter how small the parameter β in condition (2′). Indeed, observe that condition (2) holds only for a vector $\tilde{x}$ when $[\tilde{x}, x^0] \cap \Omega \neq \emptyset$. It is easy to find examples in which h and Ω admit a sequence $\{\tilde{x}^k\}$ verifying $\lim_k h(\tilde{x}^k) = 0$ with $[\tilde{x}^k, x^0] \cap \Omega = \emptyset$ for all k.

Remark 4.3. Observe that when $\lambda = \beta = 0$, $\widetilde{x}$ as in Definition 4.1 solves $VIP(T, \Omega)$.

Remark 4.4. Algorithm 1 in Example 2 provides a set $\widetilde{\Omega}$ and a point $\widetilde{x}$ verifying (1) and (2′) of the definition above. More precisely, fix $\lambda > 0$ and consider an orbit $\{x^k\}$ of Algorithm 1, applied to solve problem $VIP(T_\lambda, \Omega)$. Assume also that all problems (P_k) in Algorithm 1 are *exact*, i.e., $\varepsilon_k = 0$ for all k. Since T_λ is coercive, the assumptions of Theorem 3.4(b) hold. Therefore, $\{x^k\}$ is well defined and bounded, and every weak accumulation point is a solution. Given arbitrary $\beta > 0$, and using Remark 3.5, we conclude that for some k_0, it holds that $h(x^{k_0}) \leq \beta$. So (2′) holds for $\widetilde{x} := x^{k_0}$. By definition, $\widetilde{x} = x^{k_0} \in \Omega^{k_0} = \{x \in B \mid g(x,y) \leq 0 \forall\ y \in Y^{k_0}\} \supset \Omega$ and solves

$VIP(T_\lambda, \Omega^{k_0})$, so $\widetilde{x} = x^{k_0}$ and $\widetilde{\Omega} = \Omega^{k_0}$ verify (1). Using the approximated solutions given by Definition 4.1, we consider the following outer approximation scheme.

ALGORITHM 2.

Step 0. Initialize: Take $x^0 \in B$ and $\{\lambda_k\}, \{\beta_k\} \subset \mathbb{R}_+$.

Iteration: For $k = 1, 2, \ldots,$

Step 1. Given λ_k, β_k, find $\widetilde{x}^k = \widetilde{x}(\lambda_k, \beta_k, x^0)$ an approximated solution of $VIP(T, \Omega)$ (in the sense of Definition 4.1).

Step 2. If $\widetilde{x}^k = x^0$ and $\widetilde{x}^k \in \Omega$, stop. Otherwise,

Step 3. Set $k := k + 1$ and return to Step 1.

Remark 4.5. Regarding the stopping criterion in Step 2, if $\widetilde{x}^k = x^0$ and $\widetilde{x}^k \in \Omega$, we have by condition (1) of Definition 4.1 that

$$0 \in T_{\lambda_k}(x^0) + N_{\widetilde{\Omega}^k}(x^0) = T(x^0) + \lambda_k J(x^0 - x^0) + N_{\widetilde{\Omega}^k}(x^0) = T(x^0) + N_{\widetilde{\Omega}^k}(x^0).$$

Then x^0 is a solution of $VIP(T, \widetilde{\Omega}^k)$. Since $x^0 \in \Omega$, it is also a solution of $VIP(T, \Omega)$. Iterates as in Definition 4.1(1) always exist.

PROPOSITION 4.6. *Assume* (H_1) *holds. Given* $\widetilde{\Omega} \supset \Omega$ *and* $\lambda > 0$*, there exists a unique solution* $\widetilde{x}$ *of* $VIP(T_\lambda, \widetilde{\Omega})$.

Proof. Uniqueness of the solution follows from the fact that T_λ is strictly monotone. The existence is a consequence of the surjectivity of $T_\lambda + N_{\widetilde{\Omega}}$. Indeed, since T_λ is coercive, by Theorems 2.4 and 2.5, it is regular and onto. To use Proposition 2.2, we only have to check that $T_\lambda + N_{\widetilde{\Omega}}$ is maximal monotone. But this fact follows readily from Proposition 2.4, (H_1) and the fact that $\widetilde{\Omega} \supset \Omega$. ∎

PROPOSITION 4.7. *Assume* (H_3) *holds. Let the sequence* $\{\widetilde{x}^k\}$ *be generated by Algorithm* 2 *with approximated solutions as in Definition* 4.1(1) *and* (2). *Take* $\lambda_k, \beta_k \to 0$ *with* $\sup_k \left[\frac{\beta_k}{\lambda_k} + \frac{\lambda_k}{\beta_k}\right] \le c_0$ *for some* $c_0 > 0$. *Then every orbit of Algorithm* 2 *is bounded and* AF.

Proof. By condition (2) in Definition 4.1, there exists $y^k \in \Omega$ such that

$$\widetilde{x}^k = x^0 + (1 + \beta_k)(y^k - x^0), \tag{4.2}$$

where $x^0 \in B$ is given in Step 0 of Algorithm 2. Therefore, boundedness of the orbit will be guaranteed if we show that $\{y^k\}$ is bounded. Using condition (1) in Definition 4.1, we have that $\widetilde{x}^k$ verifies

$$u^k + \eta^k + \lambda_k J(\widetilde{x}^k - x^0) = 0 \text{ with } u^k \in T(\widetilde{x}^k),\ \eta^k \in N_{\widetilde{\Omega}^k}(\widetilde{x}^k).$$

Since $\eta^k \in N_{\widetilde{\Omega}^k}(\widetilde{x}^k)$, we can write

$$\langle u^k, x - \widetilde{x}^k \rangle \ge \lambda_k \langle J(x^0 - \widetilde{x}^k), x - \widetilde{x}^k \rangle \ \forall\ x \in \widetilde{\Omega}^k, \tag{4.3}$$

where we also used Proposition 2.6(i). Take $\overline{x} \in S^*$; then there exists $\overline{v} \in T(\overline{x})$ such that

$$\langle \overline{v}, z - \overline{x} \rangle \ge 0\ \forall z \in \Omega. \tag{4.4}$$

Using monotonicity of T and (4.3) for $x = \overline{x}$, we have that

$$\langle \overline{v}, \overline{x} - \widetilde{x}^k \rangle \ge \lambda_k \langle J(x^0 - \widetilde{x}^k), \overline{x} - \widetilde{x}^k \rangle.$$

By Proposition 2.6(ii) with $z = x^0 - \overline{x}$ and $x = x^0 - \widetilde{x}^k$ we get

$$\langle J(x^0 - \widetilde{x}^k), \overline{x} - \widetilde{x}^k \rangle \ge \frac{1}{2}(\|x^0 - \widetilde{x}^k\|^2 - \|x^0 - \overline{x}\|^2).$$

The last two expressions yield

$$\langle \overline{v}, \overline{x} - \widetilde{x}^k \rangle \geq \frac{\lambda_k}{2}(\|x^0 - \widetilde{x}^k\|^2 - \|x^0 - \overline{x}\|^2). \tag{4.5}$$

Use (4.4) for $z = y^k$ to conclude that $\langle \overline{v}, y^k - \overline{x} \rangle \geq 0$. Adding and subtracting $\widetilde{x}^k$ in the right-hand side of this inner product, we obtain

$$\beta_k \langle \overline{v}, x^0 - y^k \rangle = \langle \overline{v}, y^k - \widetilde{x}^k \rangle \geq \langle \overline{v}, \overline{x} - \widetilde{x}^k \rangle,$$

where we also used (4.2). Combine the last expression with (4.5) to get

$$\begin{array}{rcl} \beta_k \langle \overline{v}, x^0 - y^k \rangle & \geq & \frac{\lambda_k}{2}(\|x^0 - \widetilde{x}^k\|^2 - \|x^0 - \overline{x}\|^2) \\ & = & \frac{\lambda_k}{2}\left((1+\beta_k)^2\|y^k - x^0\|^2 - \|x^0 - \overline{x}\|^2\right), \end{array}$$

where we used (4.2) in the equality. Rearranging the last expression, we get

$$\begin{array}{rcl} 0 & \geq & \langle \overline{v}, y^k - x^0 \rangle + \frac{(1+\beta_k)^2}{2}\left(\frac{\lambda_k}{\beta_k}\right)\|y^k - x^0\|^2 - \frac{1}{2}\left(\frac{\lambda_k}{\beta_k}\right)\|x^0 - \overline{x}\|^2 \\ & \geq & \langle \overline{v}, y^k - x^0 \rangle + \frac{1}{2\,c_0}\|y^k - x^0\|^2 - \frac{c_0}{2}\|x^0 - \overline{x}\|^2, \end{array} \tag{4.6}$$

where we used $\frac{(1+\beta_k)^2\lambda_k}{\beta_k} \geq 1/c_0$ and $\frac{\lambda_k}{\beta_k} \leq c_0$ in the second inequality. The right-hand side of the last inequality is a convex quadratic function on y^k; thus the sequence $\{y^k\}$ must be bounded. Equivalently, $\{\widetilde{x}^k\}$ is bounded. Now we proceed to prove that every weak accumulation point is feasible. Let $\{\widetilde{x}^{k_j}\} \subseteq \{\widetilde{x}^k\}$ be a subsequence weakly convergent to x^*. By definition of $\widetilde{x}^{k_j}$ we have that $\widetilde{x}^{k_j} - x^* = y^{k_j} - x^* + \beta_{k_j}(y^{k_j} - x^0)$. Since $\lim_j \beta_{k_j} = 0$ and $\{y^{k_j}\} \subset \Omega$ is bounded, we conclude that $w - \lim_j y^{k_j} = x^*$. Since Ω is closed and convex, it is weakly closed, and hence $x^* \in \Omega$. □

Recall that a Slater condition for Ω as in (4.1) requires the existence of a point x^0 such that $g(x^0, y) < 0$ for all $y \in Y$. However, by assumptions (G_1) and (G_3), we have that the Slater condition is equivalent to the existence of some $\alpha > 0$ such that

$$g(x^0, y) \leq -\alpha \ \forall\ y \in Y. \tag{4.7}$$

So, under our hypotheses there is no loss of generality by calling (4.7) a *Slater condition* for Ω. The next result is analogous to Proposition 4.7, for the case in which the set Ω is as in (4.1), and the approximate solutions are taken as in Definition 4.1(1) and (2′).

PROPOSITION 4.8. *Assume that* (H_3) *holds. Let* Ω *be as in* (4.1) *and such that a Slater condition holds for* Ω. *Take* $\lambda_k \to 0$ *and suppose that there exists* $c_0 > 0$ *with* $\sup_k \frac{\beta_k}{\lambda_k} \leq c_0 < \infty$. *Assume also that in Step* 0 *of Algorithm* 2 *we use the* x^0 *provided by* (4.7). *Then every orbit* $\{\widetilde{x}^k\}$ *of Algorithm* 2 *is bounded and* AF.

Proof. Let x^0 and $\alpha > 0$ be such that (4.7) holds, and fix $c > c_0$. Since $\lambda_k \to 0$, there exists $k_0 \in \mathbb{N}$ such that

$$0 < \frac{c\,\lambda_k}{\alpha} < 1 \text{ for all } k \geq k_0. \tag{4.8}$$

Using that $c > c_0$, we get

$$0 < \beta_k < c\lambda_k = \frac{c\alpha\lambda_k}{\alpha} < \frac{c\alpha\lambda_k}{\alpha - c\lambda_k} \tag{4.9}$$

for all $k \geq k_0$. Define now the auxiliary sequence $x^k := \widetilde{x}^k + \frac{c\lambda_k}{\alpha}(x^0 - \widetilde{x}^k)$. We claim that $x^k \in \Omega$ for all $k > k_0$. Indeed, by convexity of h and (4.8) we have that

$$h(x^k) \leq \frac{c\lambda_k}{\alpha} h(x^0) + \left(1 - \frac{c\lambda_k}{\alpha}\right) h(\widetilde{x}^k) \ \forall \ k > k_0.$$

Using (4.7) and the fact that $\widetilde{x}^k$ is an approximated solution, we get

$$h(x^k) \leq \frac{c\lambda_k}{\alpha}(-\alpha) + \left(1 - \frac{c\lambda_k}{\alpha}\right)(\beta_k);$$

now using (4.9) for all $k > k_0$ we get

$$h(x^k) \leq -c\lambda_k + \left(1 - \frac{c\lambda_k}{\alpha}\right)\left(\frac{c\alpha\lambda_k}{\alpha - c\lambda_k}\right) = -c\lambda_k + c\lambda_k = 0. \tag{4.10}$$

Therefore $\{x^k\} \subseteq \Omega$ for all $k > k_0$. Using the fact that $\widetilde{x}^k$ verifies condition (1) in Definition 4.1, and following the same steps as in the proof of Proposition 4.7 (see equations (4.3)–(4.5)), we arrive to the inequality

$$\langle \overline{v}, \overline{x} - \widetilde{x}^k \rangle \geq \frac{\lambda_k}{2}(\|x^0 - \widetilde{x}^k\|^2 - \|x^0 - \overline{x}\|^2), \tag{4.11}$$

where $\bar{x}$ is a solution of $VIP(T, \Omega)$ and $\bar{v} \in T(\bar{x})$. Using definition of $\bar{x}$ and the fact that $x^k \in \Omega$, we get $\langle \overline{v}, x^k - \overline{x} \rangle \geq 0$. Summing and subtracting $\widetilde{x}^k$ in the right-hand side of this inner product, we obtain

$$\langle \overline{v}, x^k - \widetilde{x}^k \rangle \geq \langle \overline{v}, \overline{x} - \widetilde{x}^k \rangle. \tag{4.12}$$

Using definition of $\{x^k\}$ and combining (4.12) and (4.11), we have

$$\frac{c\lambda_k}{\alpha}\langle \overline{v}, x^0 - \widetilde{x}^k \rangle \geq \frac{\lambda_k}{2}(\|x^0 - \widetilde{x}^k\|^2 - \|x^0 - \overline{x}\|^2). \tag{4.13}$$

Dividing by $\lambda_k > 0$, we conclude that

$$\frac{c}{\alpha}\langle \overline{v}, \widetilde{x}^k - x^0 \rangle + \frac{1}{2}(\|x^0 - \widetilde{x}^k\|^2 - \|x^0 - \overline{x}\|^2) \leq 0, \tag{4.14}$$

which, in the same way as in the last part of Proposition 4.7, yields boundedness of $\{\widetilde{x}^k\}$. Now let $\{\widetilde{x}^{k_j}\} \subseteq \{\widetilde{x}^k\}$ be a subsequence weakly convergent to x^*. Using Definition 4.1(2′) we have that $h(\widetilde{x}^{k_j}) \leq \beta_{k_j}$ for all j. Since h weakly continuous and $\beta_k \to 0$, we get that $h(x^*) \leq 0$. Thus $x^* \in \Omega$, and hence the orbit $\{\widetilde{x}^k\}$ is AF. □

We proved so far that under suitable assumptions on the data, Algorithm 2 generates an orbit which is bounded and AF. We present below conditions under which these two properties guarantee optimality of weak accumulation points.

THEOREM 4.9. *Let the sequence $\{\widetilde{x}^k\}$ be generated by Algorithm 2, where the iterates $\widetilde{x}^k$ verify condition (1) of Definition 4.1 with parameters $\{\lambda_k\}$ such that $\lim_k \lambda_k = 0$. Suppose that (H_2) and (H_3) hold. If $\{\widetilde{x}^k\}$ is bounded and AF, then one of the following holds:*

(i) *Algorithm 2 has finite termination, and the last iterate is a solution of $VIP(T, \Omega)$.*
(ii) *Algorithm 2 generates an infinite orbit, and every weak accumulation point of $\{\widetilde{x}^k\}$ is a solution of $VIP(T, \Omega)$.*

Proof. Case (i) has been taken care of in Remark 4.5. Thus it is enough to consider the case in which Algorithm 2 generates an infinite orbit. Since $\widetilde{x}^k$ verifies condition (1) of Definition 4.1, we have as in (4.3),

$$\langle u^k, x - \widetilde{x}^k\rangle \geq \lambda_k \langle J(x^0 - \widetilde{x}^k), x - \widetilde{x}^k\rangle \ \forall\ x \in \widetilde{\Omega}^k. \tag{4.15}$$

Take x^* a weak accumulation point of $\{\widetilde{x}^k\}$ and a subsequence $\{\widetilde{x}^{k_j}\} \subset \{\widetilde{x}^k\}$ weakly converging to x^*. Using the last expression, together with the fact that $\Omega \subseteq \widetilde{\Omega}$, we can write for every $x \in \Omega$,

$$\begin{array}{rcl} \langle u^k, \widetilde{x}^k - x\rangle & \leq & \lambda_k \langle J(x^0 - \widetilde{x}^k), \widetilde{x}^k - x\rangle \\ & \leq & \lambda_k \|J(x^0 - \widetilde{x}^k)\|\ \|\widetilde{x}^k - x\| \\ & = & \lambda_k \|x^0 - \widetilde{x}^k\|\ \|\widetilde{x}^k - x\|, \end{array} \tag{4.16}$$

where we used the Cauchy–Schwartz inequality and definition of J. Since $\lim_k \lambda_k = 0, x^* \in \Omega$, and $\{\widetilde{x}^k\}$ is bounded, from (4.16) for $k = k_j$ we get that

$$\limsup_j \langle u^{k_j}, \widetilde{x}^{k_j} - x^*\rangle \leq 0. \tag{4.17}$$

Using $\widetilde{x}^{k_j} \xrightarrow{w} x^*$, (4.17), and pseudomonotonicity of T for $\overline{x} \in S^*$ we conclude that there exists $u^* \in T(x^*)$ such that

$$\liminf_j \langle u^{k_j}, \widetilde{x}^{k_j} - \overline{x}\rangle \geq \langle u^*, x^* - \overline{x}\rangle. \tag{4.18}$$

On the other hand, using (4.16) for $\overline{x} \in S^*$ and $k = k_j$ we get

$$\liminf_j \langle u^{k_j}, \widetilde{x}^{k_j} - \overline{x}\rangle \leq 0. \tag{4.19}$$

Using also (4.18) we get

$$\langle u^*, x^* - \overline{x}\rangle \leq 0. \tag{4.20}$$

Finally, by paramonotonicity of T and the above inequality, we can use Proposition 2.8 to guarantee that x^* is a solution of the $VIP(T, \Omega)$, as we wanted to prove. ∎

Our next convergence result requires the approximated solutions $\widetilde{x}$ to stay close enough to the original set Ω. Fix $\theta > 0$ and $r > 1$. We say that $\widetilde{x}$ is a *metric*-approximated solution if it verifies Definition 4.1(1) and the condition $(2'')$:

$$d(\widetilde{x}, \Omega) := \inf_{z \in \Omega} \|\widetilde{x} - z\| \leq \theta \lambda^r,$$

where λ is the parameter used in Definition 4.1(1).

Remark 4.10. When Ω is as in (4.1), condition $(2'')$ can be guaranteed when the *Hoffman global error bound* [25] holds:

$$\exists\, \theta > 0\ \forall\, x \notin \Omega,\ d(x, \Omega) \leq \theta \max_{y \in Y} g(x, y) = \theta\ h(x). \tag{4.21}$$

An error bound of this kind has been proved to hold in a Banach space in [15]. Namely, assume that (G_1)–(G_3) hold. Suppose also that there exist $\delta, \gamma > 0$ such that

- (G_4) $\Omega(\delta) := \{x \in B \mid h(x) \leq -\delta\} \neq \emptyset$ and
- (G_5) $Haus(\Omega, \Omega(\delta)) < \gamma$ (where $Haus(\cdot, \cdot)$ stands for the Hausdorff distance between sets).

Then [15, Proposition 1] proves that

$$\forall\, x \notin \Omega,\ d(x,\Omega) \leq \delta^{-1}\gamma\, h(x). \tag{4.22}$$

Therefore, when assumptions (G_1)–(G_5) hold for Ω, condition $(2'')$ is a consequence of Definition 4.1$(2')$ with $\beta = \lambda^r$. For alternative conditions under which a global error bound of the kind (4.21) holds in a Banach space, see [27].

In the theorem below we prove boundedness of the orbit generated by Algorithm 2, as well as optimality of all weak limit points. In the case in which B is uniformly convex, we get full strong convergence and we characterize the limit of $\{\widetilde{x}^k\}$ as the closest point to x^0 in the solution set.

THEOREM 4.11. *Let the sequence $\{\widetilde{x}^k\}$ be generated by Algorithm 2, where the iterates satisfy conditions (1) of Definition 4.1 and $(2'')$ for $\lambda = \lambda_k$. Assume that $\lambda_k \to 0$ and that (H_2) and (H_3) hold. Then the sequence $\{\widetilde{x}^k\}$ is bounded and every weak limit point is a solution of $VIP(T,\Omega)$. Moreover, if B is uniformly convex, then the sequence $\{\widetilde{x}^k\}$ converges strongly and the limit point is the unique x^* characterized by*

$$\{x^*\} = \arg\min_{y\in S^*} \|x^0 - y\|^2.$$

Proof. We prove first that $\widetilde{x}^k$ is bounded. The approximated solution $\widetilde{x}^k$ is such that

$$u^k + \eta^k + \lambda_k J(\widetilde{x}^k - x^0) = 0 \text{ with } u^k \in T(\widetilde{x}^k),\ \eta^k \in N_{\widetilde{\Omega}^k}(\widetilde{x}^k). \tag{4.23}$$

Let $y \in S^*$; then there exists $v \in T(y)$ such that

$$\langle v, x - y\rangle \geq 0\ \forall x \in \Omega. \tag{4.24}$$

Thus by Proposition 2.6(ii),

$$\|x^0 - y\|^2 \geq \|x^0 - \widetilde{x}^k\|^2 + 2\langle J(x^0 - \widetilde{x}^k), \widetilde{x}^k - y\rangle. \tag{4.25}$$

Define $A := 2\langle J(x^0 - \widetilde{x}^k), \widetilde{x}^k - y\rangle$; (4.23) implies that

$$A = \frac{2}{\lambda_k}(\langle u^k, \widetilde{x}^k - y\rangle + \langle \eta^k, \widetilde{x}^k - y\rangle).$$

Using the definition of normality operator and monotonicity of T we obtain

$$A \geq \frac{2}{\lambda_k}\langle v, \widetilde{x}^k - y\rangle. \tag{4.26}$$

Combining (4.26) and (4.25) we get

$$\|x^0 - y\|^2 \geq \|x^0 - \widetilde{x}^k\|^2 + \frac{2}{\lambda_k}\langle v, \widetilde{x}^k - y\rangle. \tag{4.27}$$

We claim that for all $y\ \in S^*$

$$\|x^0 - y\|^2 \geq \|x^0 - \widetilde{x}^k\|^2 - 2\theta{\lambda_k}^{r-1}\|v\|\ \forall\ k. \tag{4.28}$$

Indeed, assume first that $\widetilde{x}^k \in \Omega$, and by (4.24) and (4.27) we have that $\|x^0 - y\|^2 \geq \|x^0 - \widetilde{x}^k\|^2$, and hence (4.28) holds. Assume now that $\widetilde{x}^k \notin \Omega$. Let p^k be the projection of the $\widetilde{x}^k$ in Ω; we obtain

$$\begin{array}{rcl} \frac{2}{\lambda_k}\langle v, \widetilde{x}^k - y\rangle & = & \frac{2}{\lambda_k}\left[\langle v, \widetilde{x}^k - p^k\rangle + \langle v, p^k - y\rangle\right] \\ & \geq & \frac{2}{\lambda_k}\langle v, \widetilde{x}^k - p^k\rangle \geq -\frac{2}{\lambda_k}\|v\|\|\widetilde{x}^k - p^k\| \\ & \geq & -2\theta\lambda_k^{r-1}\|v\|, \end{array} \tag{4.29}$$

where we used the Cauchy–Schwartz inequality, (4.24), and condition $(2'')$. Combining (4.27) with (4.29), we conclude (4.28). Thus our claim is true and hence $\{\widetilde{x}^k\}$ is bounded. Let $\{\widetilde{x}^{k_j}\} \subseteq \{\widetilde{x}^k\}$ be a subsequence weakly convergent to x^*. Since $\lambda_{k_j} \to 0$, condition $(2'')$ readily implies the existence of a sequence $\{y^{k_j}\} \subset \Omega$ such that $w - \lim_j y^{k_j} = x^*$. Using the fact that Ω is convex and (strongly) closed, it is also weakly closed, and hence $x^* \in \Omega$. Now, using the last part of the proof of Theorem 4.9 (see equations (4.16) to (4.20)), every limit point of $\{\widetilde{x}^k\}$ is a solution of $VIP(T, \Omega)$. We proceed now to prove the last assertion of the theorem. Assume that B is uniformly convex, and take again the subsequence $\{\widetilde{x}^{k_j}\}$ weakly converging to x^*. We know that $x^* \in S^*$. By (4.28) for $k = k_j$ and $y = x^*$ we get

$$\|x^0 - x^*\|^2 \geq \|x^0 - \widetilde{x}^{k_j}\|^2 - 2\theta{\lambda_{k_j}}^{r-1}\|v\|.$$

Taking limsup for $j \to \infty$ in the expression above and using that $\lambda_k \to 0$, we have that

$$\|x^0 - x^*\|^2 \geq \limsup_{j\to\infty} \|x^0 - \widetilde{x}^{k_j}\|^2. \tag{4.30}$$

Since $\{x^0 - \widetilde{x}^{k_j}\}$ weakly converges to $x^0 - x^*$ we apply Proposition 2.13 to conclude that $\{\widetilde{x}^{k_j}\}$ converges strongly to x^*. Now, taking limits for $j \to \infty$ in (4.28) for $k = k_j$ we get

$$\|x^0 - y\|^2 \geq \|x^0 - x^*\|^2 \ \forall\ y\ \in S^*.$$

Thus,

$$x^* \in \arg\min_{y\in S^*} \|x^0 - y\|^2.$$

This point is unique by strict convexity of $\|\cdot\|^2$. □

Acknowledgment. The authors are very thankful to the two referees, whose helpful and essential corrections greatly improved an earlier version of the manuscript.

REFERENCES

[1] E. Asplund, *Positivity of duality mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 200–203.
[2] J. W. Blankenship and J. E. Falk, *Infinitely constrained optimization problems*, J. Optim. Theory Appl., 19 (1976), pp. 261–281.
[3] J. Bracken and J. F. McGill, *Mathematical programs with optimization problems in the constraints*, Oper. Res., 21 (1973), pp. 37–44.
[4] H. Brézis, *Opérateurs Monotones Maximaux et Semigroups de Contractions dans les Espaces de Hilbert*, North–Holland, Amsterdam, 1973.
[5] H. Brézis, *Analyse fonctionnele: Théorie et applications*, Masson, Paris, 1983.

[6] F. E. Browder, *Nonlinear operators and nonlinear equations of evolution in Banach spaces*, in Nonlinear Functional Analysis, AMS, Providence, RI, 1976, pp. 1–308.
[7] R. D. Bruck, *An iterative solution of a variational inequality for certain monotone operator in a Hilbert space*, Bull. Amer. Math. Soc., 81 (1975), pp. 890–892. Corrigendum in 82 (1976), p. 353.
[8] R. S. Burachik and S. M. Scheimberg, *A proximal point method for the variational inequality problem in Banach spaces*, SIAM J. Control Optim., 39 (2001), pp. 1633–1649.
[9] Y. Censor, A. Iusem, and S. Zenios, *An interior point method with Bregman functions for the variational inequality problem with paramonotone operators*, Math. Programming, 81 (1998), pp. 373–400.
[10] W. E. Cheney and A. A. Goldstein, *Newton's method for convex programming and Tchebycheff approximation*, Numer. Math., 1 (1959), pp. 253–268.
[11] P. L. Combettes, *Strong convergence of block-iterative outer approximation methods for convex optimization*, SIAM J. Control Optim., 38 (2000), pp. 538–565.
[12] G. Crombez, *Finding projections onto the intersection of convex sets in Hilbert spaces*, Numer. Funct. Anal. Optim., 16 (1995), pp. 637–652.
[13] J. M. Danskin, *The Theory of Max-Min*, Springer-Verlag, Berlin, 1967.
[14] M. A. H. Dempster and R. R. Merkovsky, *A practical geometrically convergent cutting plane algorithm*, SIAM J. Numer. Anal., 32 (1995), pp. 631–644.
[15] S. Deng, *Computable error bounds for convex inequality systems in Banach spaces*, SIAM J. Control Optim., 36 (1998), pp. 1240–1249.
[16] B. C. Eaves and W. I. Zangwill, *Generalized cutting plane algorithms*, SIAM J. Control, 9 (1971), pp. 529–542.
[17] A. V. Fiacco and K. O. Kortanek, *Semi-infinite Programming and Applications*, Lecture Notes in Economics and Math. Systems 215, Springer-Verlag, New York, 1983.
[18] C. W. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, Boston, 1984.
[19] S. A. Gustavson and K. O. Kortanek, *Numerical treatment of a class of semi-infinite programming problems*, Naval Res. Logist. Quart., 20 (1970), pp. 477–504.
[20] Y. Haugazeau, *Sur la minimisation de formes quadratiques avec contraintes*, C. R. Acad. Sci. Paris Sér. A Math., 264 (1967), pp. 322–324.
[21] Y. Haugazeau, *Sur les Inéquations Variationelles et la Minimisation de Fonctionnelles Convexes*, thesis, Université de Paris, 1968.
[22] R. Hettich, *Semi-infinite Programming*, Lecture Notes in Control and Inform. Sci. 15, Springer-Verlag, New York, 1979.
[23] R. Hettich, *A review of numerical methods for semi-infinite optimization*, in Semi-infinite Programming and Applications, Lecture Notes in Econom. and Math. Systems 215, Springer-Verlag, New York, 1983, pp. 158–178.
[24] R. Hettich and K. O. Kortanek, *Semi-infinite programming: Theory, methods, and applications*, SIAM Rev., 35 (1993), pp. 380–429.
[25] A. J. Hoffman, *On approximate solutions of system of linear inequalities*, J. Nat. Bureau Standards, 49 (1952), pp. 263–265.
[26] A. N. Iusem, *On some properties of paramonotone operators*, J. Convex Anal., 5 (1998), pp. 269–278.
[27] A. Jourani, *Hoffman's error bound, local controllability, and sensitivity analysis*, SIAM J. Control Optim., 38 (2000), pp. 947–970.
[28] A. A. Kaplan, *Determination of the extremum of a linear function on a convex set*, Soviet. Math. Dokl., 9 (1968), pp. 269–271.
[29] A. A. Kaplan and R. Tichatschke, *Stable Methods for Ill-Posed Variational Problems*, Akademie Verlag, Berlin, 1994.
[30] A. A. Kaplan and R. Tichatschke, *Variational inequalities and convex-semi-infinite programming problems*, Optimization, 26 (1992), pp. 187–214.
[31] S. Karamardian, *Complementarity problems over cones with monotone and pseudomonotone maps*, J. Optim. Theory Appl., 18 (1976), pp. 445–455.
[32] J. E. Kelley, *The cutting-plane method for solving convex programs*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 703–712.
[33] P. J. Laurent and B. Martinet, *Méthods duales pour le calcul du minimum d'une fonction convexe sur une intersection de convexes,* in Symposium on Optimization, Lecture Notes in Math. 132, Springer-Verlag, New York, 1970, pp. 159–180.
[34] F. Lui and M. Z. Nashed, *Regularization of nonlinear ill-posed variational inequalities and convergence rates*, Set-Valued Anal., 6 (1998), pp. 313–344.
[35] U. Mosco, *Convergence of convex sets and of solutions of variational inequalities*, Advances in Math., 3 (1969), pp. 510–585.

[36] W. Oettli, *Solving Optimization problems with many constraints by a sequence of subproblems containing only two constraints*, Math. Nachr., 71 (1976), pp. 143–145.
[37] D. Pascali and S. Sburlan, *Nonlinear Mappings of Monotone Type*, Ed. Academiei, Bucharest, Romania, 1978.
[38] G. Pierra, *Eclatement de contraintes en parallèle pour la minimisation d'une forme quadratique*, in Lecture Notes in Comput. Sci. 41, Springer-Verlag, New York, 1976, pp. 200–218.
[39] G. Pierra, *Decomposition through formalization in a product space*, Math. Programming, 28 (1984), pp. 96–115.
[40] R. Reemtsen and S. Goerner, *Numerical methods for semi-infinite programming: A survey*, in Semi-Infinite Programming, R. Reemtsen and J.-J. Rückmann, eds., Kluwer, Boston, 1998, pp. 195–275.
[41] R. T. Rockafellar, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75–88.
[42] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, John Wiley and Sons, New York, 1977.
[43] D. M. Topkis, *Cutting-plane methods without nested constraint sets,* Oper. Res., 18 (1970), pp. 404–413.
[44] D. M. Topkis, *A cutting-plane algorithm with linear and geometric rates of convergence*, J. Optim. Theory Appl., 36 (1982), pp. 1–22.
[45] A. F. Veinott, *The supporting hyperplane method for unimodal programming*, Oper. Res., 15 (1967), pp. 147–152.
[46] A. A. Vladimirov, Yu. E. Nesterov, and Yu. N. Cekanov, *Uniformly convex functionals*, Vestnik Moskov. Univ. Ser. XV Vychisl. Mat. Kibernet., 3 (1978), pp. 12–23.
[47] C. Zalinescu, *On uniformly convex functions*, J. Math. Anal. Appl., 94 (1983), pp. 344–374.
[48] W. I. Zangwill, *Nonlinear Programming—A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2089–2108

ASYMPTOTIC STABILITY FOR INTERMITTENTLY CONTROLLED SECOND-ORDER EVOLUTION EQUATIONS*

A. HARAUX†, P. MARTINEZ‡, AND J. VANCOSTENOBLE‡

Abstract. Motivated by several works on ordinary differential equations, we are interested in the asymptotic stability of *intermittently controlled* partial differential equations. We give a condition of asymptotic stability for second-order evolution equations uniformly damped by an on/off feedback. This result extends to the case of partial differential equations a previous result of R. A. Smith concerning ordinary differential equations.

Key words. damped wave equation, second-order evolution equations, asymptotic behavior, on-off damping

AMS subject classifications. 35L05, 35L10, 35B35, 35B40

DOI. 10.1137/S0363012903436569

1. Introduction. Motivated by several works on ordinary differential equations, we are interested in the asymptotic stability of *intermittently controlled* partial differential equations. This question has been widely studied in the case of ordinary differential equations (see, for example, [1, 9, 10, 25, 27, 28]). The typical problem is the oscillator damped by an *on/off* damping:

$$u'' + u + a(t)u' = 0, \quad t > 0, \tag{1.1}$$

where $a : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous nonnegative. For each solution u of (1.1), we define its energy by

$$\forall t \geq 0, \quad E_u(t) = \frac{1}{2}u(t)^2 + \frac{1}{2}u'(t)^2.$$

The derivative of the energy is

$$E'(t) = u(t)u'(t) + u'(t)u''(t) = -a(t)u'(t)^2,$$

hence the energy is always nonincreasing, but remains constant on the time intervals for which $a = 0$, and the decay is "very small" if a is "very small." Denote $\ell := \lim_{t\to\infty} E(t)$. Many authors (see, in particular, [1, 9, 10, 25, 27, 28]) investigated the links between the distribution of sets where a is positive and the property $\ell = 0$.

Assume that there exists a sequence $(I_n)_{n\geq 0}$ of disjoint open intervals in $(0, +\infty)$, denoted by $I_n = (a_n, b_n)$, where $b_n \leq a_{n+1}$ for all $n \in \mathbb{N}$, and such that

$$\forall t \in I_n, \quad 0 < m_n \leq a(t) \leq M_n < \infty.$$

Roughly speaking, the energy is strictly decreasing on the time intervals I_n and just nonincreasing elsewhere. It is natural to wonder whether the decay on the time

*Received by the editors October 16, 2003; accepted for publication (in revised form) July 4, 2004; published electronically April 14, 2005.
http://www.siam.org/journals/sicon/43-6/43656.html
†Laboratoire Jacques-Louis Lions, U.M.R C.N.R.S. 7598, Université Pierre et Marie Curie, Boite courrier 187, 75252 Paris Cedex 05, France (haraux@ann.jussieu.fr).
‡Laboratoire M.I.P., U.M.R. C.N.R.S. 5640, Université Paul Sabatier Toulouse III, 118 route de Narbonne, 31 062 Toulouse Cedex 4, France (martinez@mip.ups-tlse.fr, vancoste@mip.ups-tlse.fr).

intervals I_n is sufficient to drive the energy to zero. Obviously some condition on the length of the intervals I_n has to be imposed to ensure $\ell = 0$. Smith [27] proved the following *sufficient condition of asymptotic stability*:

THEOREM 1.1. [27]. *Assume that*

$$\sum_{n=0}^{\infty} m_n T_n \delta_n^2 = +\infty, \tag{1.2}$$

where m_n and M_n are the minimum and the maximum values of $a(t)$ in I_n, T_n is the length of I_n and $\delta_n = \min(T_n, (1+M_n)^{-1})$. Then (1.1) *is asymptotically stable; i.e., every solution u of* (1.1) *satisfies $E_u(t) \to 0$ as $t \to \infty$.*

For example, in the case of a damping such that $0 < m \le a(t) \le M$ for all $t \in I_n$ for all $n \in \mathbb{N}$, the condition (1.2) reduces to

$$\sum_{n=0}^{\infty} T_n^3 = +\infty. \tag{1.3}$$

It is noteworthy that (1.3) is also necessary in the following sense: given $\varepsilon > 0$ as small as we want, Pucci and Serrin [25] constructed an example for which the sequence $(T_n)_n$ satisfies

$$\sum_{n=0}^{\infty} T_n^{3-\varepsilon} = +\infty, \quad \text{while} \quad \sum_{n=0}^{\infty} T_n^3 < +\infty,$$

and suitable initial conditions such that the energy decays to some $\ell > 0$.

Note also that, under condition (1.2), *the distribution of the intervals I_n has no importance.* Only their size is important.

Condition (1.2) also requires that the damping coefficient a is not "too small" or "too large," in order to prevent "underdamping" or "overdamping." These phenomena are also a source of lack of strong stability (see [20, 22, 26], where the stability is studied for the wave equation, but always under the condition that the function a remains positive).

To our knowledge, stability properties for such "intermittently controlled" systems have not yet been studied in the case of *partial* differential equations.

In [21], we studied the effect of an on/off feedback on the wave equation. We considered the simplified case of a damping coefficient a that is $2T$-periodic and such that $a(t) = a_0 > 0$ on $(0, T)$ and $a(t) = 0$ on $(T, 2T)$. In particular, the condition (1.2) was always satisfied. And we studied the wave equation damped by a *boundary* on/off feedback or by a *locally distributed* on/off feedback. In both cases, we proved that the situation is *radically different* from the case of ordinary differential equations. Indeed, we proved that, except for a countable number of exceptional values of T, asymptotic stability occurs (and more precisely, exponential stability). But, for the exceptional values of T, asymptotic stability does not occur. This means that *the distribution of the intervals I_n is very important in the case of the locally damped wave equation.* This phenomenon is related to the optics rays propagation and the geometric condition of Bardos, Lebeau, and Rauch [2, 3]. See further comments in section 3.3.

In [21], the only case for which the situation was not different from the situation of the ordinary differential equations was the wave equation damped by an *uniformly distributed* on/off feedback. In that case, asymptotic stability occurs for any value of T. Thus the distribution of the intervals damping has no importance.

In the present work, we now study the wave equation *uniformly damped* by a *general on/off feedback* (in particular, not necessarily periodic). We prove that *"the uniformly damped wave equation behaves exactly like the oscillator"* in the sense that Theorem 1.1 is still true.

More generally, we prove this result in an abstract setting that includes both the oscillator and wavelike or platelike equations and that also includes *bounded or unbounded* and *linear and nonlinear* damping operators.

In particular, this gives for the result of Smith a new proof quite different from the original one, which was relying on monotonicity properties of the solutions of (1.1). Our method is based on a preliminary result which is interesting in itself: we provide an estimate of the energy decay on a *short* time interval (see Theorem 3.1). This estimate is true for both ordinary and partial differential equations.

The paper is organized as follows.

- In section 2, we introduce our abstract setting and we give the result of well-posedness (Theorem 2.1).
- In section 3, we provide an estimate of the energy decay on a *short* time interval (Theorem 3.1) and we deduce the asymptotic stability result (Theorem 3.2) extending the previous result of Smith. Then we make some further comments concerning the case of locally distributed dampings to explain the necessity of considering only uniformly distributed dampings.
- In section 4, we give some examples.
- In section 5, we present another application of the method to the case of a positive-negative damping (Theorem 5.1).

2. Abstract setting and well-posedness. Let H be a real Hilbert space endowed with the scalar product $(\cdot,\cdot)_H$ and the norm $|\cdot|_H$.

Assume that $A : D(A) \subset H \to H$ is a linear self-adjoint and coercive operator on H with dense domain. We define $V = D(A^{1/2})$ endowed with the scalar product $((\cdot,\cdot))_V$ and the norm $\|\cdot\|_V$ defined by

$$\forall v \in V, \quad \|v\|_V^2 = |A^{1/2}v|_H^2 = \langle \tilde{A}v, v\rangle_{V',V},$$

where $\tilde{A} \in \mathcal{L}(V, V')$ represents the extension of A.

Also let W be a Hilbert space endowed with the norm $\|\cdot\|_W$ and such that

$$V \hookrightarrow W \hookrightarrow H \equiv H' \hookrightarrow W' \hookrightarrow V'$$

with dense imbeddings. We also assume that A satisfies the following property:

$$\begin{aligned} &\exists \lambda_0,\, C_0 > 0, \ \text{ such that, } \forall \lambda \in [0, \lambda_0], \\ &\quad (I + \lambda A)^{-1} \in \mathcal{L}(W) \ \text{ and } \ \|(I + \lambda A)^{-1}\|_{\mathcal{L}(W)} \le C_0. \end{aligned} \tag{2.1}$$

Next we consider a *time-dependent* operator B such that

$$B \in L^\infty(J, \mathrm{Lip}\,(W, W')), \qquad B(t)0 = 0, \tag{2.2}$$

$$\forall t \in J,\ \forall w,\, z \in W, \quad \langle B(t)w - B(t)z, w - z\rangle_{W',W} \ \ge\ 0, \tag{2.3}$$

$$\forall t \in J,\ \forall w \in W, \quad \langle B(t)w, w\rangle_{W',W} \ \ge\ b^2(t)\|w\|_W^2, \tag{2.4}$$

$$\forall t \in J,\ \forall w,\ z \in W, \quad \|B(t)w - B(t)z\|_{W'} \ \leq\ Cb(t)^2\|w - z\|_W, \tag{2.5}$$

where $J = [0, T]$ with $T > 0$ and where $b(t) \geq 0$ with $b \in L^2(J)$. Note that $B(t)$ is *a priori unbounded and nonlinear*. (The choice $W = H$ corresponds to the particular case of a bounded operator.)

Now we consider the following second-order evolution equation

$$u'' + Au + B(t)u' = 0, \quad t > 0, \tag{2.6}$$

with the initial conditions

$$u(0) = u_0 \in V, \quad u'(0) = u_1 \in H \tag{2.7}$$

and we prove that this problem is well-posed.

THEOREM 2.1. *Under the previous assumptions, for any* $(u_0, u_1) \in V \times H$, *there exists a unique solution* $u \in L^2(0, T; V) \cap W^{1,2}(0, T; H) \cap W^{2,2}(0, T; V')$ *with* $bu' \in L^2(0, T; W)$ *and* $B(t)u' = b(t)h(t)$, $h(t) \in L^2(0, T; W')$ *of*

$$u'' + Au + B(t)u' = 0 \quad \text{in } L^2(0, T; V')$$

such that

$$u(0) = u_0 \in V, \quad u'(0) = u_1 \in H.$$

In addition $u \in C([0, T]; V) \cap C^1([0, T]; H)$ *and the energy of the solution* u *defined by*

$$\forall t \geq 0, \quad E_u(t) \ :=\ \frac{1}{2}\|u(t)\|_V^2 + \frac{1}{2}|u'(t)|_H^2,$$

is absolutely continuous on $[0, T]$ *with*

$$E_u'(t) = -\langle B(t)u'(t), u'(t)\rangle_{W',W} \quad \text{a.e on } \{t,\ b(t) > 0\},$$

and

$$E_u'(t) = 0 \quad \text{a.e on } \{t,\ b(t) = 0\}.$$

For the proof of Theorem 2.1, we first need the following lemma.

LEMMA 2.1. *Let* $b = b(t) \geq 0$ *and consider*

$$u \in L^2(0, T; V) \cap W^{1,2}(0, T; H) \cap W^{2,2}(0, T; V')$$

with

$$bu' \in L^2(0, T; W),$$

where $V \subset W \subset H$ *with continuous and dense imbeddings. Let*

$$f = bg, \ \text{with } g \in L^2(0, T; W'),$$

and assume

$$u'' + Au = f \quad \text{in } L^2(0, T; V').$$

Then, in fact, $u \in C([0,T];V) \cap C^1([0,T];H)$ *and the energy* $E_u(t)$ *is absolutely continuous on* $[0,T]$ *with*

$$E_u'(t) = \langle g(t), b(t)u'(t)\rangle_{W',W}.$$

Proof of Lemma 2.1. Let $J_\lambda = (I+\lambda A)^{-1} : H \to H$ for $\lambda > 0$. We have $J_\lambda \in \mathcal{L}(H)$, $J_{\lambda|V} \in \mathcal{L}(V)$, and $J_{\lambda|W} \in \mathcal{L}(W)$ for $\lambda \le \lambda_0$. And by (2.1), $\{J_\lambda\}_{0<\lambda\le\lambda_0}$ is uniformly equicontinuous on $W \to W$.

We claim that

$$\forall \varphi \in W, \quad J_\lambda \varphi \to \varphi \ \text{ in } W \quad \text{as } \lambda \to 0. \tag{2.8}$$

Indeed (2.8) is well-known if $\varphi \in V$, and since V is dense in W, the result follows by density.

Then we introduce

$$u_\lambda := J_\lambda u \ \text{ and } \ f_\lambda := J_\lambda(bg) = bJ_\lambda g.$$

We clearly have $u_\lambda \in L^2(0,T;D(A))$, $u_\lambda' \in L^2(0,T;D(A))$, and $u_\lambda'' \in L^2(0,T;V) \subset L^2(0,T;H)$.

Setting

$$E_\lambda(t) := \frac{1}{2}\|u_\lambda(t)\|_V^2 + \frac{1}{2}|u_\lambda'(t)|_H^2,$$

we have, a.e. on $(0,T)$,

$$E_\lambda' = ((u_\lambda, u_\lambda'))_V + (u_\lambda', u_\lambda'')_H = (Au_\lambda + u_\lambda'', u_\lambda')_H = b(J_\lambda g, u_\lambda')_H,$$

with $J_\lambda g \in L^2(0,T;V) \subset L^2(0,T;H)$.

Let α, β be two points of $[0,T]$ such that $\alpha < \beta$. Then we have

$$E_\lambda(\beta) - E_\lambda(\alpha) = \int_\alpha^\beta (bJ_\lambda g, u_\lambda')\,ds = \int_\alpha^\beta (J_\lambda g(s), b(s)u_\lambda'(s))\,ds. \tag{2.9}$$

On the other hand, we can prove

$$J_\lambda g \to g \ \text{ in } \ L^2(0,T;W') \ \text{ as } \lambda \to 0 \tag{2.10}$$

and

$$bu_\lambda' = J_\lambda b u' \to bu' \ \text{ in } \ L^2(0,T;W) \ \text{ as } \lambda \to 0. \tag{2.11}$$

Indeed, for the first property, we notice that $J_\lambda \in \mathcal{L}(W')$ for $0 < \lambda \le \lambda_0$ with a uniformly bounded norm (by duality, from (2.1)). Then $J_\lambda g(t) \to g(t)$ as $\lambda \to 0$, in W' a.e. on $(0,T)$, and

$$\|J_\lambda g(t) - g(t)\|_{W'} \le C\|g(t)\|_{W'}.$$

From Lebesgue's theorem, it follows that

$$\|J_\lambda g - g\|_{W'}^2 \to 0 \quad \text{in} \quad L^1(0,T) \ \text{ as } \lambda \to 0,$$

which gives (2.10).

Next, from (2.8), we also have $J_\lambda b(t)u' \to b(t)u'$ as $\lambda \to 0$ in W a.e. on $(0,T)$, and by (2.1), we have

$$\|J_\lambda b(t)u'(t) - b(t)u'(t)\|_W \le C\|b(t)u'(t)\|_W.$$

From Lebesgue's theorem, it follows that

$$\|J_\lambda bu' - bu'\|_W^2 \to 0 \quad \text{in} \quad L^1(0,T) \quad \text{as} \quad \lambda \to 0,$$

which gives (2.11).

Now assume for a moment that α and β are *both* such that

$$[u(\alpha), u'(\alpha)] \in V \times H \quad \text{and} \quad [u(\beta), u'(\beta)] \in V \times H.$$

Then as $\lambda \to 0$, we can pass to the limit in (2.9) to obtain

$$E_u(\beta) - E_u(\alpha) = \int_\alpha^\beta (g(s), b(s)u'(s))_{W',W}\, ds. \tag{2.12}$$

Now let α be *fixed* for a while and apply (2.12) with $\beta = \beta_n \to t \in [0,T]$ as $n \to +\infty$. We obtain that $E(\beta_n)$ is bounded and therefore (replacing if necessary $(\beta_n)_n$ by a subsequence) we have

$$(u(\beta_n), u'(\beta_n)) \rightharpoonup (\varphi, \psi) \quad \text{weakly in } V \times H \quad \text{as } n \to +\infty.$$

On the other hand, by the regularity assumptions on u, we have

$$(u(\beta_n), u'(\beta_n)) \to (u(t), u'(t)) \quad \text{strongly in } H \times V' \quad \text{as } n \to +\infty.$$

It follows that $(u(t), u'(t)) = (\varphi, \psi)$ and therefore $u(t) \in V$ and $u'(t) \in H$. Since this is valid for *any* t, (2.12) becomes true for any (α, β).

Now the vector $Y(t) = (u(t), u'(t))$ is weakly continuous on $[0,T]$ and its norm is continuous by (2.12). The remainder of the proof is obvious from (2.12). □

Proof of Theorem 2.1. (i) Uniqueness. Let u and $\tilde{u}$ be two solutions with the same initial data. We have

$$u'' + Au + B(t)u' = 0 \quad \text{and} \quad \tilde{u}'' + A\tilde{u} + B(t)\tilde{u}' = 0.$$

Then $z := \tilde{u} - u$ satisfies

$$z'' + Az = B(t)u' - B(t)\tilde{u}',$$

with

$$bz' \in L^2(0,T;W)$$

and

$$B(t)u' - B(t)\tilde{u}' = bg, \quad g \in L^2(0,T;W').$$

From Lemma 2.1, we deduce

$$E_z'(t) = \langle g(t), b(t)z'\rangle_{W',W} = \begin{cases} -\langle B(t)\tilde{u}' - B(t)u', \tilde{u}' - u'\rangle_{W',W} & \text{if } b(t) > 0, \\ 0 & \text{if } b(t) = 0. \end{cases}$$

Thus $E_z(t) = \frac{1}{2}[\|z\|_V^2 + \|z'\|_H^2]$ is nonincreasing by (2.3). Since $E_z(0) = 0$, we obtain $z \equiv 0$, i.e., $\tilde{u} \equiv u$. □

(ii) Existence. We introduce, for $\psi \in W$,

$$C(t)\psi := \begin{cases} \frac{1}{b(t)} B(t) \left(\frac{\psi}{b(t)} \right) & \text{if } b(t) > 0, \\ 0 & \text{if } b(t) = 0. \end{cases}$$

It is clear that $C(t) \in \mathrm{Lip}\,(W, W')$ for all $t \in J$ and $\|C(t)\psi_1 - C(t)\psi_2\|_{W'} \le C\|\psi_1 - \psi_2\|_W$ for all $t \in J$ and $\psi_1, \psi_2 \in W$.

Next, for (u^0, u^1) given in $V \times H$ and for $0 < \lambda \le \lambda_0$, we solve

$$\begin{cases} u_\lambda'' + Au_\lambda + J_\lambda B(t) J_\lambda u_\lambda' = 0, \quad t \in J, \\ u_\lambda(0) = u_0, \quad u_\lambda'(0) = u_1. \end{cases} \tag{2.13}$$

We have

$$u_\lambda \in \mathcal{C}^0([0,T]; V) \cap \mathcal{C}^1([0,T]; H) \cap \mathcal{C}^2([0,T]; V'),$$

and

$$\int_0^T \langle B J_\lambda u_\lambda', J_\lambda u_\lambda' \rangle_{W',W}\, ds + E_\lambda(T) = E_\lambda(0) = \frac{1}{2}[\|u_0\|_V^2 + \|u_1\|_H^2] = E(0), \tag{2.14}$$

which is fixed. Equation (2.13) can also be written as

$$u_\lambda'' + Au_\lambda + b(t) J_\lambda C(t)(b(t) J_\lambda u_\lambda') = 0. \tag{2.15}$$

From (2.14) we deduce that $bJ_\lambda u_\lambda'$ is bounded in $\mathcal{W} := L^2(J; W)$. Thus

$$B(t) J_\lambda u_\lambda' = b(t) C(t)(b(t) J_\lambda u_\lambda') = b h_\lambda,$$

where h_λ is bounded in $\mathcal{W}' := L^2(J; W')$.

Finally, u_λ is bounded in $L^\infty(J; V) \cap W^{1,\infty}(J; H)$ and we may assume that there exists a subsequence such that

$$u_{\lambda_n} \rightharpoonup u \quad \text{weakly in } L^2(J; V) \cap H^1(J; H) \quad \text{as } n \to +\infty,$$

and

$$bJ_{\lambda_n} u_{\lambda_n}' \rightharpoonup z \quad \text{weakly in } \mathcal{W} \quad \text{as } n \to +\infty.$$

Since $u_{\lambda_n}' \rightharpoonup u'$ weakly in $\mathcal{H} := L^2(J; H)$ as $n \to +\infty$, we have (taking the inner product with some test function $\varphi \in H^1(J; D(A))$, for instance)

$$bJ_{\lambda_n} u_{\lambda_n}' \rightharpoonup bu' \quad \text{weakly in } \mathcal{H} \quad \text{as } n \to +\infty.$$

Therefore, $bu' = z \in \mathcal{W}$ and

$$bJ_{\lambda_n} u_{\lambda_n}' \rightharpoonup bu' \quad \text{weakly in } \mathcal{W} \quad \text{as } n \to +\infty.$$

On the other hand, we have (taking if necessary a subsequence)

$$C(t)(bJ_{\lambda_n} u_{\lambda_n}') \rightharpoonup \psi \quad \text{weakly in } \mathcal{W}' \quad \text{as } n \to +\infty.$$

It is not difficult, using a suitable test function, to check that

$$J_{\lambda_n} C(t)(bJ_{\lambda_n} u'_{\lambda_n}) \rightharpoonup \psi \quad \text{weakly in } \mathcal{W}' \quad \text{as } n \to +\infty,$$

so that, passing to the limit in (2.15),

$$u'' + Au + b\psi = 0 \quad \text{in } L^2(J; V'). \tag{2.16}$$

To obtain $b\psi = B(t)u'$, it remains to show that

$$\psi = C(t)bu'. \tag{2.17}$$

We introduce $\mathcal{C} : \mathcal{W} \to \mathcal{W}'$ defined by

$$\forall \varphi \in \mathcal{W}, \quad (\mathcal{C}\varphi)(t) := C(t)\varphi(t) \quad \text{a.e. on } J.$$

We remark that

$$\begin{aligned}\langle \mathcal{C}(bJ_\lambda u'_\lambda), bJ_\lambda u'_\lambda \rangle_{\mathcal{W}',\mathcal{W}} &= \int_0^T \langle b\mathcal{C}(bJ_\lambda u'_\lambda), J_\lambda u'_\lambda \rangle_{W',W}\, dt \\ &= \int_0^T \langle B(t)J_\lambda u'_\lambda, J_\lambda u'_\lambda \rangle_{W',W}\, dt = E(0) - E_\lambda(T).\end{aligned}$$

Whereas, due to Lemma 2.1, we have

$$E(T) + \langle \Psi, bu' \rangle_{\mathcal{W}',\mathcal{W}} = E(0).$$

Since

$$E(T) \le \liminf_{n\to+\infty} E_{\lambda_n}(T),$$

we obtain

$$\begin{aligned}E(0) - \langle \Psi, bu' \rangle_{\mathcal{W}',\mathcal{W}} &\le \liminf_{n\to+\infty} E_{\lambda_n}(T) \\ &= \liminf_{n\to+\infty} \left(E(0) - \langle \mathcal{C}(bJ_{\lambda_n} u'_{\lambda_n}), bJ_{\lambda_n} u'_{\lambda_n} \rangle_{\mathcal{W}',\mathcal{W}}\right) \\ &= E(0) - \limsup_{n\to+\infty} \langle \mathcal{C}(bJ_{\lambda_n} u'_{\lambda_n}), bJ_{\lambda_n} u'_{\lambda_n} \rangle_{\mathcal{W}',\mathcal{W}}.\end{aligned}$$

Hence

$$\limsup_{n\to+\infty} \langle \mathcal{C}(bJ_{\lambda_n} u'_{\lambda_n}), bJ_{\lambda_n} u'_{\lambda_n} \rangle_{\mathcal{W}',\mathcal{W}} \le \langle \Psi, bu' \rangle_{\mathcal{W}',\mathcal{W}}.$$

Then we can apply the following lemma.

LEMMA 2.2. *Let $\mathcal{W}$ be a Hilbert space and let $\mathcal{C} : \mathcal{W} \to \mathcal{W}'$ be monotone and Lipschitz continuous. Assume that $(z_n)_n$ is a sequence of $\mathcal{W}$ such that $z_n \rightharpoonup z$ weakly in $\mathcal{W}$ and $\mathcal{C}z_n \rightharpoonup \Psi$ weakly in $\mathcal{W}'$ as $n \to +\infty$.*

If

$$\limsup_{n\to+\infty} \langle \mathcal{C}z_n, z_n \rangle_{\mathcal{W}',\mathcal{W}} \le \langle \Psi, z \rangle_{\mathcal{W}',\mathcal{W}},$$

then $\Psi = \mathcal{C}z$

For the proof of this lemma, let $K : \mathcal{W}' \to \mathcal{W}$ be the duality map. Then $C := K\mathcal{C} : \mathcal{W} \to \mathcal{W}$ satisfies the assumptions of Proposition 2 of [8, p. 41]. (See also Brezis [5].)

This provides (2.17) and the proof of Theorem 2.1 is finished. ∎

3. Asymptotic stability.

3.1. An energy decay estimate on a short time interval. Assume that (2.1)–(2.5) hold. In order to study the asymptotic behavior of the energy, we first prove the following result, interesting in itself, concerning the *estimate of energy decay on a short interval of time*.

THEOREM 3.1. *Let $T > 0$ be fixed and assume that there exist $M, m > 0$ such that*

$$\forall t \in (0,T), \quad \forall v \in W, \quad \langle B(t)v, v\rangle_{W',W} \geq m\|v\|_W^2, \tag{3.1}$$

and

$$\forall t \in (0,T), \quad \forall v \in W, \quad \|B(t)v\|_{W'}^2 \leq M\langle B(t)v, v\rangle_{W',W}. \tag{3.2}$$

Then there exists $c > 0$ (independent of T) such that, for all $(u_0, u_1) \in V \times H$, the solution u of (2.6)–(2.7) satisfies

$$E(T) \;\leq\; \frac{1}{1 + c\,\frac{m}{T^{-3}+T^{-1}+MmT^{-1}}}\,E(0). \tag{3.3}$$

Theorem 3.1 is interesting in itself because it provides an estimate of the decay of the energy that is valid for t *small*. In particular, $E(t) < E(0)$. It has to be noted that, in general, estimates of the decay of the energy are provided for t *large enough*, even in the case of uniformly distributed damping terms. Of course if the damping is locally distributed in the domain, it is impossible to expect that $E(t) < E(0)$ for $t > 0$ small. (See, for example, [11, 12, 15, 16, 17, 18, 19, 23, 24] for classical estimates of the energy decay when t is large enough.)

Proof of Theorem 3.1. Following [7], we set

$$\theta(t) = t^2(T-t)^2.$$

Note that

$$\forall t \in [0,T], \quad |\theta'(t)| = |2t(T-t)(T-2t)| \leq 2T\theta^{1/2}(t), \tag{3.4}$$

$$\max_{t\in[0,T]} \theta(t) = \frac{T^4}{16}, \tag{3.5}$$

and

$$\int_0^T \theta(t)\,dt = \frac{T^5}{30}. \tag{3.6}$$

We also introduce $K_W, K'_W > 0$ such that

$$\forall v \in W, \quad K'_W|v|_H \leq \|v\|_W \leq K_W\|v\|_V = K_W|A^{1/2}v|_H. \tag{3.7}$$

First note that the energy of u is nonincreasing and satisfies

$$E(0) - E(T) = \int_0^T \langle Bu', u'\rangle_{W',W} \geq 0. \tag{3.8}$$

Multiplying (2.6) by θu, we obtain

$$\begin{aligned}\int_0^T \theta|A^{1/2}u|_H^2 &= -\int_0^T \theta\langle u''+Bu',u\rangle_{V',V}\\ &= \int_0^T ((\theta u)',u'))_H - \int_0^T \theta\langle Bu',u\rangle_{W',W}\\ &= \int_0^T \theta'(u,u')_H + \int_0^T \theta|u'|_H^2 - \int_0^T \theta\langle Bu',u\rangle_{W',W}\\ &\le \varepsilon\int_0^T \theta'^2(t)|u|_H^2 + \frac{1}{4\varepsilon}\int_0^T |u'|_H^2 + \int_0^T \theta(t)|u'|_H^2\\ &\quad + \eta\int_0^T \theta\|u\|_W^2 + \frac{1}{4\eta}\int_0^T \theta\|Bu'\|_{W'}^2,\end{aligned}$$

for all $\varepsilon,\eta>0$. Using (3.7), (3.4), (3.5) and (3.2) we deduce

$$\begin{aligned}\int_0^T \theta|A^{1/2}u|_H^2 &\le 4\frac{K_W^2}{{K'_W}^2}T^2\varepsilon\int_0^T \theta|A^{1/2}u|_H^2 + \frac{1}{4\varepsilon}\int_0^T |u'|_H^2\\ &\quad + \frac{T^4}{16}\int_0^T |u'|_H^2 + K_W^2\eta\int_0^T \theta|A^{1/2}u|_H^2 + \frac{T^4}{16}\frac{M}{4\eta}\int_0^T \langle Bu',u'\rangle_{W',W}.\end{aligned}$$

We choose ε and η such that

$$4\frac{K_W^2}{{K'_W}^2}T^2\varepsilon = K_W^2\eta = 1/4,$$

hence

$$\frac{1}{4\varepsilon} = 4\frac{K_W^2}{{K'_W}^2}T^2; \quad \frac{1}{4\eta} = K_W^2.$$

Thus we obtain

$$\begin{aligned}\frac12\int_0^T \theta|A^{1/2}u|_H^2 &\le \frac{1}{4\varepsilon}\int_0^T |u'|_H^2 + \frac{T^4}{16}\int_0^T |u'|_H^2 + \frac{T^4}{16}\frac{M}{4\eta}\int_0^T \langle Bu',u'\rangle_{W',W}\\ &= \left(4\frac{K_W^2}{{K'_W}^2}T^2 + \frac{T^4}{16}\right)\int_0^T |u'|_H^2 + \frac{K_W^2T^4M}{16}\int_0^T \langle Bu',u'\rangle_{W',W}.\end{aligned}$$

Hence,

$$\int_0^T \theta|A^{1/2}u|_H^2 \le F(T)\int_0^T |u'|_H^2 + \frac{K_W^2MT^4}{8}\int_0^T \langle Bu',u'\rangle_{W',W},$$

where

$$F(T) := 8\frac{K_W^2}{{K'_W}^2}T^2 + \frac{T^4}{8}.$$

Using

$$\forall t\in[0,T], \quad 2E(t) = |A^{1/2}u(t)|_H^2 + |u'(t)|_H^2 \ge 2E(T),$$

we deduce

$$\int_0^T \theta(2E(T) - |u'|_H^2) \le F(T) \int_0^T |u'|_H^2 + \frac{K_W^2 M T^4}{8} \int_0^T \langle Bu', u' \rangle_{W',W}.$$

Hence, using (3.5) and (3.7),

$$\begin{aligned} E(T) \int_0^T \theta &\le \frac{1}{2} \left(F(T) + \frac{T^4}{16} \right) \int_0^T |u'|_H^2 + \frac{K_W^2 M T^4}{16} \int_0^T \langle Bu', u' \rangle_{W',W} \\ &\le \frac{1}{K_W'^2} \left(4 \frac{K_W^2}{{K_W'}^2} T^2 + \frac{3T^4}{32} \right) \int_0^T \|u'\|_W^2 + \frac{K_W^2 M T^4}{16} \int_0^T \langle Bu', u' \rangle_{W',W}. \end{aligned}$$

Thus, using (3.6), there exists a constant $c > 0$ (independent of T) such that

$$E(T) \le \frac{1}{c}(T^{-3} + T^{-1}) \int_0^T \|u'\|_W^2 + \frac{1}{c} M T^{-1} \int_0^T \langle Bu', u' \rangle_{W',W}.$$

Using (3.1) and (3.8), we find

$$\begin{aligned} E(T) &\le \frac{1}{cm}(T^{-3} + T^{-1}) \int_0^T \langle Bu', u' \rangle_{W',W} + \frac{1}{c} M T^{-1} \int_0^T \langle Bu', u' \rangle_{W',W} \\ &= \frac{1}{cm}(T^{-3} + T^{-1} + MmT^{-1})(E(0) - E(T)). \end{aligned}$$

Hence

$$E(T) \ \le \ \frac{1}{1 + c \frac{m}{T^{-3} + T^{-1} + MmT^{-1}}} \ E(0). \qquad \square$$

3.2. A condition for asymptotic stability. Assume that (2.1)–(2.5) hold for any $T > 0$. Then (2.6)–(2.7) is well-posed and it follows from Theorem 3.1 that the result of Smith [27] may be extended to the case of problem (2.6)–(2.7).

THEOREM 3.2. *Consider a sequence $(I_n)_{n \ge 0}$ of disjoint open intervals in $(0, +\infty)$, denoted by $I_n = (a_n, b_n)$, where $b_n \le a_{n+1}$ for all $n \in \mathbb{N}$, and assume that, for all $n \ge 0$, there exist $M_n, m_n > 0$ such that*

$$\forall t \in I_n, \quad \forall v \in W, \quad \langle B(t)v, v \rangle_{W',W} \ge m_n \|v\|_W^2, \tag{3.9}$$

and

$$\forall t \in I_n, \quad \forall v \in W, \quad \|B(t)v\|_{W'}^2 \le M_n \langle B(t)v, v \rangle_{W',W}. \tag{3.10}$$

Assume that the following condition holds:

$$\sum_{n=0}^{\infty} m_n T_n \min \left(T_n^2, \frac{1}{1 + m_n M_n} \right) = +\infty, \tag{3.11}$$

where T_n denotes the length of I_n. Then (2.6)–(2.7) is asymptotically stable; i.e., for all $(u_0, u_1) \in V \times H$, the solution u of (2.6)–(2.7) satisfies $E_u(t) \to 0$ as $t \to \infty$.

Remark 1. Moreover, the proof of Theorem 3.2 also provides an estimate of the decay of the energy: if there exists $C > 0$ such that

$$\forall n \in \mathbb{N}, \quad u_n := m_n T_n \min \left(T_n^2, \frac{1}{1 + m_n M_n} \right) \le C,$$

then there exists $\omega > 0$ such that

$$\forall n \in \mathbb{N}, \forall t \geq b_n, \quad E(t) \leq E(0) \exp\left(- \omega \sum_{p=0}^{n} u_p \right).$$

Remark 2. Note that condition (1.2) implies condition (3.11). Note also that, in the case of ordinary differential equations, Pucci and Serrin [25] improved the condition of [27] and proved asymptotic stability under the following condition:

$$\sum_{n=0}^{\infty} m_n T_n \min\left(T_n^2, \frac{1}{1 + \frac{m_n}{T_n} \int_{I_n} a} \right) = +\infty.$$

We do not know if this weaker condition is also sufficient in the case of the partial differential equations (2.6).

Proof of Theorem 3.2. For all $n \geq 0$, we denote $I_n = (a_n, b_n)$ and we apply Theorem 3.1 to the time interval I_n instead of $(0, T)$, which implies

$$E(b_n) \leq \frac{1}{1 + ck_n} E(a_n),$$

where, for all $n \geq 0$,

$$k_n := \frac{m_n}{T_n^{-3} + T_n^{-1} + M_n m_n T_n^{-1}} > 0.$$

Using that the energy is nonincreasing, we deduce, for all $n \geq 0$,

$$\begin{aligned} E(a_{n+1}) \leq E(b_n) &\leq \frac{1}{1 + ck_n} E(a_n) \\ &\leq \left(\prod_{p=0}^{n} \frac{1}{1 + ck_p} \right) E(a_0) \leq \left(\prod_{p=0}^{n} \frac{1}{1 + ck_p} \right) E(0). \end{aligned}$$

Since the energy is nonincreasing, in order to prove Theorem 3.2, it is sufficient to prove that $E(a_{n+1}) \to 0$ as $n \to \infty$. Thus it is sufficient to prove that

$$\prod_{p=0}^{+\infty} \frac{1}{1 + ck_p} = 0, \quad \text{or} \quad \sum_{p=0}^{+\infty} \ln\left(\frac{1}{1 + ck_p} \right) = -\infty.$$

If $k_p \not\to 0$ as $p \to \infty$, then the result follows, and if $k_p \to 0$ as $p \to \infty$, then it reduces to prove that $\sum_{p=0}^{+\infty} k_p = +\infty$. This condition is equivalent to (3.11) since

$$\begin{aligned} \frac{1}{2} m_n T_n \min\left(T_n^2, \frac{1}{1 + m_n M_n} \right) \leq k_n &= \frac{m_n T_n}{\frac{1}{T_n^2} + 1 + m_n M_n} \\ &\leq m_n T_n \min\left(T_n^2, \frac{1}{1 + m_n M_n} \right), \end{aligned}$$

which ends the proof of Theorem 3.2. Note also that condition (1.2) implies condition (3.11) since

$$m_n T_n \delta_n^2 \leq m_n T_n \min\left(T_n^2, \frac{1}{1 + m_n M_n} \right),$$

which proves Remark 2. It remains to prove Remark 1. Using $k_p \geq u_p/2$, we have

$$E(b_n) \leq \prod_{p=0}^{n} \frac{1}{1+cu_p/2}\, E(0) = \exp\left(-\sum_{p=0}^{n} \ln\left(1+cu_p/2\right)\right) E(0).$$

Since $\ln(1+cx/2) \geq \ln(1+cC/2)x/C$ for all $x \in (0,C)$, we obtain

$$E(b_n) \leq \exp\left(-\frac{\ln(1+cC/2)}{C} \sum_{p=0}^{n} u_p\right) E(0). \qquad \square$$

3.3. Further comments. The main restrictive assumption of our general setting is that the damping term $B(t)u'$ is assumed to be *uniformly* distributed in space. However, this restriction is crucial if we want to consider an *on/off* damping. Our result does not apply to an on/off damping that is only *locally* distributed in space, even if the geometric condition of Bardos, Lebeau, and Rauch [2, 3] is satisfied.

Let us explain why the case of a locally distributed on/off damping, for example, $B(t)u' = a(t,x)u'$, is out of reach, at least under such a general form. Indeed, even for a very simple example, the situation is complicated and the statement of the results depends on a lot of parameters.

Let us consider the one-dimensional wave equation in $(0,1)$:

$$\text{(3.12)} \qquad \begin{cases} u'' - u_{xx} = -b(t)c(x)u', & x \in (0,1),\ t>0, \\ u(t,0) = u(t,1) = 0, & t>0, \\ u(0,\cdot) = u_0 \in H_0^1(0,1),\ u'(0,\cdot) = u_1 \in L_0^2(0,1). \end{cases}$$

Here we consider $a(t,x) = b(t)c(x)$ and we can distinguish three cases.

1. *The locally distributed (non on/off) case.* Our result does not apply to this case. Actually, it was not the purpose of the present paper, since this case has been widely studied in the literature. Let us recall some well-known results.

First, we consider the time-independent case:

$$a(t,x) = c(x), \quad \text{i.e., } b(t) \equiv 1,$$

with

$$c(x) \geq c_0 > 0 \quad \text{for all } x \in \omega,$$

where ω is an open subset of $(0,1)$. If ω is nonempty, then asymptotic stability holds. More generally, this result is well known in higher dimension spaces, provided that ω satisfies the geometric condition of Bardos, Lebeau, and Rauch [2, 3]. On the other hand, this kind of result may be extended to the time-dependent case

$$a(t,x) = b(t)c(x),$$

provided that

$$0 < \sigma(t) \leq b(t) \leq 1/\sigma(t),$$

with the condition

$$\int_0^{+\infty} \sigma(\tau)\, d\tau = +\infty.$$

See, for example, [20, 22, 26]. Notice that this allows us to consider a time-dependent damping, but not an on/off damping, since the assumption $b(t) > 0$ is needed.

2. *The uniformly distributed on/off case.* In the present paper, we consider a damping that is uniformly distributed in space, but is allowed to be on/off in time. For example, we assume

$$a(t,x) = b(t)c(x) \geq c_0 b(t),$$

with $c_0 > 0$ and where $b(t) = 0$ on an infinite union of intervals. In this case, Theorem 3.2 gives a sharp condition of asymptotic stability. (See section 4 for several examples of application of Theorem 3.2.)

3. *The locally distributed on/off case.* Now let us turn to the more general case of a locally distributed on/off damping, and let us see why its study is out of reach, at least in a general setting.

We consider the following "simple" example:

$$c(x) = \chi_\omega(x), \quad \text{where } \omega = (1/2 - \lambda, 1/2 + \lambda),$$

with $0 < \lambda \leq 1/2$, and $b : \mathbb{R}_+ \to \mathbb{R}_+$ is T-periodic such that

$$b(t) = 1 \quad \text{on} \quad [0,T) \quad \text{and} \quad b(t) = 0 \quad \text{on} \quad [T,2T).$$

Notice that, for all $0 < \lambda \leq 1/2$, ω satisfies the geometric condition of Bardos, Lebeau, and Rauch [2, 3]. However, this is not sufficient to insure asymptotic stability in this case. Indeed the following result holds.

THEOREM 3.3. [21, Theorem 2.3, p. 340].

(i) *If*

$$\left(\frac{1}{T} \in 2\mathbb{N} \quad \textit{and} \quad 2\lambda < T\right),$$

then there exists initial condition $(u_0, u_1) \in H_0^1(0,1) \times L^2(0,1)$ *such that the energy of the solution of* (3.12) *remains constant with time:* $E(t) = E(0) > 0$ *for all* $t > 0$.

(ii) *If*

$$\left(\frac{1}{T} \in 2\mathbb{N} \quad \textit{and} \quad 2\lambda > T\right), \quad \textit{or} \quad \left(\frac{1}{T} \notin 2\mathbb{N}\right),$$

then the energy of the solutions of (3.12) *decays uniformly exponentially to* 0.

This result shows that the situation is much more complicated because both difficulties (coming from the fact the damping is *locally* distributed in space and is *on/off* in time) are considered. Notice that Theorem 3.3 is proved in [21] without using the notion of optic rays. However, the results can be explained with the following comments related to optic rays propagation.

First consider the case of a uniformly distributed damping, i.e., $\lambda = 1/2$. In this case, Theorem 3.3 insures that asymptotic stability holds for all $T > 0$. Notice that, in this case, it is clear that each optic ray crosses the damping space region during a period when the damping is effective.

Next consider the more interesting case of a locally distributed damping: $0 < \lambda < 1/2$. Now there are some values of T and some values of λ for which *some optic rays cross the space damping region only when the feedback is not active.* For example, take $T = 1/2$, $\lambda < T/2 = 1/4$, and consider the optic ray that leaves the

point $x = T/2 = 1/4$ and that goes to the left (toward the point $x = 0$) at time $t = 0$. This ray describes the segment $[1/4, 3/4]$ (that contains the damping region) during the time intervals $[T, 2T]$, $[3T, 4T], \dots$, thus during periods when $b(t) = 0$. The same situation occurs as soon as $1/T \in 2\mathbb{N}$ with $2\lambda < T$. In these cases, Theorem 3.3 provides negative results of stability, while it provides positive results in the other cases.

We see that all these results are coherent with the optic ray condition known for time-independent feedbacks [2, 3]. But now, the fact that the feedback depends on time has to be taken into account. And it seems to be crucial that *each optic ray crosses the damping space region during a period when the damping is effective.*

4. Some examples. We first consider the case of *ordinary differential equations* that has been widely studied (see, for example, [1, 9, 10, 25, 27, 28]).

Example 1 (the oscillator equation). Assume $a \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+)$ is a nonnegative function whose minimum and maximum values in I_n are denoted by α_n and A_n.

With $H = V = D(A) = \mathbb{R}$, $Au = u$, and $B(t) = a(t)Id$, Theorem 3.2 applies to the oscillator equation (1.1) (with $m_n := \alpha_n$ and $M_n := A_n$). Since (1.2) implies condition (3.11), this again gives Theorem 1.1 with a small improvement of the sufficient condition. (In particular, this gives a new proof of the result of Smith, very different from the original proof based on monotonicity properties.)

Taking $B(t) = a(t)f$, we may also consider the nonlinear oscillator

$$u'' + u + a(t)f(u') = 0, \qquad t > 0,$$

where we assume that $f \in \mathcal{C}^1(\mathbb{R})$ is such that $f(0) = 0$ and $0 < \beta \le f' \le B$. Hence B defined by $B(t)v := a(t)f(v)$ satisfies (3.9) and (3.10) with $m_n := \beta\alpha_n$ and $M_n := BA_n$.

Next we assume that Ω is a bounded open set of $\mathbb{R}^N$ with regular boundary and we turn to the case of *uniformly damped partial differential equations* with a *bounded* damping operator.

Example 2 (a wave equation). Let $a_1,\ a_2 \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+)$ be two nonnegative functions such that *either*

$$a_2(t) = 0 \quad \text{a.e. on } \mathbb{R}_+$$

or

$$\exists C > 0, \quad a_1(t) \le C a_2(t) \quad \text{for a.e. } t \in \mathbb{R}_+$$

and

$$\forall n \in \mathbb{N}, \quad \alpha_n \le a_1(t) + a_2(t) \le A_n \quad \text{for a.e. } t \in I_n.$$

Consider also $f \in \mathbb{C}^1(\mathbb{R})$ such that $f(0) = 0$ and $0 < \beta \le f' \le B$.

Then we study the following damped wave equation:

$$\text{(4.1)} \qquad \begin{cases} u'' - \Delta u + a_1(t)f(u') - a_2(t)\Delta u' = 0, & x \in \Omega, t > 0, \\ u = 0, & x \in \partial\Omega, t > 0, \\ u(t=0) \in H^1_0(\Omega), \quad u'(t=0) \in L^2(\Omega). \end{cases}$$

We choose $H = L^2(\Omega)$, $A = -\Delta$ (with Dirichlet boundary conditions), $D(A) = H^2 \cap H^1_0(\Omega)$, and $B(t)v := a_1(t)f(v) - a_2(t)\Delta v$. The choice of W depends on a_2: in the case $a_2 \equiv 0$, $W = H$; otherwise $W = V$.

Consider $J = (0, T)$ with $T > 0$ and let us verify that the assumptions of Theorem 2.1 are satisfied. First (2.1) and (2.2) clearly hold. Then we write, for $w, z \in W$,

$$\langle B(t)w - B(t)z, w - z\rangle_{W',W} = a_1(t)\int_\Omega (f(w) - f(z))(w - z)\,dx + a_2(t)\int_\Omega |\nabla(w - z)|^2\,dx \geq 0,$$

which gives (2.3). On the other hand, for $w \in W$, we have

$$\langle B(t)w, w\rangle_{W',W} = a_1(t)\int_\Omega f(w)w\,dx + a_2(t)\int_\Omega |\nabla w|^2\,dx \geq a_1(t)\beta\|w\|_H^2 + a_2(t)\|w\|_V^2.$$

If $a_2 \equiv 0$, then $W = H$ and we choose $b(t) = \sqrt{\beta a_1(t)}$. In the other case, $W = V$ and we choose $b(t) = \sqrt{a_2(t)}$. Thus (2.4) is also satisfied. Finally, we write, for $w, z \in W$,

$$\|B(t)w - B(t)z\|_{W'}^2 \leq 2a_1(t)^2\|f(w) - f(z)\|_{W'}^2 + 2a_2(t)^2\|\Delta w - \Delta z\|_{W'}^2 \leq 2a_1(t)^2B^2\|w - z\|_H^2 + 2a_2(t)^2\|\Delta w - \Delta z\|_{W'}^2.$$

In the case $a_2 \equiv 0$, we deduce

$$\|B(t)w - B(t)z\|_H^2 \leq 2a_1(t)^2B^2\|w - z\|_H^2 \leq 2\frac{B^2}{\beta^2}b(t)^4\|w - z\|_H^2.$$

In the case $a_2 \not\equiv 0$, we deduce

$$\|B(t)w - B(t)z\|_{H^{-1}(\Omega)}^2 \leq 2[B^2C^2a_2(t)^2\|w - z\|_{L^2(\Omega)}^2 + a_2(t)^2\|w - z\|_{H_0^1(\Omega)}^2] \leq Kb^4(t)\|w - z\|_{H_0^1(\Omega)}^2,$$

where $K > 0$ is a constant. Thus (2.5) proved in both cases. Hence Theorem 2.1 insures the well-posedness of (2.6)–(2.7).

For the study of asymptotic stability, we verify that (3.9) and (3.10) are satisfied.

First case: $a_2 \equiv 0$ ($W = H$). We write for all $n \geq 0$, $t \in I_n$, $v \in W$,

$$(B(t)v, v)_H = a_1(t)\int_\Omega f(v)v\,dx \geq \alpha_n\beta\|v\|_H^2,$$

and

$$\|B(t)v\|_H^2 = a_1(t)^2\int_\Omega f(v)^2\,dx \leq A_nBa_1(t)\int_\Omega f(v)v\,dx = A_nB(B(t)v, v)_H.$$

Thus (3.9) and (3.10) are satisfied for $m_n := \beta\alpha_n$ and $M_n := BA_n$.

Second case: $a_2 \not\equiv 0$ ($W = V = H_0^1(\Omega)$). We write for all $n \geq 0$, $t \in I_n$, $v \in W$,

$$\langle B(t)v, v\rangle_{V',V} \geq a_2(t)\|w\|_V^2 \geq \alpha_n\|w\|_V^2,$$

and

$$\begin{aligned}\|B(t)v\|^2_{H^{-1}(\Omega)} &= \|a_1(t)f(v) - a_2(t)\Delta v\|^2_{H^{-1}(\Omega)}\\ &\le 2a_1(t)^2\int_\Omega f(v)^2\,dx + 2a_2(t)^2\|\Delta v\|^2_{H^{-1}(\Omega)}\\ &\le 2Ca_2(t)a_1(t)B\int_\Omega f(v)v\,dx + 2a_2(t)^2\int_\Omega \nabla v^2\,dx\\ &\le 2(BC+1)a_2(t)\left(a_1(t)\int_\Omega f(v)v\,dx + a_2(t)\int_\Omega \nabla v^2\,dx\right)\\ &\le 2(BC+1)A_n\langle B(t)v, v\rangle_{H^{-1}(\Omega),H^1_0(\Omega)}.\end{aligned}$$

Thus (3.9) and (3.10) are satisfied for $m_n := \alpha_n$ and $M_n := 2(BC+1)A_n$.

Applying Theorem 3.2, we deduce that (3.11) is a sufficient condition of asymptotic stability for (2.6).

Note that we may also consider the more general case

$$B(t)v := a_1(t)f(v) + a_2(t)(-\Delta)^{1/2}g((-\Delta)^{1/2}v).$$

Note also that in the particular case of a linear bounded damping $B(t)v := a(t)v$, this completes the work done in [21], where we studied the case of a locally damped wave equation with a periodic on/off damping.

Example 3 (some plate equations). In the same spirit, taking $H = L^2(\Omega)$, $A = \Delta^2$, $V = H^2_0(\Omega) = \{v \in H^2(\Omega) \mid v = 0 \text{ and } \frac{\partial v}{\partial \nu} = 0 \text{ on } \partial\Omega\}$ and $D(A) = H^4 \cap H^2_0(\Omega)$, we may consider the following damped plate equation:

$$\text{(4.2)}\quad \begin{cases} u'' + \Delta^2 u + a_1(t)f(u') - a_2(t)\Delta u' + a_3(t)\Delta g(\Delta u') = 0, & x \in \Omega, t > 0,\\ u = 0, \quad \dfrac{\partial u}{\partial \nu} = 0, & x \in \partial\Omega, t > 0,\\ u(t=0) = u_0 \in H^2_0(\Omega), \quad u'(t=0) = u_1 \in L^2(\Omega). \end{cases}$$

Here the damping operator is defined by

$$B(t)v := a_1(t)f(v) - a_2(t)\Delta v + a_3(t)\Delta g(\Delta v),$$

and the choice of W depends on the assumptions on the functions a_i. There are 3 cases : $W = H$, $W = H^1_0(\Omega)$ and $W = V$. For the applications of Theorems 2.1 and 3.2, we leave the details to the reader.

5. Another result: The case of a positive-negative damping. In this part, we study the case of a "positive-negative" damping. We assume that H, $D(A)$, and V are defined as in section 2 and that A still satisfies assumption (2.1) with $W = H$. And we consider a time-dependent operator B such that $B \in L^\infty_{\text{loc}}(\mathbb{R}_+, \text{Lip}(H))$. (Note that we only consider the case of a bounded operator B, so we assume in this part that $W = H$.)

Now we assume that B is a "positive-negative" feedback: let $(t_n)_{n\in\mathbb{N}}$ be a strictly increasing sequence of $\mathbb{R}^+$ such that $t_n \to +\infty$ as $n \to +\infty$. For all $n \in \mathbb{N}$, we define $I_{2n} := (t_{2n}, t_{2n+1})$ and $I_{2n+1} := (t_{2n+1}, t_{2n+2})$, and we assume that B is positive on I_{2n} and negative on I_{2n+1}. *Hence the energy decays on the time intervals I_{2n} and increases on the time intervals I_{2n+1}.*

We also assume that, for all $n \in \mathbb{N}$, there exist three positive constants m_{2n}, M_{2n}, M_{2n+1} such that

$$\text{(5.1)}\qquad \forall t \in I_{2n}, \quad \forall v \in H, \quad (B(t)v, v)_H \ \ge\ m_{2n}\|v\|^2_H,$$

$$\forall t \in I_{2n}, \quad \forall v \in H, \quad \|B(t)v\|_H^2 \ \leq \ M_{2n}(B(t)v, v)_H, \tag{5.2}$$

$$\forall t \in I_{2n+1}, \quad \forall v \in H, \quad -M_{2n+1}\|v\|_H^2 \ \leq \ (B(t)v, v)_H \ \leq \ 0. \tag{5.3}$$

Note that the well-posedness of (2.6)–(2.7) is classical (using standard arguments on Lipschitz perturbations of contraction semigroups).

Then, from Theorem 3.2, we deduce the following sufficient condition of asymptotic stability.

THEOREM 5.1. *Assume* (2.1), (5.1), (5.2), (5.3). *Assume that the following condition holds:*

$$\begin{cases} \sum\limits_{p=0}^{+\infty} M_{2p+1}T_{2p+1} < \infty, \\ \sum\limits_{p=0}^{+\infty} m_{2p}T_{2p}\delta_{2p}^2 = +\infty, \end{cases}$$

where T_p *denotes the length of* I_p *and* $\delta_p = \min(T_p, (1 + M_p)^{-1})$. *Then equation* (2.6)–(2.7) *is asymptotically stable; i.e., for all* $(u_0, u_1) \in V \times H$, *the solution* u *of* (2.6)–(2.7) *satisfies* $E_u(t) \to 0$ *as* $t \to \infty$.

Remark. This gives a result of stability in the case of a globally distributed time-dependent feedback of indefinite sign. This completes [21], where we studied the wave equation damped by a time-dependent boundary feedback of indefinite sign. See also [4, 6] for results in the case of space-dependent feedback of indefinite sign.

Proof of Theorem 5.1. For all $n \in \mathbb{N}$, applying Theorem 3.1 on the time intervals I_{2n}, we obtain

$$E(t_{2n+1}) \leq \frac{1}{1 + c\frac{m_{2n}}{T_{2n}^{-3} + T_{2n}^{-1} + M_{2n}m_{2n}T_{2n}^{-1}}} E(t_{2n}).$$

On the other hand, on the time intervals I_{2n+1}, we can write

$$0 \leq E'(t) = -\langle B(t)u', u'\rangle_{W',W} \leq M_{2n+1}\|u'(t)\|_H^2 \leq 2M_{2n+1}E(t).$$

Thus

$$E(t_{2n+2}) \leq E(t_{2n+1})e^{2M_{2n+1}T_{2n+1}}.$$

Hence

$$E(t_{2n+2}) \leq \left(\prod_{p=0}^{n} e^{2M_{2p+1}T_{2p+1}} \frac{1}{1 + c\frac{m_{2p}}{T_{2p}^{-3} + T_{2p}^{-1} + M_{2p}m_{2p}T_{2p}^{-1}}} \right) E(0).$$

In particular, we deduce a condition of asymptotic stability:

$$\sum_{p=0}^{+\infty} \left(2M_{2p+1}T_{2p+1} - \ln\left(1 + c\frac{m_{2p}}{T_{2p}^{-3} + T_{2p}^{-1} + M_{2p}m_{2p}T_{2p}^{-1}} \right) \right) = -\infty.$$

For example, we may assume

$$\begin{cases} \sum\limits_{p=0}^{+\infty} M_{2p+1}T_{2p+1} < \infty, \\ \sum\limits_{p=0}^{+\infty} m_{2p}T_{2p}\delta_{2p}^2 = +\infty, \end{cases}$$

or we may assume

$$\begin{cases} \sum_{p=0}^{+\infty} m_{2p} T_{2p} \delta_{2p}^2 = +\infty, \\ M_{2p+1} T_{2p+1} = o(\ln(1 + m_{2p} T_{2p} \delta_{2p}^2)). \end{cases} \qquad \square$$

REFERENCES

[1] Z. Artstein and E. F. Infante, *On the asymptotic stability of oscillators with unbounded damping*, Quart. Appl. Math., 34 (1976), pp. 195–199.

[2] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[3] C. Bardos, G. Lebeau, and J. Rauch, *Un exemple d'utilisation des notions de propagation pour le contrôle et la stabilisation de problèmes hyperboliques*, Nonlinear Hyperbolic Equations in Applied Sciences, Rend. Sem. Mat. Univ. Politec. Torino 1988, Special Issue (1989) 11–31.

[4] A. Benaddi and B. Rao, *Energy decay rate of wave equations with indefinite damping*, J. Differential Equations, 161 (2000), pp. 337–357.

[5] H. Brezis, *Équations et inéquations non linéaires dans les espaces vectoriels en dualité*, Extraits des annales de l'Institut Fourier de l'Université de Grenoble, Tome XVIII, Fascicule 1, 1968.

[6] P. Freitas and E. Zuazua, *Stability results for the wave equation with indefinite damping*, J. Differential Equations, 132 (1996), pp. 338–352.

[7] A. Haraux, *On a completion problem in the theory of distributed control of wave equations*, in Nonlinear Partial Differential Equations and Their Applications, College de France Seminar 1886, H. Brezis and J. L. Lions, eds., Res. Notes Math. 220, Pitman, Boston, 1991, pp. 241–271.

[8] A. Haraux, *Nonlinear Evolution Equations—Global Behavior of Solutions*, Lecture Notes in Math., 841, Springer-Verlag, Berlin, New York, 1981.

[9] L. Hatvani, *On the stability of the zero solution of second order nonlinear differential equations*, Acta Sci. Math., 32 (1971), pp. 1–9.

[10] L. Hatvani and V. Totik, *Asymptotic stability of the equilibrium of the damped oscillator*, Differential Integral Equation, 6 (1993), pp. 835-848.

[11] V. Komornik and E. Zuazua, *A direct method for boundary stabilization of the wave equation*, J. Math. Pures Appl., 69 (1990), pp. 33–54.

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

[13] V. Komornik and S. Kouémou-Patcheu, *Well-posedness and decay estimates for a Petrovsky system with internal damping*, Adv. Math. Sci. Appl., 7 (1997), pp. 245–260.

[14] S. Kouémou-Patcheu, *Stabilisation interne de certains systèmes distribués semi-linéaires*, Ph.D. thesis, University of Strasbourg, Strasbourg, France, 1995.

[15] J. Lagnese, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.

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

[17] I. Lasiecka and R. Triggiani, *Uniform exponential decay in a bounded region with $L_2(0,T;L_2(\Sigma))$-feedback control in the Dirichlet boundary condition*, J. Differential Equations, 66 (1987), pp. 340–390.

[18] I. Lasiecka and R. Triggiani, *Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometrical conditions*, Appl. Math. Optim., 25 (1992), pp. 189–224.

[19] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[20] P. Martinez, *Precise decay rate estimates for time-dependent dissipative systems*, Israel J. Math., 119 (2000), pp. 291–324.

[21] P. Martinez and J. Vancostenoble, *Stabilization of the wave equation by on/off and positive-negative feedbacks*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 335–377

[22] M. Nakao, *On the time decay of solutions of the wave equation with a local time-dependent nonlinear dissipation*, Adv. Math. Sci. Appl., 7 (1997), pp. 317–331.

[23] P. Pucci and J. Serrin, *Precise damping conditions for global asymptotic stability for nonlinear second order systems*, Acta Math., 170 (1993), pp. 275–307.

[24] P. Pucci and J. Serrin, *Precise damping conditions for global asymptotic stability for nonlinear second order systems,* II, J. Differential Equations, 113 (1994), pp. 505–534.

[25] P. Pucci and J. Serrin, *Asymptotic stability for intermittently controlled nonlinear oscillators*, SIAM J. Math. Anal, 25 (1994), pp. 815–835.

[26] P. Pucci and J. Serrin, *Asymptotic stability for nonautonomous dissipative wave systems*, Comm. Pure Appl. Math., XLIX (1996), pp. 177–216.

[27] R. A. Smith, *Asymptotic stability of* $x''+a(t)x'+x=0$, Quart. J. Math. Oxford (2), 12 (1961), pp. 123–126.

[28] L. H. Thurston and J. S. W. Wong, *On global asymptotic stability of certain second order differential equations with integrable forcing terms*, SIAM J. Appl. Math., 24 (1973), pp. 50–61.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2109–2131

ALGORITHMS FOR COUNTABLE STATE MARKOV DECISION MODELS WITH AN ABSORBING SET*

K. HINDERER† AND K.-H. WALDMANN†

Abstract. We consider a countable state Markov decision process with bounded reward function and an absorbing set. At first we generalize known properties and derive new properties of the critical discount factor, which is roughly defined as the maximal discount factor under which for V, the maximal expected infinite-stage discounted reward, there is guaranteed existence, boundedness, and computability by the successive approximation method. The emphasis of the paper is on algorithms for computing V exactly (recursion in state space and policy iteration) or approximately (value iteration combined with an extrapolation and finite state approximation). Our extrapolation method is motivated by and based on the Perron–Frobenius theory for nonlinear operators. As a by-product we obtain an efficient algorithm for determining the distribution of the entrance time of a Markov chain into an absorbing set. Further results concern asymptotically ε-optimal policies and a new turnpike theorem. Some of the results need tightness of the transition law, which turns out to be equivalent to compactness of a nonlinear operator, which is crucial for our study.

Key words. Markov decision processes, transient Markov decision processes, extrapolation methods, Perron–Frobenius theory, expected total reward criterion, absorbing sets

AMS subject classifications. 90C40, 90C59

DOI. 10.1137/S0363012902411027

1. Introduction.

We consider a Markov decision process (MDP) with countable state space S, countable action space A, finite subsets $D(s)$ of admissible actions in state $s \in S$, constraint set $D := \{(s,a) \mid s \in S,\ a \in D(s)\}$, stochastic transition law p from D into S, bounded reward function $r : D \to \mathbb{R}$, and (constant) discount factor $\beta > 0$. (In the literature the case $\beta > 1$ is often treated by replacing $\beta \cdot p$ by a nonnegative transition law; see, e.g., Veinott (1969), Zijm (1983).) For $(s,a) \in D$, $M \subset S$ we use $p(s,a,M) := \sum_{s' \in M} p(s,a,s')$. As usual, a deterministic Markov policy $\pi = (f_0, f_1, \dots)$ is a sequence $f_0, f_1, \dots \in F := \{f : S \to A \mid f(s) \in D(s) \text{ for all } s \in S\}$ of decision rules, each specifying action $a_n = f_n(s_n)$ to be taken in state s_n at time n. Thus F^∞ is the set of all deterministic Markov policies. Mainly one is interested in stationary policies $\pi = (f, f, \dots)$ for some $f \in F$, for which we also write f.

In economical applications MDPs with $\beta \le 1$ dominate. However, there are several *cases where $\beta > 1$ is appropriate:*

(a) Many MDPs with unbounded r and $\beta < 1$ can be reduced by means of a bounding function to an MDP' with bounded r' (see Remark 3.3(a)) and a discount factor β' which easily can be larger than one.

(b) There are applications where the nonnegative matrix $\beta \cdot p$ has the interpretation of an expectation, e.g., in the control of branching processes; cf. Sladky (1980), where five examples from the literature are reviewed, and McNamara (1991).

(c) Consider an MDP which subsumes a "time factor" and a "risk factor" under a discount factor $\beta(s,a)$ depending on the actual state and action and possibly larger than one for some (s,a). As shown in section 3 this case can be reduced to an MDP'

*Received by the editors July 15, 2002; accepted for publication (in revised form) August 17, 2004; published electronically April 14, 2005.
http://www.siam.org/journals/sicon/43-6/41102.html
†Institut für Mathematische Stochastik and Institut für Wirtschaftstheorie und Operations Research, Universität Karlsruhe, D-76128 Karlsruhe, Germany (karl.hinderer@math.uni-karlsruhe.de, waldmann@wior.uni-karlsruhe.de).

with discount factor $\beta' := \max_{(s,a)} \beta(s,a) > 1$. For an economic interpretation, cf. Branger (2001).

(d) Consider an MDP with discount factor $\alpha \leq 1$ which is nonstationary only because rewards earned in later periods have a larger utility, expressed by a "utility factor" $\gamma > 1$ per period. Then the reward $r(s_n, a_n)$ in period n enters the balance as $(\alpha\gamma)^n \cdot r(s_n, a_n)$, which of course can be treated as a stationary MDP with discount factor $\beta := \alpha\gamma$, possibly larger than one.

An important role will play *absorbing sets* $J_0 \subset S$, $J_0 \neq S$, defined by the requirement that $p(s, a, S - J_0) = 0$, $r(s,a) = 0$ for $s \in J_0$, $a \in D(s)$. Note that the empty set is always absorbing, but mainly we are interested in $J_0 \neq \emptyset$. An MDP may have several nonempty absorbing sets and the union of all absorbing sets is the largest absorbing set (provided that $r \not\equiv 0$). Throughout the paper J_0 denotes an arbitrary absorbing set. $J := S - J_0$ is called the *essential state space.*

Denote by $V(s)$ $[v_n(s)]$ the maximal expected infinite-stage [n-stage] reward when the process starts in $s \in S$ (and when a bounded terminal reward function v_0 is used), provided V exists.

The following property of our MDP with an absorbing set J_0 is of major interest.

(Pβ) *There exists $V(s)$ for $s \in S$, V is bounded, and successive approximation holds*, i.e., V can be approximated uniformly by v_n, starting with any bounded v_0 with $v_0 = 0$ on J_0.

One of the earliest results for standard MDPs (i.e., when J_0 is empty) says that (Pβ) holds whenever $\beta < 1$, and simple examples show that (Pβ) may fail for certain (not necessarily for all) r if $\beta \geq 1$. However, it has also been known for a long time that (Pβ) holds for $\beta = 1$ provided J_0 is nonempty and reached by the process with probability one for each initial state s; cf. Proposition 2.6(c_7).

Given an absorbing set J_0, a useful notion for a systematic study of the validity of (Pβ) in case $\beta \geq 1$ is the so-called *critical discount factor* $\beta^* = \beta^*(J_0)$, which we define as the supremum of those $\beta \in (0, \infty)$, for which (Pβ) holds for each choice of r (bounded and $r(s,a) = 0$ for $s \in J_0$, $a \in D(s)$). Due to Proposition 2.2(iii) below, this definition is consistent with the use in Hinderer and Waldmann (2003) for finite S and A (pp. 2 and 3) and Borel S and A (p. 13). (Note that in the latter paper V and v_∞ are denoted by V_∞ and V, respectively.) Obviously, $\beta^* = 1$ if $J_0 = \emptyset$. Furthermore, the largest absorbing set J_0 has the largest β^* (see Proposition 2.2(i)).

For finite S and A the number $\beta^* \in [0, \infty]$ is studied in detail in Hinderer and Waldmann (2003). In particular, it turned out that $1/\beta^* = \lambda^*$, the spectral radius of the nonlinear operator H associated with the transition law (cf. section 2) and, using a result in Rieder (1976), that this holds even for Borel S and A.

Characterizations of β^* are contained in many further papers on MDPs with an absorbing set. For finite S and A they are extensively studied in the classical paper of Veinott (1969). Extensions to a countable state space, Borel state space, compact sets of admissible actions, and/or unbounded rewards can be found, e.g., in Hordijk (1974), Pliska (1978), and Hernández-Lerma and Lasserre (1999). We also refer to Whittle (1983), Bertsekas (2000/2001), and the references given there.

The *main contributions* in the present paper (sections 4 through 10) concern algorithms for computing V, exactly or approximately, in case of $\beta < \beta^*$. This task motivated our choice of a countable state and action space: most numerical methods work only for countable or even finite S and A (e.g., turnpike theorems). For this reason we also did not present possible extensions of our results in sections 2 and 3 for more general S and A.

Our main results and the organization of the paper are as follows. Section 2 contains properties of H and β^* needed later on. In particular, characterizations of β^* and equivalent conditions for $\beta < \beta^*$, $\beta^* = 1$, and $\beta^* > 1$ are given. Some are generalizations to countable S and A and some are new even in the finite case. In particular, the special case of transient MDPs is characterized. Some of these and later results require tightness of the transition law (condition (A0), which turns out to be equivalent to compactness of H; cf. Proposition 2.1). Sometimes (e.g., for the characterization of β^* in Proposition 2.2(vi) as the largest of the maximal eigenvalues of linear operators H_f) we use spectral properties of H resulting from Ogiwara (1995). A brief derivation of standard results (optimality equation, optimality of a stationary policy, value iteration) and of the contraction of the optimal reward operator U follow in section 3. Our algorithms begin (section 4) with a recursion in the state space for obtaining V on a subset $\tilde{I}_k$ of J. We give a necessary and sufficient condition for $\tilde{I}_k = J$ and introduce an MDP' with (smaller) essential state space $J - \tilde{I}_k$ to which the methods of sections 5 and 7 through 9 can be applied. Our extrapolation methods are introduced and studied in great detail in section 5. Such methods have a long tradition in solving the standard MDP with $\beta < 1$. For MDPs having a transition law with unequal row sums there are a few references (including Hübner (1980)) only. One reason may be that the unequal row sums make a natural extension of the classical MacQueen extrapolation inefficient. On the other hand, an extrapolation method due to Schellhaas (1974), which is scarcely noticed in the literature, is more able to deal with unequal row sums. It forms the basis of our approach. Our results include sufficient conditions for the existence, monotonicity, and convergence of lower and upper bounds for V and lower bounds for "good" decision rules f, using again Perron–Frobenius results from Ogiwara (1995) and Bohl (1974). Our lower bounds are similar to those obtained in Hübner (1980); our upper bounds are new, also for $\beta < 1$. For V_f, the expected infinite-stage reward when using the stationary policy f, we obtain simpler bounds and insight into the speed of convergence of the bounds, as observed in numerical work. As a by-product we obtain (Example 5.6) an efficient method for determining the distribution $P_{fs}(\tau > n)$, $n \in \mathbb{N}$, of the entrance time of a Markov chain into an absorbing set. The paper proceeds with some numerical results in section 6. A slight improvement of the policy iteration is given in section 7. A finite state approximation, suggested in a natural way by our tightness assumption (A0), is considered in section 8. In particular, bounds and convergence results are given. In section 9 we deal with asymptotically ε-optimal policies, which can be combined with our extrapolation method. The paper closes with a turnpike theorem, which improves earlier results.

Notation. We use $\mathbb{N}_0$ ($\mathbb{N}$) to denote the set of all nonnegative (positive) integers. For $v, v' \in \mathbb{R}^J$ we write $v \le v'$ (resp., $v < v'$) if $v(s) \le v'(s)$ (resp., $v(s) < v'(s)$) holds for all $s \in J$. We write $\sum_i A_i$ for the union of disjoint sets A_i, and $A - B$ to denote $\{a \in A \mid a \notin B\}$.

2. The critical discount factor. Let $\tau := \inf\{n \in \mathbb{N} \mid \zeta_n \in J_0\} \le \infty$ denote the entrance time into J_0, i.e., the first time the state process (ζ_n) is in the absorbing set J_0, when the process starts in some state in J. Given a policy $\pi \in F^\infty$ and an initial state $s \in J$, the distribution of τ can be obtained by evaluating $P_{\pi s}(\tau > n)$, $n \in \mathbb{N}_0$, recursively, where $P_{\pi s}$ denotes the distribution of (ζ_n) and $E_{\pi s}$ the associated expectation.

The expected infinite-stage discounted reward up to absorption starting in $s \in J$

and following $\pi \in F^\infty$ is defined by

$$V_\pi(s) := E_{\pi s} \sum_{n=0}^{\tau} \beta^n r(\zeta_n, f_n(\zeta_n)) = E_{\pi s} \sum_{n=0}^{\infty} \beta^n r(\zeta_n, f_n(\zeta_n)),$$

provided that the expectation exists. Then a policy π^* is called optimal if $V_{\pi^*}(s) = V(s) := \sup_{\pi \in F^\infty} V_\pi(s)$, $s \in S$. We also say that a decision rule f^* is optimal if the associated stationary policy is optimal.

Let $\mathfrak{V}_0$ denote the set of all bounded functions on S which vanish on J_0. Note that $V_\pi \in \mathfrak{V}_0$, $\pi \in F^\infty$, and thus $V \in \mathfrak{V}_0$.

We also consider the expected N-stage discounted reward up to absorption, defined by

$$v_{N,\pi}(s) := E_{\pi s} \left(\sum_{n=0}^{N-1} \beta^n r(\zeta_n, f_n(\zeta_n)) 1_{[\tau > n]} + \beta^N v_0(\zeta_N) 1_{[\tau > N]} \right), \quad s \in S,$$

for $\pi \in F^\infty$ and terminal reward $v_0 \in \mathfrak{V}_0$, and put $v_N := \sup_{\pi \in F^\infty} v_{N,\pi}$.

Let $\mathfrak{V}$ denote the Banach space of all bounded functions on J (w.r.t. the supremum norm $\|v\| := \sup_{s \in J} |v(s)|$), $\mathfrak{V}_+ := \{v \in \mathfrak{V} \mid v \geq 0\}$, and $\mathfrak{W} := \{v \in \mathfrak{V} \mid \inf v > 0\}$. If $v \in \mathfrak{V}$ and $w \in \mathfrak{W}$, then $v/w \in \mathfrak{V}$ and the norm $\|v\|_w := \|v/w\|$ induced by w is equivalent to the supremum norm, since $\|v\| \leq \|v\|_w \cdot \|w\| \leq \|v\| \cdot (\|w\| / \inf w)$.

Where no confusion is possible, we also denote $V|J$, the restriction of V to J, by V, and analogously for V_f, $v_{N,\pi}$, and v_N.

On $\mathfrak{V}$ define the operators H, H_f, $f \in F$, well known from the literature, by

$$Hv(s) := \sup_{f \in F} H_f v(s) := \sup_{f \in F} \sum_{s' \in J} p(s, f(s), s') v(s'), \quad s \in J.$$

The sup is attained since $D(s)$ is finite for all s. H and H_f map $\mathfrak{V}$ and $\mathfrak{V}_+$ into itself. The operator H has the desirable properties of being monotone (i.e., $Hv \leq Hv'$ for $v \leq v'$), positively homogenous (i.e., $H(\lambda v) = \lambda H v$ for $\lambda \geq 0$), and subadditive (i.e., $H(v + v') \leq Hv + Hv'$). Moreover, H_f, $f \in F$, are linear and continuous.

For an arbitrary operator T from $\mathfrak{V}$ into $\mathfrak{V}$ (with $Tv \in \mathfrak{V}_+$ for $v \in \mathfrak{V}_+$) we use the following definitions according to Ogiwara (1995, pp. 47–49): (i) The operator norm is $\|T\| := \sup\{\|Tv\| \mid v \geq 0,\ \|v\| \leq 1\}$; (ii) $\lambda \in \mathbb{R}_+$ is called an eigenvalue of T if $Tv = \lambda v$ for some $v \in \mathfrak{V}_+$, $v \not\equiv 0$; (iii) T is compact if T is continuous and if the image under T of each bounded subset of $\mathfrak{V}_+$ is relatively compact. The latter property holds if and only if each sequence $(v_n)_1^\infty \subset \mathfrak{V}_+$ with $\|v_n\| \leq 1$ for all n has a subsequence (v_{n_k}) such that Tv_{n_k} converges in norm to some $v \in \mathfrak{V}_+$. (i) and (iii) are equivalent to the usual definition for continuous linear operators, while (ii) is narrower than the usual definition.

Let $D|_M := \{(s,a) \in D \mid s \in M\}$, $M \subset S$. Some of our results are based on the following assumption.

(A0) The set $(p(s,a,.),\ (s,a) \in D|_J)$ of substochastic measures on J is (uniformly) tight, i.e., for each $\varepsilon > 0$ there exists a finite set $K \subset J$ such that $\sup_{(s,a) \in D|_J} \{p(s,a,J-K)\} \leq \varepsilon$.

If S is finite, then (A0) holds trivially. Thus our results extend those in Hinderer and Waldmann (2003). For the classical problem of selling an asset (cf., e.g., DeGroot (1970)), (A0) can easily be shown to hold. A cash management system (cf., e.g., Hinderer and Waldmann (2001)), where the cash balance $s + a$ after transferring

money at the beginning of a period (always) belongs to a finite set $\{m^-, \dots, m^+\}$ (with $m^\pm \in \mathbb{Z}$), is a simple example with $J_0 = \emptyset$, for which (A0) holds. Similar in structure is the following more technical example with $J = \mathbb{N}_0$, $J_0 = \{-1\}$, $A = \mathbb{Z}$, $s' = \max\{-1, s + a + x\}$, where $P(X = x) > 0$, $x \in \mathbb{Z}$. Let $D(s) = \{\underline{a} - s, \dots, \overline{a} - s\}$, $s \in J$ (with $\underline{a}, \overline{a} \in \mathbb{N}$). Then, for all $\varepsilon > 0$, there exists $x_\varepsilon \in \mathbb{N}$ such that $P(X > x_\varepsilon) \le \varepsilon$. Finally, selecting $K_\varepsilon := \{0, \dots, \overline{a} + x_\varepsilon\}$, it follows that $p(s, a, J - K_\varepsilon) = P(s + a + X > \overline{a} + x_\varepsilon) \le P(X > x_\varepsilon) \le \varepsilon$, $(s, a) \in D|_J$.

If $J_0 = \emptyset$ (hence $p(s, a, .)$ are probability measures), condition (A0) can be expressed via metrics on the space of probability measures on J; see Dudley (1989, Theorem 11.5.4).

Given (A0), the operators H and H_f, $f \in F$, are compact. Indeed, building on Lemma 11.2 in Hordijk (1974), we obtain the next result. Its proof shows, by the way, that Lemma 11.2 cited before also holds for substochastic matrices $(p(i, j))$, provided tightness is defined as in (A0).

PROPOSITION 2.1. *H is compact if and only if* (A0) *holds. Then also H_f, $f \in F$, is compact.*

Proof. First note that H is continuous on $\mathfrak{V}$ (hence on $\mathfrak{V}_+$) whether or not (A0) is true, since $\|Hv - Hw\| \le \|v - w\|$, $v, w \in \mathfrak{V}$. (a) Assume (A0). Select $(v_n) \subset \mathfrak{V}_+$ and $\|v_n\| \le 1$, $n \in \mathbb{N}$. It follows as in the proof of Lemma 11.2 in Hordijk (1974) that given $\varepsilon > 0$ there exist $v_{n_k} \subset \mathfrak{V}_+$, $v^* \in \mathfrak{V}_+$ and $N \in \mathbb{N}$ such that for all $f \in F$ there holds $\|H_f v_{n_k} - H_f v^*\| \le 3\varepsilon$ for $k \ge N$. In particular, H_f is compact. Moreover, it follows that $\|Hv_{n_k} - Hv^*\| \le \sup_{f \in F} \|H_f v_{n_k} - H_f v^*\| \le 3\varepsilon$ for $k \ge N$; hence also H is compact.

(b) Assume that H is compact. Select finite sets $K_n \subset K_{n+1}$, $n \in \mathbb{N}$, with $\cup_n K_n = J$ and put $v_n = 1_{J - K_n}$. Then $v_n \in \mathfrak{V}_+$ and $\|v_n\| \le 1$, $n \in \mathbb{N}$, and $Hv_n(s) = \max_{a \in D(s)} p(s, a, J - K_n)$, $s \in S$, decreases to zero since $D(s)$ is finite. The convergence is uniform since H is compact. Thus, if $\varepsilon > 0$, there exists $n \in \mathbb{N}$ such that $\sup_{(s,a) \in D|_J} \{p(s, a, J - K_n)\} \le \varepsilon$. □

Let $H^{n+1} v := H(H^n v)$, $v \in \mathfrak{V}$, $n \in \mathbb{N}_0$. Analogously for H_f, $f \in F$. Let $e_0 := 1$ and

$$e_n(s) := H^n 1(s) = \sup_{\pi \in F^\infty} P_{\pi s}(\tau > n) \ge P_{fs}(\tau > n) = H_f^n 1(s) =: e_{n,f}(s), \quad s \in J.$$

Obviously $\|e_n\| = \|H^n 1\|$ is an upper bound for the probability that the process has not yet entered the absorbing set J_0 at time n. The asymptotic behavior of $\|e_n\|$ plays a key role in determining the critical discount factor β^*.

To begin with, let $\lambda_f^* := \inf_{k \in \mathbb{N}} \|e_{k,f}\|^{1/k}$, $f \in F$, be the spectral radius of the continuous linear operator H_f. Applied to H, introduce $\lambda^* := \inf_{k \in \mathbb{N}} \|e_k\|^{1/k}$.

PROPOSITION 2.2. *There holds the following:*

(i) $0 \le \lambda^* = \lim_{n \to \infty} \|e_n\|^{1/n} \le \|e_1\| \le 1$.

(ii) $\lambda^* = \lim_{n \to \infty} (\|H^n v\|_w)^{1/n}$ *for all* $v, w \in \mathfrak{W}$.

(iii) *The critical discount factor β^* is equal to $1/\lambda^*$.*

(iv) $0 \le \lambda_f^* \le \lambda^*$ *for all* $f \in F$.

Under (A0) *we additionally have the following:*

(v) *Let $\lambda^* > 0$. Then λ^* is an eigenvalue of H. Moreover, it is the largest one.*

(vi) $\lambda^* = \sup_{f \in F} \lambda_f^*$, *and the supremum is attained.*

Proof. (i) Based on the inequalities $0 \le \|e_{n+m}\| \le \|e_n\| \|e_m\| \le 1$ for $n, m \in \mathbb{N}_0$, existence of $\lim_{n \to \infty} \|e_n\|^{1/n}$ and equality with $\inf_{k \in \mathbb{N}} \|e_k\|^{1/k}$ can be shown as in the case of a linear operator (cf., e.g., Kato (1980, p. 27)). Thus (i) holds.

(ii) Part (ii) is an immediate consequence of $\|e_n\| \le \|H^n v\|/\inf v \le \|H^n v\|_w \cdot (\|w\|/\inf v) \le \|e_n\| \cdot (\|v\|/\inf v)(\|w\|/\inf w)$, $n \in \mathbb{N}$, and (i).

(iii) Note that $1/\lambda^*$ is the radius of convergence of the power series $\sum_{n=0}^{\infty} \|e_n\| x^n$. Thus $c := \sum_{n=0}^{\infty} \beta^n \|e_n\| < \infty$ for $\beta < 1/\lambda^*$. Now we get, using $\|r\| \le M$ for some $M < \infty$,

$$E_{\pi s} \sum_{n=0}^{\infty} \beta^n |r(\zeta_n, f_n(\zeta_n))| \le Mc < \infty.$$

Thus V exists and $|V(s)| \le Mc$; hence $V \in \mathfrak{V}_0$. Moreover,

$$|V_\pi(s) - v_{N,\pi}(s)| \le \alpha_N := M \cdot \sum_{n=N}^{\infty} \beta^n \|e_n\| + \beta^N \|e_N\| \|v_0\|.$$

Hence $\|V - v_N\| \le \alpha_N \to 0$ as $N \to \infty$, which proves that $1/\lambda^* \le \beta^*$. The proof of $1/\lambda^* \ge \beta^*$ is a copy of the proof of Theorem 2.7(a) in Hinderer and Waldmann (2003) with $b = 1_J$.

(iv) Part (v) results from the compactness, positive homogeneity, and monotonicity of H and Proposition 3.1.5 in Ogiwara (1995), observing that $\mathfrak{V}_+$ is a closed convex cone with interior point $v \equiv 1$ and that $0 \le v \le w$ implies $\|v\| \le \|w\|$.

(v) Fix $f \in F$. Recall that $e_{n,f} \le e_n$, $n \in \mathbb{N}$. Together with (i) we then obtain $0 \le \lambda_f^* \le \lambda^*$. Hence $0 \le \sup_{f \in F} \lambda_f^* \le \lambda^*$.

Now, if $\lambda^* = 0$, then (vi) holds trivially. Otherwise, using (v), λ^* is an eigenvalue of H and thus of H_g for some $g \in F$. Together with Proposition 3.1.7(iii) in Ogiwara (1995) for $T := H_g$ we then obtain $\lambda^* \le \lambda_g^*$, which completes the proof of Proposition 2.2(vi). □

It follows from Proposition 2.2(iii) that $\beta^* = 1$ if the MDP has (besides J_0) another absorbing set disjoint from J_0.

PROPOSITION 2.3. *The following are equivalent:*

(a_1) $\beta < \beta^*$.
(a_2) $\beta^m \|e_m\| < 1$ *for some* $m \in \mathbb{N}$.
(a_3) $\beta^n \|e_n\| \to 0$ *as* $n \to \infty$.
(a_4) $\sum_{n=0}^{\infty} \beta^n \|e_n\| < \infty$.
(a_5) $\beta^n \|e_n\| \le c \cdot \delta^n$ *for all* $n \in \mathbb{N}$ *and some* $c \in \mathbb{R}_+$ *and* $\delta \in (0,1)$.
(a_6) $\lim_{n\to\infty} \beta^n \|H^n v\|_w = 0$ *for all* $v \in \mathfrak{V}$, $w \in \mathfrak{W}$.

If (A0) *holds, then* (a_1) *is equivalent to*

(a_7) $\beta \lambda_f^* < 1$ *for all* $f \in F$.

Proof. Let $\beta < \beta^* = 1/\lambda^*$. Then, using Proposition 2.2(i), $\beta \|e_n\|^{1/n} \le \delta < 1$ for $n \ge N$, say. Hence (a_1) $\Rightarrow$ (a_5) $\Rightarrow$ (a_4) $\Rightarrow$ (a_3) $\Rightarrow$ (a_2). (a_2) implies $\beta\lambda^* = \inf_{k\in\mathbb{N}} \beta \|e_k\|^{1/k} < 1$. Hence ($a_2$) $\Rightarrow$ (a_1). Since $0 \le \beta^n \|H^n v\|_w \le (\|v\|/\inf w)\beta^n \|e_n\|$, ($a_3$) $\Rightarrow$ (a_6). Choosing $v = w = 1$, (a_6) $\Rightarrow$ (a_3). Under (A0) (a_1) $\Leftrightarrow$ (a_7) results from Proposition 2.2(vi). □

For later use denote by MDP_f, $f \in F$, the special MDP with $D(s) = \{f(s)\}$, $s \in S$. Let $r_f(s) := r(s, f(s))$, $f \in F$, $s \in S$.

Remark 2.4. (a) As in Hinderer, Waldmann (2003), condition (a_7) is the link to many additional equivalent properties. For example, under (A0) property (a_4) is equivalent to each of the following properties:

(a_4') $\sum_{n=0}^{\infty} \beta^n \|e_{n,f}\| < \infty$ for all $f \in F$.
(a_4'') $\sum_{n=0}^{\infty} \beta^n e_{n,f}(s) < \infty$ for all $f \in F$, $s \in J$.

Indeed, applying Proposition 2.3 to MDP_f, we find $\sum_{n=0}^{\infty} \beta^n \|e_{n,f}\| < \infty$ is equivalent to $\beta\lambda_f^* < 1$. Hence $(a_4) \Leftrightarrow (a_4')$. Clearly $(a_4') \Rightarrow (a_4'')$. On the other hand, fix $f \in F$ and let $\sum_{n=0}^{\infty} \beta^n e_{n,f} < \infty$. Then $\beta^n e_{n,f}$ converges to zero pointwise. Let $\gamma := \max\{1, \beta\}$. Given (A0), for all $0 < \varepsilon < 1/\gamma$, there exists a finite set K such that $p(s, f(s), J - K) \le \varepsilon$, $s \in J$. Moreover, there exists $n \in \mathbb{N}$ such that $\beta^{n+j} e_{n+j,f} \le \varepsilon$, $j \in \mathbb{N}_0$, on K. It follows for all $s \in J$ that $\beta^{n+1} e_{n+1,f}(s) = \beta H_f(\beta^n e_{n,f})(s) \le \beta[(1-\varepsilon)\varepsilon + \varepsilon\gamma^n] \le \gamma\varepsilon[(1-\varepsilon)(1-\gamma\varepsilon)^{-1} + \gamma^n]$. Induction on j then gives $\beta^{n+j} e_{n+j,f} \le \gamma\varepsilon(1-\varepsilon)(1-\gamma\varepsilon)^{-1} + (\gamma\varepsilon)^j \gamma^n$, $j \in \mathbb{N}$. Hence $\beta^n \|e_{n,f}\| \to 0$, which is equivalent to $\sum_{n=0}^{\infty} \beta^n \|e_{n,f}\| < \infty$ by Proposition 2.3 (applied to MDP_f). Thus $(a_4'') \Rightarrow (a_4')$.

(b) Pliska (1978) treats a semicontinuous model and does not suppose (A0). He shows $\sup_{f \in F} \sum_{n=0}^{\infty} \beta^n \|e_{n,f}\| < \infty \Leftrightarrow \sup_{f \in F} \beta^m \|e_{m,f}\| < 1$ for some $m \in \mathbb{N} \Leftrightarrow \sup_{f \in F} \beta^n \|e_{n,f}\| \le \alpha\gamma^n$, $n \in \mathbb{N}$, for some positive numbers α and γ with $\gamma < 1$. Moreover, each of these properties is shown to imply $\sup_{f \in F} \beta\lambda_f^* < 1$.

(c) Veinott (1969), Hordijk (1974), and Pliska (1978) show for different state and action spaces that $\sup_{f \in F} \|\sum_{n=0}^{\infty} \beta^n H_f^n 1\| < \infty$ implies that $\sup_{\pi \in F^\infty} \|\sum_{n=0}^{\infty} \beta^n H_\pi^n 1\| < \infty$, where $H_\pi^0 v := v$, $H_\pi^n := H_{f_0}, \dots, H_{f_{n-1}}$, $n \in \mathbb{N}$, for $\pi = (f_0, f_1, \dots) \in F^\infty$, which for our model also follows from $(a_4'') \Leftrightarrow (a_4)$. Hernández–Lerma and Lasserre (1999) suppose $\sup_{\pi \in F^\infty} \|b^{-1} \sum_{n=0}^{\infty} H_\pi^n b\| < \infty$ in a transient model with unbounded reward function and weight function b as in Remark 3.3(a). Their assumption is equivalent to $\sup_{\pi \in F^\infty} \|\sum_{n=0}^{\infty} \hat{\gamma}^n \hat{H}_\pi^n 1\| < \infty$ in the reduced model (provided that $0 < \hat{\gamma} < \infty$). Further, given (A0), the last property is equivalent to (a_4) (in the reduced model with $\beta = \hat{\gamma}$).

(d) Proposition 2.2(iii) as well as the equivalence of (a_1)–(a_4) of Proposition 2.3 are obtained in Hinderer and Waldmann (2003), too, for a model with standard Borel state and action spaces.

Next we distinguish between $\beta^* = 1$ and $\beta^* > 1$.

PROPOSITION 2.5. *Assume* (A0). *Then the following are equivalent:*

(b_1) $\beta^* = 1$.

(b_2) *There exists* $s_0 \in J$ *such that* $e_n(s_0) = 1$ *for all* $n \in \mathbb{N}$.

(b_3) *There exist* $M \subset J$, $M \neq \emptyset$, *and* $f \in F$ *such that* $p(s, f(s), M) = 1$ *for all* $s \in M$.

(b_4) *There exist* $f \in F$ *and* $s_0 \in J$ *such that* $P_{fs_0}(\tau = \infty) = 1$.

Proof. (i) Let $\beta^* = 1$. Then $\|e_n\| = 1$, $n \in \mathbb{N}$, by Proposition 2.2(iii). Since λ^* is an eigenvalue of H (cf. Proposition 2.2(v)), there exists $e^* \ge 0$ with $\|e^*\| = 1$ and $He^* = e^*$. By Proposition 2.1 there exists a finite $K \subset J$ such that $p(s, a, J-K) \le 1/2$ for all $(s, a) \in D|_J$. $K \neq \emptyset$, since otherwise $e_1(s) = \max_{a \in D(s)} p(s, a, J) \le 1/2$ for $s \in J$, which contradicts $\|e_1\| = 1$. Assume $e^* \le \alpha$ on K for some $\alpha \in (0, 1)$. Then, for all $f \in F$, $s \in J$, we get $H_f e^*(s) \le \alpha p(s, f(s), K) + p(s, f(s), J - K) \le \alpha(1 - p(s, f(s), J-K)) + p(s, f(s), J-K) \le \alpha + (1-\alpha)/2$. Thus $e^*(s) = \sup_{f \in F} H_f e^*(s) \le \alpha + (1 - \alpha)/2$, which contradicts $\|e^*\| = 1$. Thus $e^*(s_0) = 1$ for some $s_0 \in K$ and, since $e^* \le e_n \le 1$ for $n \in \mathbb{N}$, $e_n(s_0) \ge e^*(s_0) = 1$, $n \in \mathbb{N}$. This verifies $(b_1) \Rightarrow (b_2)$.

(ii) Assume $(b2)$. Then $\|e_n\| = 1$, $n \in \mathbb{N}$. Hence we can use e^* from (i), which implies that $M := \{s \in J \mid e^*(s) = 1\}$ is nonempty. There exists $f \in F$ such that $1 = e^*(s) = He^*(s) = H_f e^*(s)$, $s \in M$. Fix $s \in M$ and assume $p(s, f(s), M) < 1$ in contrast to $(b3)$. By (A0) there exists a finite $K \subset J$ such that $p(s, f(s), M) + p(s, f(s), J - K) < 1$. Put $\alpha := \max\{e^*(s') \mid s' \in (J - M) \cap K\}$ if $(J - M) \cap K \neq \emptyset$

and $\alpha := 0$ else. Then

$$\begin{aligned} 1 = H_f e^*(s) &= p(s, f(s), M) + \sum_{s' \in J-M} p(s, f(s), s') e^*(s') \\ &\le p(s, f(s), M) + \alpha p(s, f(s), (J-M) \cap K) + p(s, f(s), (J-M) \cap (J-K)) \\ &\le 1 - (1-\alpha) p(s, f(s), (J-M) \cap K). \end{aligned}$$

This implies $p(s, f(s), (J-M) \cap K) = 0$, since $\alpha < 1$, and hence $1 \le p(s, f(s), M) + p(s, f(s), J-K) < 1$, which is a contradiction. Hence $(b_2) \Rightarrow (b_3)$.

(iii) Let $M \subset J$, $f \in F$ such that $p(s, f(s), M) = 1$, $s \in M$. Then $e_{1,f}(s) \ge p(s, f(s), M) = 1$, $s \in M$. Thus suppose $e_{n,f} = 1$ on M to hold for some $n \in \mathbb{N}$. Then $e_{n+1,f}(s) = \sum_{s' \in M} p(s, f(s), s') e_{n,f}(s') + \sum_{s' \in J-M} p(s, f(s), s') e_{n,f}(s') \ge p(s, f(s), M) = 1$ for all $s \in M$, which gives $e_{n,f} = 1$ on M (and thus $\|e_{n,f}\| = 1$) for all $n \in \mathbb{N}$. Hence $P_{fs}(\tau = \infty) = \lim_{n\to\infty} P_{fs}(\tau > n) = \lim_{n\to\infty} e_{n,f}(s) = 1$, $s \in M$. This proves $(b_3) \Rightarrow (b_4)$. Assume (b_4). From Proposition 2.2(i) applied to MDP_f, we then obtain $\lambda_f^* = 1$. Finally, using Proposition 2.2(iv) and $\lambda^* \le 1$ (by Proposition 2.2(i)), we have $\lambda^* = 1$. Hence $(b_4) \Rightarrow (b_1)$, which completes the proof. □

PROPOSITION 2.6. *The following are equivalent:*

(c_1) $\beta^* > 1$.

(c_2) *Each of the statements* (a_2)–(a_6) *of Proposition* 2.3 *with* $\beta = 1$.

Moreover, under (A0), (c_1) *is equivalent to each of the following conditions:*

(c_3) *There is a finite partition* $J_1, \dots, J_m$ *of* J *with nonempty sets*

$$J_k := \{s \in J \mid e_k(s) < 1 = e_{k-1}(s)\}, \quad 1 \le k \le m.$$

(Thus $e_m < 1$, *and* $m \le |J|$ *if* J *is finite.)*

(c_4) $E_{fs}(\beta^\tau) < \infty$ *for all* $f \in F$, $s \in J$, *and some* $\beta > 1$.

(c_5) $E_{fs}(\tau^k) < \infty$ *for all* $f \in F$, $s \in J$, $k \in \mathbb{N}$.

(c_6) $E_{fs}(\tau) < \infty$ *for all* $f \in F$, $s \in J$.

(c_7) $P_{fs}(\tau < \infty) = 1$ *for all* $f \in F$, $s \in J$.

(c_8) $\lambda_f^* < 1$ *for all* $f \in F$.

(c_9) *The MDP is transient, i.e.,* $\sum_{n=0}^\infty \sup_{s\in J} E_{\pi s} 1_J(\zeta_n) < \infty$ *for* $\pi \in F^\infty$.

Proof. (i) $(c_1) \Leftrightarrow (c_2)$ and, given (A0), $(c_1) \Leftrightarrow (c_8)$ are obvious from Proposition 2.3.

(ii) If $\beta^* > 1$, then $\|e_n\| < 1$ (hence $e_n < 1$) by Proposition 2.3(a_2) with $\beta = 1$ for some $n \in \mathbb{N}$. Put $M_\nu := \{s \in J \mid e_\nu(s) = 1\}$, $\nu \in \mathbb{N}_0$. Then $M_n = J - \sum_{k=1}^n J_k = \emptyset$ and since $M_{\nu+1} \subset M_\nu$, $\nu \in \mathbb{N}_0$, there exists a smallest $m \le n$ such that $M_n = \emptyset$. Then $J_1, \dots, J_m$ is a partition of J. Moreover, $J_k \ne \emptyset$ for $1 \le k \le m$. For a proof assume that $J_k = M_{k-1} - M_k = \emptyset$ for some $1 \le k \le m$. First $M_{k-1} \ne \emptyset$. Then an argument similar to part (ii) of Proposition 2.5 with M and e^* replaced by M_{k-1} and e_{k-1}, respectively, and exploiting $1 = H_f e_{k-1}(s)$, $s \in M_{k-1}$, for some $f \in F$ shows that $p(s, f(s), M_{k-1}) = 1$ for all $s \in M_{k-1}$. Then $\beta^* = 1$ by Proposition 2.5(b_3), which is a contradiction. Altogether we have $(c_1) \Rightarrow (c_3)$.

(iii) Assume (c_3). We prove that $\|e_{m+1}\| < 1$. (Then (c_1) holds by Proposition 2.3(a_2) with $\beta = 1$.) By (A0) there exists a finite $K \subset J$ such that $p(s, f(s), J - K) \le 1/2$ for $f \in F$, $s \in J$. Since $\alpha := \max_{s' \in K}\{e_m(s')\} < 1$ we get as in part (ii) of the proof of Proposition 2.5 $H_f e_m(s) \le \alpha p(s, f(s), K) + p(s, f(s), J-K) \le (\alpha+1)/2$. Thus $e_{m+1}(s) = \sup_{f\in F} H_f e_m(s) \le (\alpha+1)/2 < 1$ for $s \in J$.

(iv) Assume (c_1). Then there exists some $\beta \in (1, \beta^*)$. Using Proposition 2.3(a_4), we get (cp. (2.9) in Hinderer, Waldmann (2003))

$$\infty > \sum_{n=0}^{\infty} \beta^n \|e_n\| \geq \sum_{n=0}^{\infty} \beta^n e_{n,f}(s) = P_{fs}(\tau = \infty) + (\beta - 1)^{-1} \left[E_{fs}(\beta^\tau) - 1\right]$$

for $s \in J$, $f \in F$, which implies $E_{fs}(\beta^\tau) < \infty$. Hence $(c_1) \Rightarrow (c_4)$. Since $E_{fs}(\beta^\tau) \geq ((\ln \beta)^k / k!) E_{fs}(\tau^k)$ for $k \in \mathbb{N}$, we next obtain $(c_4) \Rightarrow (c_5) \Rightarrow (c_6) \Rightarrow (c_7)$.

(v) (c_7) implies $\lim_{n\to\infty} P_{fs}(\tau > n) = \lim_{n\to\infty} e_{n,f}(s) = 0$, $s \in J$, $f \in F$. H_f is compact by Proposition 2.1. Thus $(e_{n_k+1,f})_1^\infty = (H_f e_{n_k,f})_1^\infty$ converges in norm to 0 for some $(n_k) \subset \mathbb{N}$. Since $(\|e_{n,f}\|)$ is decreasing, we get $\lim_{n\to\infty} \|e_{n,f}\| = 0$. Applying Proposition 2.3(a_3) and (a_7) with $\beta = 1$ to MDP_f, it follows that $\lambda_f^* < 1$. Finally, from Proposition 2.3(a_7) and (a_1), $\beta^* > 1$. Hence $(c_7) \Rightarrow (c_1)$.

(vi) Condition (a_4) from Proposition 2.3 with $\beta = 1$ implies (c_9), since $E_{\pi s} 1_J(\zeta_n) \leq \|e_n\|$. Moreover, $(c_9) \Rightarrow (a_4') \Rightarrow (a_4)$ by Remark 2.4(a) with $\beta = 1$, since $\sup_{s\in J} E_{fs} 1_J(\zeta_n) = \|e_{n,f}\|$, which completes the proof. ∎

3. The optimality equation and related results. For all $s \in J$, $a \in D(s)$, $v \in \mathfrak{V}$ introduce

$$Lv(s,a) := r(s,a) + \beta \sum_{s' \in J} p(s,a,s') v(s'),$$
$$Uv(s) := \sup_{f \in F} U_f v(s) := \sup_{f \in F} Lv(s, f(s)).$$

The supremum is attained, since $D(s)$ is finite for all $s \in S$. Moreover, $U_f v = r_f + \beta H_f v$ and $U(v + w) \leq Uv + \beta H w$ for $v, w \in \mathfrak{V}$.

THEOREM 3.1. *Assume $\beta < \beta^*$. Then the following occur.*

(i) *Value iteration works; i.e., for all $v_0 \in \mathfrak{V}$ it holds that $v_n := U^n v_0$ converges in norm to V.*

(ii) *V is the unique bounded solution of the optimality equation*

$$V(s) = \max_{a \in D(s)} \left\{ r(s,a) + \beta \sum_{s' \in J} p(s,a,s') V(s') \right\}, \quad s \in J.$$

(iii) *f is optimal if and only if f is a maximizer of LV (i.e., $UV(s) = LV(s, f(s))$, $s \in J$); thus there exists an optimal stationary policy.*

Proof. (i) is obvious from Proposition 2.2(iii).

(ii) Let $\beta < \beta^*$. Then $\beta^m \|e_m\| < 1$ for some $m \geq 1$ by Proposition 2.3, hence U^m is contractive (w.r.t. the sup-norm) on the Banach space $\mathfrak{V}$. Moreover, the unique fixed point v^*, say, of U^m is also the unique fixed point of U. First, $Uv^* = U^{m+1} v^* = U^m (Uv^*)$ implies $Uv^* = v^*$. Second, $Uw^* = w^*$ results in $U^m w^* = w^*$ and thus $w^* = v^*$.

(iii) follows (i) as usual, using (ii) for MDP_f. ∎

If $\beta < \beta^*$, then there exists $m \in \mathbb{N}$ such that $\eta := 1 - \beta^m \|e_m\| > 0$ by Proposition 2.3(a_2). Then $w := \eta^{-1} \sum_{j=0}^{m-1} \beta^j e_j \geq 1/\eta$, hence $w \in \mathfrak{W}$, and we have $1 - 1/w(s) \leq M$, $s \in J$, for some $M \in (0,1)$. Now, $1 + \beta H w - w \leq e_0 + \eta^{-1}(\beta^m e_m - e_0) = \beta^m (e_m - \|e_m\|)/\eta \leq 0$ by subadditivity of H, which implies $\beta H w / w \leq 1 - 1/w \leq M$. Hence $\rho := \beta \|Hw\|_w < 1$. We then get the contraction property $\|Uv - Uv'\|_w \leq \beta \sup_{f\in F} \|H_f(v - v')/w\| \leq \beta \|Hw\|_w \|v - v'\|_b$ for $v, v' \in \mathfrak{V}$.

As a by-product, we get a short constructive proof for a result due to van Hee and Wessels (1978) in case of $\beta \le 1 < \beta^*$; see also Zijm (1978).

PROPOSITION 3.2. *Let $\beta < \beta^*$ and $\varepsilon > 0$ such that $\beta\lambda^* + \varepsilon < 1$. Then there exists some $\bar{w} \in \mathfrak{W}$ such that U is contractive with module $\beta\|H\bar{w}\|_{\bar{w}} \le \beta\lambda^* + \varepsilon$.*

Proof. Since $\gamma := \beta/(\beta\lambda^* + \varepsilon) < \beta^*$, we can replace β by γ in the construction of w above. Then we get $\gamma\|H\bar{w}\|_{\bar{w}} < 1$ and thus $\beta\|H\bar{w}\|_{\bar{w}} < \beta\lambda^* + \varepsilon$. □

Now introduce $\gamma^{\pm} := \pm\sup\{\pm r(s,a)/w(s) \mid (s,a) \in D\}$. Note that $\gamma^- \le 0 \le \gamma^+$. Then, based on w, $u^{\pm} := \gamma^{\pm}(1-\rho)^{-1}w$ fulfill $\pm Uu^{\pm} \le \pm u^{\pm}$ and thus initiate monotone sequences of iterates converging to V by Theorem 3.1(i). Hence they define rough bounds $u^- \le Uu^- \le V \le Uu^+ \le u^+$ for V on J and, additionally, may be used as an initial value v_0 for a monotone sequence of iterates.

If J is finite, using $\tilde{J}_0 := \emptyset$, $\tilde{J}_k := J_1 \cup \cdots \cup J_k$, $1 \le k \le m$, the partition $J_1, \ldots, J_m$ of J can be rewritten as

$$J_k = \{s \in J - \tilde{J}_{k-1} \mid p(s,a,J_0 + \tilde{J}_{k-1}) \ge \varepsilon_k \text{ for all } a \in D(s)\}$$

for suitable constants $\varepsilon_1, \ldots, \varepsilon_m \in (0,1)$. This fact can be used to construct a constant

$$\rho := 1 - \min_{1\le k\le m} \frac{\varepsilon_1 \ldots \varepsilon_m(1-\varepsilon_k)}{1-\varepsilon_1\ldots\varepsilon_k} \in (0,1) \tag{3.1}$$

and a function $w \in \mathbb{R}^J$, $w(s) := 1 - \prod_{j=1}^{k}\varepsilon_j$ for all $s \in J_k$, $1 \le k \le m$, for which it holds that $w > 0$ and $Hw \le \rho w$.

For a proof first note that w is constant on J_k and $0 < w(s) \le w(s') \le 1$ for $s \in J_k$, $s' \in J_{k'}$, $1 \le k \le k' \le m$. Hence, for $s \in J_k$, $1 \le k \le m$, we have

$$\begin{aligned}
Hw(s) &= \max_{a\in D(s)} \left\{ \sum_{\nu=1}^{k-1} (1-\varepsilon_1\ldots\varepsilon_\nu) p(s,a,J_\nu) + \sum_{\nu=k}^{m} (1-\varepsilon_1\ldots\varepsilon_\nu) p(s,a,J_\nu) \right\} \\
&\le \max_{a\in D(s)} \{(1-\varepsilon_1\ldots\varepsilon_{k-1}) p(s,a,J_0+\tilde{J}_{k-1}) \\
&\qquad + (1-\varepsilon_1\ldots\varepsilon_m)(1-p(s,a,J_0+\tilde{J}_{k-1}))\} \\
&\le (1-\varepsilon_1\ldots\varepsilon_{k-1})\varepsilon_k + (1-\varepsilon_1\ldots\varepsilon_m)(1-\varepsilon_k) \\
&= \frac{1-\varepsilon_1\ldots\varepsilon_k - \varepsilon_1\ldots\varepsilon_m(1-\varepsilon_k)}{1-\varepsilon_1\ldots\varepsilon_k} \cdot (1-\varepsilon_1\ldots\varepsilon_k) \\
&\le \max_{1\le k\le m} \frac{1-\varepsilon_1\ldots\varepsilon_k - \varepsilon_1\ldots\varepsilon_m(1-\varepsilon_k)}{1-\varepsilon_1\ldots\varepsilon_k} \cdot w(s) \\
&= \rho w(s).
\end{aligned}$$

Furthermore, since $\max_{1\le k\le m}\{1-\varepsilon_1\ldots\varepsilon_m(1-\varepsilon_k)(1-\varepsilon_1\ldots\varepsilon_k)^{-1}\} \ge 1-\varepsilon_1\ldots\varepsilon_m$, we additionally have $\rho \in (0,1)$. Thus, for finite J and $\beta = 1$ (or applied to $\beta H \le \beta\rho w$ provided that $\beta\rho < 1$) also these ρ and w can be used to construct $u^{\pm}$ as above.

A similar approach in Tseng (1990) for a transient model with $\beta = 1$ yields $w' > 0$ and $\rho' \in (0,1)$, say, with $Hw' \le \rho' w'$, where $\rho' := (1-\varepsilon^{2m-1})(1-\varepsilon^{2m})^{-1}$. Our contraction factor ρ, however, is smaller. Indeed, choosing $\varepsilon = \min\{\varepsilon_1, \ldots, \varepsilon_m\}$, then ρ reduces to $1 - \varepsilon^m(1-\varepsilon)(1-\varepsilon^m)^{-1}$, which is less than ρ'.

Remark 3.3. (a) Consider the more general model MDP, which has a bounding function $b : S \to \mathbb{R}_+$, $b \not\equiv 0$ (i.e., using $0/0{:=}0$, $\sup_{(s,a)\in D} |r(s,a)|/b(s) < \infty$, $\sup_{s\in S} |v_0(s)|/b(s) < \infty$, $\gamma := \sup_{(s,a)\in D} b^{-1}(s) \sum_{s'\in S} p(s,a,s')b(s') < \infty$) and a variable discount factor $\beta(s,a) \ge 0$ with $0 < \bar{\beta} := \sup_{(s,a)\in D} \beta(s,a) < \infty$.

This model can be reduced to an MDP' with constant discount factor β' and bounded r' and v'_0, provided that $\gamma > 0$, as follows: $S' := S + \{\hat{s}\}$, $A' := A$, $D'(s) := D(s)$, $s \in S$, $D'(\hat{s}) := A$, $p'(s,a,s') := \beta(s,a)p(s,a,s')b(s')/[\bar{\beta}\gamma b(s)]$, $(s,a) \in D$, $s' \in S$, $r'(s,a) := r(s,a)/b(s)$, $v'_0(s) := v_0(s)/b(s)$, $(s,a) \in D$, $\beta' := \bar{\beta}\gamma$. The missing values of p', r', and v'_0 are determined by the requirement that $\{\hat{s}\}$ is absorbing.

The asserted reduction holds since one easily verifies that for all $v : S \in \mathbb{R}$ with $\|v\|_b < \infty$ there holds $Lv(s,a)/b(s) = L'((v/b) \cdot 1_S)(s,a)$, $(s,a) \in D$. It follows that $v_n(s)/b(s) = v'_n(s)$, $s \in S$, $n \in \mathbb{N}$, and that $f' : S \to A$ is a maximizer of $L'v'_{n-1}$, if and only if $f'|S$ is a maximizer of Lv_{n-1}.

(b) The reduction in (a) is useful in cases where $\gamma > 1$, where the discount factor is a constant $\beta \in (1/\gamma, 1)$ (hence $\beta' = \beta\gamma > 1$), and where $\beta'^* \geq \gamma$, in the following way: While the standard result about MDPs with bounding functions yields, e.g., the existence of V^β only for $0 < \beta < 1/\gamma$, Proposition 2.2(i) and Theorem 3.1 show that $V'^{(\beta\gamma)}$, and hence also V^β exists for $0 < \beta < 1$.

A simple example is the classical asset selling problem (cf., e.g., DeGroot (1970)) with $\mathbb{N}_0$-valued offers having expectation $\mu > 0$, with state space $S := \mathbb{N}_0 + \{\tilde{s}\}$ for an absorbing state $\tilde{s}$, and with bounding function b, where $b(s) = 1 + s$ for $s \in \mathbb{N}_0$ and $b(\tilde{s}) = 1$. Then $\beta'^* = \gamma = 1 + \mu > 1$.

(c) A terminal reward $\tilde{r}_{J_0}(s)$, say, when entering an absorbing state $s \in J_0$, can be easily handled by considering the equivalent model with the one-stage rewards $r'(s,a) := r(s,a) + \beta \sum_{s' \in J_0} p(s,a,s')\tilde{r}_{J_0}(s')$, $(s,a) \in D|_J$, and $r'(s,a) := 0$ otherwise.

4. Recursion in the state space and the model MDP′. Observe that $\beta p(s,a,s) < 1$ for all $\beta < \beta^*$, $s \in J$, and $a \in D(s)$, which follows from Proposition 2.3(a_4) and $\sum_n (\beta p(s,a,s))^n \leq \sum_n \beta^n \|e_n\|$. Introduce

$$I_1 := \{s \in J \mid p(s,a,s) = p(s,a,J) \text{ for all } a \in D(s)\}$$

and, using $\tilde{I}_0 := \emptyset$, $\tilde{I}_{\nu-1} := \sum_{\ell=1}^{\nu-1} I_\ell$, for $\nu > 1$ define

$$I_\nu := \{s \in J - \tilde{I}_{\nu-1} \mid p(s,a,s) + p(s,a,\tilde{I}_{\nu-1}) = p(s,a,J) \text{ for all } a \in D(s)\}.$$

Obviously $I_\nu = \emptyset$ for some $\nu \in \mathbb{N}$ implies $I_j = \emptyset$ for $j \geq \nu$. Then, for all $k \in \mathbb{N}$, one easily sees that V can be obtained by recursion in the state space on the subsets $I_1, \ldots, I_k$ of J according to

$$V(s) = \max_{a \in D(s)} \left\{ \frac{1}{1 - \beta p(s,a,s)} \left[r(s,a) + \beta \sum_{s' \in \tilde{I}_{\nu-1}} p(s,a,s')V(s') \right] \right\}, \quad s \in I_\nu,\ 1 \leq \nu \leq k.$$

If $I_\nu = \emptyset$ for some $\nu \in \mathbb{N}$, then the recursion in the state space stops with the largest k such that $I_k \neq \emptyset$. *If $\tilde{I}_k = J$, then V can be completely determined by a finite recursion in the state space.* To give a necessary and sufficient condition for this case in Proposition 4.1 below, define an operator $\tilde{H}$ on J by $\tilde{H}v(s) := \max_{a \in D(s)} \sum_{s' \in J - \{s\}} p(s,a,s')v(s')$ for all $s \in J$, $v \in \mathfrak{V}$. Set $\tilde{e}_t := \tilde{H}^t 1$, $t \in \mathbb{N}_0$.

PROPOSITION 4.1. *If $\beta < \beta^*$, then recursion in the state space stops with $\tilde{I}_k = J$ if and only if $\tilde{e}_k \equiv 0$.*

Proof. (a) We prove by induction on $j \in \mathbb{N}$ that $I_j = [\tilde{e}_j = 0 < \tilde{e}_{j-1}] := \{s \in J \mid \tilde{e}_j(s) = 0 < \tilde{e}_{j-1}(s)\}$. First, $\tilde{e}_0 \equiv 1$; hence $s \in I_1 \Leftrightarrow \tilde{e}_1(s) = 0$. Let $I_\ell = [\tilde{e}_\ell = 0 < \tilde{e}_{\ell-1}]$ hold for some $j \geq 1$ and all $1 \leq \ell \leq j$. Then $\tilde{I}_j = [\tilde{e}_j = 0]$; hence,

$\tilde{e}_{j+1}(s) = \max_{a\in D(s)} \sum_{s'\in[\tilde{e}_j>0]-\{s\}} p(s,a,s')\tilde{e}_j(s')$, $s \in J$. Now $s \in I_{j+1} \Leftrightarrow (\tilde{e}_j(s) > 0) \wedge (p(s,a,[\tilde{e}_j > 0] - \{s\}) = 0$ for all $a \in D(s)) \Leftrightarrow (\tilde{e}_j(s) > 0) \wedge (\tilde{e}_{j+1}(s) = 0)$. Now the induction is complete.

(b) From (a) it follows that $\tilde{I}_k = J$ if and only if $\tilde{e}_k \equiv 0$, which is the assertion. □

If $J' := J - \tilde{I}_k \neq \emptyset$ and $I_\nu = \emptyset$ for $\nu \geq k+1$, then V coincides on J' with the value function V', say, of an MDP' with essential state space J' and absorbing set $J'_0 := J_0 + \tilde{I}_k$, where $r'(s,a) := r(s,a) + \beta \sum_{s'\in J-J'} p(s,a,s')V(s')$ for $(s,a) \in D|_{J'}$ (and $r' = 0$ on $D|_{J'_0}$). Note that $(\beta^*)' \geq \beta^*$ holds for the critical discount factor of the reduced model MDP'.

As far as we know, recursion in the state space occurs in the literature only for specific models; cf., e.g., Hinderer (1971). Related general results can be found in Bertsekas (2001, pp. 97 and 133).

5. An extrapolation method. In what follows we consider a sequence (v_n) of successive approximations $v_n = Uv_{n-1}$, $n \in \mathbb{N}$, starting with some $v_0 \in \mathfrak{V}$. It is convenient to rewrite v_n as $v_n = U_{f_n}v_{n-1}$ for some maximizer f_n of Lv_{n-1}. Furthermore, let $d_n := v_n - v_{n-1}$ denote the difference of two successive approximations. Of course, $d_n \in \mathfrak{V}$.

Value iteration is now combined with an extrapolation, giving upper and lower bounds for V of the form $v_n + c_n^{\pm} d_n$ at each step n of iteration.

Our extrapolation method is based on the assumption that $d_1 \geq 0$. Then $d_n \geq 0$ for all $n \geq 1$. If $r \geq 0$, then $d_1 \geq 0$ trivially holds using initial value $v_0 \equiv 0$. Otherwise, we have to construct $d_1 \geq 0$ in an initial step. More details will be given below.

For $n \geq 2$ let

$$\alpha_n^- := \inf_{s\in J} \{d_n(s)/d_{n-1}(s) \mid d_{n-1}(s) > 0\}$$

(with $\alpha_n^- := 0$, if $d_{n-1} \equiv 0$). Note that $0 \leq \alpha_n^- d_{n-1} \leq d_n \leq \beta H d_{n-1}$. Together with Lemma 3.1.7(iii) in Ogiwara (1995) (which holds without (A0)) we then get $\alpha_n^- \leq \beta\lambda^*$. Using monotonicity of H_{f_n} we further get $\beta H_{f_n} d_n \geq \beta H_{f_n}(\alpha_n^- d_{n-1}) = \alpha_n^-[U_{f_n}v_{n-1} - U_{f_n}v_{n-2}] \geq \alpha_n^-[v_n - Uv_{n-2}] = \alpha_n^- d_n$ and thus $d_{n+1} \geq \beta H_{f_n} d_n \geq \alpha_n^- d_n$, which implies $\alpha_n^- \leq \alpha_{n+1}^-$. Moreover, if $\alpha_n^- > 0$ and $d_n > 0$ for some $n \geq 2$, then $d_{n+j} \geq (\alpha_n^-)^j d_n > 0$ for all $j \in \mathbb{N}$, which leads to $\alpha_{n+j}^- = \inf_{s\in J}\{d_{n+j}/d_{n+j-1}\} > 0$, $j \in \mathbb{N}$.

Next, for $k \geq 1$ select $\delta_k \in \mathfrak{V}$ such that $\hat{d}_k := d_k + \delta_k \in \mathfrak{W}$. Let

$$\hat{\alpha}_{k,m}^+ := \sup_{s\in J}\{\beta^m H^m \hat{d}_k(s)/\hat{d}_k(s)\}, \quad m \in \mathbb{N}.$$

Obviously H^m is monotone and positively homogeneous. It follows from Proposition 2.2(i) that H^m has the spectral radius $(\lambda^*)^m$. Using $\beta^m H^m \hat{d}_k \leq \hat{\alpha}_{k,m}^+ \hat{d}_k$, we obtain $\hat{\alpha}_{k,m}^+ \geq (\beta\lambda^*)^m$ (without (A0)) by Lemma 3.1.7(ii) in Ogiwara (1995) for $T := H^m$. Furthermore, by applying Proposition 2.3(a_6) with $v = w = \hat{d}_k$, we get $\hat{\alpha}_{k,m}^+ \to 0$ as $m \to \infty$, if $\beta < \beta^*$. In particular, $\hat{\alpha}_{k,m}^+ < 1$ for large enough m. More details on the choice of δ_k will be given below.

For fixed $f \in F$ we call the operator H_f *primitive* if for each $v \in \mathfrak{V}_+$, $v \not\equiv 0$, $v \notin \mathfrak{W}$ there exists $m = m(v) \in \mathbb{N}$ such that $H_f^m v \in \mathfrak{W}$. For finite sets J this definition is consistent with the usual one (i.e., for some $k \in \mathbb{N}$ the matrix associated with H_f^k has positive entries only).

THEOREM 5.1. *Let $d_1 \geq 0$ and $\beta < \beta^*$. Then,*

(i) *For all $n > 1$ it holds that $\alpha_n^- \leq \beta\lambda^* < 1$ and*

$$V \geq V_{f_n} \geq w_n^- := v_n + \alpha_n^-(1-\alpha_n^-)^{-1}d_n \geq v_n.$$

(ii) *Let $\hat{d}_k \in \mathfrak{W}$ for some $k \in \mathbb{N}$, and $m \in \mathbb{N}$ such that $\hat{\alpha}_{k,m}^+ < 1$. Then for all $n \geq k$*

$$V - v_n \leq \sup_{s\in J}\left\{\frac{d_n(s)}{\hat{d}_k(s) - \beta^m H^m \hat{d}_k(s)}\right\} \cdot \sum_{j=1}^m \beta^j H^j \hat{d}_k \leq \frac{\|d_n/\hat{d}_k\|}{1-\hat{\alpha}_{k,m}^+} \cdot \sum_{j=1}^m \beta^j H^j \hat{d}_k.$$

In particular, if $d_{n-1} \in \mathfrak{W}$, $d_n = \beta H d_{n-1}$, and $\alpha_n^+ := \sup_{s\in J}\{d_n(s)/d_{n-1}(s)\} < 1$ hold for some $n \in \mathbb{N}$, then

$$V \leq w_n^+ := v_n + \alpha_n^+(1-\alpha_n^+)^{-1}d_n.$$

(iii) *The weights $c_n^- := \alpha_n^-/(1-\alpha_n^-)$ are increasing in n, $\lim_{n\to\infty} c_n^- \leq \beta\lambda^*/(1-\beta\lambda^*)$. Furthermore, the lower bounds w_n^- are increasing in n and converge in norm to V as $n \to \infty$.*

(iv) *Under the assumptions in* (ii) *we get $\hat{w}_{n+m,k,m}^+ \leq \hat{w}_{n,k,m}^+ := v_n + \|d_n/\hat{d}_k\|(1-\hat{\alpha}_{k,m}^+)^{-1} \cdot \sum_{j=1}^m \beta^j H^j \hat{d}_k$. Furthermore, both upper bounds in* (ii) *converge in norm to V as $n \to \infty$.*

(v) *Let J be finite. Assume that LV has the unique maximizer f and let H_f be primitive. Then the procedure either stops with $d_n \equiv 0$ (hence $V = v_n$) for some $n \in \mathbb{N}$ or the weights $c_n^+ := \alpha_n^+/(1-\alpha_n^+)$ are decreasing in $n \geq n_0$ and both $c_n^\pm$ converge to $\beta\lambda_f^*/(1-\beta\lambda_f^*)$. If λ_f^* is also an eigenvalue of H with positive eigenvector, then $\lambda_f^* = \lambda^*$.*

Proof. By assumption, $\beta < \beta^*$ and thus $\beta\lambda^* < 1$ by Proposition 2.2(iii). Furthermore, remember that $\alpha_n^- \leq \beta\lambda^*$ (hence $\alpha_n^- < 1$) and $d_{n+1} \geq \beta H_{f_n} d_n \geq \alpha_n^- d_n$. Together with $U_{f_n} v_{n-1} = v_n$,

$$U_{f_n} w_n^- - w_n^- = \beta H_{f_n} d_n + \alpha_n^-(1-\alpha_n^-)^{-1}\beta H_{f_n} d_n - \alpha_n^-(1-\alpha_n^-)^{-1}d_n \geq 0,$$

which implies $V_{f_n} \geq w_n^-$ (and thus (i)) by Theorem 3.1(i) for MDP_f and the monotonicity of U_{f_n}. Additionally exploiting $0 \leq \alpha_n^- \leq \alpha_{n+1}^-$, we get

$$w_{n+1}^- - w_n^- = d_{n+1} + \alpha_{n+1}^-(1-\alpha_{n+1}^-)^{-1}d_{n+1} - \alpha_n^-(1-\alpha_n^-)^{-1}d_n \geq 0,$$

which gives monotonicity of w_n^- in n. Since $V \geq w_n^- \geq v_n$, $n > 1$, convergence in norm of the lower bounds follows from Theorem 3.1(i). The properties of c_n^- immediately follow from those of α_n^-. Thus (iii) holds.

We continue with the first part of (ii). By construction and since $\hat{\alpha}_{k,m}^+ < 1$, $\eta := \hat{d}_k - \beta^m H^m \hat{d}_k \in \mathfrak{W}$. Set $c := \sup_{s\in J}\{d_n(s)/\eta(s)\}$. Then, for some $f \in F$,

$$\begin{aligned} U\left(v_{n-1} + c\sum_0^{m-1}\beta^j H^j \hat{d}_k\right) - \left(v_{n-1} + c\sum_0^{m-1}\beta^j H^j \hat{d}_k\right) \\ = U_f v_{n-1} + c\sum_0^{m-1}\beta^{j+1} H_f H^j \hat{d}_k - v_{n-1} - c\sum_0^{m-1}\beta^j H^j \hat{d}_k \\ \leq d_n - c(\hat{d}_k - \beta^m H^m \hat{d}_k) = d_n - \eta \cdot \|d_n/\eta\| \leq 0. \end{aligned}$$

Next $V \le U(v_{n-1} + c\sum_0^{m-1} \beta^j H^j \hat{d}_k) \le v_n + c\sum_1^m \beta^j H^j \hat{d}_k$, which follows from Theorem 3.1(i) and from $U(v+w) \le Uv + \beta Hw$ for $v, w \in \mathfrak{V}$. It gives the left-hand inequality. The right-hand inequality then follows from the definition of $\hat{\alpha}^+_{k,m}$. The second part of (ii) follows from the first part by choosing $k := n-1$, $m := 1$, and $\delta_{n-1} :\equiv 0$; hence $\hat{\alpha}^+_{n-1,1} = \alpha^+_n$.

From $d_{n+j} \le \beta^j H^j d_n \le \|d_n/\hat{d}_k\| \beta^j H^j \hat{d}_k$, $1 \le j \le m$, we obtain $\|d_{n+m}/\hat{d}_k\| \le \hat{\alpha}^+_{k,m} \|d_n/\hat{d}_k\|$ and then, using $v_{n+m} - v_n = \sum_{j=1}^m d_{n+j}$, that $\hat{w}^+_{n+m,k,m} - \hat{w}^+_{n,k,m} \le \|d_n/\hat{d}_k\| [1 + \hat{\alpha}^+_{k,m}(1-\hat{\alpha}^+_{k,m})^{-1} - (1-\hat{\alpha}^+_{k,m})^{-1}] \times \sum_{j=1}^m \beta^j H^j \hat{d}_k = 0$. From $d_n \le \beta H d_{n-1}$ we further obtain $\|d_n\| \le \|d_1\| \beta^{n-1} \|e_{n-1}\|$. Now Proposition 2.3($a_3$) completes (iv).

Theorem 5.1(v). Since f is the unique maximizer of LV, we obtain from Theorem 10.1 (below) that for some $n_0 \ge 1$ there holds $d_n = \beta^{n-n_0} H_f^{n-n_0} d_{n_0}$ for $n > n_0$. Put $k := n_0$. Assume $d_k \equiv 0$. Then $V = v_k$. Otherwise $d_k \not\equiv 0$. If $d_k \notin \mathfrak{W}$, primitivity of βH_f (applied to $v = d_k$) gives $d_{k+m} \in \mathfrak{W}$ for some $m \in \mathbb{N}$. Hence, if $d_k \not\equiv 0$, there is $l \ge k$ such that $d_l \in \mathfrak{W}$. Now we apply Satz III.2.7 in Bohl (1974) with the following choices (and observing that primitive linear operators are called "streng-monoton" there): $X := \mathfrak{V}$, $K := \mathfrak{V}_+$, $\bar{e} :\equiv 1$, $S = T := H_f$, $e_0 := d_l$. Then it follows that $\alpha^\pm_n$, $n > l$, are monotone in n and that $\lim_{n\to\infty} \alpha^\pm_n = \beta\lambda^*_f$. The properties of $c^\pm_n$ now easily follow from those of $\alpha^\pm_n$. If λ^*_f is an eigenvalue of H with positive eigenvector, then $\lambda^*_f = \lambda^*$ by Ogiwara (1995, Lemma 3.1.7(ii) and (iii)). □

The asymmetry in the bounds results from the sublinearity of H. The lower bounds are easy to calculate and need $d_1 \ge 0$ only. The upper bounds are based on $\hat{d}_k \in \mathfrak{W}$ and need some additional effort due to the computation of $\beta^j H^j \hat{d}_k$, $1 \le j \le m$, up to some $m \in \mathbb{N}$ such that $\hat{\alpha}^+_{k,m} < 1$. To reduce the computational effort, we may use these fixed objects for more than one step of iteration. In this case, we only have to adapt v_n and $\|d_n/(\hat{d}_k - \beta^m H^m \hat{d}_k)\|$ (resp., $\|d_n/\hat{d}_k\|$) by switching from n to $n+1$, which can be easily realized.

We may use $\delta_1 = e_0 - d_1$ in order to obtain $\hat{d}_1 = e_0 \in \mathfrak{W}$. Then, for $m \in \mathbb{N}$ such that $\beta^m \|e_m\| < 1$, it follows from Theorem 5.1(ii) for $n \ge 1$ that

$$V - v_n \le \sup_{s\in J} \left\{ \frac{d_n(s)}{1 - \beta^m e_m(s)} \right\} \cdot \sum_{j=1}^m \beta^j e_j \le \frac{\|d_n\|}{1 - \beta^m \|e_m\|} \cdot \sum_{j=1}^m \beta^j e_j.$$

Mainly, however, we are interested in applying (ii) with $\delta_k = 0$, provided that $d_k \in \mathfrak{W}$. Extensive numerical results show that the extrapolation method works very well and, in particular, is highly superior to an extrapolation method of the MacQueen type, which cannot balance out the unequal row sums in the defining matrices of H. Some numerical results are displayed in section 6. Finally, if $d_k \notin \mathfrak{W}$, we may use $\hat{d}_k = d_k + \varepsilon e_0$ for some $\varepsilon > 0$.

Due to $d_n \ge \alpha^-_n d_{n-1}$, Theorem 5.1(i) improves the lower bound in Schellhaas (1974, Satz 4.1), which may be written as $v_{n-1} + \alpha^-_n (1-\alpha^-_n)^{-1} d_{n-1}$ and which is derived under the additional assumptions that J is finite, $\beta < 1$, and $0 < d_n < d_{n-1}$. His upper bounds $w^+_n := v_{n-1} + \alpha^+_n (1-\alpha^+_n)^{-1} d_{n-1}$ are constructed in the same way and need the verification of $Uw^+_n \le w^+_n$.

Theorem 5.1 can be combined with the following proposition.

PROPOSITION 5.2. *Let $\beta < \beta^*$, $d_1 \ge 0$. If $Hd_n = 0$ on $M_n := \{s \in J \mid d_n(s) = 0\}$ for some $n \in \mathbb{N}$, then*

(i) *$p(s, a, J - M_n) = 0$ for $(s,a) \in D|_{M_n}$,*

(ii) *$V = v_k = v_n$ on M_n, $k \ge n$.*

Proof. (i) is a consequence of $d_n = 0$ on M_n, $d_n > 0$ on $J - M_n$, and

$$0 = Hd_n(s) = \max_{a \in D(s)} \left\{ \sum_{s' \in M_n} p(s,a,s')d_n(s') + \sum_{s' \in J - M_n} p(s,a,s')d_n(s') \right\}, \quad s \in M_n.$$

(ii) Observe that $0 \le d_{n+k} \le \beta^k H^k d_n$, $k \in \mathbb{N}$. Together with (i), $H^k d_n = 0$, $k \in \mathbb{N}$, on M_n. Hence $d_{n+k} = 0$ on M_n and thus $v_n = V$ on M_n by Proposition 2.2(iii). □

On the basis of Proposition 5.2, value iteration can be continued after the first n steps with $MDP^{(n)}$, say, with essential state space $J - M_n$ and one-stage rewards $r(s,a) + \beta \sum_{s' \in M_n} p(s,a,s')v_n(s')$, $(s,a) \in D|_{J-M_n}$.

Fix $n \in \mathbb{N}$ and assume $d_1 \ge 0$, $\beta < \beta^*$. Let $e_n = 0$ on some $\hat{M}_n \subset J$. Then, using $e_{n+1} \le e_n$ and $0 \le d_{n+k} \le \beta^{n+k} H^{n+k} d_1 \le \|d_1\| \beta^{n+k} e_{n+k}$, $k \in \mathbb{N}$, a simple argument leads to $v_n = V$ on $\hat{M}_n$. Furthermore, if $\hat{M}_n = J$, then $v_n \equiv V$, and value iteration stops after n steps of iteration. See also Proposition 4.1.

Hübner (1980) derives bounds for general S, A under an assumption "$d_n \in \mathcal{T}$," which includes "$d_n \ge 0$" and weakens "$d_n > 0$" similar to Proposition 5.2. He has found the lower bound for V given in Theorem 5.1(i) under the assumptions $V = \lim_{N \to \infty} v_N$, and $\alpha_n^- < 1$ (cf. Theorem 3 and Remark 7 there). The lower bound also follows from Example 3.2 in Waldmann (1985) (within a general framework for constructing extrapolation functions) if $V = \lim_{N \to \infty} v_N$. Additional assumptions are not needed there.

Let $d_{n-1} \in \mathfrak{W}$ and $\beta \|Hd_{n-1}/d_{n-1}\| < 1$. Then, by applying Theorem 5.1(ii) with $k = n-1$ and $m = 1$, we get the (right-hand) upper bound $v_n + \|d_n/d_{n-1}\|(1 - \beta\|Hd_{n-1}/d_{n-1}\|)^{-1}\beta Hd_{n-1}$ for V, which is superior to the upper bound given in Hübner (1980, Theorem 2 and Remark 7), due to $\beta Hd_{n-1} \le \beta \|Hd_{n-1}/d_{n-1}\| d_{n-1}$.

Example 5.3. Consider the standard Markov decision model with finite state space S and $J_0 = \emptyset$. Then $e_n \equiv e_{n,f} \equiv 1$, $n \in \mathbb{N}$, and $\beta^* = \lambda^* = \lambda_f^* = 1$ for all $f \in F$ by Proposition 2.2. Assume f to be the unique maximizer of LV and H_f to be primitive. Then, from Theorem 5.1(v), either both $c_n^{\pm}$ converge to $\beta/(1-\beta)$ as $n \to \infty$ or the value iteration stops with $V = v_n$ for some n.

Analogous bounds for v_N, which are based on v_n and d_n for $n \le N$, exist and belong to $\mathfrak{V}$ for all $\beta > 0$. The initial value v_0 (terminal reward) is now fixed and cannot be chosen such that $d_1 \ge 0$ holds.

THEOREM 5.4. *Let $N \in \mathbb{N}$, $1 < n < N$, and α_n^- and $\hat{\alpha}_{k,m}^+$ be as in Theorem* 5.1.

(i) *If $d_{n-1} \ge 0$, then*

$$v_n + \sum_{j=1}^{N-n} (\alpha_n^-)^j d_n \le v_{n+1} + \sum_{j=1}^{N-n-1} (\alpha_{n+1}^-)^j d_{n+1} \le v_N.$$

(ii) *Let $\hat{d}_k \in \mathfrak{W}$ for some $k \le n$. Then, for $\ell, m \in \mathbb{N}$ such that $N - n \le \ell m$,*

$$v_N \le v_n + \|d_n/\hat{d}_k\| \cdot \sum_{i=0}^{\ell-1} (\hat{\alpha}_{k,m}^+)^i \sum_{j=1}^{m} \beta^j H^j \hat{d}_k.$$

Proof. (i) uses $\alpha_n^- d_n \le d_{n+1}$, $\alpha_n^- \le \alpha_{n+1}^-$, and $v_N - v_n = \sum_{j=1}^{N-n} d_{n+j}$. For the proof of (ii) one uses $d_{n+j} \le \|d_n/\hat{d}_k\| \cdot \beta^j H^j \hat{d}_k$ for $j \ge 1$, derives $\beta^{mi} H^{mi} \hat{d}_k \le$

$(\hat\alpha^+_{k,m})^i \hat d_k$ by induction on $i \in \mathbb{N}$, and finally uses $v_N - v_n \le \sum_{j=1}^{lm} d_{n+j} \le \|d_n/\hat d_k\| \cdot \sum_{i=0}^{l-1}\sum_{\nu=1}^{m} \beta^\nu H^\nu(\beta^{mi} H^{mi} \hat d_k)$. □

Theorem 5.4(i) verifies the lower bounds presented in Theorem 3 of Hübner (1980), which have been derived there under the stronger assumption $d_{n-1} \in \mathcal{T}$. Applying Theorem 5.4(ii) with $\hat d_k = d_{n-1}$, $m = 1$, $\ell = N - n$, we get the upper bound $v_n + \|d_n/d_{n-1}\| \cdot \sum_{i=0}^{N-n-1} (\hat\alpha^+_{n-1,1})^i \beta H d_{n-1}$, which is smaller than the bound $v_n + \|d_n/d_{n-1}\| \cdot \sum_{i=1}^{N-n} (\hat\alpha^+_{n-1,1})^i d_{n-1}$ given in Theorem 2 there (due to $\beta H d_{n-1} \le \hat\alpha^+_{n-1,1} d_{n-1}$).

The need of an increasing sequence (v_n) of value iterates may be seen as a restriction of our method. We therefore give some hints for choosing a suitable initial value v_0 in order to destroy any doubts.

(a) Choose $v_0 \equiv 0$ in case of $r_f \ge 0$ for some $f \in F$. Then $v_0 \le U v_0$ trivially holds.

(b) Select some $f \in F$. Iterate up to some $m \in \mathbb{N}$ such that $\beta^m \|e_{m,f}\| < 1$. Then $v_0 := \inf_{s\in S}\{r_f(s)/(1 - \beta^m e_{m,f}(s))\} \sum_{j=0}^{m-1} \beta^j e_{j,f} \in \mathfrak{V}$ and we have $U_f v_0 \ge v_0$ and thus $U v_0 \ge v_0$. Moreover, for any $\varepsilon \in \mathfrak{W}$, replacing r_f by $r_f - \varepsilon$ in the definition of v_0, we additionally have $d_1 = U v_0 - v_0 \in \mathfrak{W}$. Finally, if $e_{1,f}$, $f \in F$, and d_1 belong to $\mathfrak{W}$ (or alternatively $\alpha_2^- > 0$, $d_1 \in \mathfrak{W}$), then $d_n \in \mathfrak{W}$, $n \in \mathbb{N}$.

(c) Select $f \in F$. Based on MDP_f we may choose w and ρ as in section 3. Then for $v_0 := u^-$ it holds that $U v_0 \ge v_0$.

(d) Since $V_f = U_f V_f \le U V_f$ for all $f \in F$, we may also compute V_f for some $f \in F$ and start with $v_0 = V_f$.

In Example 5.6 below we present an important case where it is useful to have bounds for V_f for a single $f \in F$. First note that Theorem 5.1 can be applied to MDP_f, replacing H by H_f, d_n by $d_{n,f}$, etc. Note that $e_{n,f} \le e_n$; hence $\beta_f^* \ge \beta^*$. Several results, e.g., in Theorem 5.1(ii), simplify since $d_{n+1,f} = \beta H_f d_{n,f}, n \in \mathbb{N}$, by linearity of H_f. In particular, we obtain the following:

(i) For $n \ge 2$ there holds $\alpha^-_{n,f} \le \beta\lambda_f^* < 1$, the weights $c^-_{n,f} := \alpha^-_{n,f}/(1-\alpha^-_{n,f}) d_{n,f}$ and the lower bounds $w^-_{n,f} := v_{n,f} + c^-_{n,f} d_{n,f} \le V_f$ increase, and $w^-_{n,f}$ converges in norm to V_f.

(ii) The upper bounds for V_f, obtainable from Theorem 5.1(ii), can be rewritten in a simpler way. In particular, $V_f - v_{n,f} \le \|d_{n,f}/(d_{n-m,f} - d_{n,f})\|(v_{n,f} - v_{n-m,f}) \le \hat\alpha^+_{n-m,m,f}(1-\hat\alpha^+_{n-m,m,f})^{-1}(v_{n,f} - v_{n-m,f})$ for $k = n-m$ and all $1 \le m < n$ such that $d_{n-m,f} \in \mathfrak{W}$ and $\hat\alpha^+_{n-m,m,f} := \|d_{n,f}/d_{n-m,f}\| < 1$. Actually the two bounds coincide since $\|v/(1-v)\| = \|v\|/(1-\|v\|)$ for $\|v\| < 1$.

Since finiteness of J is only needed in Theorem 5.1(v) to apply the turnpike Theorem 10.1 below, we additionally obtain the following proposition.

PROPOSITION 5.5. *Fix $f \in F$. Assume $\beta < \beta_f^*$ (hence $\beta\lambda_f^* < 1$), $d_{1,f} \ge 0$, that H_f is primitive, and that H_f^ν is compact for some $\nu \in \mathbb{N}$. (The latter holds under* (A0)*, in particular if J is finite.) Then either $d_{n_0,f} \equiv 0$ for some $n_0 \in \mathbb{N}$ (and then $V_f = v_{n_0,f}$) or $d_{n,f} \in \mathfrak{W}$ for $n \ge n_1$. In the latter case the following hold for $n > n_1$:*

(i) *The weights $c^+_{n,f} = \alpha^+_{n,f}/(1-\alpha^+_{n,f})$ decrease in n, and, as well as $c^-_{n,f}$, converge to $\beta\lambda_f^*/(1-\beta\lambda_f^*)$ as $n \to \infty$.*

(ii) *The bounds $w^+_{n,f} := v_{n,f} + c^+_{n,f} d_{n,f}$ decrease in n and converge in norm to V_f.*

Example 5.6. *Mean entrance time of a Markov chain into an absorbing set.* Applied to $r_f = 1_J$ and $\beta = 1$, the total expected reward $V_f(s)$ equals the expected entrance time $E_{fs}(\tau)$ into J_0 and we obtain $v_{n,f} = \sum_{j=0}^n e_{j,f}$ (with $v_{0,f} = e_{0,f}$). Hence, if $d_{n-1,f} = e_{n-1,f} \in \mathfrak{W}$ and $\|e_{n,f}/e_{n-1,f}\| < 1$ (hence $\beta_f^* > 1$ by Proposition

2.6), then by Theorem 5.1(i) and (ii)

$$\sum_{j=0}^{n} e_{j,f}(s) + \frac{\alpha_{n,f}^{-}}{1-\alpha_{n,f}^{-}}\, e_{n,f}(s) \le E_{fs}(\tau) \le \sum_{j=0}^{n} e_{j,f}(s) + \frac{\alpha_{n,f}^{+}}{1-\alpha_{n,f}^{+}}\, e_{n,f}(s), \quad s \in J,$$

where $\alpha_{n,f}^{\pm} = \pm \sup_{s\in J}\{\pm e_{n,f}(s)/e_{n-1,f}(s)\}$. Additionally, if H_f is primitive and H_f^{ν} is compact for some $\nu \in \mathbb{N}$, then, by Proposition 5.5(i), the weights $c_{n,f}^{\pm} = \alpha_{n,f}^{\pm}/(1-\alpha_{n,f}^{\pm})$ are monotone in n and converge to $\lambda_f^*/(1-\lambda_f^*)$ as $n \to \infty$ or value iteration stops with $e_{n,f} \equiv 0$ for some n.

Remember that $P_{fs}(\tau > n) = e_{n,f}(s)$, $s \in J$. Note that $e_{n+1,f} = H_f e_{n,f} \le \alpha_{n,f}^{+} e_{n,f}$. Hence $\alpha_{n+1,f}^{+} \le \alpha_{n,f}^{+}$ (additionally using $e_{n+1,f} \le e_{n,f}$ and putting $0/0 = 0$). Analogously, $\alpha_{n+1,f}^{-} \ge \alpha_{n,f}^{-}$. Thus we also obtain bounds

$$(\alpha_{n,f}^{-})^k e_{n,f} \le (\alpha_{n+1,f}^{-})^{k-1} e_{n+1,f} \le e_{n+k,f} \le (\alpha_{n+1,f}^{+})^{k-1} e_{n+1,f} \le (\alpha_{n,f}^{+})^k e_{n,f}, \quad k \in \mathbb{N},$$

for the (whole) distribution of τ at each step n of iteration.

Applied to the mean run length of a quality control scheme the procedure works very well (cf. Waldmann (1986a, 1986b) for details). Bearing in mind that λ_f^* is nearly one and that the second largest eigenvalue is small compared with λ_f^*, this is no longer surprising looking at the asymptotic properties (Proposition 5.5(i)) of our extrapolation method.

6. A numerical example. Our numerical results are based on an MDP (Problem I) with essential state space $J = \{1, 2, 3\}$, discount factor $\beta = 1$, transition law p, and one-stage reward r as in Table 6.1.

Remember that $\hat{\alpha}_{k,m}^{+} \ge (\beta\lambda^*)^m$. Thus $\beta^* > 1$ can be verified by the extrapolation itself. In particular, by applying Theorem 5.1(ii) with $k = n$, $\hat{d}_n = d_n$, and $m = 1$, at the first step n with $\hat{\alpha}_{n,1}^{+} < 1$ we know that $\beta^* > 1$ and thus Theorem 5.1 works. Since $r \ge 0$, we start with $v_0 \equiv 0$. The resulting bounds for V are displayed in Table 6.2. Missing upper bounds due to $\hat{\alpha}_{n,1}^{+} \ge 1$ are marked by "∞."

Changing the sign of $r(2,1)$ (Problem II), we can still apply Theorem 5.1 with $v_0 \equiv 0$ and get the same results as in Table 6.2, since $r_f \ge 0$ for $f \equiv 2$. For $f \equiv 1$,

TABLE 6.1
Definition of p and r.

		$p(s,a,s')$			
s	a	$s'=1$	$s'=2$	$s'=3$	$r(s,a)$
1	1	8/16	4/16	0	8
	2	1/16	12/16	0	11/4
	3	4/16	2/16	0	17/4
2	1	8/16	0	8/16	16
	2	1/16	14/16	1/16	15
3	1	4/16	4/16	0	7
	2	2/16	12/16	1/16	4
	3	12/16	1/16	1/16	9/2

TABLE 6.2
Upper and lower bounds for V, starting with $v_0 \equiv 0$.

n	$v_n(1)$	$v_n(2)$	$v_n(3)$	$w_n^-(1)$	$\hat{w}_{n,n,1}^+(1)$	$w_n^-(2)$	$\hat{w}_{n,n,1}^+(2)$	$w_n^-(3)$	$\hat{w}_{n,n,1}^+(3)$
1	8.0	16.0	7.0						
2	16.0	29.9	17.4	70.1	∞	124.1	∞	88.0	∞
3	26.2	43.3	29.5	257.1	∞	345.3	∞	303.5	∞
4	36.9	56.4	41.6	544.6	859.5	679.8	1066.4	615.8	972.0
5	47.3	69.2	53.5	664.0	679.1	826.4	845.0	753.6	770.8
6	57.6	81.9	65.2	670.2	678.2	834.3	844.1	760.9	770.0
7	67.8	94.3	76.7	677.7	678.2	843.5	844.1	769.4	770.0
8	77.7	106.5	88.0	678.2	678.2	844.1	844.1	769.9	769.9

however, we have to construct an initial value v_0. Based on MDP_f we then get (cf. section 3) a partition $J_1 = \{1, 3\}$, $J_2 = \{2\}$ of J with $\varepsilon_1 = 1/4$, for example, $\varepsilon_2 = 1/2$, $\rho = 13/14$, $w(1) = w(3) = 3/4$, $w(2) = 7/8$, and, finally, $v_0 = (-192, -224, -192)^T$. It is not surprising that the worse initial value v_0 leads to an increase in the number of iterations (due to a later stabilization of the upper bounds). Compared with an extrapolation of the MacQueen type, however, these bounds are strongly superior.

7. The policy iteration method. The policy iteration is a standard procedure in obtaining V and an optimal policy. The special case of primitive operators H_f, however, has not been studied explicitly to the best of our knowledge.

THEOREM 7.1 (A0). *Assume $\beta < \beta^*$. Let $f \in F$.*

(i) *If $UV_f = V_f$, then $V = V_f$.*

(ii) *Otherwise, there is some $g \in F$ such that $U_g V_f \geq V_f$, $U_g V_f \neq V_f$, and $V_g \geq V_f$, $V_g \neq V_f$. Additionally, if H_g is primitive, then $V_g \geq V_f + \varepsilon$ for some $\varepsilon > 0$.*

Proof. (i) and the first part of (ii) follow by standard arguments. For the second part of (ii) we first observe that $\beta^n H_g^n(v - v') = U_g^n v - U_g^n v'$ holds for all $v, v' \in \mathfrak{V}$, $n \in \mathbb{N}$. Since H_g is primitive, there is some $m \in \mathbb{N}$ such that $H_g^m(U_g V_f - V_f) \in \mathfrak{W}$. Hence $U_g(U_g^m V_f) - U_g^m V_f \in \mathfrak{W}$. Furthermore, using monotonicity of U_g, we obtain from Theorem 3.1(iii), applied to MDP_g, $U_g^n V_f \uparrow V_g$ as $n \to \infty$. Hence $V_g - V_f \geq U_g^{m+1} V_f - U_g^m V_f \in \mathfrak{W}$. □

At each step of iteration, V_f can be obtained by solving the infinite linear system of equations $V_f = U_f V_f$ by known methods or alternatively by Theorem 5.1 and Proposition 5.5.

8. Finite state approximations of the MDP. Finite state approximations have a long tradition in dynamic programming. See Sennott (1999) for recent work on this subject and the references given there.

Applied to the MDP, assumption (A0) suggests an approximation scheme in a natural way. Thus, depending on a finite set $K \subset J$ with $\rho(K) := \sup_{(s,a) \in D|_J} \{p(s, a, J - K)\}$ and $c \in \mathbb{R}$, introduce $MDP^{(K,c)}$, say, with finite essential state space $J^{(K)} := K \subset J$, extended absorbing set $J_0^{(K)} := J_0 + (J - K)$, transition law $p^{(K)}(s, a, s') = p(s, a, s')$, $(s, a) \in D|_K$, and reward function $r^{(K,c)} := r + \beta c \rho(K)$ on $D|_K$ ($r^{(K,c)} = 0$ otherwise).

Since $p^{(K)}$ is independent of c, then also $e_n^{(K)}$ (resp., $e_{n,f}^{(K)}$, $f \in F$), which is defined in analogy to e_n (resp., $e_{n,f}$), does not depend on c. One easily verifies that $e_n^{(K)} \leq e_n$, $n \in \mathbb{N}$, on K. Hence $\lambda^{*(K)} := \inf_{n \in \mathbb{N}} \|e_n^{(K)}\|^{1/n} \leq \inf_{n \in \mathbb{N}} \|e_n\|^{1/n} = \lambda^*$.

Now Proposition 2.2 implies $\beta^{*(K)} = 1/\lambda^{*(K)} \geq 1/\lambda^* = \beta^*$. Thus, for $\beta < \beta^*$, let $V^{(K,c)}$ (resp., $V_f^{(K,c)}$, $f \in F$) be defined in analogy to V (resp., V_f).

$MDP^{(K,c)}$ behaves as an absorbing MDP with finite state and action spaces, which has been studied in detail in Hinderer and Waldmann (2003). In this section we look at bounds for V based on $V^{(K,c)}$, are interested in ε-optimal policies for MDP resulting from extensions of optimal ones for $MDP^{(K,c)}$, and look at the convergence of a pure truncation method. For the last approach let $D^*(s)$, $s \in J$, be the set of all maximum points of $LV(s,\cdot)$. For the model $MDP^{(K,c)}$ define $D^{*(K,c)}(s)$ analogously.

LEMMA 8.1. *For $\beta < \beta^*$, $c \in \mathbb{R}$, and $\delta > 0$ we have on finite $K \subset J$*

(i) $V_f^{(K,c+\delta)} - V_f^{(K,c)} = \beta\delta\rho(K) \cdot \sum_{n=0}^{\infty} \beta^n e_{n,f}^{(K)}$ *for all $f \in F$,*

(ii) $0 \leq V^{(K,c+\delta)} - V^{(K,c)} \leq \beta\delta\rho(K) \cdot \sup_{f\in F} \sum_{n=0}^{\infty} \beta^n e_{n,f}^{(K)} \leq \beta\delta\rho(K) \cdot \sum_{n=0}^{\infty} \beta^n e_n^{(K)}$.

Proof. (i) results from $r^{(K,c+\delta)} = r^{(K,\delta)} + \beta\delta\rho(K)$ on $D|_K$ and the linearity of the expectation operator; (ii) is an immediate consequence of (i), since $V^{(K,c)} = \sup_{f\in F} V_f^{(K,c)}$ by Theorem 3.1(iii) and $e_{n,f}^{(K)} \leq e_n^{(K)}$. □

In the following theorem we make use of rough bounds $c^{\pm}$ of V on $J - K$, e.g., $c^{\pm} := \pm\max\{0, \sup_{s\in J}\{\pm\hat{u}^{\pm}(s)\}\}$ with $\hat{u}^{\pm} := u^{\pm}$ or $\hat{u}^{\pm} := Uu^{\pm}$ and $u^{\pm}$ as defined in section 3.

THEOREM 8.2. *Let K be a finite subset of J, $\beta < \beta^*$, and $c^- \leq 0 \leq c^+$ such that $c^- \leq V(s) \leq c^+$ for all $s \in J - K$. Then the following hold:*

(i) $V^{(K,c^-)} \leq V \leq V^{(K,c^+)}$ *on K.*

(ii) *Let $g \in F$ be optimal for $MDP^{(K,c^-)}$. Then, for all $f \in F$ such that $f = g$ on K and $c^- \leq V_f(s)$, $s \in J - K$, it holds that $V_f \geq V - \varepsilon$ on K, where $\varepsilon := \beta(c^+ - c^-)\rho(K) \cdot \sum_{n=0}^{\infty} \beta^n e_n^{(K)}$.*

(iii) *Assume* (A0). *Then there exist $K_n \subset K_{n+1}$, $n \in \mathbb{N}$, with $\cup_n K_n = J$ such that $\rho(K_n) \to 0$ as $n \to \infty$. Furthermore, for all $s \in J$, $\lim_{n\to\infty} V^{(K_n,0)}(s) = V(s)$, and the cluster points of all (a_n), where $a_n \in D^{*(K_n,0)}(s)$, belong to $D^*(s)$.*

Proof. (i) Since $p(s,a,J-K) \leq \rho(K)$, $(s,a) \in D$, and $c^+ \geq 0$ hold by assumption, we have, for $s \in K$,

$$V(s) = UV(s) \leq \max_{a\in D(s)} \left\{ r(s,a) + \beta \sum_{s'\in K} p(s,a,s')V(s') + \beta c^+\rho(K) \right\} =: U^{(K,c^+)}V(s).$$

Monotonicity of $U^{(K,c^+)}$ and Theorem 3.1(i) (applied to $MDP^{(K,c^+)}$) then give $V \leq V^{(K,c^+)}$ on K, since V is bounded on K and since $\beta^{*(K)} \geq \beta^*$, and analogously for the lower bound.

(ii) Let g be optimal for $MDP^{(K,c^-)}$ and $f = g$ on K. By a similar argument as in (i) we get $V_f = U_f V_f \geq U_f^{(K,c^-)} V_f = U_g^{(K,c^-)} V_f$. Using Theorem 3.1(iii) applied to $MDP^{(K,c^-)}$, it follows that $V_f \geq (U_g^{(K,c^-)})^n V_f \to V_g^{(K,c^-)}$. Thus $V_f \geq V_g^{(K,c^-)} = V^{(K,c^-)}$ on K. Together with (i) and Lemma 8.1(ii) it then follows that $V - V_f \leq V^{(K,c^+)} - V^{(K,c^-)} \leq \varepsilon$ on K.

(iii) The existence of the sets K_n is obvious. Put $V^{(n,c)} := V^{(K_n,c)}$. Fix $s \in J$. Then $s \in K_n$ for all $n \geq n_0$ and some $n_0 \in \mathbb{N}$. Recall that $c^- \leq 0 \leq c^+$. Then, using (i) and $V^{(n,c^-)} \leq V^{(n,0)} \leq V^{(n,c^+)}$, we have $|V(s) - V^{(n,0)}(s)| \leq V^{(n,c^+)}(s) - V^{(n,c^-)}(s)$, which converges to zero as $n \to \infty$ by $\rho(K_n) \to 0$ and Lemma 8.1(ii).

For $n \geq m$, using (i), $e_j^{(K_m)} \leq e_j$, and Lemma 8.1(ii), $V^{(n,0)} - V^{(m,0)} \leq |V - V^{(n,0)}| + |V - V^{(m,0)}| \leq 2\beta(c^+ - c^-)\rho(K_m) \sum_{j=0}^{\infty} \beta^j \|e_j\| =: \eta(m)$ on K_m and $V^{(n,0)} \leq$

$\|u^+\|$ on K_n (e.g., by verifying $U^{(K_n,0)}u^+ \le u^+$ on K_n) and thus on $K_n - K_m$. Fix $s \in J$. Let $N \in \mathbb{N}$ such that $s \in K_N$. Then, for $N \le m \le n$, $w_n \le w_m + \varepsilon_m$, where $w_n(a) := r(s,a) + \beta \sum_{s' \in K_n} p(s,a,s')V^{(n,0)}(s')$, $a \in D(s)$, and $\varepsilon_n := \beta\eta(n) + \beta\|u^+\|\rho(K_n)$. Since $\varepsilon_n \to 0$ as $n \to \infty$, Proposition 10.1 in Schäl (1975) completes the proof of (iii). □

9. Asymptotically ε-optimal policies. The infinite-stage model is often used as an approximation of a finite-stage model with a large horizon. We say that a decision rule $f \in F$ is asymptotically optimal if $\|v_N - v_{N,f}\| \to 0$ as $N \to \infty$. Furthermore, f is said to be ε-optimal for some $\varepsilon \in \mathfrak{V}_+$ if $V_f \ge V - \varepsilon$, and asymptotically ε-optimal if $\|v_N - v_{N,f} - \varepsilon\| \to 0$ as $N \to \infty$.

THEOREM 9.1. *Let $\beta < \beta^*$. Then the following hold:*

(i) *f is asymptotically optimal if and only if f is a maximizer of LV.*

(ii) *Let f be ε-optimal. Then f is also asymptotically ε-optimal. Furthermore, for $c^\pm \in \mathbb{R}$ such that $c^- \le v_0 - V \le c^+$ on J, and all $N \in \mathbb{N}$ it holds on J that*

$$0 \le v_N - v_{N,f} \le \max\{0, c^+\}\beta^N e_N - c^-\beta^N e_{N,f} + \varepsilon. \tag{9.1}$$

Proof. (i). Let f be asymptotically optimal. Then $\|V - U_f V\| \le \|V - U_f^N V\| \le \|V - v_N\| + \|v_N - v_{N,f}\| + \|U_f^N v_0 - V_f\| + \|V_f - U_f^N V\|$, which converges to zero as $N \to \infty$ by Theorem 3.1 and the definition of asymptotical optimality. On the other hand, if f is a maximizer of LV, it is optimal by Theorem 3.1; hence $\|v_N - v_{N,f}\| \le \|v_N - V\| + \|V_f - U_f^N v_0\|$, which also converges to zero as $N \to \infty$ by Theorem 3.1. Thus (i) holds. If f is ε-optimal, then $V \le V_f + \varepsilon = U_f^N V_f + \varepsilon \le U_f^N V + \varepsilon$ for $N \in \mathbb{N}$. Furthermore, for $N \in \mathbb{N}$, $U^N v_0 - U^N V \le \beta^N H^N(v_0 - V)$ and $U_f^N v_0 - U_f^N V = \beta^N H_f^N(v_0 - V)$. Applied to $v_N - v_{N,f} \le v_N - V - (v_{N,f} - U_f^N V) + \varepsilon$ we then get (9.1), which implies (ii), since $\beta^N e_{N,f} \le \beta^N e_N \le \beta^N \|e_N\| \to 0$ by Proposition 2.3(a_3). □

Combining the value iteration with an extrapolation as in Theorem 5.1, we get $w_n^- \le V_{f_n} \le V \le w_n^+$ for suitable $w_n^\pm \in \mathfrak{V}$ at step n of iteration. Then Theorem 9.1(ii) works with $c^\pm := \pm \sup_{j \in J}\{\pm[v_0(s) - w_n^\mp(s)]\}$, and $\varepsilon := w_n^+ - w_n^-$. The upper bound in (9.1) can be simplified to $[\max\{0, c^+\} - c^-]\beta^N e_N + \varepsilon$, since $c^- \le \inf_{s \in S}\{v_0(s) - V(s)\} \le 0$. Using the following Proposition 9.2, e_N can be bounded from above by $\|e_m\|^{\lfloor N/m \rfloor}$, and $\|e_m/e_{m-1}\|^{N-m} e_m$, respectively, for $N \ge m$.

PROPOSITION 9.2. *For $m, n \in \mathbb{N}$ it holds that*

(i) $\beta^n e_n \le \max_{0 \le t < m}\{\beta^t e_t\}(\beta^m \|e_m\|)^{\lfloor n/m \rfloor}$,

(ii) $e_n \le \|e_m/e_{m-1}\|^{n-m} e_m$, $m \le n$.

Proof. (i) Rewrite n as $km + \nu$ with $k := \lfloor n/m \rfloor$, $\nu := n - m\lfloor n/m \rfloor$. Then, since $e_{t+\nu} \le \|e_t\| \cdot e_\nu$ for $t \ge 0$, we get $\beta^n e_n \le \beta^\nu e_\nu \beta^{mk}\|e_{mk}\| \le \beta^\nu e_\nu (\beta^m \|e_m\|)^k \le \max_{0 \le t < m}\{\beta^t e_t\}(\beta^m \|e_m\|)^k$.

(ii) Note that $e_m \le e_{m-1}$. Let $c := \|e_m/e_{m-1}\|$. Then monotonicity of H implies $e_n = H^{n-m} e_m \le H^{n-m}(c \cdot e_{m-1}) = cH^{n-m-1} e_m \le \cdots \le c^{n-m} e_m$. □

10. A turnpike theorem. Next we are interested in a turnpike theorem on the basis of ε-optimal decision rules. For $\beta < \beta^*$, let F^* (resp., F_n^*) be the set of all maximizers of LV (resp., Lv_{n-1}). Then the number

$$N^* := \inf\{N \in \mathbb{N} \mid F_n^* \subset F^* \text{ at all stages } n \ge N\}$$

is called the turnpike horizon of F^*, provided N^* is finite. It is shown in the following Theorem 10.1 that a turnpike horizon N^* exists for an MDP with a finite state space S. The proof is constructive and leads to an upper bound for N^*. Then, applied

to $MDP^{(K,c)}$ with countable S, the extension of each maximizer $f_n^{(K,c)} \in F_n^{*(K,c)}$, $n \geq N^{*(K,c)}$, is an ε-optimal decision rule in MDP in the sense of Theorem 8.2(ii).

We exclude the trivial case $F^* = F$ since then $N^* = 1$, and suppose S to be finite. Let $\mathrm{sp}v := \max_{s\in S} v(s) - \min_{s\in S} v(s)$ denote the span of v extended to S with $v = 0$ on J_0. Furthermore, let $D^*(s) \subset D(s)$, $s \in J$, be the set of all maximum points of $LV(s,\cdot)$. Put $B := \{s \in J \mid D^*(s) \neq D(s)\}$. Since B is finite, there exists $\gamma \in (0,\infty)^B$

$$\gamma(s) := V(s) - \max_{a\in D(s)-D^*(s)} \{LV(s,a)\}, \quad s \in B.$$

To give a finite upper bound N, say, for N^*, we use

$$\delta(s) := 1 - \max_{a'\in D^*(s)} \min_{a\in D(s)-D^*(s)} \sum_{j\in S} \min\{p(s,a,j), p(s,a',j)\}, \quad s \in B,$$

and $\mu := \beta \cdot \max_{s\in B} \delta(s)/\gamma(s)$. Then, since $\lim_{k\to\infty} \mathrm{sp}(v_k - V) = 0$, we have

$$N := \inf\{n \in \mathbb{N} \mid \mathrm{sp}(v_{k-1} - V) \cdot \mu < 1 \text{ for all } k \geq n\} < \infty.$$

To simplify the computation of N, note that $\mathrm{sp}(U^t v_{k-1} - U^t V) \leq \beta^t \|e_t\| \mathrm{sp}(v_{k-1} - V)$ for all $t, k \in \mathbb{N}$. Thus, if $\beta < \beta^*$, we have some $m \in \mathbb{N}$ such that $\beta^m \|e_m\| < 1$, and it holds that

$$N = \inf\{n \in \mathbb{N} \mid \mathrm{sp}(v_{k-1} - V) \cdot \mu < 1 \text{ for all } n \leq k < n + m\}. \tag{10.1}$$

Put

$$\kappa := 1 - \min_{s'\in S} \max_{a'\in D^*(s')} \min_{(s,a)\in D} \sum_{j\in S} \min\{p(s,a,j), p(s',a',j)\}.$$

If $\beta\kappa < 1$, then, for $n \in \mathbb{N}$,

$$\bar{N}(n) := \inf\left\{k \geq n \mid \mathrm{sp}(v_{n-1} - V) \cdot (\beta\kappa)^{k-n}\mu < 1\right\} < \infty.$$

Alternatively we may use $n \in \mathbb{N}$ and $m \in \mathbb{N}$ such that $\beta\|e_m\|^{1/m} < 1$ in order to get

$$\hat{N}(m,n) := \inf\left\{k \geq n \mid \mathrm{sp}(v_{n-1} - V) \cdot \max_{0\leq t<m} \{\beta^t \|e_t\|\} (\beta^m \|e_m\|)^{\lfloor (k-n)/m \rfloor} \mu < 1\right\} < \infty.$$

THEOREM 10.1. *Assume S to be finite. Let $\beta < \beta^*$ and $F^* \neq F$. Then $N^* \leq N < \infty$. Furthermore,*

(i) *if $\beta\kappa < 1$, then $N = \inf\{t \in \mathbb{N} \mid \mathrm{sp}(v_{t-1} - V) \cdot \mu < 1\} \leq \bar{N}(n) < \infty$, $n \in \mathbb{N}$;*

(ii) *if $m \in \mathbb{N}$ such that $\beta\|e_m\|^{1/m} < 1$, then $N \leq \hat{N}(m,n) < \infty$ for $n \in \mathbb{N}$.*

Proof. (a) Extend $v \in \mathbb{R}^J$ to $v \in \mathbb{R}^S$ with $v = 0$ on J_0. Let $f, h \in F$. Fix $s, s' \in S$. For $j \in S$ set $x_j := p(s, f(s), j)$, $y_j := p(s', h(s'), j)$. Then, using $(x-y)^+ = x - x \wedge y$, $(x-y)^- = y - x \wedge y$ for $x, y \in \mathbb{R}$,

$$\begin{aligned} H_f v(s) - H_h v(s') &= \sum_{j\in S} (x_j - y_j)^+ v(j) - \sum_{j\in S} (x_j - y_j)^- v(j) \\ &\leq \sum_{j\in S} x_j \cdot \max v - \sum_{j\in S} y_j \cdot \min v - \sum_{j\in S} x_j \wedge y_j \cdot \mathrm{sp}v \\ &= \left[1 - \sum_{j\in S} p(s, f(s), j) \wedge p(s', h(s'), j)\right] \cdot \mathrm{sp}v. \end{aligned}$$

(b) Fix $k \geq N$. We have to show that $F_k^* \subset F^*$. Assume $F_k^* \not\subset F^*$. Then there exists a maximizer f_k of Lv_{k-1} with $f_k(s) \notin D^*(s)$ for some $s \in J$. By definition, $s \in B$. Now, by applying (a) with $s = s'$, $v = v_{k-1} - V$, $f = f_k$, and $h \in F^*$ such that $\delta(s) = 1 - \min_{a \in D(s) - D^*(s)} \sum_{j \in S} p(s,a,j) \wedge p(s,h(s),j)$, we get

$$\begin{aligned}
&U_h v_{k-1}(s) - U_{f_k} v_{k-1}(s) \\
&\quad = U_h v_{k-1}(s) - U_h V(s) + U_h V(s) - U_{f_k} V(s) - (U_{f_k} v_{k-1}(s) - U_{f_k} V(s)) \\
&\quad = V(s) - U_{f_k} V(s) - [\beta H_{f_k}(v_{k-1} - V)(s) - \beta H_h(v_{k-1} - V)(s)] \\
&\quad \geq \gamma(s) - \beta\delta(s) \cdot \operatorname{sp}(v_{k-1} - V) \\
&\quad \geq \gamma(s)[1 - \mu \cdot \operatorname{sp}(v_{k-1} - V)].
\end{aligned}$$

Hence $U_h v_{k-1}(s) - U_{f_k} v_{k-1}(s) > 0$, which is the desired contradiction that f_k is a maximizer of Lv_{k-1}.

(c) We prove $N \leq \bar{N}(n)$. Put $Q(s',a') := 1 - \min_{(s,a) \in D} \sum_{j \in S} p(s,a,j) \wedge p(s',a',j)$. Select $h(s') \in D^*(s')$ as the minimum point of $Q(s',\cdot)$, $s' \in S$. Applying (a) with $v = v_{k-1} - V$, and $f = f_k$, we have for suitable $s, s' \in S$

$$\begin{aligned}
\operatorname{sp}(v_k - V) &= (v_k - V)(s) - (v_k - V)(s') \\
&\leq \beta H_{f_k}(v_{k-1} - V)(s) - \beta H_h(v_{k-1} - V)(s') \\
&\leq \beta Q(s', h(s')) \cdot \operatorname{sp}(v_{k-1} - V) \\
&\leq \beta \max_{s' \in S} \min_{a' \in D^*(s')} Q(s', h(s')) \cdot \operatorname{sp}(v_{k-1} - V) \\
&= \beta\kappa \cdot \operatorname{sp}(v_{k-1} - V).
\end{aligned}$$

This verifies, since $\beta\kappa < 1$, formula (10.1) and that $N \leq \bar{N}(n) < \infty$.

(d) Exploiting $\operatorname{sp}(U^t v_{n-1} - U^t V) \leq \beta^t \|e_t\| \cdot \operatorname{sp}(v_{n-1} - V)$ and Proposition 9.2(i), the proof of (ii) is similar to (i). □

Theorem 10.1 improves (in case of $\beta < \beta^*$) Theorems 2.3, 2.4, and 3.1 in Hinderer and Hübner (1977) in three respects: a larger γ, smaller δ and κ, and a slightly improved upper bound for $\|e_{n-1}\|$, which results from Proposition 9.2(ii).

REFERENCES

D. P. Bertsekas, *Dynamic Programming and Optimal Control*, Vol. I/II, Athena Scientific, Belmont, MA, 2000/2001.

E. Bohl, *Monotonie: Lösbarkeit und Numerik bei Operatorgleichungen*, Springer-Verlag, Berlin, 1974.

N. Branger, *Bewertung nicht redundanter Finanzderivate mittels Entropie und Cross-Entropie*, Dissertation, Fak. Wirtschaftsw., Univ. Karlsruhe, Karlsruhe, Germany, 2001.

M. H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, New York, 1970.

R. M. Dudley, *Real Analysis and Probability*, Wadsworth and Brooks, Pacific Grove, CA, 1989.

K. M. van Hee and J. Wessels, *Markov decision processes and strongly excessive functions*, Stochastic Process. Appl., 8 (1978/79), pp. 59–76.

O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.

K. Hinderer, *On the application of dynamic programming to economic lot size problems*, Oper. Res. Verf., 10, (1970), pp. 57–75.

K. Hinderer and G. Hübner, *An improvement of J. F. Shapiro's turnpike theorem for the horizon of finite stage discrete dynamic programs*, in Transactions of the 7th Prague Conference 1974, Reichel, Dordrecht, 1977, pp. 245–255.

K. Hinderer and K.-H. Waldmann, *Cash management in a randomly varying environment*, Eur. J. Oper. Res., 130 (2001), pp. 468–485.

K. Hinderer and K.-H. Waldmann, *The critical discount factor for finite Markovian decision processes with an absorbing set*, Math. Methods Oper. Res., 57 (2003), pp. 1–19.
A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tracts 51, Mathematisch Centrum, Amsterdam, 1974.
G. Hübner, *Bounds and good policies in stationary finite-stage Markovian decision problems*, Adv. Appl. Probab., 12 (1980), pp. 154–173.
T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.
J. M. McNamara, *Optimal life histories: A generalization of the Perron-Frobenius theorem*, Theor. Popul. Biol., 40 (1991), pp. 230–245.
T. Ogiwara, *Nonlinear Perron-Frobenius problem on an ordered Banach space*, Jap. J. Math. New Ser., 21 (1995), pp. 43–103.
S. R. Pliska, *On the transient case for Markov decision chains with general state spaces*, in Dynamic Programming and Its Applications, M. L. Puterman, ed., Academic Press, New York, 1978, pp. 335–349.
U. Rieder, *On Dynamic Programming with Unbounded Reward Functions*, Working Paper, Universität Hamburg, Hamburg, Germany, 1976.
M. Schäl, *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 32 (1975), pp. 179–196.
H. Schellhaas, *Zur Extrapolation in Markoffschen Entscheidungsmodellen mit Diskontierung*, Z. Oper. Res. Ser. A., 18 (1974), pp. 91–104.
L. I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*, Wiley, New York, 1999.
K. Sladky, *Bounds on discrete dynamic programming recursions* I. *Models with non–negative matrices*, Kybernetika, 16 (1980), pp. 526–547.
P. Tseng, *Solving H-horizon stationary Markov decision problems in time proportional to* $\log(H)$, Oper. Res. Lett., 9 (1990), pp. 287–297.
A. F. Veinott, *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist., 40 (1969), pp. 1635–1660.
K.-H. Waldmann, *On bounds for dynamic programs*, Math. Oper. Res., 10 (1985), pp. 220–232.
K.-H. Waldmann, *Bounds for the distribution of the run length of one-sided and two-sided CUSUM quality control schemes*, Technometrics, 28 (1986a), pp. 61–67.
K.-H. Waldmann, *Bounds for the distribution of the run length of geometric moving average charts*, J. Roy. Statist. Soc. Ser. C, 35 (1986b), pp. 151–158.
P. Whittle, *Optimization Over Time*, Vol. II. Dynamic Programming and Stochastic Control, John Wiley, Chichester, 1983.
W. H. M. Zijm, *Bounding Functions for Markov Decision Processes in Relation to the Spectral Radius*, Oper. Res. Verf. 33, 1978, pp. 461–472.
W. H. M. Zijm, *Nonnegative Matrices in Dynamic Programming*, Mathematical Centre Tracts 167, Mathematisch, Centrum, Amsterdam, 1983.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2132–2156

A NEW APPROACH TO DETECTABILITY OF DISCRETE-TIME INFINITE MARKOV JUMP LINEAR SYSTEMS*

EDUARDO F. COSTA†, JOÃO B. R. DO VAL‡, AND MARCELO D. FRAGOSO§

Abstract. This paper deals with detectability for the class of discrete-time Markov jump linear systems (MJLS) with the underlying Markov chain having countably infinite state space. The formulation here relates the convergence of the output with that of the state variables, and due to the rather general setting, a novel point of view toward detectability is required. Our approach introduces invariant subspaces for the autonomous system and exhibits the role that they play. This allows us to show that detectability can be written equivalently in term of two conditions: stability of the autonomous system in a certain invariant space and convergence of general state trajectories to this invariant space under convergence of input and output variables. This, in turn, provides the tools to show that detectability here generalizes uniform observability ideas as well as previous detectability notions for MJLS with finite state Markov chain, and allows us to solve the jump-linear-quadratic control problem. In addition, it is shown for the MJLS with finite Markov state that the second condition is redundant and that detectability retrieves previously well-known concepts in their respective scenarios. Illustrative examples are included.

Key words. detectability, stochastic systems, Markov jump systems, infinite Markov state space, optimal control

AMS subject classifications. 93E03, 93B07, 93E20, 60J05, 34A30, 93B12

DOI. 10.1137/S036301290342992X

1. Introduction.

Structural concepts such as observability and detectability have a solid ground in system theory, as the imposing literature for linear and linear-Gaussian systems conveys (see, e.g., [15]). For instance, in control problems, detectability firmly associates the solution for the optimal problems with stability of the corresponding controlled system, whereas, for filtering, it makes the system observations meaningful for state estimates by connecting convergence of the output with convergence of the state. Although the theory involving these concepts is quite developed and a number of results are available in the context of linear deterministic systems, there is still a great deal of research activity in this area (see, e.g., [13, 17] and references therein).

Among the most important properties of detectability for the linear deterministic scenario, we mention that

(i) detectability can be expressed in terms of the parameters of the autonomous version of the system, e.g., by requiring that nonobserved modes of the autonomous system are stable.

(ii) Detectability generalizes observability.

Another important but less acknowledged property is that

*Received by the editors June 13, 2003; accepted for publication (in revised form) August 6, 2004; published electronically April 14, 2005. Research supported in part by FAPESP, Grant 01/12233-2, by CNPq, Grant 300721/86-2, by PRONEX Grant 015/98 "Control of Dynamical Systems" and by the IM-AGIMB grant.

http://www.siam.org/journals/sicon/43-6/42992.html

†USP-Inst. de Ciências Mathemáticas e de Computação, Depto. de Ciências de Comp. e Estatística, C.P. 668, 13560-970, São Carlos, SP, Brazil (efcosta@icmc.usp.br).

‡UNICAMP—Fac. de Engenharia Elétrica e de Computação, Depto. de Telemática, C.P. 6101, CEP 13081-970, Campinas, SP, Brazil (jbosco@dt.fee.unicamp.br).

§Lab. Nacional de Computação Científica—LNCC/CNPq, Av. Getúlio Vargas 333, CEP 25651-075, Petrópolis, RJ, Brazil (frag@lncc.br).

(iii) detectability is a necessary and sufficient condition to guarantee convergence of the state from convergence of the output (under regular nonsingular linear state feedback controls).

Property (iii) ensures that the optimal control solution is stabilizing and makes output observations meaningful in filtering problems.

Due to its generic formulation, these properties constitute a paradigm for more general contexts. The challenge then is how to devise a detectability concept for a certain class of systems that allows one to employ the structure of the system to retrieve properties (i)–(iii).

In this spirit, the authors have recently developed a notion of detectability (called weak detectability) that generalizes previous detectability ideas for MJLS with finite Markov chain state, retrieves the properties (i)–(iii), and allows an associate observability matrix, in an extension to the well-known deterministic concepts, see [1] and [2]. In this process all but one[1] of the linear deterministic concepts are retrieved.

However, as far as the authors are aware, these ideas have no parallel in more complex scenarios such as the MJLS with countably infinite state space of the Markov chain. This is a rather general class of systems that includes the classes of finite MJLS and linear deterministic systems, as well as deterministic time varying systems. Previous works dealing with infinite MJLS are [7, 8, 9, 10].For this class of systems, up to this date there is no detectability concept that retrieves properties (i)–(iii) above. For instance, the stochastic notion in [7] can be expressed in terms of the autonomous system data, thus satisfying (i), but (ii) does not hold and only the sufficiency part of (iii) holds; in [4] we derive a detectability notion in the perspective of (iii) for which (ii) holds, but it does not satisfy (i).

These shortfalls come, in part, from the analytical complexity inherent to the infinite many Markov state contexts, and the loss of some friendlier structures of the simpler cases. In particular, the main difficulty arises from the fact demonstrated in this paper that converging input and output *do not* ensure convergence of state trajectory to the observed space; see Example 2 in connection. In the simpler case of finitely many Markov states, the above convergence relation holds, and apart from ensuring stability within the observed space, with detectability it guarantees convergence of the state trajectory to the origin. This is the mechanism that fails here, and in this regard we can conclude that any detectability concept with the perspective of (i) (stable nonobserved modes) by itself cannot provide the property in (iii) and thus, it cannot ensure that the optimal control is stabilizing.

In this paper, with the aim of studying detectability for MJLS with countably infinite state space of the Markov chain and to retrieve (i)–(iii), we introduce a novel point of view toward detectability by considering the paradigmatic property in (iii) as a general, direct, and intuitive notion of detectability, which relates the convergence of the input and output with that of the state variables. Then we introduce certain invariant subspaces for the autonomous system, which play a key role to relate detectability with stability and convergence of the state trajectory; this allows us to show that detectability here generalizes uniform observability ideas as well as previous detectability notions for MJLS with finite state Markov chain, and to solve the jump-linear-quadratic control problem. In order to show some nuances of the approach developed here, and to clarify the role of some tools, we also analyze the MJLS with finite state Markov chain and present illustrative examples.

[1]The observability idea that after a number of observations that equals the system dimension, the initial state value can be precisely retrieved. This is inherently a nonstochastic idea.

An outline of the content of this paper is as follows. In section 2 we provide the bare essential of notations, state the model, and discuss the general ideas of the paper. Section 3 provides some preliminaries. Necessary and sufficient conditions for detectability are treated in section 4, and some sufficient conditions are presented in section 5. The finite MJLS is analyzed in section 6, and the control problem is studied in section 7. Some illustrative examples are exhibited in section 8. Finally, section 9 presents some conclusions.

2. Problem formulation and general ideas. Let $\mathbb{R}^n$ represent the usual linear space of all n-dimensional vectors and $\mathcal{R}^{r,n}$ (respectively, $\mathcal{R}^n$) the normed linear space formed by all $r \times n$ real matrices (respectively, $n \times n$). For $V \in \mathcal{R}^{n,r}$, V' denotes the transpose of V. $\sigma^+(V)$ and $\sigma^-(V)$ stand, respectively, for the largest and smallest singular value of V and $\|V\| = \sigma^+(V)$. For $V, W \in \mathcal{R}^n$, $V > W$ ($V \geq W$) indicates that $V - W$ is positive definite (semidefinite).

Let $\mathcal{H}_\infty^{r,n}$ denote the linear space formed by sequences of matrices $H = \{H_i \in \mathcal{R}^{r,n}; i \in \mathcal{Z}\}$ such that $\sup_{i\in\mathcal{Z}} ||H_i|| < \infty$; also, $\mathcal{H}_\infty^n \equiv \mathcal{H}_\infty^{n,n}$ and $\|H\|_\infty = \sup_{i\in\mathcal{Z}} \|H_i\|$. For $H, V \in \mathcal{H}_\infty^n$, $H \geq V$ indicates that $H_i \geq V_i$ for each $i \in \mathcal{Z}$; similarly, for $H \in \mathcal{H}_\infty^{r,n}$ and $V \in \mathcal{H}_\infty^{n,s}$, the "product" HV indicates the element of $\mathcal{H}_\infty^{r,s}$ formed by the sequence $\{H_iV_i, i \in \mathcal{Z}\}$, and equivalent understanding should apply to any basic mathematical operation involving elements of $\mathcal{H}_\infty^{r,n}$. In what follows, capital letters denote elements of $\mathcal{H}_\infty^{r,n}$, and capital letters with an index denote elements of $\mathcal{R}^{r,n}$.

The system we deal with is the discrete-time MJLS with infinite countably Markov chain, defined in a fixed stochastic basis $(\Omega, \mathfrak{F}, (\mathfrak{F}_k), \mathcal{P})$ by

$$\Psi : \begin{cases} x(k+1) = A_{\theta(k)}x(k) + B_{\theta(k)}u(k), & k \geq 0, \\ y(k) = C_{\theta(k)}x(k) + D_{\theta(k)}u(k), & x(0) = x,\ \theta(0) = \theta, \end{cases} \tag{1}$$

where y is the output process and u is the input, an $(\mathfrak{F}_k)$-adapted process. The mode θ is the state of an underlying discrete-time Markov chain $\Theta = \{\theta(k); k \geq 0\}$ taking values in $\mathcal{Z} = \{1, 2, \dots\}$ and having a stationary transition probability matrix $\mathbb{P} = [p_{ij}]$, $i, j \in \mathcal{Z}$. The state of the system is the compound variable $(x(k), \theta(k))$. The matrices A_i belong to the sequence of matrices $A \in \mathcal{H}_\infty^n$, and similarly for $B \in \mathcal{H}_\infty^{n,r}$, $C \in \mathcal{H}_\infty^{q,n}$, and $D \in \mathcal{H}_\infty^{q,r}$. In addition, without loss of generality, we also assume that $C'D = 0$.

In this paper we deal with detectability for systems described by (1). The departure point is the following concept of detectability that follows from property (iii) of section 1. We emphasize that the specific notion of convergence is not relevant; the essence of the concept is the relation among convergence of state, input and output, and a particular sense of convergence is adopted later in connection with the choice of the cost functional.

DEFINITION 1 (detectability). *The system Ψ is detectable if the state converges provided that the output and the input converge.*

With the detectability concept above at hands, which trivially embraces property (iii) in the introduction, the issues pursued here are primarily summarized as follows:

(I) Relate the concept with the autonomous version of the system, aiming at mimicking item (i) mentioned in the introduction.

(II) Show that it retrieves property (ii) mentioned in the introduction.

(III) Investigate the extent to which the above concept is related to the weak detectability concept in [1] and [2] for MJLS, and the usual concept for deterministic linear systems.

We consider a cost functional that is an ℓ_2-measurement of the output (the expected accumulated energy in the output process path),

$$\mathcal{Y}_u(x,\theta) = E_{x,\theta}\left\{\sum_{k=0}^{\infty} |y(k)|^2\right\}, \tag{2}$$

defined for an admissible control u whenever $x(0) = x$ and $\theta(0) = \theta$. We also denote for the autonomous system obtained from Ψ with $u \equiv 0$,

$$\mathcal{Y}_0(x,\theta) = \mathcal{Y}_{u\equiv 0}(x,\theta). \tag{3}$$

In agreement with (2), we adopt the corresponding ℓ_2-convergence notion for each Ψ-processes, namely, we say that the output y converges whenever $\mathcal{Y}_u(\cdot,\cdot) < \infty$; similar notion holds for u and x.

Our approach starts from a novel point of view, which hinges on the following steps. We first locate an invariant linear subspace for the autonomous system, in the sense that the trajectories remain almost surely confined to it. Then we indicate the role that the invariant space plays in the convergence of an arbitrary state trajectory, showing that the existence of an invariant space for which the autonomous system is stable, together with the convergence to this set of an arbitrary trajectory, is equivalent to convergence to the origin of such a trajectory (see section 4 and Theorem 12).

Note that the announced result reduces to a tautology if the invariant space is taken to be the origin, and to make the above result suitable to deal with (I), we seek the largest of such an invariant space. It turns out to be the linear subspace $\mathcal{F} = \{(x,\theta) : \mathcal{Y}_0(x,\theta) < \infty\}$, and in Theorem 18 we state that detectability according to Definition 1 is equivalent to requiring that

(A1) the autonomous system is stable in $\mathcal{F}$,

(A2) the state x converges to $\mathcal{F}$ provided that both y and u converge.

Notice that condition (A1) accounts for the autonomous version of system Ψ only, and it is consistent with the notion of detectability for finite dimensional linear deterministic systems. Together with condition (A2) for system Ψ (not only the autonomous version), they build the essentials to complete the aforementioned mechanism yielding (iii). Due to (A2), a complete counterpart for property (i) is not viable in the present setup (see Example 3 in connection), and any attempt to enlarge $\mathcal{F}$ is worthless, as we show in Lemma 17.

Section 5 addresses (II), where we show that detectability according to Definition 1 generalizes uniform observability as in [1, 3, 5, 12], which, by its turn, generalizes previous observability concepts for MJLS, like those in [11]. We also show that an earlier ℓ_2-detectability concept in [7] is stricter than detectability. Moreover, we introduce a notion of uniform observability in the invariant space $\mathcal{F}$ that serves as a sufficient condition for (A2). See Proposition 28 for a summary.

Regarding (III), in section 6 we show that $\mathcal{F}^{\perp}$ is uniformly observable in the finite Markov chain case, which renders condition (A2) always true. Thus, we have that detectability is equivalent to (A1) in the finite case, allowing us to show that the weak detectability in [2] and the usual detectability concept in the deterministic linear case are necessary and sufficient conditions (in their particular contexts) for detectability according to Definition 1 (see Remark 3). The fact that (A2) holds true for the case in which the Markov chain is finite explains why no such condition appears in those simpler scenarios. By contrast, (A2) may fail in the infinite Markov chain case, as illustrated in Example 2.

Another important feature of the setting and results here is that, unlike previous ones, the focus is not constrained (i.e., is not *ad hoc*) to the optimal jump-linear-quadratic (JLQ) control and/or controls in linear feedback form, where detectability appears as a dual notion to stabilizability. It covers any $(\mathfrak{F}_k)$-adapted converging control that induces a finite cost $\mathcal{Y}_u$ for each initial state, assuring that it is stabilizing, and clearly encompassing the optimal solution (see Remark 1). In particular for the JLQ control, we show that the solution to the associated infinite coupled algebraic Riccati equation is unique (see section 7).

3. Preliminaries. In this section we introduce some basic machinaries, which will allow us to devise our approach toward detectability for (1). We consider the autonomous version of (1), which will be essential to relate detectability with stability and convergence of the state trajectory (see (A.1) and (A.2) in section 2). We define the various essential elements such as invariant space, some notions of convergence, some useful spaces, operators, and some preliminary results.

We consider the autonomous version of system (1):

$$\Psi_0 : \begin{cases} x_0(k+1) = A_{\theta(k)}x_0(k), & k \geq 0, \\ y_0(k) = C_{\theta(k)}x_0(k), & x_0(0) = x, \theta(0) = \theta. \end{cases}$$

Sometimes we refer to the autonomous system by the pair $(A, \mathbb{P})$ or by the triplet $(A, C, \mathbb{P})$. In addition, in what follows, for each $i \in \mathcal{Z}$, let $\mathcal{S}_i \subset \mathbb{R}^n$ stand for a vector subspace and let $\mathcal{S} = \{\mathcal{S}_i, i \in \mathcal{Z}\}$.

DEFINITION 2 (Ψ_0-invariant space). *Consider the autonomous system Ψ_0. We say that $\mathcal{S}$ is an invariant space if $x_0(k) \in \mathcal{S}_{\theta(k)}$ implies that $x_0(t) \in \mathcal{S}_{\theta(t)}$ almost surely (a.s.) for each $t \geq k$.*

DEFINITION 3 (projections onto $\mathcal{S}^{\perp}$). *For each $i \in \mathcal{Z}$, let $P_i \in \mathcal{R}^n$ denote the orthogonal projection onto $\mathcal{S}_i^{\perp}$. Clearly, $P = \{P_i, i \in \mathcal{Z}\} \in \mathcal{H}_\infty^n$.*

DEFINITION 4 (Ψ_0-convergence). *We say that $x(\cdot)$ converges (in the ℓ_2 sense) to the Ψ_0-invariant space $\mathcal{S}$ if*

$$\sum_{k=0}^{\infty} E_{x,\theta}\{|P_{\theta(k)}x(k)|^2\} < \infty.$$

We say that $x(\cdot)$ converges if it converges to the trivial Ψ_0-invariant space $\mathcal{S} = 0$.

DEFINITION 5 (ℓ_2-stability). *Consider the autonomous system Ψ_0. We say that $(A, \mathbb{P})$ is ℓ_2-stable in the invariant space $\mathcal{S}$ if $x_0(\cdot)$ converges for each initial condition $\theta \in \mathcal{Z}$ and $x \in \mathcal{S}_\theta$. We say that $(A, \mathbb{P})$ is ℓ_2-stable if it is ℓ_2-stable in $\mathcal{S}$ with $\mathcal{S}_i = \mathbb{R}^n$, $i \in \mathcal{Z}$*

Notice that $x(\cdot)$ converges if and only if $\sum_{k=0}^{\infty} E\{|x(k)|^2\} < \infty$, since $P = I$ whenever $\mathcal{S}$ is trivial. Also, ℓ_2-stability of $(A, \mathbb{P})$ is equivalent to convergence of $x_0(\cdot)$ for each initial condition $\theta \in \mathcal{Z}$ and $x \in \mathbb{R}^n$.

We will need the following property related with the concept of ℓ_2-stability in $\mathcal{S}$ and the projections P.

LEMMA 6. *Assume that $(A, \mathbb{P})$ is ℓ_2-stable in $\mathcal{S}$. Then, $(A - AP, \mathbb{P})$ is ℓ_2-stable.*

Proof. Consider the following version of system Ψ:

$$x_P(k+1) = (A_{\theta(k)} - A_{\theta(k)}P_{\theta(k)})x_P(k), \quad x_P(0) = x,\ \theta(0) = \theta. \tag{4}$$

Let us employ the trajectory of system Ψ_0, $x_0(k) = A_{\theta(k-1)} \cdots A_{\theta(0)}x_0$, with initial condition x_0 being the projection of x into $\mathcal{S}_\theta$, i.e., $x_0 = (I - P_\theta)x$. Since $\mathcal{S}$ is a invariant space, we have that $x_0(k) \in \mathcal{S}_{\theta(k)}$, $k \geq 0$.

We start by showing inductively that $x_P(\cdot)$ evolves as an open-loop trajectory according to $x_P(k) = x_0(k)$, for $k \geq 1$, for all $x \in \mathbb{R}^n$. For $k = 1$ we have that

$$x_P(1) = (A_\theta - A_\theta P_\theta)x = A_\theta(I - P_\theta)x = x_0(1).$$

From the induction assumption, we have that $x_P(k) = x_0(k)$ for $k \geq 1$, and recalling that $x_0(k) \in \mathcal{S}_{\theta(k)}$ we evaluate

$$x_P(k+1) = (A_{\theta(k)} - A_{\theta(k)}P_{\theta(k)})x_0(k) = A_{\theta(k)}x_0(k) = x_0(k+1)$$

and the induction is completed. Due to the facts that (i) $(A, \mathbb{P})$ is ℓ_2-stable in $\mathcal{S}$, (ii) $x_P(1) = x_0(1) \in \mathcal{S}_{\theta(1)}$ a.s., and (iii) $x_P(k)$, $k \geq 1$ evolves as a trajectory of the autonomous system for any $x_P(0) = x$, $\theta(0) = \theta$, we have from the definition of ℓ_2-stability in $\mathcal{S}$ that $x_P(\cdot)$ converges and, thus, $(A - AP, \mathbb{P})$ is ℓ_2-stable. □

In what follows, we introduce a certain space $\mathcal{H}_F^n$, an element $X(k)$ related with the second moment of the state, an operator $\mathcal{L}$ related with the evolution of $X(k)$, and some associated results which will be useful to present the results of the paper in a concise manner.

Let $\mathcal{H}_1^n$ denote the linear space formed by sequences of matrices $H = \{H_i = H_i' \geq 0; i \in \mathcal{Z}\}$ such that $\sum_{i\in\mathcal{Z}} tr\{H_i\} < \infty$. Let $\mathcal{H}_F^n \subset \mathcal{H}_1^n$ denote the closed cone formed by sequences of symmetric positive semidefinite matrices $H = \{H_i = H_i' \geq 0; i \in \mathcal{Z}\}$. For $H, V \in \mathcal{H}_F^n$ we define the inner product

$$\langle H, V\rangle = \sum_{i\in\mathcal{Z}} tr\{H_i'V_i\}$$

and the Frobenius norm

$$\|H\|_F = \langle H, I\rangle. \tag{5}$$

Recall from the definition of the Ψ_0-invariant subspace $\mathcal{S}$ that $\mathcal{S}_i = \{x : P_i x = 0\}$. In connection, we define the spaces $\bar{\mathcal{S}} = \{H \in \mathcal{H}_F^n : PHP' = 0\} \subset \mathcal{H}_F^n$ and $\bar{\mathcal{S}}^\perp = \{H \in \mathcal{H}_F^n : H - PHP' = 0\}$. PHP' is the orthogonal projection of H onto $\bar{\mathcal{S}}^\perp$; indeed, P inherits from P_i the property that $P^2 = P$, and it is easy to check that $\langle PHP', H - PHP'\rangle = \langle H, PHP' - P^2HP^{2\prime}\rangle = 0$.

Definition 7 (convergence in $\mathcal{H}_F^n$). *We refer to convergence of sequences in $\mathcal{H}_F^n$ in the ℓ_1 sense: we say that a sequence $H(\cdot) \in \mathcal{H}_F^n$ converges to the space $\bar{\mathcal{S}}$ whenever $\sum_{k=0}^\infty \|PH(k)P'\|_F < \infty$; we say that $H(\cdot)$ converges if it converges to the trivial space $\bar{\mathcal{S}} = 0$.*

We define $X(\cdot) \in \mathcal{H}_F^n$ and $U(\cdot) \in \mathcal{H}_F^r$ as

$$\begin{aligned} X_i(k) &= E\{x(k)x(k)'1_{\{\theta(k)=i\}}\} \\ U_i(k) &= E\{u(k)u(k)'1_{\{\theta(k)=i\}}\}\ \forall i \in \mathcal{Z}, k \geq 0, \end{aligned} \tag{6}$$

where $1_{\{\cdot\}}$ is the Dirac indicator function. We write $X_0(\cdot)$ when we refer to the autonomous system. We define $\mathcal{Y}_u^{t,T}$ similarly to the functional $\mathcal{Y}$ in (2) as follows:

$$\mathcal{Y}_u^{t,T}(x,\theta) = E_{x,\theta}\left\{\sum_{k=t}^{t+T-1} |y(k)|^2\right\} = \sum_{k=t}^{t+T-1} \Big(\langle X(k), C'C\rangle + \langle U(k), D'D\rangle\Big) \tag{7}$$

whenever $x(0) = x, \theta(0) = \theta$; for simplicity we write $\mathcal{Y}_u^{t=0,T}(x,\theta) = \mathcal{Y}_u^T(x,\theta)$ and also $\mathcal{Y}_{u\equiv 0}^T(x,\theta) = \mathcal{Y}_0^T(x,\theta)$.

Using the notation above we can write $E_{x,\theta}\{|x(k)|^2\} = \|X(k)\|_F$ and this provides a connection between convergence in the ℓ_1 sense of $X(\cdot) \in \mathcal{H}_F^n$ with the ℓ_2 convergence of $x(\cdot)$. A further connection is presented in the next lemma; the proof is presented in Appendix 9.

LEMMA 8. *$x(\cdot)$ converges to $\mathcal{S}$ if and only if $X(\cdot)$ converges to $\bar{\mathcal{S}}$.*

Now, let us define for $V \in \mathcal{H}_\infty^{n,r}$ the linear operator $\mathcal{L}_V : \mathcal{H}_F^r \to \mathcal{H}_F^n$

$$\mathcal{L}_{Vi}(H) = \sum_{j\in\mathcal{Z}} p_{ji} V_j H_j V_j'. \tag{8}$$

It is shown in [7] that the limit in (8) is well defined. We denote $\mathcal{L}^0(H) = H$, and for $k \geq 1$, we can define $\mathcal{L}^k(H)$ recursively by $\mathcal{L}^k(H) = \mathcal{L}(\mathcal{L}^{k-1}(H))$. Also, $r_\sigma(\mathcal{L})$ denotes the spectral radius of $\mathcal{L}$. Operator $\mathcal{L}$ is related to system Ψ as follows; the result is adapted from [7].

PROPOSITION 9. *The following assertions hold:*

(i) $X_0(k+1) = \mathcal{L}_A(X_0(k)),\ k \geq 0$;

(ii) $(A, \mathbb{P})$ *is ℓ_2-stable if and only if* $r_\sigma(\mathcal{L}_A) < 1$.

For the nonautonomous system Ψ, the evolution of X is still related to the operator $\mathcal{L}$, as follows. See Appendix 9 for the proof.

LEMMA 10. *Let $\alpha \neq 0$. Then,*

$$X(k+1) \leq (1+\alpha^2)\mathcal{L}_A(X(k)) + (1+1/\alpha^2)\mathcal{L}_B(U(k)),\ \ k \geq 0.$$

The following basic properties concerning the operator $\mathcal{L}$, which are easy to check by inspection, will be useful.

PROPOSITION 11. *The following properties hold, for $V, W \in \mathcal{H}_\infty^n$ and $H, Y \in \mathcal{H}_F^n$:*

(i) $\mathcal{L}_{VW}(H) = \mathcal{L}_V(WHW')$;

(ii) $\mathcal{L}_{V+W}(H) \geq (1-\alpha^2)\mathcal{L}_V(H) + (1-1/\alpha^2)\mathcal{L}_W(H)\ \forall \alpha \neq 0$;

(iii) $\mathcal{L}_{V+W}(H) \leq (1+\alpha^2)\mathcal{L}_V(H) + (1+1/\alpha^2)\mathcal{L}_W(H)\ \forall \alpha \neq 0$;

(iv) $\mathcal{L}_V(H) \geq \mathcal{L}_V(Y)$ *whenever* $H \geq Y$;

(v) $||\mathcal{L}_V(H)||_F \leq ||V||_\infty^2 ||H||_F$.

We finish the section with the following facts that we believe are worth mentioning. $\bar{\mathcal{S}}$ inherits from $\mathcal{S}$ the property that it is a Ψ_0-invariant subspace, that is, $PX_0(k)P' = 0$, $k \geq 0$, implies that $PX_0(t)P' = 0$, $t \geq k$. The notion of convergence in $\mathcal{H}_F^n$ is usual, in the sense that a sequence $H(\cdot) \in \mathcal{H}_F^n$ converges to the space $\bar{\mathcal{S}}$ if and only if $\sum_{k=0}^\infty \inf_{V\in\bar{\mathcal{S}}} \|H(k) - V\|_F < \infty$. Actually, the proof follows immediately from the fact that for each $H(k)$ there exists $V \in \bar{\mathcal{S}}$ for which $\|H(k) - V\|_F = \|PH(k)P'\|_F$ (indeed, $V = H(k) - PH(k)P'$).

4. A necessary and sufficient condition for detectability. We show in section 4.1 that a general state trajectory $x(\cdot)$ converges if and only if there exists an invariant space $\mathcal{S}$ for which: (i) $(A, \mathbb{P})$ is ℓ_2-stable in $\mathcal{S}$ and (ii) $x(\cdot)$ converges to $\mathcal{S}$. In section 4.2 we introduce the Ψ_0-invariant space $\mathcal{F}$ and we show the appropriateness of $\mathcal{F}$ to formulate the equivalence between detectability and conditions (A1) and (A2).

4.1. Conditions for state convergence. In this section we examine the state convergence of system Ψ and its interplay with the Ψ_0-invariant subspace $\mathcal{S}$. The first requirement is the existence of a Ψ_0-invariant space $\mathcal{S}$ in which $(A, \mathbb{P})$ is ℓ_2-stable, namely, the condition (A1) holds.

Notice that (A1) does not impose any condition on the system Ψ (or, even on Ψ_0) in the subspaces orthogonal to $\mathcal{S}$. From this perspective, one can infer that (A1) can be employed only to establish the convergence of Ψ_0-trajectories that remain a.s. in $\mathcal{S}$ or, at most, of Ψ_0-trajectories that converge to $\mathcal{S}$. Surprisingly, the combination of (A1) with convergence to $\mathcal{S}$ of Ψ-trajectories guarantee the convergence of any trajectory of Ψ with converging inputs, as the next theorem shows. Clearly, if $\mathcal{S}$ is trivial, Theorem 12 becomes a tautology.

THEOREM 12. *Consider system Ψ and assume that the input converges. The state $x(\cdot)$ converges if and only if there exists an invariant space $\mathcal{S}$ such that the following conditions hold:*

(i) *$(A, \mathbb{P})$ is ℓ_2-stable in $\mathcal{S}$;*
(ii) *$x(\cdot)$ converges to $\mathcal{S}$.*

Proof. (Necessity.) Since $x(\cdot)$ converges to the origin, $\mathcal{S} = 0$ trivially satisfies (i) and (ii).

(Sufficiency.) We show that $x(\cdot)$ converges provided (i) and (ii) hold and $u(\cdot)$ converges. Recall from Lemma 6 that condition (i) provides that $(A - AP, \mathbb{P})$ is ℓ_2-stable for P the projection in $\mathcal{S}$ and from Proposition 9 (ii) we have that $r_\sigma(\mathcal{L}_{A-AP}) < 1$. Let $\alpha \neq 0$ be such that $(1+\alpha^2)^2 r_\sigma(\mathcal{L}_{A-AP}) < 1$. For ease of notation, we define the operators $\hat{\mathcal{L}}, \tilde{\mathcal{L}}, \bar{\mathcal{L}} : \mathcal{H}_F^n \to \mathcal{H}_F^n$ as

$$\begin{aligned} \hat{\mathcal{L}}(H) &= (1+\alpha^2)^2 \mathcal{L}_{A-AP}(H), \\ \tilde{\mathcal{L}}(H) &= (1+\alpha^2)(1+1/\alpha^2)\mathcal{L}_{AP}(H), \text{ and} \\ \bar{\mathcal{L}}(H) &= (1+1/\alpha^2)\mathcal{L}_B(H) \end{aligned}$$

for $H \in \mathcal{H}_F^n$. We also define the series $Z(\cdot)$ with $Z(k) \in \mathcal{H}_F^n$, $k \geq 0$, by

$$\begin{cases} Z(k+1) = \hat{\mathcal{L}}(Z(k)) + \tilde{\mathcal{L}}(X(k)) + \bar{\mathcal{L}}(U(k)), \ k \geq 0, \\ Z(0) = X(0). \end{cases}$$

Noticing that

$$Z(m) = \hat{\mathcal{L}}^m(X(0)) + \sum_{k=0}^{m-1} \hat{\mathcal{L}}^k\Big(\tilde{\mathcal{L}}(X(-k+m-1)) + \bar{\mathcal{L}}(U(-k+m-1))\Big),$$

we write

$$\begin{aligned} \sum_{m=0}^{\infty} \langle Z(m), I\rangle &= \sum_{m=0}^{\infty} \langle \hat{\mathcal{L}}^m(X(0)), I\rangle \\ &\quad + \sum_{m=0}^{\infty}\sum_{k=0}^{m-1} \Big\langle \hat{\mathcal{L}}^k\Big(\tilde{\mathcal{L}}(X(-k+m-1)) + \bar{\mathcal{L}}(U(-k+m-1))\Big), I\Big\rangle. \end{aligned} \tag{9}$$

For the second term in the right-hand side of (9) we evaluate

$$
\begin{aligned}
(10)\qquad &\sum_{m=0}^{\infty}\sum_{k=0}^{m-1}\left\langle \hat{\mathcal{L}}^k\Big(\tilde{\mathcal{L}}(X(m-k-1))+\bar{\mathcal{L}}(U(m-k-1))\Big), I\right\rangle\\
&=\sum_{k=0}^{\infty}\sum_{m=k+1}^{\infty}\langle \hat{\mathcal{L}}^k\Big(\tilde{\mathcal{L}}(X(m-k-1))+\bar{\mathcal{L}}(U(m-k-1))\Big), I\Big\rangle\\
&=\sum_{k=0}^{\infty}\left\langle \hat{\mathcal{L}}^k\left(\sum_{m=0}^{\infty}\tilde{\mathcal{L}}(X(m))+\bar{\mathcal{L}}(U(m))\right), I\right\rangle\\
&=\sum_{k=0}^{\infty}\langle \hat{\mathcal{L}}^k(\Upsilon), I\rangle,
\end{aligned}
$$

where we set

$$
\Upsilon=\sum_{m=0}^{\infty}\tilde{\mathcal{L}}(X(m))+\hat{\mathcal{L}}(U(m))=(1+\alpha^2)(1+1/\alpha^2)\sum_{m=0}^{\infty}\mathcal{L}_{AP}(X(m))
+(1+1/\alpha^2)\sum_{m=0}^{\infty}\mathcal{L}_B(U(m)).
$$

We need to show that Υ is well defined, i.e., that $\Upsilon\in\mathcal{H}_F^n$; the result is presented in the next lemma.

LEMMA 13. *$\|\sum_{k=0}^{\infty}\mathcal{L}_B(U(k))\|_F<\infty$ if $u(\cdot)$ converges.*
$\|\sum_{k=0}^{\infty}\mathcal{L}_{AP}(X(k))\|_F<\infty$ if condition (ii) *in Theorem* 12 *holds.*

Proof. From Proposition 11 (v), we obtain

$$
(11)\qquad \left\|\sum_{k=0}^{\infty}\mathcal{L}_B(U(k))\right\|_F\le\sum_{k=0}^{\infty}\|\mathcal{L}_B(U(k))\|_F\le\|B\|_{\infty}^2\sum_{k=0}^{\infty}\|U(k)\|_F<\infty.
$$

For the second assertion, we employ Proposition 11 (i), (v), to evaluate

$$
\begin{aligned}
(12)\qquad \left\|\sum_{k=0}^{\infty}\mathcal{L}_{AP}(X(k))\right\|_F&\le\sum_{k=0}^{\infty}\|\mathcal{L}_{AP}(X(k))\|_F=\sum_{k=0}^{\infty}\|\mathcal{L}_A(PX(k)P')\|_F\\
&\le||A||_{\infty}\sum_{k=0}^{\infty}\|PX(k)P'\|_F.
\end{aligned}
$$

Since from the assumption $x(\cdot)$ converges to $\mathcal{S}$, we have that $\sum_{k=0}^{\infty}\|PX(k)P'\|_F<\infty$ from Lemma 8 and the result follows. □

Proof of sufficiency of Theorem 12 continued. Lemma 13 provides, under the assumptions of the theorem, that Υ is well defined, and from (9) and (10) we obtain

$$
(13)\qquad \sum_{k=0}^{\infty}\|Z(k)\|_F=\sum_{k=0}^{\infty}\langle\hat{\mathcal{L}}^k(X(0)+\Upsilon), I\rangle,
$$

and recalling that $r_\sigma(\hat{\mathcal{L}})<1$, we have that

$$
(14)\qquad \sum_{k=0}^{\infty}\|Z(k)\|_F<\infty.
$$

Now we show by induction that

$$X(k) \leq Z(k) \; \forall k \geq 0. \tag{15}$$

For $k = 0$ we defined $Z(0) = X(0)$; assuming $X(k) \leq Z(k)$ one can check employing Lemma 10 and Proposition 11 (iii), (iv), that

$$\begin{aligned} X(k+1) &\leq (1+\alpha^2)\mathcal{L}_{A+AP-AP}(X(k)) + (1+1/\alpha^2)\mathcal{L}_B(U(k)) \\ &\leq (1+\alpha^2)^2\mathcal{L}_{A-AP}(X(k)) + (1+\alpha^2)(1+1/\alpha^2)\mathcal{L}_{AP}(X(k)) \\ &\qquad + (1+1/\alpha^2)\mathcal{L}_B(U(k)) \\ &= \hat{\mathcal{L}}(X(k)) + \tilde{\mathcal{L}}(X(k)) + \bar{\mathcal{L}}(U(k)) \\ &\leq \hat{\mathcal{L}}(Z(k)) + \tilde{\mathcal{L}}(X(k)) + \bar{\mathcal{L}}(U(k)) = Z(k+1) \end{aligned}$$

and the induction is complete. From (15) we obtain that for each $k \geq 0$, $Z(k) - X(k) \geq 0$ in such a manner that $Z(k) - X(k) \in \mathcal{H}_F^n$ and $\langle Z(k) - X(k), I \rangle \geq 0$. This leads to $\|Z(k)\|_F \geq \|X(k)\|_F$ and from (14) we obtain

$$\sum_{k=0}^{\infty} ||X(k)||_F \leq \sum_{k=0}^{\infty} ||Z(k)||_F < \infty,$$

and Lemma 8 with trivial $\mathcal{S} = 0$ provides that $x(\cdot)$ converges. □

4.2. The main result. The first result of this section follows in a straightforward manner from Theorem 12 and the definition of detectability. We omit the proof.

LEMMA 14. *System Ψ is detectable if and only if there exists an invariant space $\mathcal{S}$ such that the following conditions hold:*

(i) *$(A, \mathbb{P})$ is ℓ_2-stable in $\mathcal{S}$;*

(ii) *$x(\cdot)$ converges to $\mathcal{S}$ provided that $y(\cdot)$ and $u(\cdot)$ converge.*

Notice that, for $\mathcal{S}$ trivial, Lemma 14 becomes a tautology; indeed, item (i) holds trivially and item (ii) reduces to the definition of detectability. The larger the invariant space $\mathcal{S}$ is, the more significant the result will be. Along this line, in this section we introduce the set $\mathcal{F} = \{\mathcal{F}_i, i \in \mathcal{Z}\}$ as

$$\mathcal{F}_i = \{x \in \mathbb{R}^n : \mathcal{Y}_0(x, i) < \infty\} \quad \forall i \in \mathcal{Z} \tag{16}$$

and we show that $\mathcal{F}$ is the largest of such Ψ_0-invariant space.

The first step is to show that $\mathcal{F}$ is indeed a Ψ_0-invariant space. We need the following preliminary result, adapted from [7].

PROPOSITION 15. *For each $t \geq 0$, there exists $H \in \mathcal{H}_F^n$ such that $\mathcal{Y}_0^t(x, i) = x' H_i x$.*

LEMMA 16. *$\mathcal{F}$ is a Ψ_0-invariant space.*

Proof. ($\mathcal{F}_i$ is a vector subspace.) For $x_1, x_2 \in \mathcal{F}_i$ and $\alpha, \beta \in \mathbb{R}$, from Proposition 15 it is simple to check that

$$\begin{aligned} \mathcal{Y}_0^T(\alpha x_1 + \beta x_2, i) &= (\alpha x_1 + \beta x_2)' H_i (\alpha x_1 + \beta x_2) \\ &\leq 2\alpha^2 x_1' H_i x_1 + 2\beta^2 x_2' H_i x_2 \\ &= 2\alpha^2 \mathcal{Y}_0^T(x_1, i) + 2\beta^2 \mathcal{Y}_0^T(x_1, i) \quad \forall T \geq 0. \end{aligned}$$

Taking limits, we obtain

$$\mathcal{Y}_0(\alpha x_1 + \beta x_2, i) \leq 2\alpha^2 \mathcal{Y}_0(x_2, i) + 2\beta^2 \mathcal{Y}_0(x_2, i) < \infty$$

in such a manner that $\alpha x_1 + \beta x_2 \in \mathcal{F}_i$.

($\mathcal{F}$ is invariant.) We set $x_0(k) \in \mathcal{F}_{\theta(k)}$ and assume without loss that $k = 0$. Let us deny the assertion and assume that $x_0(s) \notin \mathcal{F}_{\theta(s)}$ for some $s > 0$, with probability (w.p.) $\epsilon > 0$. In this situation, it is simple to check that

$$\forall \gamma > 0, \exists t_\gamma > 0 : \sum_{k=s}^{s+t_\gamma} \|y_0(k)\|^2 \geq \gamma \quad \text{w.p. } \epsilon. \tag{17}$$

Employing the Tchebychev inequality, we evaluate

$$E\left\{\sum_{k=s}^{s+t_\gamma} \|y_0(k)\|^2\right\} \geq \gamma P\left\{\sum_{k=s}^{s+t_\gamma} \|y_0(k)\|^2 > \gamma\right\} = \gamma\epsilon, \tag{18}$$

and we conclude that $\mathcal{Y}_0(x(0), \theta(0)) \geq \gamma\epsilon$ for all $\gamma > 0$, which is a contradiction in view of the fact that $x_0(0) \in \mathcal{F}_{\theta(0)}$. □

Next we show that $\mathcal{F}$ is the largest Ψ_0-invariant space that possibly meets the condition (i) in Lemma 14.

LEMMA 17. *If $\mathcal{S}$ is such that $(A, \mathbb{P})$ is ℓ_2-stable in $\mathcal{S}$, then $\mathcal{S} \subset \mathcal{F}$.*

Proof. Let us deny the assertion of the lemma and assume that there exists $i \in \mathcal{Z}$ for which $\mathcal{F}_i \subset \mathcal{S}_i$ strictly. We have that there exists $x \in \mathcal{S}_i$ with $x \notin \mathcal{F}_i$ and from the definition of $\mathcal{F}$ we conclude that $\mathcal{Y}_0(x, i) = \infty$, which provides that the associated output does not converge. Then, $(A, \mathbb{P})$ is not ℓ_2-stable in $\mathcal{S}$. □

Lemmas 14 and 17 allow us to derive the main result of the paper.

THEOREM 18. *System Ψ is detectable if and only if the following conditions hold:*

(A1) *$(A, \mathbb{P})$ is ℓ_2-stable in $\mathcal{F}$;*

(A2) *$x(\cdot)$ converges to $\mathcal{F}$ provided $y(\cdot)$ and $u(\cdot)$ converge.*

Proof. (Sufficiency.) (A1) and (A2) satisfy the conditions for detectability in Lemma 14.

(Necessity.) Since $(A, C, \mathbb{P})$ is detectable, from Lemma 14 we have that there exists $\mathcal{S}$ for which $(A, \mathbb{P})$ is ℓ_2-stable in $\mathcal{S}$ and Lemma 17 provides that $\mathcal{S} \subset \mathcal{F}$. Lemma 14 also yields that $x(\cdot)$ converges to $\mathcal{S}$ provided $y(\cdot)$ and $u(\cdot)$ converges; this fact together with the fact that $\mathcal{S} \subset \mathcal{F}$ lead immediately to (A2).

Now, notice from the concept of detectability that, in particular for the autonomous system Ψ_0, $x_0(\cdot)$ converges whenever the corresponding output $y(\cdot)$ converges or, equivalently, whenever $x(0) \in \mathcal{F}_{\theta(0)}$. This means that $(A, \mathbb{P})$ is ℓ_2-stable in $\mathcal{F}$ and (A1) holds. □

5. Sufficient conditions for (A1) and (A2). In this section we deal with other detectability and observability concepts that appear in the literature of MJLS and we present the role that they play as sufficient conditions (expressed entirely in terms of the autonomous version of the system) for (A1) and (A2), and therefore for the detectability concept here.

Initially, we introduce a concept of uniform observability related to the Ψ_0-invariant space $\mathcal{S}$. From (7), recall that we set $\mathcal{Y}_0^T(x, \theta) = E\{\sum_{k=0}^{T-1} |y_0(k)|^2\}$, where $y_0(\cdot)$ denotes the output trajectory of the autonomous system Ψ_0 with $x_0(0) = x$.

DEFINITION 19 (uniform observability w.r.t. $\mathcal{S}$). *Consider the autonomous system Ψ_0. We say that $(A, C, \mathbb{P})$ is uniformly observable with respect to (w.r.t.) $\mathcal{S}$ if*

there exists $T, \epsilon > 0$ *such that* $\mathcal{Y}_0^T(x,\theta) \geq \epsilon\|x\|^2$ *whenever* $x \in \mathcal{S}_\theta^\perp$. *We say that* $(A, C, \mathbb{P})$ *is uniformly observable if it is uniformly observable w.r.t.* 0.

A particular case of this concept with trivial $\mathcal{S}$ appears in [1, 3, 5, 12], and it generalizes previous observability concepts for MJLS, like the ones appearing in [11].

In Lemma 22 in what follows, we show that uniform observability is a sufficient condition for the state convergence to $\mathcal{S}$ and, in particular, for (A2) to hold when $\mathcal{S} \subset \mathcal{F}$. For the proof, we need the next two lemmas; their proofs are presented in Appendix 9. Recall that P_i denotes the orthogonal projection onto $\mathcal{S}_i^\perp$, $i \in \mathcal{Z}$.

LEMMA 20. *If* $(A, C, \mathbb{P})$ *is uniformly observable w.r.t.* $\mathcal{S}$, *then there exist* $T, \epsilon > 0$ *such that* $\mathcal{Y}_0^T(x,\theta) \geq \epsilon|P_\theta x|^2$ *for each* $x \in \mathbb{R}^n$ *and* $\theta \in \mathcal{Z}$.

LEMMA 21. *There exist* $\delta_1, \delta_2 > 0$ *for which*

$$\mathcal{Y}_u^{t,T}(x,\theta) \geq \delta_1 E\{\mathcal{Y}_0^T(x(t),\theta(t))\} - \delta_2 \sum_{k=t}^{T+t-1} E\{|u(k)|^2\} \; \forall x \in \mathbb{R}^n \text{ and } \theta \in \mathcal{Z}.$$

LEMMA 22. *If* $(A, C, \mathbb{P})$ *is uniformly observable w.r.t.* $\mathcal{S}$, *then* $x(\cdot)$ *converges to* $\mathcal{S}$ *provided that* $y(\cdot)$ *and* $u(\cdot)$ *converge. In addition, if* $\mathcal{S} \subset \mathcal{F}$, *then* (A2) *holds.*

Proof. Provided that $u(\cdot)$ and the output $y(\cdot)$ converge, i.e., $E[\sum_{k=0}^\infty \|u(k)\|^2] < \infty$ and $\mathcal{Y}_u(x,\theta) < \infty$, respectively, we show that $(A, C, \mathbb{P})$ uniformly observable w.r.t. $\mathcal{S}$ suffices for convergence of the state to $\mathcal{S}$, namely, $E[\sum_{k=0}^\infty \|P_{\theta(k)}x(k)\|^2] < \infty$. For each $t \geq 0$, we employ Lemmas 20 and 21 to evaluate

$$\begin{aligned}(19)\quad \mathcal{Y}_u(x,\theta) \geq \mathcal{Y}_u^{t,\infty}(x,\theta) &= \sum_{k=0}^\infty \mathcal{Y}_u^{t+kT,T}(x,\theta) \\ &\geq \sum_{k=0}^\infty \left(\delta_1 E\{\mathcal{Y}_0^T(x(t+kT),\theta(t+kT))\} - \delta_2 \sum_{\ell=t+kT}^{t+(k+1)T-1} E\{|u(k)|^2\} \right) \\ &\geq \delta_1 \epsilon \sum_{k=0}^\infty E\{|P_{\theta(t+kT)}x(t+kT)|^2\} - \delta_2 \sum_{k=0}^\infty E\{|u(k)|^2\}.\end{aligned}$$

Summing (19) for $t = 0, \ldots, T-1$, we obtain

$$T\mathcal{Y}_u(x,\theta) \geq \delta_1 \epsilon \sum_{m=0}^\infty E\{|P_{\theta(m)}x(m)|^2\} - T\delta_2 \sum_{k=0}^\infty E\{|u(k)|^2\},$$

which leads to

$$\sum_{m=0}^\infty E\{|P_{\theta(m)}x(m)|^2\} \leq \frac{T}{\delta_1\epsilon}\mathcal{Y}_u(x,\theta) + \frac{T\delta_2}{\delta_1\epsilon}\sum_{k=0}^\infty E\{|u(k)|^2\} < \infty$$

and the first assertion is proven.

Now, from Definition 19 we obtain that if $(A, C, \mathbb{P})$ is uniformly observable w.r.t. $\mathcal{S} \subset \mathcal{F}$, then it is uniformly observable w.r.t. $\mathcal{F}$. The result then follows immediately from the first assertion. ∎

Notice from Lemma 22 with trivial $\mathcal{S}$ that, if $(A, C, \mathbb{P})$ is uniformly observable, then $x(\cdot)$ converges to $\mathcal{S} = 0$, provided $y(\cdot)$ and $u(\cdot)$ converges, i.e., $(A, C, \mathbb{P})$ is detectable, which implies the following corollary.

COROLLARY 23. *If* $(A, C, \mathbb{P})$ *is uniformly observable, then* Ψ *is detectable.*

Next, we are concerned with an earlier ℓ_2-detectability sense; see [7] in a setting similar to the one of this paper, or [8] in the continuous time case, or [6] and [12] in the finite dimensional case.

DEFINITION 24 (ℓ_2-detectability). *Consider the autonomous system Ψ_0. We say that $(A, C, \mathbb{P})$ is ℓ_2-detectable if there exists $L \in \mathcal{H}_\infty^{q,n}$ for which $(A + LC, \mathbb{P})$ is ℓ_2-stable.*

LEMMA 25. *If $(A, C, \mathbb{P})$ is ℓ_2-detectable, then system Ψ is detectable.*

Proof. We assume that $(A, C, \mathbb{P})$ is ℓ_2-detectable and $y(\cdot)$ and $u(\cdot)$ converge, and we show that $x(\cdot)$ converges in a similar manner to the proof of sufficiency of Theorem 12. Here we only point out the differences. For $L \in \mathcal{H}_\infty^{q,n}$ as in the ℓ_2-detectability definition, $r_\sigma(\mathcal{L}_{A+LC}) < 1$, see Proposition 9 (ii) in connection. We chose $\alpha \neq 0$ in such a way that $(1+\alpha^2)^2 r_\sigma(\mathcal{L}_{A+LC}) < 1$ holds. The operators $\hat{\mathcal{L}}, \tilde{\mathcal{L}}, \bar{\mathcal{L}} : \mathcal{H}_F^n \to \mathcal{H}_F^n$ are defined as

$$\begin{aligned}
\hat{\mathcal{L}}(H) &= (1+\alpha^2)^2 \mathcal{L}_{A+LC}(H), \\
\tilde{\mathcal{L}}(H) &= (1+\alpha^2)(1+1/\alpha^2)\mathcal{L}_{LC}(H) \text{ and} \\
\bar{\mathcal{L}}(H) &= (1+1/\alpha^2)\mathcal{L}_B(H) \quad \text{for } H \in \mathcal{H}_F^n.
\end{aligned}$$

In parallel with Lemma 13, we also need to show that $\sum_{k=0}^\infty \mathcal{L}_{LC}(X(k)) < \infty$. In fact, since $y(\cdot)$ converges for the autonomous system $(A+LC, \mathbb{P})$, we get that $\mathcal{Y}_0(x, \theta) < \infty$ and from (7) we evaluate

$$\infty > \mathcal{Y}_0(x,\theta) \geq \sum_{k=0}^\infty \langle X(k), C'C \rangle = \sum_{k=0}^\infty \|C'X(k)C\|_F$$

and employ Proposition 11 (i), (v), to obtain

$$\begin{aligned}
(20) \qquad \left\| \sum_{k=0}^\infty \mathcal{L}_{LC}(X(k)) \right\|_F &\leq \sum_{k=0}^\infty \|\mathcal{L}_{LC}(X(k))\|_F = \sum_{k=0}^\infty \|\mathcal{L}_L(CX(k)C')\|_F \\
&\leq ||L||_\infty \sum_{k=0}^\infty \|CX(k)C'\|_F < \infty. \qquad \square
\end{aligned}$$

Remark 1. In [7] it is shown that ℓ_2-detectability (together with ℓ_2-stabilizability) ensures that the optimal linear state feedback control that arises in the JLQ problem is ℓ_2-stabilizing, considering an additional assumption on matrices D_i as in Lemma 35. Lemma 25 generalizes this result in the sense that $x(\cdot)$ converges provided that the output and input converge; here, neither optimality nor linear state feedback is required.

The next concept is named W_S-detectability, and it was introduced in [5] as an attempt to deal with detectability in the present context. It can be seen as a particularization of detectability with $u \equiv 0$.

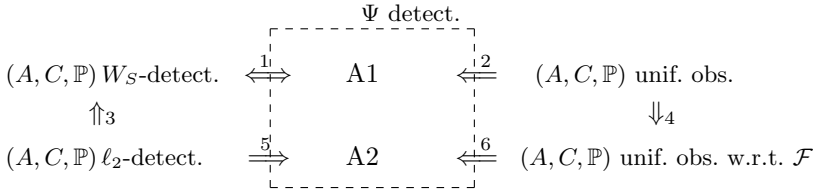
DEFINITION 26 (W_S-detectability). *Consider the autonomous system Ψ_0. We say that $(A, C, \mathbb{P})$ is W_S-detectable provided that $x_0(\cdot)$ converges whenever the output $y_0(\cdot)$ converges.*

It follows directly from the definitions that the concept is equivalent to ℓ_2-stability in $\mathcal{F}$.

PROPOSITION 27. *$(A, C, \mathbb{P})$ is W_S-detectable if and only if $(A, \mathbb{P})$ is ℓ_2-stable in $\mathcal{F}$ ((A1) holds).*

We finish with a summary of the main results of this section. For ease of reference, the relations are numbered. The relation (1) follows from Proposition 27, (2) follows from Corollary 23 and Theorem 18, (3) and (5) follows from Lemma 25, (4) follows immediately from definition, and (6) follows from Lemma 22.

PROPOSITION 28. *The following relations hold:*



Remark 2. In principle the detectability concept depends on the nonautonomous system (made explicit by assumption (A2)). However, for systems that are ℓ_2-detectable, uniformly observable, or uniformly observable w.r.t. $\mathcal{F}$, (A2) always holds true, as indicated in Proposition 28. In section 6 we show that this is also the situation for MJLS with finite Markov state.

6. Finite MJLS. Recall from the main result of the paper, Theorem 18, that the system is detectable if and only if (A1) and (A2) hold. In this section, we show that (A2) is made redundant when the Markov state space is finite, $\mathcal{Z} = \{1, \dots, N\}$. This leads to the main result of the section: (A1) is a necessary and sufficient condition for detectability, in parallel with detectability notions for linear deterministic systems and previous concepts for MJLS [2]. The result here also generalizes previous results in the literature, which require that the control is in the linear state feedback form.

We start showing that uniform observability w.r.t. $\mathcal{S}$ always holds with $\mathcal{S} = \mathcal{N}$, where the set $\mathcal{N} = \{\mathcal{N}_i, i \in \mathcal{Z}\}$ is defined as

$$\mathcal{N}_i = \{x \in \mathbb{R}^n : \mathcal{Y}_0(x, i) = 0\}, \quad i \in \mathcal{Z}. \tag{21}$$

Notice by inspection of (16) and (21) that $x \in \mathcal{F}_\rangle$ whenever $x \in \mathcal{N}_\rangle$, thus yielding that $\mathcal{N} \subset \mathcal{F}$. One can also check that $\mathcal{N}$ is an invariant space, in a similar manner to the proof of Lemma 16. We state this property formally.

PROPOSITION 29. *$\mathcal{N}$ is an invariant space, $\mathcal{N} \subset \mathcal{F}$.*

The preliminary results of Proposition 30 and Lemma 31 in what follows will be needed. First, let us generalize the definition of the cost functional $\mathcal{Y}_0$, as follows. Suppose that the initial conditions (x, θ) are random variables with x a second order random variable. In this situation, we set $X(0) \in \mathcal{H}_F^n$ such that $X_i(0) = E\{xx' 1_{\{\theta=i\}}\}$. Conversely, given any $X \in \mathcal{H}_F^n$, there exists a second order r.v. x and some distribution for θ in such a way that we can represent $X_i = E\{xx' 1_{\{\theta=i\}}\}$. These considerations allow us to generalize the definition of $\mathcal{Y}_0$ by writing, for each $X \in \mathcal{H}_F^n$,

$$\mathcal{Y}_0^T(X) = \sum_{k=0}^{T-1} \langle X_0(k), C'C \rangle \tag{22}$$

whenever $X(0) = X$. Notice that $\mathcal{Y}_0^T(x, \theta) = \mathcal{Y}_0^T(X)$ whenever X is defined as above with $X_\theta = xx'$ and $X_i = 0$, $i \neq \theta$.

The next preliminary result is adapted from [3, Prop. 1].

PROPOSITION 30. *Assume that $\mathcal{Z} = \{1, \dots, N\}$. If $\mathcal{Y}_0^{n^2 N}(X) = 0$, then $\mathcal{Y}_0^t(X) = 0$ for all $t \geq 0$.*

LEMMA 31. *Let P be the projection onto $\mathcal{N}^\perp$. The following assertions hold:*

(i) $\mathcal{Y}_0(X) = 0$ *if and only if* $P'XP = 0$*;*

(ii) $(A, C, \mathbb{P})$ *is uniformly observable w.r.t.* $\mathcal{N}$ *if there exists* $\epsilon > 0$ *for which* $\mathcal{Y}_0^T(X) \geq \epsilon \|X\|_F$ *whenever* $X - PXP' = 0$.

For the proof of Lemma 31, see Appendix 9.

LEMMA 32. *Assume that* $\mathcal{Z} = \{1, \ldots, N\}$. *Then,* $(A, C, \mathbb{P})$ *is uniformly observable w.r.t.* $\mathcal{N}$.

Proof. Let P be the projection onto $\mathcal{N}^\perp$ and recall that $\bar{\mathcal{N}} = \{H \in \mathcal{H}_F^n : PHP' = 0\}$ and $\bar{\mathcal{N}}^\perp = \{H \in \mathcal{H}_F^n : H - PHP' = 0\}$. Let us show that there exist $\epsilon > 0$ such that $\mathcal{Y}_0^T(X) \geq \epsilon \|X\|_F$ whenever $X \in \bar{\mathcal{N}}^\perp$. Let us deny this assertion and assume that there exists a sequence $X_m \in \mathcal{N}_C$, $m = 1, 2, \ldots$, for which

$$\mathcal{Y}_0^{n^2N}(X_m) \leq m^{-1}, \tag{23}$$

where

$$\mathcal{N}_C = \{X \in \bar{\mathcal{N}}^\perp : \|X\|_F = 1\} \subset \bar{\mathcal{N}}^\perp.$$

For the countably finite case, one can check that $\mathcal{N}_C$ is a compact set and this leads to the fact that there exists a subsequence X_{m_k} that converges to

$$\bar{X} \in \mathcal{N}_C \subset \bar{\mathcal{N}}^\perp. \tag{24}$$

Moreover, it is not difficult to check that $\mathcal{Y}_0^{n^2N}(\cdot)$ is continuous; this fact and (23) allow us to write that $\mathcal{Y}_0^{n^2N}(\bar{X}) = \lim_{k\to\infty} \mathcal{Y}_0^{n^2N}(X_{m_k}) \leq \lim_{k\to\infty} k^{-1} = 0$. Then, Proposition 30 yields that $\mathcal{Y}_0(\bar{X}) = 0$ and from Lemma 31 (i) we conclude that $\bar{X} \in \bar{\mathcal{N}}$, which is a contradiction, in view of (24). We have shown that there exist $\epsilon > 0$ such that $\mathcal{Y}_0^T(X) \geq \epsilon \|X\|_F$ whenever $X \in \bar{\mathcal{N}}^\perp$; Lemma 31 (ii) completes the proof. □

The result of Lemma 32 cannot be extended to the countably infinite case, as we show in the following counterexample. In connection, note that the set $\mathcal{N}_C$ in the proof of Lemma 32 is no longer compact.

Example 1. Let $n = 1$, $p_{i\,i+1} = 1$, $A_i = 1$, $C_i = r^i$, $|r| < 1$. It is simple to check that $\mathcal{Y}_0(x, i) = r^i(1 - r)|x|^2$, in such a manner that for each $\gamma > 0$ there exists i such that $\mathcal{Y}_0(x, i) < \gamma|x|^2$, which implies that $(A, C, \mathbb{P})$ is not uniformly observable w.r.t. $\mathcal{N}$.

The next result follows from Lemmas 22 and 32 and the fact that $\mathcal{N} \subset \mathcal{F}$ (see Proposition 29); the proof is omitted.

LEMMA 33. *Assume that* $\mathcal{Z} = \{1, \ldots, N\}$. *Then,* (A2) *holds.*

The next result is immediate from Lemma 33 and Theorem 18. We state the result in terms of the triplet $(A, C, \mathbb{P})$ to emphasize that the detectability concept depends only on the autonomous version Ψ_0 of the system.

THEOREM 34. *Assume that* $\mathcal{Z} = \{1, \ldots, N\}$. $(A, C, \mathbb{P})$ *is detectable if and only if* (A1) *holds.*

Remark 3. The relation between detectability and other detectability concepts for finite scenarios is discussed here. The weak detectability concept for MJLS with finite Markov state was introduced in [2]. It requires that $x(\cdot)$ converges provided $\mathcal{Y}_0(x, \theta) = 0$. In [5] it was shown that this concept is equivalent to W_S-detectability when reduced to the finite case. Assuming $\mathcal{Z} = \{1, \ldots, N\}$, Proposition 27, Theorem 34, and the aforementioned facts provide the following relations:

$$\text{weak detectability} \Leftrightarrow W_S\text{-detectability} \Leftrightarrow \text{A1} \Leftrightarrow \text{detectability}. \tag{25}$$

For (finite dimensional) linear deterministic systems, it was shown in [2] that the weak detectability concept retrieves the usual detectability concept. Then, from the relations in (25) we conclude that, for linear deterministic systems,

usual detectability concept $\Leftrightarrow$ *detectability.*

7. Detectability and the jump linear quadratic problem. In this section we are concerned with the JLQ problem, which consists of obtaining the control $u(\cdot)$ that minimizes the cost functional $\mathcal{Y}_u(x,\theta)$. We also consider the related infinite coupled algebraic Riccati equations (ICARE).

We assume here with no loss of generality that the control is in linear state feedback form, $u(k) = G_{\theta(k)}x(k)$, $G \in \mathcal{H}_\infty^{r,n}$. Indeed, it is a well-known fact that the optimal control is in this form; see, e.g., [7]. In connection, we denote $\mathcal{Y}_G(\cdot) = \mathcal{Y}_u(\cdot)$ to emphasize the dependence on G.

Also a standard assumption in the JLQ problem, that $\inf_{i\in\mathcal{Z}} \sigma^-(D_i'D_i) = \xi > 0$, is in force here. In this situation, the convergence of the input and the output are directly connected and the condition in (A2) (e.g., in Theorem 18) related to the input is not essential; the following lemma formalizes the result.

LEMMA 35. *If* $\inf_{i\in\mathcal{Z}} \sigma^-(D_i'D_i) = \xi > 0$ *and* $\mathcal{Y}_u(x,\theta) < \infty$, *then* $u(\cdot)$ *converges.*

Proof. Employing (7) and the assumptions in the lemma, we evaluate $\infty > \mathcal{Y}_u(x,\theta) \geq \sum_{k=0}^{\infty}\langle U(k), D'D\rangle \geq \xi \sum_{k=0}^{\infty} \|U(k)\|_F$. □

The next result establishes that a linear state feedback control is stabilizing whenever the associated cost is bounded.

LEMMA 36. *Assume that* $(A, C, \mathbb{P})$ *is detectable. If* $G \in \mathcal{H}_\infty^{r,n}$ *is such that* $\mathcal{Y}_G(x,\theta) < \infty$ $\forall x \in \mathbb{R}^n, \theta \in \mathcal{Z}$, *then* $(A + BG, \mathbb{P})$ *is* ℓ_2*-stable.*

Proof. Consider the system Ψ in closed loop form with $u(k) = G_{\theta(k)}x(k)$,

$$\begin{cases} x(k+1) = (A_{\theta(k)} + B_{\theta(k)}G_{\theta(k)})x(k), \quad k \geq 0, \\ y(k) = (C_{\theta(k)} + D_{\theta(k)}G_{\theta(k)})x(k). \end{cases} \tag{26}$$

For each initial condition $x \in \mathbb{R}^n$ and $\theta \in \mathcal{Z}$ we have from the lemma that $\mathcal{Y}_u(x,\theta) = \mathcal{Y}_G(x,\theta) < \infty$, which means that the associated output $y(\cdot)$ converges; moreover, Lemma 35 provides that $u(\cdot)$ converges. In this situation, detectability yields that $x(\cdot)$ converges, and we conclude that $(A + BG, \mathbb{P})$ is ℓ_2-stable. □

In what follows, we consider the following ICARE in the unknown $R \in \mathcal{H}_F^n$ that arises in the JLQ problem (see, e.g., [7]):

$$0 = \big(A_i + B_iG_i\big)' \sum_{j\in\mathcal{Z}} p_{ij}R_j\big(A_i + B_iG_i\big) + C_i'C_i + G_i'D_i'D_iG_i, \tag{27}$$

$$G_i = -\left(D_i'D_i + B_i' \sum_{j\in\mathcal{Z}} p_{ij}R_jB_i\right)^{-1} B_i' \sum_{j\in\mathcal{Z}} p_{ij}R_jA_i, \quad i \in \mathcal{Z}. \tag{28}$$

The following results are adapted from [7].

PROPOSITION 37. *Assume that* $R \in \mathcal{H}_F^n$ *satisfies the ICARE* (27)–(28). *The following assertions hold:*

(i) $\mathcal{Y}_G(x,\theta) \leq x'R_\theta x$;

(ii) *If* $(A+BG, \mathbb{P})$ *is* ℓ_2*-stable, then* $R \in \mathcal{H}_\infty^n$ *is the unique solution of the ICARE. Moreover, the solution of the JLQ problem is* $u(k) = G_{\theta(k)}x(k)$, *where* G *is given by* (28).

THEOREM 38. *Assume that* $(A, C, \mathbb{P})$ *is detectable according to Definition* 1. *Then, the ICARE has at most one solution. Moreover, if* $R \in \mathcal{H}_F^n$ *is the solution of the ICARE, then* $(A + BG, \mathbb{P})$ *is* ℓ_2*-stable with the optimal control* (28).

Proof. Let $R \in \mathcal{H}_F^n$ be a solution of the ICARE. From Proposition 37 (i) we have that $\mathcal{Y}_G(x, \theta) \leq x' R_\theta x$, for each x, θ, and Lemma 36 provides that $(A + BG, \mathbb{P})$ is ℓ_2-stable. Hence, Proposition 37 (ii) yields that R is the unique solution of the ICARE and the optimal control is given by (28). □

Remark 4. The results in this section generalize previous result in [7] from the fact that detectability here generalizes the ℓ_2-detectability notion employed there; see Lemma 25.

8. Examples. We start this section with an example showing that (A2) does not necessarily hold for MJLS with infinite countably Markov chain. Then Example 3 shows that the detectability notion according to Definition 1 depends on the collections of matrices B and D, and thus it cannot be related to the autonomous version Ψ_0 only.

We also show, via Example 4, that the detectability concept here generalizes the earlier ℓ_2-detectability and uniform observability concepts, in the sense that the converse relations of Proposition 28 involving those concepts does not hold.

Example 2. This example illustrates that (A2) does not necessarily hold true for MJLS with infinite countably Markov chain. Indeed, we present a system for which the state trajectory does not converge to $\mathcal{F}$ under converging input and output.

Assume that $p_{i\,i+1} = 1$, $i \in \mathcal{Z}$, in such a manner that $\theta(k) = k + i$ a.s. whenever $\theta(0) = i$. Let $n = 1$, $A_i = B_i = 1$, $D_i = 0$, $i \in \mathcal{Z}$. As regards to $C \in \mathcal{H}_F^1$, we set $C_1 = 0$ and $C_i = (i-1)^{-1/2}$, $i \geq 2$, in order to get that $C_{\theta(k)} = (k+i-1)^{-1/2}$, $k \geq 1$.

It is simple to check for the autonomous system that $\mathcal{Y}_0(x, \theta) = \sum_{k=0}^{\infty} x^2/(k + i - 1)$, which converges if and only if $x = 0$, thus leading to

$$\mathcal{F} = 0.$$

Now, for simplicity, we consider fixed initial conditions $x = 1$ and $\theta = 1$. Consider the control given by $u(0) = 0$ and $u(k) = (k+1)^{-1/2} - k^{-1/2}$, $k \geq 1$. We get that $x(k) = k^{-1/2}$, $k \geq 1$ is the corresponding trajectory. It is a simple matter to check that (see [16, Chap. 2.6])

$$E_{x,\theta}\left\{\sum_{k=0}^{\infty} |u(k)|^2\right\} = \sum_{k=1}^{\infty} \frac{(k^{1/2} - (k+1)^{1/2})^2}{k(k+1)} \leq \sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1, \tag{29}$$

and we have that the input converges. As regards to the output, we first evaluate

$$E\left\{\sum_{k=0}^{\infty} x(k)' C'_{\theta(k)} C_{\theta(k)} x(k) | x = \theta = 1\right\} = \sum_{k=1}^{\infty} \frac{1}{k^{4\times 1/2}} = \sum_{k=1}^{\infty} \frac{1}{k^2} \leq \sum_{k=0}^{\infty} \frac{1}{2^k} \leq 2,$$

where, in the last inequality, we employed the evaluation in [16, Chap. 3.1]). Together with (29), they provide that

$$\mathcal{Y}(1, 1) \leq 3,$$

which means that the output converges. However, we can also write that

$$E_{x,\theta}\left\{\sum_{k=0}^{\infty} |x(k)|^2\right\} = \sum_{k=0}^{\infty} \frac{1}{k} = \infty,$$

and the state does not converge to the trivial $\mathcal{F}$. Thus, (A2) does not hold and the system is not detectable. It is also interesting to mention that from Proposition 29 we have that $\mathcal{N} = 0$ and, thus, the example also illustrates that the state trajectory does not converge to the nonobserved space, despite the fact that both input and output converge.

Example 3. In the system of Example 2, let $B_i = 0$, $i \in \mathcal{Z}$. In this case, $\mathcal{Y}(x,i) = \mathcal{Y}_0(x,i) = \sum_{k=0}^{\infty} x^2/(k+i-1)$, and the output converges if and only if $x(0) = x = 0$. Then, the system is trivially detectable in the sense of Definition 1. On the other hand, recall that the system in Example 2 is not detectable. This makes clear the dependence of the detectability concept on $B \in \mathcal{H}^{n,r}$ and $D \in \mathcal{H}^{q,r}$ and not only on the parameters of the autonomous system, which shows that the class of systems studied here share with general nonlinear systems the characteristic that observability and detectability in general depends on features of the input class.

Example 4. Consider the following version of the JLQ problem of section 7:

$$\min_u \mathcal{Y}_u^{k_0}(x,\theta), \quad \text{where } \mathcal{Y}_u^{k_0}(x,\theta) = E\left\{\sum_{k=0}^{k_0} |y(k)|^2\right\}, \tag{30}$$

y is the output of system Ψ defined in (1) and k_0 is a $\mathfrak{F}_k$-stopping time defined as the time that the Markov chain $\Gamma = \{\gamma(k), k \geq 0\}$ taking values on the set $\{n, f\}$ first enters the state f, i.e., $k_0 = \inf\{k : \gamma(k) = f\}$. We assume that the transition probabilities are given by $q_{nn} \geq 0$, $0 \leq q_{nf} \leq \nu < 1$ and $q_{ff} = 1$ (f is a cemetery state).

A possible physical interpretation is that the cemetery state f represents a critical failure of the system, which forces the system to stop for maintenance at time k_0; n and f stand for *normal* and *failure,* respectively.

We start showing that the problem (30) can be cast as the JLQ problem for the MJLS defined as

$$\bar{\Psi}: \begin{cases} \bar{\theta} \text{ takes values on } \bar{\mathcal{Z}} = \{(f), (1,n), (2,n), \dots\}; \\ \bar{\mathbb{P}}: \ \bar{p}_{ff} = q_{ff} = 1, \ \bar{p}_{(i,n)(j,n)} = (1-q_{nf})p_{ij}, \ \bar{p}_{(i,n)f} = q_{nf}; \\ \text{system matrices: } \bar{A}_f = 0, \ \bar{A}_{(i,n)} = A_i, \ i \neq f, \text{ and similarly for } \bar{B}, \bar{C}, \bar{D}; \\ \text{initial conditions: } \begin{array}{l} x(0) = x; \\ \bar{\theta}(0) = f \text{ if } \gamma(0) = f; \ \bar{\theta}(0) = (\theta(0), n) \text{ otherwise.} \end{array} \end{cases} \tag{31}$$

LEMMA 39. *The problem* (30) *is equivalent to the JLQ problem for system* $\bar{\Psi}$.

Proof. Let us define the $\mathfrak{F}_k$-stopping time $t_0 = \inf\{k : \bar{\theta}(k) = f\}$. It is a simple matter to check that the random variables t_0 and k_0 have the same statistics and similarly for the variables of system Ψ and $\bar{\Psi}$ for $k \leq t_0$ (e.g., $y(k)$ and $\bar{y}(k)$ are statistically identical for $k \leq t_0$). Let $\bar{Y}_u(\cdot)$ represent the cost functional in (2) associated with system $\bar{\Psi}$. Then, we can write

$$\begin{aligned} \bar{Y}_u(x,\theta) &= \sum_{k=0}^{\infty} E\{|\bar{y}(k)|^2 1_{\{k \leq t_0\}}\} + \sum_{k=0}^{\infty} E\{|\bar{y}(k)|^2 1_{\{k > t_0\}}\} \\ &= \sum_{k=0}^{\infty} E\{|\bar{y}(k)|^2 1_{\{k \leq t_0\}}\} = \sum_{k=0}^{\infty} E\{|y(k)|^2 1_{\{k \leq k_0\}}\} = \mathcal{Y}_u^{k_0}(x,\theta). \end{aligned} \tag{32}$$

∎

The next lemma establishes that system $(\bar{A}, \bar{C}, \bar{\mathbb{P}})$ is a counterexample for the converse of Corollary 23; namely, it shows that the uniformly observable systems

form a strictly subset of the set of detectable systems. The proof is presented in Appendix 9.

LEMMA 40. *If* $(A, C, \mathbb{P})$ *is uniformly observable, then system* $\bar{\Psi}$ *is detectable.* $(\bar{A}, \bar{C}, \bar{\mathbb{P}})$ *is not uniformly observable.*

Example 5. Let us consider systems that present Markov chains with distinct communicating classes $\mathcal{Z}_j = \{i_{n_{j-1}}, \ldots, i_{n_j}\}$, $n_0 = 1$, $j = 1, \ldots, N$, for which $P\{\theta(k+1) \in \mathcal{Z}_j | \theta(k) \in \mathcal{Z}_i\} = 0$ for all $i \neq j$. Let us denote such a system by Ψ_c; we also denote $A^j = (A_i)$, $i \in \mathcal{Z}_j$ and similarly for C^j and $\mathbb{P}^j$. We refer to the system associated with a class $\mathcal{Z}_j$ as a subsystem $(A^j, C^j, \mathbb{P}^j)$.

The following result is adapted from [5].

PROPOSITION 41. *Consider system* Ψ_c. *Assumption* (A1) *holds for* $(A, C, \mathbb{P})$ *if and only if* (A1) *holds for each subsystem* $(A^j, C^j, \mathbb{P}^j)$, $j = 1, \ldots, N$.

Let us construct in this example a system Ψ_c composed of one uniformly observable subsystem plus one finite dimensional subsystem, composed by the two classes $\mathcal{Z}_1 = \{1, 2\}$ and $\mathcal{Z}_2 = \{3, 4, \ldots\}$, and probability matrix

$$\mathbb{P} = \begin{bmatrix} p_{11} & (1-p_{11}) & 0 & \ldots & \\ (1-p_{22}) & p_{22} & 0 & \ldots & \\ 0 & 0 & p_{33} & p_{34} & \ldots \\ \vdots & \vdots & p_{43} & p_{44} & \ldots \\ & & \vdots & \vdots & \end{bmatrix},$$

and data $A_1 = a_1$, $A_2 = a_2$, $C_1 = 1$, $C_2 = 0$, $p_{11} > 0$, and $p_{22}a_2^2 > 1$. We assume that $(A^2, C^2, \mathbb{P}^2)$ is uniformly observable.

It is shown in [1] that $(A^1, C^1, \mathbb{P}^1)$ is uniformly observable. Since uniform observability implies (A1) (see Proposition 28), we have that (A1) holds for both $(A^1, C^1, \mathbb{P}^1)$ and $(A^2, C^2, \mathbb{P}^2)$ and Proposition 41 yields that (A1) holds for the overall system Ψ_c.

Uniform observability also implies uniform observability w.r.t. $\mathcal{F}$ for each subsystem (see Proposition 28), and it is simple to check that the overall system is uniformly observable w.r.t. $\mathcal{F}$. In this situation, Proposition 28 yields that (A2) holds for the overall system.

Then, Theorem 18 implies that system Ψ_c is detectable. However, for this simple example, it was shown in [5] that the overall system is not ℓ_2-detectable. This implies that the converse of Lemma 25 does not hold, and we conclude that detectability here generalizes ℓ_2-detectability.

9. Conclusions. This paper deals with detectability for discrete-time Markov jump linear systems with countably infinite Markov state. Beginning with Definition 1, which expresses an idea that at same time is purposeful and captures the abstract notion of detectability, we show that it can be written down in terms of conditions (A1) and (A2). Condition (A1) alone refers to the autonomous systems and its behavior within the invariant space $\mathcal{F}$. It is reminiscent of detectability concepts related with finite dimensional linear systems. Condition (A2) refers to the complete system Ψ and its behavior within set $\mathcal{F}^\perp$. It comes as an essential condition, connected to the fact that the observed part of the autonomous system, represented by $\mathcal{F}^\perp$, may not be uniformly observable, contrary to the finite dimensional case. Example 2 shows that (A2) may fail in the infinite Markov state case. This clarifies that, unlike the finite dimensional contexts, the detectability notion yielding property (i) (stated in section 1) cannot be expressed in terms of the parameters of the autonomous version Ψ_0;

thus, (iii) cannot be completely reproduced. Exceptions are pointed out in Remark 2.

Regarding the issues (I)–(III), note that (I) is accomplished in the sense that we show that $\mathcal{F}$ is the largest Ψ_0 invariant space for which (A1) and (A2) together possibly hold. Along this line, a remarkable feature is that (A2) is weaker than the natural extension of the finite dimensional case, that the trajectories converge to the nonobservable space $\mathcal{N} \subset \mathcal{F}$. In addition, (II) and (III) are accomplished by showing that the notion of detectability generalizes previous notions of ℓ_2-detectability and uniform observability, as well as detectability notions for the finite Markov state case and the usual detectability concept for linear deterministic systems, in their respective scenarios. Moreover, these relations provide a generalization for earlier results concerning stability of trajectories with associated finite cost, in the sense that here we are not constrained to linear feedback form nor optimal control; see Remark 1). A particularization of the results for the JLQ optimal control problem, which was the initial motivation for this work, provides that the JLQ control is stabilizing and the solution to the associated ICARE is unique.

Finally, although the analysis here concludes a circle of ideas toward detectability of MJLS, which has began in [1, 3], we believe that the approach via invariant subspaces proposed here may be useful elsewhere, in contexts such as nonlinear systems or other infinite dimensional systems.

Appendix: Proof of Lemmas 8 and 10.

Proof of Lemma 8. It is simple to check that

$$\begin{aligned} E[|P_{\theta(k)}x(k)|^2] &= tr\{E[P_{\theta(k)}x(k)x(k)'P'_{\theta(k)}]\} \\ &= \sum_{i\in\mathcal{Z}} tr\{P_iE[x(k)x(k)'1_{\{\theta(k)=i\}}]P_i'\} \\ &= \sum_{i\in\mathcal{Z}} tr\{P_iX_i(k)P_i'\} = \|PX(k)P'\|_F, \end{aligned}$$

which provides that $\sum_{k=0}^{\infty} E[|P_{\theta(k)}x(k)|^2] < \infty$ if and only if $\sum_{k=0}^{\infty} \|PX(k)P'\|_F < \infty$. □

Proof of Lemma 10. For any scalar $\alpha \neq 0$, we have that

$$\begin{aligned} X_i(k+1) &= E\{[A_{\theta(k)}x(k) + B_{\theta(k)}u(k)][A_{\theta(k)}x(k) + B_{\theta(k)}u(k)]'1_{\{\theta(k+1)=i\}}\} \\ &\le E\{[(1+\alpha^2)(A_{\theta(k)}x(k)x(k)'A'_{\theta(k)}) \\ &\qquad + (1+1/\alpha^2)(B_{\theta(k)}u(k)u(k)'B'_{\theta(k)})]1_{\{\theta(k+1)=i\}}\} \\ &= \sum_{j\in\mathcal{Z}} E\{[(1+\alpha^2)(A_jx(k)x(k)'A'_j) \\ &\qquad + (1+1/\alpha^2)(B_ju(k)u(k)'B'_j)]1_{\{\theta(k+1)=i,\theta(k)=j\}}\} \\ &= (1+\alpha^2)\sum_{j\in\mathcal{Z}} p_{ji}E\{A_jx(k)x(k)'A'_j1_{\{\theta(k)=j\}}\} \\ &\qquad + (1+1/\alpha^2)\sum_{j\in\mathcal{Z}} p_{ji}E\{B_ju(k)u(k)'B'_j1_{\{\theta(k)=j\}}\} \\ &= (1+\alpha^2)\mathcal{L}_A(X(k)) + (1+1/\alpha^2)\mathcal{L}_B(U(k)), \quad k \ge 0. \end{aligned}$$

□

Appendix: Proofs of Lemmas 20 and 21.

Proof of Lemma 20. Since $P_i x$ is the projection of x onto $\mathcal{S}_i^{\perp}$, we have that $P_i x \in \mathcal{S}_i^{\perp}$ and from the hypothesis of the lemma we have that there exists $T, \epsilon > 0$ such that $\mathcal{Y}_0^T(P_i x, \theta) > \epsilon |P_i x|^2$. Employing this fact and Proposition 15 we evaluate, for $\alpha > 0$,

$$\begin{aligned}\mathcal{Y}_0^T(x,i) &= x'Hx = (x - P_i x + P_i x)'H(x - P_i x + P_i x)\\ &\geq (1-\alpha^2)(x - P_i x)'H(x - P_i x) + (1 - 1/\alpha^2)(P_i x)'H(P_i x)\\ &\geq (1 - 1/\alpha^2)(P_i x)'H(P_i x) = (1 - 1/\alpha^2)\mathcal{Y}_0^T(P_i x, i)\\ &> (1 - 1/\alpha^2)\epsilon |P_i x|^2. \qquad \square\end{aligned}$$

The next results are needed for the proof of Lemma 21.

LEMMA 42. *$E_{x,\theta}\{\mathcal{Y}_0^T(x(t), \theta(t))\} = \mathcal{Y}_0^T(X(t))$.*

Proof. Using (7) we can write that

$$\mathcal{Y}_0^{t,T}(x,\theta) = \sum_{k=t}^{t+T-1} \langle X(k), C'C\rangle = E_{x,\theta}\{\mathcal{Y}_0^T(x(t),\theta(t))\}.$$

However,

$$\sum_{k=t}^{t+T-1} \langle X(k), C'C\rangle = \sum_{\ell=0}^{T-1} \langle \tilde{X}(\ell), C'C\rangle = \mathcal{Y}_0^T(\tilde{X}(0)),$$

where $\tilde{X}_i(\ell) = X_i(t+\ell) = E_{x,\theta}\{x(t+\ell)x(t+\ell)' 1_{\{\theta(t+\ell)=i\}}\}$ for $\ell = 0, \ldots, T-1$, which shows the result. $\square$

LEMMA 43. *Let $V \in \mathcal{H}_\infty^n$. The following inequality holds:*

$$\langle \mathcal{L}_V(X(k)), C'C\rangle \geq (1-\alpha^2)\langle \mathcal{L}_V(\mathcal{L}_A(X(k-1))), C'C\rangle - \kappa \|U(k-1)\|_F \tag{33}$$

for some $0 < \alpha < 1$ and $\kappa > 0$.

Proof. From Lemma 10 we evaluate

$$\begin{aligned}\langle \mathcal{L}_V(X(k)), C'C\rangle \geq{}& (1-\alpha^2)\langle \mathcal{L}_V(\mathcal{L}_A(X(k-1))), C'C\rangle\\ &+ (1 - 1/\alpha^2)\langle \mathcal{L}_V(\mathcal{L}_B(U(k-1)), C'C\rangle\end{aligned} \tag{34}$$

and for the second term on the right-hand side of (34) we employ Proposition 11 (v) to obtain, for $0 < \alpha < 1$,

$$\begin{aligned}&(1 - 1/\alpha^2)\langle \mathcal{L}_V(\mathcal{L}_B(U(k-1))), C'C\rangle\\ &\qquad \geq (1 - 1/\alpha^2)\|C\|_\infty^2 \|V\|_\infty^2 \|B\|_\infty^2 \langle U(k-1), I\rangle\\ &\qquad = -\kappa\langle U(k-1), I\rangle,\end{aligned} \tag{35}$$

where $\kappa = -(1 - 1/\alpha^2)\|C\|_\infty^2 \|V\|_\infty^2 \|B\|_\infty^2 > 0$. The result follows immediately from (34) and (35). $\square$

Proof of Lemma 21. From (33) with $V = I$, we get that

$$\langle X(m), C'C\rangle \geq (1-\alpha^2)\langle \mathcal{L}_A(X(m-1)), C'C\rangle - \kappa \|U(m-1)\|_F. \tag{36}$$

For the first term on the right-hand side of (36), we employ (33) with $V = A$, and we repeat this step recursively for $m = k-1, \dots, t+1$, to obtain

$$\begin{aligned}\langle X(k), C'C\rangle &\geq (1-\alpha^2)^{k-t}\langle \mathcal{L}_A^{k-t}(X(t)), C'C\rangle \\ &\quad - \kappa \sum_{\ell=t}^{k-1} (1-\alpha^2)^{-\ell+k-1}\|U(\ell)\|_F.\end{aligned} \tag{37}$$

Noticing that $(1-\alpha^2)^{k-t} \geq (1-\alpha^2)^T$ for $k-t \leq T$ and $-(1-\alpha^2)^k \geq -1$ for all $k \geq 0$, we get that

$$\langle X(k), C'C\rangle \geq (1-\alpha^2)^T \langle \mathcal{L}_A^{k-t}(X(t)), C'C\rangle - \kappa \sum_{\ell=t}^{k-1} \|U(\ell)\|_F \tag{38}$$

for $t \leq k \leq T+t-1$. Then, from (7) and (38) we evaluate

$$\begin{aligned}\mathcal{Y}_u^{t,T}(x,\theta) &= \sum_{k=t}^{T+t-1} \langle X(k), C'C\rangle + \langle U(k), D'D\rangle \\ &\geq (1-\alpha^2)^T \sum_{k=t}^{T+t-1} \langle \mathcal{L}_A^{k-t}(X(t)), C'C\rangle - \kappa \sum_{k=t}^{T+t-1}\sum_{\ell=t}^{k-1} \|U(\ell)\|_F \\ &\geq (1-\alpha^2)^T \sum_{k=0}^{T-1} \langle \mathcal{L}_A^{k}(X(t)), C'C\rangle - \kappa T \sum_{k=t}^{T+t-1} \|U(k)\|_F\end{aligned}$$

or, equivalently,

$$\mathcal{Y}_u^{t,T}(x,\theta) \geq \delta_1 \mathcal{Y}_0^T(X(t)) - \delta_2 \sum_{k=t}^{T+t-1} \|U(k)\|,$$

where $\delta_1 = (1-\alpha^2)^T$ and $\delta_2 = \kappa T$. Lemma 42 completes the proof. □

Appendix: Proof of Lemma 31. For the proof of Lemma 31, we write X in the following form [14, Them. 7.5.2],

$$X_i = v_i^1 v_i^{1\prime} + \cdots + v_i^{r_i} v_i^{r_i\prime}, \tag{39}$$

where $v_i^m \in \mathbb{R}^n, m = 1, \dots, r$, and $r_i = \operatorname{rank}(X_i) \leq n$, and we write the trajectory $X(k)$ as a linear combination of trajectories $X^{i,m}(k)$ associated with initial condition v_i^m, as follows.

Let $x^{i,m}(0) = v_i^m \in \mathcal{N}(\mathcal{O}_i)$. Let $x^{i,m}(k) \in \mathbb{R}^n$, $m = 1, \dots, r_i$, be given by the difference equation $x^{i,m}(k+1) = A_{\theta(k)} x^{i,m}(k)$, $\theta(0) = i$. Let $X^{i,m}(k) \in \mathcal{H}_F^{n0}$ be the second moment matrix $X_j^{i,m}(k) = E\{x^{i,m}(k)x^{i,m}(k)' 1_{\{\theta(k)=j\}}\}$, $j \in \mathcal{Z}$. Notice that $X_i^{i,m}(0) = v_i^m v_i^{m\prime}$ and $X_j^{i,m}(0) = 0$ for $j \neq i$, and we can write $X_i = \sum_{j\in\mathcal{Z}} \sum_{m=1}^{r_j} X_i^{j,m}(0)$. Then, from the linearity of the operator $\mathcal{L}$ we have that, provided $X(0) = X$,

$$X_{0,i}(k) = \sum_{j\in\mathcal{Z}} \sum_{m=1}^{r_j} X_i^{j,m}(k)$$

which leads to

$$\begin{aligned}\mathcal{Y}_0(X) &= \sum_{k=0}^{\infty}\langle X_0(k), C'C\rangle = \sum_{k=0}^{\infty}\left\langle \sum_{j\in\mathcal{Z}}\sum_{m=1}^{r_j} X^{j,m}(k), C'C\right\rangle \\ &= \sum_{j\in\mathcal{Z}}\sum_{m=1}^{r_j}\sum_{k=0}^{\infty}\langle X^{j,m}(k), C'C\rangle = \sum_{j\in\mathcal{Z}}\sum_{m=1}^{r_j}\mathcal{Y}_0(v_j^m, j). \qquad \square\end{aligned} \tag{40}$$

Proof of Lemma 31. (i) Notice that (39) provides that

$$\begin{aligned}PXP' = 0 &\Leftrightarrow P_i v_i^m v_i^{m\prime} P_i' = 0 \quad \forall i, m \\ &\Leftrightarrow v_i^m \in \mathcal{N}_i \quad \forall i, m \Leftrightarrow \mathcal{Y}_0(v_i^m, i) = 0 \quad \forall i, m.\end{aligned} \tag{41}$$

From (40), $\mathcal{Y}_0(X) = 0$ is equivalent to $\mathcal{Y}_0(v_i^m, i) = 0$, for each i and m, and, from (41), this is equivalent to $PXP' = 0$.

(ii) We shall show that $\mathcal{Y}_0(x, i) \geq \epsilon|x|^2$ whenever $x \in \mathcal{N}_i^{\perp}$. Let $X \in \mathcal{H}_F^n$ be defined as $X_i = xx'$ and $X_j = 0$, $j \neq i$. Since $x \in \mathcal{N}_i^{\perp}$, we have that $P_i x = x$ and it is simple to check that $PXP' = X$ and $X \in \bar{\mathcal{N}}^{\perp}$. Then, from the assumption in the lemma, we obtain $\mathcal{Y}_0(X) \geq \epsilon\|X\|_F$ and it follows that $\mathcal{Y}_0(x, i) = \mathcal{Y}_0(X) \geq \epsilon\|X\|_F = \epsilon|x|^2$. $\square$

Appendix: Proof of Lemma 40.

Proof of Lemma 40. First we show that (A2) holds for $(\bar{A}, \bar{C}, \bar{\mathbb{P}})$. Notice that the requirements in Assumption (A2) hold trivially whenever $\bar{\theta}(0) = f$; indeed, $\bar{x}(k) = 0$, $k > 0$, a.s..

Now consider $\bar{\theta}(0) \neq f$. From the uniform observability of $(A, C, \mathbb{P})$ we have that there exists $T, \epsilon > 0$ for which $\sum_{k=0}^{T-1} E\{|y(k)|^2 | T < k_0\} > \epsilon|x_0|^2$. Then, we define the stopping time $t_0 = \inf\{k : \bar{\theta}(k) = f\}$ and similarly to (32) we evaluate

$$\begin{aligned}\mathcal{Y}_0^T(x, \theta) &\geq \sum_{k=0}^{T-1} E\{|\bar{y}(k)|^2 1_{\{t_0 \geq T\}}\} = \sum_{k=0}^{T-1} E\{|y(k)|^2 1_{\{k_0 \geq T\}}\} \\ &= P\{k_0 \geq T\}\sum_{k=0}^{T-1} E\big\{|y(k)|^2 \big| k_0 \geq T\big\} \geq P\{k_0 \geq T\}\epsilon|x|^2.\end{aligned}$$

Since $q_{\theta f} \leq \nu < 1$, we have that $P\{k_0 \geq T\} > 0$ whenever $\theta(0) = \theta \neq f$ ($\gamma(0) \neq f$). This allows us to write

$$\mathcal{Y}_0^T(x, \theta) \geq \epsilon_2|x|^2, \quad \theta \neq f, \tag{42}$$

where $\epsilon_2 = \epsilon P\big\{k_0 \geq T \big| \theta \neq f\big\} > 0$.

Now we show that (A1) holds for $(\bar{A}, \bar{C}, \bar{\mathbb{P}})$, that is, we show that $\mathcal{Y}_0(x, \theta) = \infty$ provided that $\sum_{k=0}^{\infty} E\{|\bar{x}(k)|^2\} = \infty$. Note that we can assume that $\theta(0) \neq f$, otherwise we have that $\sum_{k=0}^{\infty} E\{|\bar{x}(k)|^2\} = |x|^2 < \infty$. We evaluate, for any $\ell \geq 0$,

$$\begin{aligned}\mathcal{Y}_0(x, \theta) &\geq \sum_{k=\ell}^{\infty} E\{|\bar{y}(k)|^2\} = \sum_{m=0}^{\infty}\sum_{k=mT+\ell}^{(m+1)T+\ell-1} E\{|\bar{y}(k)|^2\} \\ &\geq \sum_{m=0}^{\infty}\sum_{k=mT+\ell}^{(m+1)T+\ell-1} E\{|\bar{y}(k)|^2 1_{\{t_0 > mT+\ell\}}\} \\ &= \sum_{m=0}^{\infty} P(t_0 > mT+\ell)\sum_{k=mT+\ell}^{(m+1)T+\ell-1} E\big\{|\bar{y}(k)|^2 \big| t_0 > mT+\ell\big\},\end{aligned}$$

and since $t_0 > mT+\ell$ implies in particular that $\theta(mT+\ell) \neq f$, we get from (42) that

$$\begin{aligned}\mathcal{Y}_0(x,\theta) &\geq \sum_{m=0}^{\infty} P(t_0 > mT+\ell)\epsilon_2 E\big\{|\bar{x}(mT+\ell)|^2 \big|\, t_0 > mT+\ell\big\} \\ &= \epsilon_2 \sum_{m=0}^{\infty} E\{|\bar{x}(mT+\ell)|^2 1_{\{t_0 > mT+\ell\}}\}.\end{aligned} \tag{43}$$

Summing up (43) for $\ell = 0,\ldots,T-1$, we obtain

$$T\mathcal{Y}_0(x,\theta) \geq \epsilon_2 \sum_{k=0}^{\infty} E\{|\bar{x}(k)|^2 1_{\{t_0>k\}}\}. \tag{44}$$

Now, recalling that $\theta(0) \neq f$, we evaluate

$$\begin{aligned}E\{|\bar{x}(k)|^2 1_{\{t_0>k\}}\} &= P\{t_0 > k\} E\big\{|\bar{x}(k)|^2 \big|\, t_0 > k\big\} \\ &= (1-q_{nf})^k E\big\{|\bar{x}(k)|^2 \big|\, t_0 > k\big\} = (1-q_{nf})^k E\big\{|\bar{x}(k)|^2 \big|\, t_0 = k\big\} \\ &= \frac{(1-q_{nf})}{q_{nf}}(1-q_{nf})^{k-1} q_{nf} E\big\{|\bar{x}(k)|^2 \big|\, t_0 = k\big\} \\ &= \frac{(1-q_{nf})}{q_{nf}} P\{t_0 = k\} E\big\{|\bar{x}(k)|^2 \big|\, t_0 = k\big\} = \frac{(1-q_{nf})}{q_{nf}} E\big\{|\bar{x}(k)|^2 1_{\{t_0=k\}}\big\} \\ &= \epsilon_3 E\{|\bar{x}(k)|^2 1_{\{t_0=k\}}\},\end{aligned} \tag{45}$$

where $\epsilon_3 = (1-q_{nf})/q_{nf}$. Finally, (44), (45), and the fact that $\bar{x}(k) = 0$ a.s. for each $k > t_0$ lead to

$$\begin{aligned}T\mathcal{Y}_0(x,\theta) &\geq \epsilon_2 \sum_{k=0}^{\infty} E\{|\bar{x}(k)|^2 1_{\{t_0>k\}}\} \\ &= \epsilon_2 \frac{\epsilon_3}{1+\epsilon_3} \frac{1+\epsilon_3}{\epsilon_3} \sum_{k=0}^{\infty} E\{|\bar{x}(k)|^2 1_{\{t_0>k\}}\} \\ &= \epsilon_2 \frac{\epsilon_3}{1+\epsilon_3} \left(\sum_{k=0}^{\infty} E\{|\bar{x}(k)|^2 1_{\{t_0>k\}}\} + \sum_{k=0}^{\infty} E\{|\bar{x}(k)|^2 1_{\{t_0=k\}}\} \right) \\ &= \epsilon_2 \frac{\epsilon_3}{1+\epsilon_3} \sum_{k=0}^{\infty} E\{|\bar{x}(k)|^2\} = \infty.\end{aligned}$$

We have shown that (A1) and (A2) holds for $(\bar{A}, \bar{C}, \bar{\mathbb{P}})$; Theorem 18 provides the first assertion in the lemma. For the second assertion, note that $\mathcal{Y}_0(x,\theta) = 0$ whenever $\theta = f$. □

Acknowledgment. The authors would like to express their gratitude to the referees for their suggestions and helpful comments. We are particularly grateful to one of the referees for the comment that we have added to Example 3, besides many others.

REFERENCES

[1] E. F. Costa and J. B. R. do Val, *On the detectability and observability of discrete-time Markov jump linear systems*, Systems Control Lett., 44 (2001), pp. 135–145.

[2] E. F. Costa and J. B. R. do Val, *On the detectability and observability of continuous-time Markov jump linear systems*, SIAM J. Control Optim., 41 (2002), pp. 1295–1314.
[3] E. F. Costa and J. B. R. do Val, *Weak detectability and the linear quadratic control problem of discrete-time Markov jump linear systems*, Int. J. Control, 16/17 (2002), pp. 1282–1292.
[4] E. F. Costa, J. B. R. do Val, and M. D. Fragoso, *On a detectability concept of discrete-time infinite Markov jump linear systems*, Stochastic Analysis and Applications, to appear.
[5] E. F. Costa, J. B. R. do Val, and M. D. Fragoso, *On a detectability concept of discrete-time infinite Markov jump linear systems*, in Proceedings of the 15th Triennal World Congress IFAC, 2002, pp. 2660–2665.
[6] O. L. V. Costa, *Discrete-time coupled Riccati equations for systems with Markov switching parameters*, J. Math. Anal. Appl., 194 (1995), pp. 197–216.
[7] O. L. V. Costa and M. D. Fragoso, *Discrete-time LQ-optimal control problems for infinite Markov jump parameter systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 2076–2088.
[8] M. D. Fragoso and J. Baczynski, *Optimal control for continuous time problems with infinite Markov jump parameters*, SIAM J. Control Optim., 40 (2001), pp. 270–297.
[9] M. D. Fragoso and J. Baczynski, *Lyapunov coupled equations for continuous-time infinite Markov jump linear systems*, J. Math. Anal. Appl., 274 (2002), pp. 319–335.
[10] M. D. Fragoso and J. Baczynski, *Stochastic versus mean square stability in continuous time linear infinite Markov jump parameter systems*, Stochastic Anal. Appl., 20 (2002), pp. 347–356.
[11] Y. Ji and H. J. Chizeck, *Controllability, stabilizability and continuous time Markovian jump linear quadratic control*, IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.
[12] T. Morozan, *Stability and control for linear systems with jump Markov perturbations*, Stochastic Anal. Appl., 13 (1995), pp. 91–110.
[13] W. S. Gray and J. P. Mesko, *Observability functions for linear and nonlinear systems*, Systems Control Lett., 38 (1999), pp. 99–113.
[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990.
[15] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[16] K. Knopp, *Infinite Sequences and Series*, Dover, New York, 1956.
[17] I. R. Petersen, *Notions of observability for uncertain linear systems with structured uncertainty*, SIAM J. Control Optim., 41 (2002), pp. 345–361.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2157–2165

SOME RESULTS ON TWO-PERSON ZERO-SUM LINEAR QUADRATIC DIFFERENTIAL GAMES*

PINGJIAN ZHANG†

Abstract. Considered in this paper are the two-person zero-sum linear quadratic differential games. It is shown that the value of the game exists if and only if both the upper and lower values exist. As a consequence, we prove that another necessary and sufficient condition for the existence of the values of the game is the existence of an open loop–open loop saddle point. An example is also given in which the lower value exists but the upper value does not exist.

Key words. differential games, value of games, saddle points, integral equations, optimal controls

AMS subject classifications. 90D05, 90D25, 45B05

DOI. 10.1137/S036301290342560X

1. Introduction.

Since the 1980s, much work has been contributed to linear quadratic differential games due to their essential role in modern robust control and H^∞-optimization design; see, e.g., [1], [3], [7]. In this paper, we consider the two-person, zero-sum linear quadratic differential games on a finite horizon. It is well known that the solvability of the corresponding Riccati differential equations is equivalent to the existence of the H^∞-optimal control [1]. In [2], the relationship between saddle points of the game and the solvability of various Riccati differential equations is studied. It is shown that

(a) if the Riccati differential equation admits a solution, then the game admits a closed loop–closed loop saddle point;
(b) if both the Riccati differential equation and the lower Riccati differential equation (for definitions, see [2]) admit a solution, then the game admits a closed loop–open loop saddle point;
(c) if both the Riccati differential equation and the upper Riccati differential equation (for definitions, see [2]) admit a solution, then the game admits an open loop–closed loop saddle point.

Note that nothing is said about the open loop–open loop saddle point. This paper is to address this issue. Obviously, the existence of open loop–open loop saddle points guarantees the existence of the value of the game; we shall show that this is also necessary. This follows easily from our main result, which states that a necessary and sufficient condition for the existence of the value of the game is that both the lower value and the upper value of the game exist. Examples show that, except for the open loop–open loop saddle point, existence of any other type of saddle point cannot guarantee the existence of the open loop value of the game.

Consider the following dynamic system:

$$\dot{x} = Ax + Bu + Gv \qquad \text{in} \quad [0, T], \tag{1.1}$$

$$x(0) = x_0 \tag{1.2}$$

*Received by the editors April 5, 2003; accepted for publication (in revised form) August 29, 2004; published electronically May 27, 2005.
http://www.siam.org/journals/sicon/43-6/42560.html
†School of Computer Science and Technology, South China University of Technology, Guangzhou 510640, China (pjzhang@scut.edu.cn).

with quadratic index

$$J(u,v)=\int_0^T (|u|^2-|v|^2+x'Qx)dt+x(T)'Wx(T), \tag{1.3}$$

where $T>0$ is a given final time, and where A,B,G,Q,W are matrices of suitable dimension such that Q and W are symmetric but not necessarily nonnegative. Moreover, without loss of generality, cross terms in (u,v,x) are not present in the quadratic form because they can be eliminated by appropriate transformations. Let

$$v^-(x_0)=\sup_{v\in L^2(0,T;\mathbb{R}^m)}\inf_{u\in L^2(0,T;\mathbb{R}^l)} J(u,v)$$

be the open loop lower value of the game and

$$v^+(x_0)=\inf_{u\in L^2(0,T;\mathbb{R}^l)}\sup_{v\in L^2(0,T;\mathbb{R}^m)} J(u,v)$$

be the open loop upper value of the game. If both $v^-(x_0)$ and $v^+(x_0)$ exist and $v^-(x_0)=v^+(x_0)$, then we say that the open loop value of the game exists and is $v(x_0)=v^-(x_0)=v^+(x_0)$. Hereafter, unless stated otherwise, the value, the lower value, and the upper value of the game mean the open loop value, the open loop lower value, and the open loop upper value of the game, respectively . Obviously, we always have

$$v^-(x_0)\le v^+(x_0).$$

Natural questions are, Do the lower value and the upper value exist? Are they equal to each other if both of them exist? How are they related to the existence of saddle points? In section 3, we discuss an example in which the lower value exists while the upper value does not. We then prove the main theorem of the paper in section 4, using results established on Fredholm integral equations in section 2. Some final remarks are given in section 5.

2. Solution set of Fredholm integral equations. It is convenient to introduce some technical notation that will be used throughout the paper. Let $M^+=BB', M^-=GG', M=M^+-M^-$ define the following linear operators:

$$\begin{aligned}
&L: L^2(0,T;\mathbb{R}^n)\to L^2(0,T;\mathbb{R}^n), && (Lx)(t)=\textstyle\int_0^t e^{A(t-s)}x(s)ds,\\
&\hat{L}: L^2(0,T;\mathbb{R}^n)\to \mathbb{R}^n, && (\hat{L}x)(t)=\textstyle\int_0^T e^{A(T-s)}x(s)ds,\\
&S:\mathbb{R}^n\to L^2(0,T;\mathbb{R}^n), && (Sy)(t)=e^{At}y,\\
&\hat{S}:\mathbb{R}^n\to\mathbb{R}^n, && \hat{S}y=e^{AT}y
\end{aligned}$$

with the adjoint operators given by

$$\begin{aligned}
&L^*: L^2(0,T;\mathbb{R}^n)\to L^2(0,T;\mathbb{R}^n), && (L^*x)(t)=\textstyle\int_t^T e^{A'(s-t)}x(s)ds,\\
&\hat{L}^*:\mathbb{R}^n\to L^2(0,T;\mathbb{R}^n), && (\hat{L}^*y)(t)=e^{A'(T-t)}y,\\
&]2pt]S^*: L^2(0,T;\mathbb{R}^n)\to\mathbb{R}^n, && S^*x=\textstyle\int_0^T e^{A't}x(t)dt,\\
&\hat{S}^*:\mathbb{R}^n\to\mathbb{R}^n, && \hat{S}^*y=e^{A'T}y.
\end{aligned}$$

Moreover, define operators

$$K=L^*QL+\hat{L}^*W\hat{L},\qquad \hat{K}=L^*QS+\hat{L}^*W\hat{S}.$$

Using these operators, the system (1.1)–(1.2) can be rewritten as

$$x = Sx_0 + LBu + LGv$$

with

$$x(T) = \hat{S}x_0 + \hat{L}Bu + \hat{L}Gv$$

and the index (1.3) can be written as a bilinear form in Hilbert space $L^2(0,T;\mathbb{R}^l) \times L^2(0,T;\mathbb{R}^m)$:

$$\begin{aligned} J(u,v) \quad &= ((I + B'KB)u, u) - ((I - G'KG)v, v) + 2(B'KGv, u) \\ &\quad + 2(B'(L^*QS + \hat{L}^*W\hat{S})x_0, u) + 2(G'(L^*QS + \hat{L}^*W\hat{S})x_0, v) \\ &\quad + \langle (S^*QS + \hat{S}^*W\hat{S})x_0, x_0\rangle, \end{aligned}$$

where $(\cdot,\cdot)$ and $\langle\cdot,\cdot\rangle$ denote the inner product in L^2 and the Euclidean space, respectively. Therefore, the game problem is converted into the min-max problem of bilinear forms in Hilbert space.

Remark 1. Under coercivity assumptions on $I + B'KB$ and $I - G'KG$, the game problems admits a unique open loop–open loop saddle point which can be constructed explicitly. In this paper, no such coercivity conditions are imposed, and the usual Riccati equation approach cannot be applied.

Remark 2. The problem setting of this paper can be easily generated for the time-varying systems, where the transition matrix $\Phi(t,s)$ replaces $e^{A(t-s)}$ in the operators defined above.

We will first establish some preliminary results.

LEMMA 2.1. *Let $\mathcal{N}(A)$ denote the kernel of operator A, and let $C, D \in \mathcal{L}(X)$ be linear bounded operators on some Hilbert space X. Then $\mathcal{N}(I+CD) = C\mathcal{N}(I+DC)$.*

Proof. Let $(I+CD)x = 0$. Then $x = -CDx$ and $(I+DC)Dx = D(I+CD)x = 0$, that is, $Dx \in \mathcal{N}(I + DC)$, and hence $x = -CDx \in C\mathcal{N}(I + DC)$. Conversely, let $x = Cp$ for some $p \in \mathcal{N}(I+DC)$. Then $(I+CD)x = (I+CD)Cp = C(I+DC)p = 0$, and hence $x \in \mathcal{N}(I + CD)$. □

Consider the Fredholm integral equation with parameter v:

$$(I + KM^+)p = Gv + \hat{K}x_0. \tag{2.1}$$

Let $V(x_0) = \{v \in L^2(0,T;\mathbb{R}^m) : v$ be such that (2.1) has solution.$\}$. If $V(x_0)$ is not empty, then we can pick a $v \in V(x_0)$ and let $P(v,x_0)$ be the solution set of (2.1) corresponding to v. If we choose one $p \in P(v,x_0)$, then $P(v,x_0)$ can be expressed as

$$P(v,x_0) = p + N, \tag{2.2}$$

where $N = \mathcal{N}(I + KM^+)$. The following theorem characterizes the set $V(x_0)$.

THEOREM 2.2. *$V(x_0)$ is given by*

$$V(x_0) = \{v \in L^2(0,T;\mathbb{R}^m) : (v, G'q) + \langle x_0, q(0)\rangle = 0 \ \forall q \in N\}.$$

Proof. Let $q \in N$; then $q = -KM^+q$ and $q(0) = -\hat{K}^*M^+q$ by simple verification. Therefore, by virtue of Lemma 2.1 and the fact that operator K is compact; hence

$I + KM^+$ has closed range. Thus

$$\begin{aligned} & (v, G'q) + \langle x_0, q(0)\rangle = 0 \ \forall q \in N \\ \Leftrightarrow \quad & (v, G'KM^+q) + (\hat{K}x_0, M^+q) = 0 \ \forall q \in N \\ \Leftrightarrow \quad & (KGv + \hat{K}x_0, M^+q) = 0 \ \forall q \in N \\ \Leftrightarrow \quad & KGv + \hat{K}x_0 \perp \mathcal{N}(I + M^+K) \\ \Leftrightarrow \quad & KGv + \hat{K}x_0 \in \mathcal{R}(I + KM^+). \qquad \square \end{aligned}$$

COROLLARY 2.3. *The following statements hold:*

(1) $V(0) = (G'N)^\perp$.

(2) *If $V(x_0)$ is not empty. Then $V(x_0) = v + V(0)$ for arbitrary $v \in V(x_0)$.*

THEOREM 2.4. *Suppose $V(x_0)$ is not empty. Then for every $v \in V(x_0)$, $G'P(v, x_0) \cap V(x_0)$ contains exactly one element.*

Proof. Fix $v \in V(x_0)$. Then $G'P(v, x_0) = G'p + G'N$ for some $p \in N$. Let $\mathbf{P}_{G'N}$ and $\mathbf{P}_{(G'N)^\perp}$ be projections onto $G'N$ and $(G'N)^\perp$, respectively. Decompose $v = v_1 + v_2$ with $v_1 = \mathbf{P}_{G'N}v, v_2 = \mathbf{P}_{(G'N)^\perp}v$, and $G'p = q_1 + q_2$ with $q_1 = \mathbf{P}_{G'N}G'p, q_2 = \mathbf{P}_{(G'N)^\perp}G'p$. Then, since $q_2 \in (G'N)^\perp$, $v_1 + q_2 \in V(x_0)$ by Corollary 2.3. Moreover, since $q_2 = G'p - q_1 \in G'p + G'N$ and $v_1 \in G'N$, $v_1 + q_2 \in G'p + G'N = G'P(v, x_0)$. Thus, we have verified that

$$\mathbf{P}_{G'N}v + \mathbf{P}_{(G'N)^\perp}G'p = v_1 + q_2 \in G'P(v, x_0) \cap V(x_0).$$

To prove the uniqueness, let $u_1, u_2 \in G'P(v, x_0) \cap V(x_0)$. Then $u_1 - u_2 \in G'N$; on the other hand, $u_1 - u_2 \in V(0) = (G'N)^\perp$, and hence $u_1 = u_2$. $\square$

Now, denote the only element in $G'P(v, x_0) \cap V(x_0)$ by $G'p_*$; although p_* might not be unique in $P(v, x_0)$, $G'p_*$ is uniquely determined. Therefore, we are able to define the operator

$$D^{x_0} : V(x_0) \to V(x_0), \qquad D^{x_0}v = G'p_*.$$

In particular, we can define

$$D : V(0) \to V(0), \qquad Dv = G'p_*^0.$$

To characterize these operators, let $(I + KM^+)_{N^\perp}$ be the restriction of $I + KM^+$ on $N^\perp$. Then $(I + KM^+)_{N^\perp}$ is a bijection from $N^\perp$ to $\mathcal{R}(I + KM^+)$ and hence has a bounded inverse. Now choose

$$p_* = ((I + KM^+)_{N^\perp})^{-1}(KGv + \hat{K}x_0).$$

Then

$$D^{x_0}v = G'((I + KM^+)_{N^\perp})^{-1}(KGv + \hat{K}x_0) \tag{2.3}$$

and

$$Dv = G'((I + KM^+)_{N^\perp})^{-1}KGv. \tag{2.4}$$

THEOREM 2.5. *D is a compact, self-adjoint operator on $V(0)$.*

Proof. Since K is compact, so is D. To show that D is self-adjoint, let $v_i \in V(0), p_i \in P(v_i, 0), i = 1, 2$; then

$$\begin{aligned} & (Dv_2, v_1) - (Dv_1, v_2) \\ = \ & (G'p_2, v_1) - (G'p_1, v_2) \\ = \ & (Gv_1, p_2) - (Gv_2, p_1) \\ = \ & (Gv_1, KGv_2 - KM^+p_2) - (Gv_2, KGv_1 - KM^+p_1) \\ = \ & -(KGv_1, M^+p_2) + (KGv_2, M^+p_1) \\ = \ & -((I + KM^+)p_1, M^+p_2) + ((I + KM^+)p_2, M^+p_1) \\ = \ & 0. \end{aligned}$$

This completes the proof. □

3. A game in which value does not exist. Consider the following dynamics:

$$\dot{x} = x + u + v, \quad x(0) = 0,$$

with index

$$J(u, v) = \int_0^1 (u^2 - v^2 + 2x^2)dt.$$

We claim that the lower value of the game is 0, whereas the upper value does not exist. To prove $v^-(0) = 0$ is easy: for each control action v, $J(-v, v) = 0$, and thus $\inf J(u, v) \leq 0$, and hence $v^-(0) \leq 0$. On the other hand, $v^-(0) \geq \inf J(u, 0) = 0$. Now let's deal with the hard part; we need to show $v^+(0) = \infty$. Fix control action u. We write the index as

$$J(u, v) = -((I - 2L^*L)v, v) + 4(L^*Lu, v) + ((I + 2L^*L)u, u),$$

where L is the integral operator defined in section 2. Since the underlying Riccati differential equation blows up in interval [0, 1], by virtue of results in [4] and [6], $I - 2L^*L$ is not positive definite; i.e., there exists a nonzero $\bar{v}$ such that $((I-2L^*L)\bar{v}, \bar{v}) \leq 0$. In fact, strict inequality holds. To see this, take $\bar{v}(t) = 1$ on [0, 1]. We calculate $((I - 2L^*L)\bar{v}, \bar{v}) \approx 1 - 2 * 0.757762 < 0$, and thus $\sup\{J(u, v); v \in L^2(0, 1)\} = \infty$ for every u given, which shows that the upper value of the game does not exist.

Remark 3. It is easy to construct similar examples in which the upper value exist but the lower value does not exist and examples in which neither the lower value nor the upper value exists.

Remark 4. The example game admits an open loop–closed loop saddle point as well as a closed loop–closed loop saddle point because both the Riccati differential equation and the lower Riccati differential equation admit a solution on [0, 1] (see [2]).

4. Main results and proof. The main results in this paper are contained in the following.

THEOREM 4.1. *Consider the game problem* (1.1)–(1.3). *The following statements are equivalents:*

(1) *There exists an open loop–open loop saddle point* $(u^*, v^*) \in L^2(0, T; \mathbb{R}^m) \times L^2(0, T; \mathbb{R}^l)$ *such that*

$$J(u^*, v) \leq J(u^*, v^*) \leq J(u, v^*) \quad \forall u \in L^2(0, T; \mathbb{R}^m), v \in L^2(0, T; \mathbb{R}^l).$$

(2) *The value of the game exists.*

(3) *Both the lower value and the upper value exist.*

To prove the theorem, some lemmas are needed. Using the notation of section 2, we rewrite the index as

$$J(u,v) = ((I+E)u,u) + 2(f(v,x_0),u) + J(0,v),$$

where

$$E = B'L^*QLB + B'\hat{L}^*W\hat{L}B = B'KB$$

and

$$f(v,x_0) = B'KGv + B'\hat{K}x_0.$$

Obviously, the lower value exists if and only if both of the following conditions hold:

(A) There exists $v \in L^2(0,T;\mathbb{R}^m)$ such that

$$\hat{J}(v) = \inf_u J(u,v) > -\infty. \tag{4.1}$$

(B) Let $\hat{V}(x_0) = \{v \in L^2(0,T;\mathbb{R}^m) : v \text{ is such that (4.1) holds true}\}$. Then

$$\sup\{\hat{J}(v) : v \in \hat{V}(x_0)\} < \infty. \tag{4.2}$$

By standard extremal theory (see, e.g., [5]), (4.1) holds if and only if $I+E \geq 0$ and

$$f(v,x_0) \in \mathcal{N}(I+E)^{\perp} = \mathcal{R}(I+E). \tag{4.3}$$

Moreover, u is an optimizer if and only if

$$f(v,x_0) + (I+E)u = 0. \tag{4.4}$$

Our first observation is the following.

LEMMA 4.2. *Suppose the lower value exists. Then $\hat{V}(x_0)$ coincides with $V(x_0)$.*

Proof. Let $v \in \hat{V}(x_0)$. Then, by definition and (4.3),

$$(f(v,x_0),u) = (B'KGv + B'\hat{K}x_0, u) = 0 \forall u \in \mathcal{N}(I+E). \tag{4.5}$$

Notice that $B\mathcal{N}(I+E) = \mathcal{N}(I+M^+K)$ by Lemma 4.2, and that (4.5) is equivalent to

$$KGv + \hat{K}x_0 \in (\mathcal{N}(I+M^+K))^{\perp} = \mathcal{R}(I+KM^+); \tag{4.6}$$

hence $v \in V(x_0)$. Since the above procedure is invertible, the opposite inclusion also holds. □

LEMMA 4.3. *Suppose the lower value exists. Then for $v \in V(x_0)$,*

$$\hat{J}(v) = -(v,v) + (Gv,p) + \langle x_0, p(0)\rangle, \tag{4.7}$$

where $p \in P(v,x_0)$ and $\hat{J}(v)$ is independent of the choice of p.

Proof By (4.4), the optimizer u can be expressed as

$$u = -B'(KBu + KGv + \hat{K}x_0). \tag{4.8}$$

Let $p = KBu + KGv + \hat{K}x_0$. Then $u = -B'p$ and $p = KB(-B'p) + KGv + \hat{K}x_0$, and thus $p \in P(v, x_0)$. Conversely, let $p \in P(v, x_0)$, $u = -B'p$ must be the optimizer for $J(u, v)$. Therefore,

$$\begin{aligned}
\hat{J}(v) &= ((I+E)u, u) + 2(f(v, x_0), u) + J(0, v) \\
&= (f(v, x_0), u) + J(0, v) \\
&= (B'KGv + B'\hat{K}x_0, -B'p) + J(0, v) \\
&= -(Q(LGv + Sx_0), LM^+p) - \langle W(\hat{L}Gv + \hat{S}x_0), \hat{L}M^+p\rangle \\
&\quad -\|v\|^2 + (Q(LGv + Sx_0), LGv + Sx_0) + \langle W(\hat{L}Gv + \hat{S}x_0), \hat{L}Gv + \hat{S}x_0\rangle \\
&= -\|v\|^2 + (Gv, L^*Qx + \hat{L}^*Wx(T)) + \langle x(0), S^*Qx + \hat{S}^*Wx(T)\rangle \\
&= -\|v\|^2 + (Gv, p) + \langle x_0, p(0)\rangle
\end{aligned}$$

and $\hat{J}(v)$ does not depend on the choice of p. □

LEMMA 4.4. *Suppose the lower value exists. Then it equals* $\langle x_0, p(0)\rangle$, *and is independent of the choice of solution* p *to the following Fredholm equation:*

$$(I + KM)p = \hat{K}x_0. \tag{4.9}$$

Proof. We want to calculate $\sup\{\hat{J}(v); v \in V(x_0)\}$. Fix $\bar{v} \in V(x_0)$ and $\bar{p} \in P(\bar{v}, x_0)$; then $v \in V(x_0)$ can be written as $v = \bar{v} + v^0$ and $p \in P(v, x_0)$ can be written as $p = \bar{p} + p^0$ for some $v^0 \in V(0)$ and $p^0 \in P(v^0, 0)$. Similar to Theorem 2.5, it can be shown that

$$-(G\bar{v}, p^0) + (Gv^0, \bar{p}) = \langle x_0, p^0(0)\rangle.$$

Then

$$\begin{aligned}
\hat{J}(v) &= -\|\bar{v} + v^0\|^2 + (G(\bar{v} + v^0), \bar{p} + p^0) + \langle x_0, \bar{p}(0) + p^0(0)\rangle \\
&= -\|\bar{v}\|^2 + (G\bar{v}, \bar{p}) + \langle x_0, \bar{p}(0)\rangle - \|v^0\|^2 + (G\bar{v}p^0) \\
&\quad + \langle x_0, p^0(0)\rangle + (Gv^0, \bar{p}) + (Gv^0, p^0) - 2(\bar{v}mv^0) \\
&= \hat{J}(\bar{v}) - \|v^0\|^2 + (v^0, G'p^0) - 2(v^0, \bar{v} - G'\bar{p}).
\end{aligned}$$

Hence,

$$\sup_{v \in V(x_0)} \hat{J}(v) = \hat{J}(\bar{v}) + \sup_{v^0 \in V(0)} (-((I - D)v^0, v^0) - 2(v^0, \bar{v} - D^{x_0}\bar{v})). \tag{4.10}$$

Again, by standard extremal theory, $\sup_{v \in V(x_0)} \hat{J}(v)$ is finite if and only if

$$I - D \geq 0$$

and

$$\bar{v} - D^{x_0}\bar{v} \in \mathcal{N}(I - D)^\perp = \mathcal{R}(I - D).$$

Moreover, if $v^0 \in V(0)$ is such that

$$\bar{v} - D^{x_0}\bar{v} + (I - D)v^0 = 0,$$

then $v^* = \bar{v} + v^0$ is the optimizer for $\hat{J}(v)$. Recalling definitions (2.3)–(2.4), we have

$$v^* = D^{x_0}\bar{v} + Dv^0 = D^{x_0}(\bar{v} + v^0) = D^{x_0}v^*.$$

Now let $v = v^*$, and let p be any $p^* \in P(v^*, x_0)$ in (4.7). Then we have

$$\sup_{v \in V(x_0)} \hat{J}(v) = \hat{J}(v^*) = -\|v^*\|^2 + (v^*, G'p^*) + \langle x_0, p^*(0)\rangle = \langle x_0, p^*(0)\rangle,$$

which is independent of the choice of p^*.

Finally, we verify that p^* satisfies the Fredholm equation (4.9). In fact, since $p^* \in P(v^*, x_0)$, we have $(I + KM^+)p^* = KGv^* + \hat{K}x_0$; moreover, $v^* = D^{x_0}v^* = G'p^*$, and hence $(I + KM)p^* = \hat{K}x_0$. □

LEMMA 4.5. *Suppose the upper value exists. Then it equals* $\langle x_0, p(0)\rangle$ *and is independent of the choice of solution* p *to the Fredholm equation* (4.9).

Proof. Notice that

$$\inf_u \sup_v J(u, v) = -\sup_v \inf_u (-J(u, v)).$$

Hence, if the upper value exists, then, by Lemma 4.4,

$$\sup_v \inf_u (-J(u, v)) = \langle x_0, q(0)\rangle, \tag{4.11}$$

where q satisfies

$$(I + KM)q = -\hat{K}x_0$$

and (4.11) is independent of the choice of q, or equivalently, the upper value equals $\langle x_0, p(0)\rangle$ and is independent of the choice of solution p to the Fredholm equation (4.9). □

Proof of Theorem 4.1. 1) ⇒ 2) and 2) ⇒ are obvious. We need only prove 3) ⇒ 1).

By virtue of Lemma 4.4, if the lower value exists, then there exists $v^* = G'p^* \in V(x_0)$, where p^* satisfies (4.9). Let u^* be the optimizer of $J(u, v^*)$. Then $u^* = -B'p^*$ and

$$J(u^*, v^*) \le J(u, v^*).$$

Similarly, if the upper value exists, the same (u^*, v^*) verifies

$$-J(u^*, v^*) \le -J(u^*, v).$$

Thus, (u^*, v^*) constitutes the desired open loop–open loop saddle point. □

5. Concluding remarks. This paper studies the two-person, zero-sum linear quadratic differential games on a finite horizon. Some necessary and sufficient conditions for the existence of the value of the game are derived. Although we obtain the open loop–open loop saddle points whenever the value of the game exists, nothing is said about their synthesis as state feedback. In subsequent papers, we shall further investigate the relationship among open loop saddle points, closed loop saddle points, value of the game, and the Riccati differential equations. Another future research will discuss infinite horizontal differential game problems.

Acknowledgment. The author would like to thank the referees for their constructive suggestions which improved this work.

REFERENCES

[1] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minmax Design Problems: A Dynamical Game Approach*, Birkhaüser, Boston, 1995.

[2] P. Bernhard, *Linear-quadratic, two-person, zero-sum differential game: Necessary and sufficient conditions*, J. Optim. Theory Appl., 27 (1979), pp. 51–69.

[3] K. Glover, D. J. N. Limebeer, J. C. Doyle, E. M. Kasenally, and M. G. Safonov, *A characterization of all solutions to the four block general distance problems*, SIAM J. Control Optim., 29 (1991), pp. 283–324.

[4] D. Hinrichsen and A. J. Pritchard, *The Riccati Equation*, unpublished lecture notes, 1991.

[5] A. D. Ioffe and V. M. Tikhomirov, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.

[6] B. Jacob, *Linear quadratic optimal control of time-varying systems with indefinite costs on Hilbert spaces: The finite horizon problem*, J. Math. Systems Estim. Control, 8 (1998), pp. 261–288.

[7] K. Zhou and P. Khargonekar, *An algebraic Riccati equation approach to H^∞ optimization*, Systems Control Lett., 11 (1988), pp. 85–91.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2166–2190

ON THE ATTAINABLE SET FOR TEMPLE CLASS SYSTEMS WITH BOUNDARY CONTROLS*

FABIO ANCONA† AND GIUSEPPE MARIA COCLITE‡

Abstract. Consider the initial-boundary value problem for a strictly hyperbolic, genuinely nonlinear, Temple class system of conservation laws

$$u_t + f(u)_x = 0, \qquad u(0,x) = \overline{u}(x), \qquad \begin{cases} u(t,a) = \widetilde{u}_a(t), \\ u(t,b) = \widetilde{u}_b(t), \end{cases} \tag{1}$$

on the domain $\Omega = \{(t,x) \in \mathbb{R}^2 : t \geq 0,\ a \leq x \leq b\}$. We study the mixed problem (1) from the point of view of control theory, taking the initial data $\overline{u}$ fixed and regarding the boundary data $\widetilde{u}_a$, $\widetilde{u}_b$ as control functions that vary in prescribed sets $\mathcal{U}_a$, $\mathcal{U}_b$, of $\mathbf{L}^\infty$ boundary controls. In particular, we consider the family of configurations

$$\mathcal{A}(T) \doteq \{u(T,\cdot);\ u \text{ is a sol. to (1)},\ \widetilde{u}_a \in \mathcal{U}_a,\ \widetilde{u}_b \in \mathcal{U}_b\}$$

that can be attained by the system at a given time $T > 0$, and we give a description of the attainable set $\mathcal{A}(T)$ in terms of suitable Oleinik-type conditions. We also establish closure and compactness of the set $\mathcal{A}(T)$ in the $\mathbf{L}^1$ topology.

Key words. hyperbolic systems, conservation laws, Temple class systems, boundary control, attainable set

AMS subject classifications. 35L65, 35B37

DOI. 10.1137/S0363012902407776

1. Introduction.

Consider the initial-boundary value problem for a strictly hyperbolic, genuinely nonlinear, system of conservation laws in one space dimension

$$u_t + f(u)_x = 0, \tag{1.1}$$

$$u(0,x) = \overline{u}(x), \tag{1.2}$$

$$u(t,a) = \widetilde{u}_a(t), \tag{1.3}$$

$$u(t,b) = \widetilde{u}_b(t) \tag{1.4}$$

on the strip $\Omega = \{(t,x) \in \mathbb{R}^2;\ t \geq 0,\ x \in [a,b]\}$. Here, $u = u(t,x) \in \mathbb{R}^n$ is the vector of the conserved quantities, $\widetilde{u}_a, \widetilde{u}_b$ are measurable, bounded boundary data, and the flux function $f : U \mapsto \mathbb{R}^n$ is a smooth vector field defined on some open set $U \subseteq \mathbb{R}^n$ that belongs to a class of fields introduced by Temple [29, 28] for which rarefaction and Hugoniot curves coincide. We recall that, for problems of this type, classical solutions may develop discontinuities in finite time, regardless of the regularity of the initial and boundary data. Hence, it is natural to consider weak solutions in the sense of distributions. Moreover, since, in general, the Dirichlet conditions (1.3)–(1.4) cannot be fulfilled pointwise a.e. (see [7, 19]), different weaker formulations of the boundary condition have been considered in the literature (see [1, 20, 27] and references therein). Here, following Dubois and LeFloch [19], we will adopt a formulation of (1.3)–(1.4)

*Received by the editors May 15, 2002; accepted for publication (in revised form) October 20, 2003; published electronically May 27, 2005.
http://www.siam.org/journals/sicon/43-6/40777.html

†Dipartimento di Matematica and C.I.R.A.M., Università di Bologna, P.zza Porta S. Donato 5, 40123 Bologna, Italy (ancona@ciram3.ing.unibo.it).

‡Dipartimento di Matematica, Università di Bari, Via E. Orabona 4, 70125 Bari, Italy (giusepc@math.uio.no).

based on the definition of a time-dependent set of *admissible boundary data* that is related to the notion of the Riemann problem.

In the present paper, having in mind applications of Temple systems to problems of oil reservoir simulation, multicomponent chromatography, and traffic flow models, we study the effect of the boundary conditions (1.3)–(1.4) on the solution of (1.1)–(1.2) from the point of view of control theory. Namely, following the same approach adopted in [4, 5, 21] for scalar conservation laws, we fix an initial data $\overline{u} \in \mathbf{L}^\infty([a,b])$ and consider the family of configurations

$$\mathcal{A}(T;\mathcal{U}_a,\mathcal{U}_b) \doteq \{u(T,\cdot);\ u \text{ sol. to (1.1)–(1.4)},\ \widetilde{u}_a \in \mathcal{U}_a,\ \widetilde{u}_b \in \mathcal{U}_b\} \tag{1.5}$$

that can be attained at a given time $T > 0$ by solutions to (1.1)–(1.4), with boundary data $\widetilde{u}_a$, $\widetilde{u}_b$ that vary in prescribed sets $\mathcal{U}_a$, $\mathcal{U}_b \subset \mathbf{L}^\infty(\mathbb{R}^+)$ of *admissible boundary controls.* In the case of scalar, convex conservation laws, it was proved in [4], by using the theory of generalized characteristics [17], that the profiles $w(x)$ which can be attained at a fixed time $T > 0$ are only those for which the map $x \mapsto \frac{f'(w(x))}{x}$ is nonincreasing. Under the assumption that $f'(u) \geq 0$ for all u, and for solutions of the mixed problem (1.1)–(1.4) on the region Ω, this condition is equivalent to the Oleinik-type inequalities

$$D^+ w(x) \leq \frac{f'(w(x))}{(x-a)f''(w(x))} \quad \text{for a.e. } x \in [a,b] \tag{1.6}$$

($D^+ w$ denoting the upper Dini derivative of w). For general $n \times n$ systems, a complete characterization of the attainable set does not seem possible, due to the complexity of repeated wave-front interactions. However, in the particular case of Temple systems, wave interactions can change only the speed of wave-fronts without modifying their amplitudes, due to the special geometric features of such systems. Therefore, the only restriction to boundary controllability is the decay due to genuine nonlinearity. We then consider here a convex, compact set $\Gamma \subset U$ and provide a description of the attainable set

$$\mathcal{A}(T) \doteq \mathcal{A}(T;\mathcal{U}^\infty,\mathcal{U}^\infty), \qquad \mathcal{U}^\infty \doteq \mathbf{L}^\infty([0,T],\Gamma)$$

in terms of certain Oleinik-type conditions. We also establish the compactness of $\mathcal{A}(T)$ in the $\mathbf{L}^1$ topology, as was shown in [4] for the scalar case. These results are useful in applications to calculus of variations and optimal control problems where the cost functional depends on the profile of the solution to (1.1)–(1.4) at a fixed time T. An example is given by a model of two-component chromatography that describes a liquid flowing through a tube packed with solid particles that absorbs (with different rates of adsorption) two interacting chemical substances dissolved in the liquid. In case one is interested in producing the separation of the two substances, the controller acts by varying the concentration of the solutes entering the tube to maximize the concentration of each substance in the liquid phase on opposite sides of the tube.

The paper is organized as follows. Section 2 contains the basic definitions and the statement of the main results, together with a discussion of the two-component chromatography problem. We also provide in this section a review of the existence and well-posedness theory for the mixed problem (1.1)–(1.4), and a description of a front tracking algorithm that will be used throughout the paper. In section 3 we establish some preliminary estimates and a regularity result concerning the global structure of solutions to the mixed problem (1.1)–(1.4) generated by a front tracking algorithm. The proof of the main results is contained in section 4.

2. Preliminaries and statement of the main results.

2.1. Formulation of the problem. Let $f : U \mapsto \mathbb{R}^n$ be the flux function of the strictly hyperbolic system (1.1) defined on a neighborhood of the origin $U \subseteq \mathbb{R}^n$. Denote by $\lambda_1(u) < \cdots < \lambda_n(u)$ the eigenvalues of the Jacobian matrix $Df(u)$, and let $\{r_1(u), \ldots, r_n(u)\}$ be a basis of right eigenvectors of $Df(u)$. By possibly considering a sufficiently small restriction of the domain U, we may assume that the following *uniform* strict hyperbolicity condition holds.

(SH) For every $u, v \in U$, the characteristic speeds at these points satisfy

$$\lambda_i(u) < \lambda_j(v) \quad \forall 1 \le i < j \le n. \tag{2.1}$$

We also assume that there is a fixed set of characteristic lines entering the interior of the strip $[a, b] \times \mathbb{R}^+$ at the boundaries $x = a$, $x = b$, i.e., that for some index $p \in \{1, \ldots, n\}$, there holds

$$\lambda_p(u) < 0 < \lambda_{p+1}(u) \quad \forall u \in U, \tag{2.2}$$

and we let $\lambda^{\min}$, $\lambda^{\max}$ denote the minimum and maximum characteristic speed so that there holds

$$0 < \lambda^{\min} \le |\lambda_i(u)| \le \lambda^{\max} \quad \forall u \in U. \tag{2.3}$$

Moreover, we assume that each ith characteristic field r_i is *genuinely nonlinear* in the sense of Lax [22], and that system (1.1) is of Temple class according to the following definition.

DEFINITION 2.1. *A system of conservation laws is of Temple class if there exists a system of coordinates $w = (w_1, \ldots, w_n)$ consisting of Riemann invariants, and such that the level sets $\{u \in U;\ w_i(u) = \text{constant}\}$ are hyperplanes (see* [28]*).*

By possibly performing a translation of coordinates, it is not restrictive to assume that the Riemann invariants are chosen so that $\partial_i \lambda_i(w) > 0$, $i = 1, \ldots, n$, for all $w = w(u)$, $u \in U$. Throughout the paper, we will often write $w_i(t, x) \doteq w_i(u(t, x))$ to denote the ith Riemann coordinate of a solution $u = u(t, x)$ to (1.1). We recall that, for a Temple class system, Hugoniot curves and rarefaction curves coincide and are straight lines [29]. Moreover, as observed in [3], thanks to the existence of Riemann coordinates one can show that the assumption (SH) implies the invertibility of the map $f : U \mapsto f(U)$.

We next introduce a definition of weak solution to (1.1)–(1.4) which includes an entropy admissibility condition of Oleinik type on the decay of positive waves, to achieve uniqueness. The boundary conditions (1.3)–(1.4) are formulated in terms of the weak trace of the flux $f(u)$ at the the boundaries $x = a$, $x = b$ and are related to the notion of the Riemann problem in the same spirit of [19]. To this purpose, letting $u(t, x) = W(\xi = x/t; u_L, u_R)$, $u_L, u_R \in U$, denote the self-similar solution of the Riemann problem for (1.1) with initial data

$$u(0, x) = \begin{cases} u_L & \text{if } x < 0, \\ u_R & \text{if } x > 0 \end{cases}$$

for any given boundary state $\widetilde{u} \in U$, we define the set of *admissible states at the boundaries*

$$\begin{aligned} \mathcal{V}_a(\widetilde{u}) &\doteq \{W(0+; \widetilde{u}, u_R);\ u_R \in U\}, \\ \mathcal{V}_b(\widetilde{u}) &\doteq \{W(0-; u_L, \widetilde{u});\ u_L \in U\}. \end{aligned} \tag{2.4}$$

DEFINITION 2.2. *A function $u : [0,T] \times [a,b] \mapsto U$ is an entropy weak solution of the initial-boundary value problem (1.1)–(1.4) on $\Omega_T \doteq [0,T] \times [a,b]$ if it is continuous as a function from $]0,T]$ into $\mathbf{L}^1$, and the following properties hold:*

(i) *u is a distributional solution to the Cauchy problem (1.1)–(1.2) on Ω_T in the sense that, for every test function $\phi \in \mathcal{C}_c^1$ with compact support contained in the set $\{(t,x) \in \mathbb{R}^2;\ a < x < b,\ t < T\}$, there holds*

$$\int_0^T \int_a^b (u(t,x) \cdot \phi_t(t,x) + f(u(t,x)) \cdot \phi_x(t,x)) dx\, dt + \int_a^b \overline{u}(x) \cdot \phi(0,x) dx = 0;$$

(ii) *the flux $f(u)$ admits weak* traces at the boundaries $x = a$, $x = b$, i.e., there exist two measurable functions $\Psi_a, \Psi_b : [0,T] \mapsto \mathbb{R}^n$ such that*

$$f(u(\cdot,x)) \xrightarrow[x \to a^+]{*} \Psi_a, \qquad f(u(\cdot,x)) \xrightarrow[x \to b^-]{*} \Psi_b \quad \text{in } \mathbf{L}^\infty([0,T]), \tag{2.5}$$

and the boundary conditions (1.3)–(1.4) are satisfied in the following sense:

$$\Psi_a(t) \in f(\mathcal{V}_a(\widetilde{u}_a(t))), \qquad \Psi_b(t) \in f(\mathcal{V}_b(\widetilde{u}_b(t))) \quad \text{for a.e. } 0 \le t \le T; \tag{2.6}$$

(iii) *u satisfies the following entropy conditions on the decay of positive waves in time and space. There exists some constant $C > 0$, depending only on the system (1.1), so that*

(a) *for any $0 < t \le T$, and for a.e. $a < x < y < b$, there holds*

$$w_i(t,y) - w_i(t,x) \le C \cdot \left\{ \frac{y-x}{t} + \log\left(\frac{y-b}{x-b}\right) \right\} \quad \text{if } i \in \{1,\dots,p\}, \tag{2.7}$$

$$w_i(t,y) - w_i(t,x) \le C \cdot \left\{ \frac{y-x}{t} + \log\left(\frac{y-a}{x-a}\right) \right\} \quad \text{if } i \in \{p+1,\dots,n\}; \tag{2.8}$$

(b) *for a.e. $a < x < b$, and for a.e. $0 < \tau_1 < \tau_2 \le T$, there holds*

$$w_i(\tau_2,x) - w_i(\tau_1,x) \le C \cdot \left\{ \frac{\tau_2-\tau_1}{x-b} + \log\left(\frac{\tau_2}{\tau_1}\right) \right\} \quad \text{if } i \in \{1,\dots,p\}, \tag{2.9}$$

$$w_i(\tau_2,x) - w_i(\tau_1,x) \le C \cdot \left\{ \frac{\tau_2-\tau_1}{x-a} + \log\left(\frac{\tau_2}{\tau_1}\right) \right\} \quad \text{if } i \in \{p+1,\dots,n\}. \tag{2.10}$$

Remark 2.3. The set of admissible flux values at the boundaries $x = a$, $x = b$, can be expressed in Riemann coordinates as

$$\begin{aligned} f(\mathcal{V}_a(\widetilde{u})) &= \{f(u);\ w_i(u) = w_i(\widetilde{u}) \qquad \forall i = p+1,\dots,n\}, \\ f(\mathcal{V}_b(\widetilde{u})) &= \{f(u);\ w_i(u) = w_i(\widetilde{u}) \qquad \forall i = 1,\dots,p\}. \end{aligned} \tag{2.11}$$

Hence, by the invertibility of the map $f : U \mapsto f(U)$, the above boundary conditions (2.6) are equivalent to the set of equalities

$$\begin{aligned} &w_i(f^{-1}(\Psi_a(t))) = w_i(\widetilde{u}_a(t)) \quad \text{for a.e. } 0 \le t \le T,\ di = p+1,\dots,n, \\ &w_i(f^{-1}(\Psi_b(t))) = w_i(\widetilde{u}_b(t)) \quad \text{for a.e. } 0 \le t \le T,\ i = 1,\dots,p. \end{aligned} \tag{2.12}$$

This means that the boundary conditions (2.6) guarantee that, at almost every time $t \in [0,T]$, the solution to the Riemann problem for (1.1), having left and right initial

states $u^L = \widetilde{u}_a(t)$, $u^R = f^{-1}(\Psi_a(t))$, contains only waves with negative speeds, while the solution to the Riemann problem with initial states $u^L = f^{-1}(\Psi_b(t))$, $u^R = \widetilde{u}_b(t)$, contains only waves with positive speeds. Thus, in particular, such solutions do not contain any front entering the domain $[t, +\infty[\times]a, b[$.

In the present paper we regard the boundary data as admissible controls and, in connection with a fixed convex, compact set $\Gamma \subset U$ having the form

$$\Gamma = \{u \in U;\ w_i(u) \in [\alpha_i, \beta_i],\ i = 1, \dots, n\}, \tag{2.13}$$

we study the basic properties of the *attainable set* for (1.1)–(1.2), i.e., of the set

$$\mathcal{A}(T) \doteq \{u(T, \cdot);\ u \text{ is a sol. to (1.1)–(1.4)},\ \widetilde{u}_a, \widetilde{u}_b \in \mathbf{L}^\infty([0, T], \Gamma)\} \tag{2.14}$$

which consists of all profiles that can be attained at a fixed time $T > 0$ by entropy weak solutions of (1.1)–(1.4) (according to Definition 2.2) with a fixed initial data $\overline{u} \in \mathbf{L}^\infty([a, b], \Gamma)$ and boundary data $\widetilde{u}_a$, $\widetilde{u}_b$ that vary in

$$\mathcal{U}_T^\infty \doteq \mathbf{L}^\infty([0, T], \Gamma). \tag{2.15}$$

We will establish a characterization of (2.14) in terms of certain Oleinik type estimates on the decay of positive waves, and we will prove the compactness of (2.14) in the $\mathbf{L}^1$ topology.

2.2. Statements of the main results. For any $\rho > 0$, consider the set of maps

$$K^\rho \doteq \left\{ \varphi \in \mathbf{L}^\infty([a, b], \Gamma);\ \begin{array}{l} \dfrac{w_i(\varphi(y)) - w_i(\varphi(x))}{y - x} \le \dfrac{\rho}{x - a} \left\{ \begin{array}{l} \text{for a.e. } a < x < y < b, \\ \text{if } i \in \{p+1, \dots, n\} \end{array} \right. \\ \dfrac{w_i(\varphi(y)) - w_i(\varphi(x))}{y - x} \le \dfrac{\rho}{b - y} \left\{ \begin{array}{l} \text{for a.e. } a < x < y < b, \\ \text{if } i \in \{1, \dots, p\} \end{array} \right. \end{array} \right\}. \tag{2.16}$$

The inequalities in (2.16) reflect the fact that positive waves entering through the boundaries $x = a$, $x = b$ decay in time. Therefore, their density (expressed in terms of Riemann coordinates) is inversely proportional to their distance from their entrance point on the boundary.

THEOREM 2.4. *Let* (1.1) *be a system of Temple class with all characteristic fields genuinely nonlinear, and assume that the strict hyperbolicity condition* (SH) *is verified. Then, for every fixed* $\overline{\tau} > 0$, *there exists* $\rho = \rho(\overline{\tau}) > 0$ *such that*

$$\mathcal{A}(\tau) \subseteq K^\rho \quad \forall \tau \ge \overline{\tau}. \tag{2.17}$$

Moreover, taking $T \doteq \frac{4(b-a)}{\lambda^{\min}}$, *there exists* $\rho' < \rho(T)$ *such that*

$$K^{\rho'} \subseteq \mathcal{A}(\tau) \quad \forall \tau > T. \tag{2.18}$$

Remark 2.5. Observe that, given $\varphi \in K^\rho$, any map $x \mapsto w_i(\varphi(x))$, $i \in \{1, \dots, n\}$, is essentially bounded and has finite total increasing variation on subsets of $[a, b]$ bounded away from the end points a, b. Hence, any map $x \mapsto w_i(\varphi(x))$, $i \in \{1, \dots, n\}$, also has finite total variation on such sets and, in particular, it admits left and right limits in any point $x \in]a, b[$. Moreover, since an element φ of K^ρ is defined up to $\mathbf{L}^1$

equivalence, we may always assume that there is a right continuous representative of $w_i(\varphi)$, $i \in \{1, \dots, n\}$, that satisfies the inequalities appearing in the definition of K^ρ.

Remark 2.6. It can be easily seen that, no matter how small we choose ρ, it is not possible, in general, to reach at a time $T < \frac{2(b-a)}{\lambda^{\max}}$ any prescribed profile $\varphi \in K^\rho$ with an entropy weak solution u to (1.1)–(1.4) that starts with a fixed initial data $\overline{u}$. This is due to the fact that in this case the domains of determinacy of the line segments $t = 0$, $a \le x \le b$, and $t = T$, $a \le x \le b$ have a nonempty intersection and hence the determination of the possible solution u induced by $u(T, x) = \varphi(x)$, $x \in [a, b]$, and $u(0, x) = \overline{u}(x)$, $x \in [a, b]$ would be, in general, inconsistent. We expect that the same type of noncontrollability result holds for time larger than $\frac{2(b-a)}{\lambda^{\max}}$. This raises the question concerning the identification of a minimum time $T_m < \frac{4(b-a)}{\lambda^{\min}}$ for which the inclusion (2.18) of Theorem 2.4 is verified, although a definite answer to such a problem does not seem possible due to the complexity of the wave-front structure of a solution to the mixed problem (1.1)–(1.4).

THEOREM 2.7. *Under the same assumptions of Theorem* 2.4*, the set* $\mathcal{A}(T)$ *is a compact subset of* $\mathbf{L}^1([a, b], \Gamma)$ *for each* $T > 0$.

2.3. An application: Chromatography of two solutes. Chromatography is a process used by chemists and engineers to separate two (or more) chemical species in a fluid phase by selective adsorption on a solid medium. We consider here the case of a mixture of two interacting solutes S_1 and S_2 dissolved in a liquid with concentrations c_1 and c_2 which passes through the interstices of a solid bed of particles packed in a tube. The solid surface of the filtering bed absorbs different amounts of the two solutes, while it is possible for the particles of each substance to pass from the fluid to the solid phase, and vice versa. The different rates of adsorption of the two solutes causes the less strongly adsorbed solute, say S_2, to move ahead of the more strongly adsorbed one S_1, thus inducing a separation of the two chemical substances. If we make the assumption of local equilibrium in the tube between the liquid and solid phase for each substance, one can express the solid concentrations of the two substances n_1, n_2 as functions of both liquid concentrations. In the case of the Langmuir isotherm equilibrium, the solid concentrations take the form (see [24])

$$n_i \doteq \frac{N_i k_i c_i}{1 + k_1 c_1 + k_2 c_2}, \quad i = 1, 2, \tag{2.19}$$

where k_1, k_2 are constitutive constants depending on the temperature, while N_i denotes the limiting value of n_i (representing the maximum concentration of the solute that can be adsorbed by the solid medium). After performing a suitable transformation of the independent variables, one can express the mass conservation equations for the two species as (see [26])

$$\begin{aligned} [c_1]_x + \left[\frac{\gamma c_1}{1 + c_1 + c_2}\right]_t &= 0, \\ [c_2]_x + \left[\frac{c_2}{1 + c_1 + c_2}\right]_t &= 0 \end{aligned} \tag{2.20}$$

for some constant $\gamma \in]0, 1]$. Notice that the mathematical roles played by the space-like variable x and by the timelike variable t in (2.20) is the opposite of the typical roles played by such variables in most physical hyperbolic systems. Because of the particular nonlinearity relation (2.19) of the Langmuir isotherm, (2.20) enjoy the special geometric properties of Temple systems. Moreover, by direct computations,

one can verify that (2.20) satisfies all the other assumptions made in Theorems 2.4–2.7, namely the following hold (see [26]):

- system (2.20) is strictly hyperbolic if we assume that the state variables c_i take values in some fixed interval $[\overline{c}_i, \overline{d}_i]$ (which corresponds to requiring that the concentrations are nonnegative and do not approach infinity);
- both characteristic speeds of (2.20) are positive and genuinely nonlinear if $\gamma \in]0,1[$ (while, in the case $\gamma = 1$, the first characteristic speed is genuinely nonlinear and the second is linearly degenerate).

A general introduction to the mathematical modeling of chromatography and a detailed analysis of the equilibrium system in the special case of Langmuir isotherm is provided by Rhee, Aris, and Amundson in [25, 26].

We are concerned here with the problem of controlling the feed data c_1^f, c_2^f, i.e., the concentrations of the two solutes S_1, S_2 entering the tube in a time interval $[0,T]$, to maximize the separation of the two substances at time T. Namely, we are interested in maximizing the difference between the distributions of S_1 and S_2 on opposite sides of the tube at a fixed time T. Let $x = 0$ and $x = L$ denote the locations, respectively, of the inlet and outlet of the tube, and suppose that $\overline{c}_1$, $\overline{c}_2$ are the initial distributions in the liquid of S_1 and S_2 at time $t = 0$. Then, consider the functional

$$J(x, c^f) \doteq \int_0^x (c_1(T,\xi) - c_2(T,\xi))d\xi + \int_x^L (c_2(T,\xi) - c_1(T,\xi))d\xi, \tag{2.21}$$

where $(c_1, c_2)(t,x)$ denotes the solution to the mixed problem for (2.20) on the strip $[0,\infty[\times[0,L]$, with initial data $\overline{c} = (\overline{c}_1, \overline{c}_2)$ and boundary data $c^f = (c_1^f, c_2^f)$. Notice that, since both characteristic speeds are positive, the set $\mathcal{V}_L(\widetilde{c})$ of admissible states at the right boundary $x = L$ (defined as in (2.4)) turns out to be equal to the whole space $\mathbb{R}^2$ for every $\widetilde{c} \in \mathbb{R}^2$. Hence, the only significant boundary condition is assigned at the left boundary $x = 0$. Then, assuming that C_i denotes the maximum amount of concentration of the solute S_i that can be introduced in the tube in the time interval $[0,T]$, we are led to study the maximization problem

$$\max_{x\in[0,1], c^f \in \mathcal{U}_T^\infty} J(x, c^f), \tag{2.22}$$

where the set of admissible boundary controls is given by

$$\mathcal{U}_T^\infty \doteq \left\{ \widetilde{c} \in \mathbf{L}^\infty([0,T]) : \widetilde{c}_i(t) \in [\overline{c}_i, \overline{d}_i],\ \int_0^T \widetilde{c}_i(t) \le C_i \right\}. \tag{2.23}$$

With the same arguments we will use to establish Theorem 2.7 one can prove the compactness of the attainable set $\mathcal{A}(T)$ in the case of a system like (2.20) where the mathematical roles played by the x-t variables is reversed w.r.t. (1.1) and in connection with a set of admissible boundary controls that satisfy an additional integral constraint as in (2.23). Then, observing that the map

$$(x, c) \mapsto \int_0^x (c_1(\xi) - c_2(\xi))\, d\xi + \int_x^L (c_2(\xi) - c_1(\xi))\, d\xi \tag{2.24}$$

is continuous as a functional from $[0,L] \times \mathbf{L}^1([0,T])$ to $\mathbb{R}$, we deduce that the maximization problem (2.22) admits a solution.

2.4. Existence and uniqueness of solutions. We describe here a front tracking algorithm that generates approximate solutions to (1.1) on the strip $[a,b]\times\mathbb{R}^+$ continuously depending on the initial and boundary data, which represents a natural extension of [3, 12]. Fix an integer $\nu\geq 1$ and consider the discrete set of points in Γ whose coordinates are integer multiples of $2^{-\nu}$:

$$\Gamma^\nu \doteq \{u\in\Gamma;\ w_i(u)\in 2^{-\nu}\mathbb{Z},\ i=1,\dots,n\}. \tag{2.25}$$

Moreover, consider the domain

$$\mathcal{D}^\nu \doteq \{(u,u',u'');\ u\in\mathbf{L}^\infty([a,b],\Gamma^\nu),\ u',u''\in\mathbf{L}^\infty(\mathbb{R}^+,\Gamma^\nu),\ u,u',u'' \text{ piecew. const.}\}. \tag{2.26}$$

On $\mathcal{D}^\nu$ we now construct a flow map E^ν whose trajectories are front tracking approximate solutions of (1.1). To this end, we first describe how to solve a Riemann problem with left and right initial states $u^L,u^R\in\Gamma^\nu$. In Riemann coordinates, assume that

$$w(u^L)\doteq w^L=(w_1^L,\dots,w_n^L),\qquad w(u^R)\doteq w^R=(w_1^R,\dots,w_n^R).$$

Consider the intermediate states

$$z^0\doteq u^L,\dots,\qquad z^i\doteq u(w_1^R,\dots,w_i^R,w_{i+1}^L,\dots,w_n^L),\dots,\qquad z^n\doteq u^R. \tag{2.27}$$

The solution to the Riemann problem (u^L,u^R) is constructed by piecing together the solutions to the simple Riemann problems (z^{i-1},z^i), $i=1,\dots,n$. If $w_i^R<w_i^L$, the solution of the Riemann problems (z^{i-1},z^i) will contain a single i-shock, connecting the states z^{i-1}, z^i and traveling with the Rankine–Hugoniot speed $\lambda_i(z^{i-1},z^i)$. Here and in what follows, by $\lambda_i(u,u')$ we denote the ith eigenvalue of the averaged matrix

$$A(u,u')\doteq\int_0^1 Df(\theta u+(1-\theta)u')\,d\theta. \tag{2.28}$$

If $w_i^R>w_i^L$, the exact solution of the Riemann problem (z^{i-1},z^i) would contain a centered rarefaction wave. This is approximated by a rarefaction fan as follows. If $w_i^R=w_i^L+p_i2^{-\nu}$ we insert the states

$$z^{i,\ell}\doteq(w_1^R,\dots,w_i^L+\ell 2^{-\nu},w_{i+1}^L,\dots,w_n^L),\qquad \ell=0,\dots,p_i, \tag{2.29}$$

so that $z^{i,0}=z^{i-1}$, $z^{i,p_i}=z^i$. Our front tracking solution will then contain p_i fronts of the ith family, each connecting a couple of states $z^{i,\ell-1}$, $z^{i,\ell}$ and traveling with speed $\lambda_i(z^{i,\ell-1},z^{i,\ell})$.

For any given triple of (piecewise constant) initial and boundary data $(\overline{u},\widetilde{u}_a,\widetilde{u}_b)\in\mathcal{D}^\nu$, the approximate solution $u(t,\cdot)\doteq E_t^\nu(\overline{u},\widetilde{u}_a,\widetilde{u}_b)$ is now constructed as follows. At time $t=0$, for $a<x<b$ we solve the initial Riemann problems determined by the jumps in $\overline{u}$ according to the above procedure, while at $x=a$ we construct the solution to the Riemann problem with left and right initial states $u^L=\widetilde{u}_a(0+)$, $u^R=\overline{u}(a+)$ and take its restriction to the interior of the domain Ω. In the same way, at $x=b$ we take the restriction to the interior of Ω of the solution to the Riemann problem with initial states $u^L=\overline{u}(b-)$, $u^R=\widetilde{u}_b(0+)$. This yields a piecewise constant function with finitely many fronts, traveling with constant speeds. The solution is then prolonged up to the first time t_1 at which one of the following events takes place:

(a) two or more discontinuities interact in the interior of Ω;

(b) one or more discontinuities hit the boundary of Ω;

(c) the boundary data $\widetilde{u}_a$ has a jump;
(d) the boundary data $\widetilde{u}_b$ has a jump.

If case (a) occurs, we then solve the resulting Riemann problems applying again the above procedure, while in cases (b), (c), and (d) we construct the solution to the Riemann problem with left and right initial states $u^L = \widetilde{u}_a(t_1+)$, $u^R = u(t_1, a+)$, or $u^L = u(t_1, b-)$, $u^R = \widetilde{u}_b(t_1+)$ and take its restriction to the interior of the domain Ω. This determines the solution $u(t,\cdot)$ until the time $t_2 > t_1$ where one of the events (a), (b), or (c) again takes place, etc. Notice that at any time where case (b) occurs but (c) or (d) do not take place, no new wave is generated. Therefore, waves entering the domain Ω at the boundaries $x = a$, $x = b$ are produced only by the jumps of the boundary data $\widetilde{u}_a$, $\widetilde{u}_b$.

As in [3, 12], one checks that the approximate solution u constructed with this algorithm is well defined for all times $t \geq 0$. Indeed, the following properties hold.

- The total variation of $u(t,\cdot)$, measured w.r.t. the Riemann coordinates $w_1(t,\cdot), \dots, w_n(t,\cdot)$ is nonincreasing in time.
- The number of wave-fronts in $u(t,\cdot)$ is nonincreasing at each interaction. Hence, the total number of wave-fronts in $u(t,\cdot)$ remains finite.

It is then possible to define a flow map

$$\mathbf{p} \mapsto E^\nu_t \mathbf{p}, \quad \mathbf{p} \doteq (\overline{u}, \widetilde{u}_a, \widetilde{u}_b) \in \mathcal{D}^\nu, \quad t \geq 0, \tag{2.30}$$

of approximate solutions of (1.1). By construction, each trajectory $t \mapsto E^\nu_t \mathbf{p}$ is a weak solution of (1.1) (because all fronts of $u(t,\cdot) \doteq E^\nu_t \mathbf{p}$ satisfy the Rankine–Hugoniot conditions) but may contain discontinuities that do not satisfy the usual Lax stability conditions (due to the presence of rarefaction fronts). On the other hand, one can verify as in [3, Lemma 4.4] that, due to genuine nonlinearity, the amount of positive waves in $u(t,\cdot)$, measured w.r.t. the Riemann coordinates $w_1(t,\cdot), \dots, w_n(t,\cdot)$, decays in time and space. Hence, for a.e. $a < x < y < b$, one obtains the Oleinik-type estimates

$$\begin{aligned} w_i(t,y) - w_i(t,x) &\leq C \cdot \left\{ \frac{y-x}{t} + \log\left(\frac{y-b}{x-b}\right) \right\} + N_\nu 2^{-\nu} \quad \text{if } i \in \{1,\dots,p\}, \\ w_i(t,y) - w_i(t,x) &\leq C \cdot \left\{ \frac{y-x}{t} + \log\left(\frac{y-a}{x-a}\right) \right\} + N_\nu 2^{-\nu} \quad \text{if } i \in \{p+1,\dots,n\}, \end{aligned} \tag{2.31}$$

where N_ν denotes the maximum number of shocks of each family present in the initial data $\overline{u}$, and in the boundary data $\widetilde{u}_a$, $\widetilde{u}_b$. Similarly, one can check that along the x-sections, for a.e. $0 < \tau_1 < \tau_2$, there holds

$$\begin{aligned} w_i(\tau_2,x) - w_i(\tau_1,x) &\leq C \cdot \left\{ \frac{\tau_2-\tau_1}{x-b} + \log\left(\frac{\tau_2}{\tau_1}\right) \right\} + N_\nu 2^{-\nu} \quad \text{if } i \in \{1,\dots,p\}, \\ w_i(\tau_2,x) - w_i(\tau_1,x) &\leq C \cdot \left\{ \frac{\tau_2-\tau_1}{x-a} + \log\left(\frac{\tau_2}{\tau_1}\right) \right\} + N_\nu 2^{-\nu} \quad \text{if } i \in \{p+1,\dots,n\}. \end{aligned} \tag{2.32}$$

Remark 2.8. Observe that if $u(t,x)$ is a front tracking solution of the Cauchy problem for (1.1) (with initial data $\overline{u}(x) \doteq u(0,x)$) constructed by the algorithm in [12] on the upper half plane $\mathbb{R}^+ \times \mathbb{R}$, then the restriction of $u(t,\cdot)$ to the interval $[a,b]$ coincides with the front tracking solution $E^\nu_t(\overline{u}, \widetilde{u}_a, \widetilde{u}_b)$ of the mixed problem for (1.1), with boundary data $\widetilde{u}_a(t) \doteq u(t,a)$, $\widetilde{u}_b(t) \doteq u(t,b)$.

As $\nu \to \infty$, the domains $\mathcal{D}^\nu$ become dense in

$$\mathcal{D} \doteq \{(\overline{u}, \widetilde{u}_a \widetilde{u}_b);\ \overline{u} \in \mathbf{L}^\infty([a,b],\Gamma),\ \widetilde{u}_a, \widetilde{u}_b \in \mathbf{L}^\infty(\mathbb{R}^+,\Gamma)\}. \tag{2.33}$$

Thus, following the same technique adopted in [3], one can define a flow map E_t on $\mathcal{D}$ as a suitable limit of the flows E_t^ν in (2.30) that depends Lipschitz continuously on the initial and boundary data. Namely, the following holds.

THEOREM 2.9. *Let* (1.1) *be a system of Temple class with all characteristic fields genuinely nonlinear, and assume that the strict hyperbolicity condition* (SH) *holds. Then, there exists a continuous map*

$$(t, \overline{u}, \widetilde{u}_a, \widetilde{u}_b) \mapsto E_t(\overline{u}, \widetilde{u}_a, \widetilde{u}_b), \quad t \geq 0,\ (\overline{u}, \widetilde{u}_a, \widetilde{u}_b) \in \mathcal{D}, \tag{2.34}$$

and some constant $C > 0$ depending only on the system (1.1) *and on the domain Γ, so that, for every fixed $0 < \delta < (b-a)/2$ and for all $\mathbf{p}_1 \doteq (\overline{u}, \widetilde{u}_a, \widetilde{u}_b)$, $\mathbf{p}_2 \doteq (\overline{v}, \widetilde{v}_a, \widetilde{v}_b) \in \mathcal{D}$, letting $L_t \doteq L_t(\delta) = C(1 + \log(t/\delta))$, there holds*

$$\begin{aligned} &\|E_t\mathbf{p}_1 - E_t\mathbf{p}_2\|_{\mathbf{L}^1([a+\delta, b-\delta])} \\ &\quad \leq L_t \cdot \Big\{ \|\overline{u} - \overline{v}\|_{\mathbf{L}^1([a,b])} + \|\widetilde{u}_a - \widetilde{v}_a\|_{\mathbf{L}^1([0,t])} + \|\widetilde{u}_b - \widetilde{v}_b\|_{\mathbf{L}^1([0,t])} \Big\} \end{aligned} \tag{2.35}$$

for all $t \geq \delta$. Moreover, the map $(t,x) \mapsto E_t(\overline{u}, \widetilde{u}_a, \widetilde{u}_b)(x)$ yields an entropy weak solution (in the sense of Definition 2.2*) to the initial-boundary value problem* (1.1)–(1.4) *on Ω that admits strong $\mathbf{L}^1$ traces at the boundaries $x = a$ and $x = b$, i.e., there exist two measurable maps $\psi_a, \psi_b : \mathbb{R}^+ \mapsto U$ such that*

$$\begin{aligned} \lim_{x \to a^+} \int_0^\tau |E_t(\overline{u}, \widetilde{u}_a, \widetilde{u}_b)(x) - \psi_a(t)|\, dt = 0, \\ \lim_{x \to b^-} \int_0^\tau |E_t(\overline{u}, \widetilde{u}_a, \widetilde{u}_b)(x) - \psi_b(t)|\, dt = 0 \end{aligned} \tag{2.36}$$

$\forall \tau \geq 0$.

The proof of Theorem 2.9 can be obtained with arguments entirely similar to those used to establish [3, Theorem 2.1], where a continuous flow of solutions to (1.1) is constructed in the case of a mixed problem on the quarter of plane $\{(t,x) \in \mathbb{R}^2;\ t \geq 0,\ x \geq 0\}$, with a single boundary at $x = 0$.

Concerning uniqueness, with the same arguments in [3] one obtains the following result which is the extension of [3, Theorem 2.2] to the present case of a domain Ω with two boundaries at $x = a$ and at $x = b$.

THEOREM 2.10. *Let* (1.1) *be a system of Temple class satisfying the same assumptions as in Theorem* 2.9. *Let $u = u(t,x)$ be an entropy weak solution to the mixed problem* (1.1)–(1.4) *on the region $\Omega_T \doteq [0,T] \times [a,b]$ (in the sense of Definition* 2.2*). Assume that the following conditions hold.*

(i) *The map $(t,x) \to (u(t,\cdot), u(\cdot,x))$ takes values within the domain*

$$\mathcal{D}_T \doteq \{(\overline{u}, \widetilde{u}_a, \widetilde{u}_b);\ \overline{u} \in \mathbf{L}^\infty([a,b],\Gamma),\ \widetilde{u}_a, \widetilde{u}_b \in \mathbf{L}^\infty([0,T],\Gamma)\}. \tag{2.37}$$

(ii) *There holds*

$$\operatorname*{ess\,sup}_{t \to 0^+} \int_a^b |u(t,x) - \overline{u}(x)|\, dx = 0. \tag{2.38}$$

(iii) *There holds*

$$\operatorname*{ess\,sup}_{x\to a^+}\int_0^T |w_i(u(t,x))-w_i(\widetilde{u}_a(t))|\,dt=0 \quad \forall i=p+1,\dots,n, \tag{2.39}$$

$$\operatorname*{ess\,sup}_{x\to b^-}\int_0^T |w_i(u(t,x))-w_i(\widetilde{u}_b(t))|\,dt=0 \quad \forall i=1,\dots,p. \tag{2.40}$$

Then, u coincides with the corresponding trajectory of the flow map E_t provided by Theorem 2.9*, namely one has*

$$u(t,\cdot)=E_t(\overline{u},\widetilde{u}_a,\widetilde{u}_b)(\cdot) \quad \forall 0\le t\le T. \tag{2.41}$$

The next result shows that the conditions (2.38)–(2.40) are certainly satisfied by entropy weak solutions to the mixed problem (1.1)–(1.4) obtained as the limit of front tracking approximations.

THEOREM 2.11. *Let* (1.1) *be a system of Temple class satisfying the same assumptions as in Theorem* 2.9*. Consider a sequence $u^\nu(t,\cdot):[a,b]\mapsto\Gamma^\nu$ of wave-front tracking approximate solutions of the mixed problem for* (1.1) *(constructed with the above algorithm) that converges in $\mathbf{L}^1$, as $\nu\to\infty$, to some function $u(t,\cdot):[a,b]\mapsto\Gamma$, for every $t\in[0,T]$, and assume that the corresponding sequences of boundary data $\widetilde{u}_a^\nu$, $\widetilde{u}_b^\nu$ converge in $\mathbf{L}^1$ to $\widetilde{u}_a\doteq u(\cdot,a),\widetilde{u}_b\doteq u(\cdot,b)$. Then, there exist the right limit at $x=a$ and the left limit at $x=b$ of the map $x\to u(t,x)$ for every $t\in[0,T]$, and the right limit at $t=0$ of the map $t\to u(t,x)$ for every $x\in[a,b]$. Moreover, there is a countable set $\mathcal{N}\subset\mathbb{R}$ such that $u(t,a)=u(t,a+),u(t,b)=u(t,b-)$ for all $t\in[0,T]\setminus\mathcal{N}$, and $u(0,x)=u(0+,x)$ for all $x\in[a,b]\setminus\mathcal{N}$, and setting $\overline{u}\doteq u(0,\cdot)$, there holds* (2.41)*.*

Remark 2.12. It was shown in [3, Lemma 2.1] that an alternative way to prove the essential limits (2.39)–(2.40) is to employ the distributional entropy inequalities associated with the "boundary entropy pairs" for (1.1), introduced by Chen and Frid in [14, 15].

In order to prove Theorem 2.11, we will show in the next section that, for Temple systems, solutions of the mixed problem (1.1)–(1.4) with possibly unbounded variation enjoy the same regularity property (of being continuous outside a countable number of Lipschitz curves) possessed by solutions with small total variation of a general system, thus extending the regularity results obtained under the smallness assumption of the total variation by DiPerna [18] and Liu [23] (for solutions constructed by the Glimm scheme) and by Bressan and LeFloch [13] (for solutions generated by a front tracking algorithm).

PROPOSITION 2.13. *In the same setting as Theorem* 2.11*, consider a sequence $u^\nu(t,\cdot):[a,b]\mapsto\Gamma^\nu$ of wave-front tracking approximate solutions of the mixed problem for* (1.1) *(constructed with the above algorithm) that converges in $\mathbf{L}^1$, as $\nu\to\infty$, to some function $u(t,\cdot):[a,b]\mapsto\Gamma$, for every $t\in[0,T]$. Then, there exist a countable set of interaction points $\Theta\doteq\{(\tau_l,x_l);\ l\in\mathbb{N}\}\subset\Omega_T\doteq[0,T]\times[a,b]$, and a countable family of Lipschitz continuous shock curves $\Upsilon\doteq\{x=y_m(t);\ t\in]r_m,s_m[,\ m\in\mathbb{N}\}$, such that the following hold.*

(i) *For each $m\in\mathbb{N}$, and for any $\tau\in]r_m,s_m[$ with $(\tau,y_m(\tau))\notin\Theta$, there exist the derivative $\dot{y}_m(\tau)$ and the left and right limits*

$$\lim_{(s,y)\to(\tau,y_m(\tau)),\,y<y_m(\tau)} u(s,y)\doteq u^-, \qquad \lim_{(s,y)\to(\tau,y_m(\tau)),\,y>y_m(\tau)} u(s,y)\doteq u^+. \tag{2.42}$$

Moreover, these limits satisfy the Rankine–Hugoniot relations

$$\dot{y}_m(\tau)\cdot(u^+-u^-)=f(u^+)-f(u^-) \tag{2.43}$$

and for some $i\in\{1,\dots,n\}$ there hold the Lax entropy inequalities

$$\lambda_i(u^+)<\dot{y}_m(t)<\lambda_i(u^-). \tag{2.44}$$

(ii) *The map u is continuous outside the set $\Theta\cup\Upsilon$.*

3. Preliminary results. In this section we first provide some estimates on the distance between two rarefaction fronts of a front tracking solution (constructed by the algorithm described in section 2.4) similar to [12, Lemma 4], [8, Proposition 4.5]. We next show how to approximate the profile $u(t,\cdot)$ of a solution of the mixed problem (1.1)–(1.4), with a function taking values in the discrete set Γ^ν defined at (2.25), which enjoys the same type of estimates on the positive waves as $u(t,\cdot)$. We conclude the section establishing the regularity result stated in Proposition 2.13 on the global structure of solutions to the mixed problem for (1.1), which in turn yields Theorem 2.11.

LEMMA 3.1. *There exists some constant $C_1>0$ depending only on the system* (1.1) *such that the following holds. Consider a front tracking solution $u(t,x)$ with values in Γ^ν, constructed by the algorithm of section* 2.4 *on the region $[\tau,\tau']\times[a,b]$. Then, given any two adjacent rarefaction fronts of u located at $x(t)\le y(t)$, $t\in[\tau,\tau']$, and belonging to the same family, there holds*

$$|y(\tau')-x(\tau')|\le|y(\tau)-x(\tau)|+C_1(\tau'-\tau)2^{-\nu}. \tag{3.1}$$

Proof. Consider two adjacent rarefaction fronts of the kth family $x(t)\le y(t)$, $t\in[\tau,\tau']$, and let $\tau_1<\cdots<\tau_N$ be the interaction times of $x(t)$ in the interval $[\tau,\tau']$. Set $\tau_0\doteq\tau$, $\tau_{N+1}\doteq\tau'$, and fix $\alpha\in\{0,\dots,N\}$. Let $t\to z(t;s,x)$ be the characteristic curve of the kth family starting at (s,x), i.e., the solution to the ODE

$$\dot{z}=\lambda_k(u(t,z)),\qquad z(s;s,x)=x.$$

Notice that, although the above ODE has a discontinuous right-hand side (because of the discontinuities in the front tracking solution u), its solution $z(\cdot;s,x)$ is unique and depends Lipschitz continuously on the initial data x since it crosses only a finite number of jumps (see [10]). Choose $t_0<t_1<\tau_{\alpha+1}$ so that the characteristic curve $z(\cdot;t_0,x(t_0))$ does not cross any wave-front of the other families in the interval $[t_0,t_1]$, and then, by induction, define a sequence of times $\{t_i\}_{i\in\mathbb{Z}}\subset]\tau_\alpha,\tau_{\alpha+1}[$ so that

$$\begin{gathered}\tau_\alpha<t_{-i-1}<t_{-i}\le t_0\le t_i<t_{i+1}<\tau_{\alpha+1},\quad i\in\mathbb{N},\\ \lim_{i\to-\infty}t_i=\tau_\alpha,\qquad\lim_{i\to+\infty}t_i=\tau_{\alpha+1},\end{gathered} \tag{3.2}$$

with the properties that the characteristic curve of the kth family starting at $(t_i,x(t_i))$, does not cross any wave-front of the other families in the interval $[t_i,t_{i+1}]$, for each $i\in\mathbb{Z}$. Thus, setting

$$u_i^+\doteq u(t_i,x(t_i)+),\qquad u_i^-\doteq u(t_i,x(t_i)-)$$

and observing that, by construction, one has $|w(u_i^+)-w(u_i^-)|<2^{-\nu}$, we derive

$$\begin{aligned}|z(t_{i+1};t_i,x(t_i))-x(t_{i+1})|&\le(t_{i+1}-t_i)\cdot|\lambda_k(u_i^+)-\lambda_k(u_i^+,u_i^-)|\\&\le c\cdot(t_{i+1}-t_i)\cdot|w(u_i^+)-w(u_i^-)|\\&\le c\cdot(t_{i+1}-t_i)\cdot2^{-\nu}\end{aligned} \tag{3.3}$$

for some constant $c > 0$ depending only on the system. Relying on (3.3), and since $z(\tau'; t_{i+1}, x)$ depends Lipschitz continuously on the initial data x, we deduce that there exists some other constant $c' > 0$, depending only on the system and on the set Γ, so that there holds

$$\begin{aligned} |z(\tau'; t_i, x(t_i)) - z(\tau'; t_{i+1}, x(t_{i+1}))| &\leq c' \cdot |z(t_{i+1}; t_i, x(t_i)) - x(t_{i+1})| \\ &\leq c' \cdot c \cdot (t_{i+1} - t_i) \cdot 2^{-\nu} \end{aligned} \tag{3.4}$$

for any $i \in \mathbb{Z}$. Thus, by (3.2) and thanks to (3.4), we obtain

$$\begin{aligned} |z(\tau'; \tau_\alpha, x(\tau_\alpha)) - z(\tau'; \tau_{\alpha+1}, x(\tau_\alpha))| &\leq \sum_{i \in \mathbb{Z}} |z(\tau'; t_i, x(t_i)) - z(\tau'; t_{i+1}, x(t_{i+1}))| \\ &\leq c' \cdot c \cdot (\tau_{\alpha+1} - \tau_\alpha) \cdot 2^{-\nu}. \end{aligned} \tag{3.5}$$

Repeating this computation for every interval $]\tau_\alpha, \tau_{\alpha+1}[$, $\alpha \in \{0, \dots, N\}$, we get

$$\begin{aligned} |z(\tau'; \tau, x(\tau)) - x(\tau')| &\leq \sum_{\alpha=0}^{N} |z(\tau'; \tau_\alpha, x(\tau_\alpha)) - z(\tau'; \tau_{\alpha+1}, x(\tau_\alpha))| \\ &\leq c' \cdot c \cdot (\tau' - \tau) \cdot 2^{-\nu}. \end{aligned} \tag{3.6}$$

Clearly, one obtains the same type of estimate as (3.6) for the other rarefaction front $y(t)$, i.e., there holds

$$|z(\tau'; \tau, y(\tau)) - y(\tau')| \leq c' \cdot c \cdot (\tau' - \tau) \cdot 2^{-\nu}. \tag{3.7}$$

On the other hand, by (2.3), we have

$$|z(\tau'; \tau, x(\tau)) - z(\tau'; \tau, y(\tau))| \leq |x(\tau) - y(\tau)| + 2\lambda^{\max} \cdot (\tau' - \tau). \tag{3.8}$$

Thus, (3.6)–(3.8) together yield (3.1), concluding the proof. □

In the following, in connection with any (right continuous) piecewise constant map $\psi : [a, b] \mapsto 2^{-\nu}\mathbb{Z}$, we will let $\pi(\psi) = \{x_0 = a < x_1 < \cdots < x_{\overline{\ell}} = b\}$ denote the partition of $[a, b]$ induced by ψ, in the sense that $\psi(x)$ is constant on every interval $[x_\ell, x_{\ell+1}[$, $0 \leq \ell < \overline{\ell}$. Then, given $\rho > 0$, for any $\nu \geq 1$, consider the set of piecewise constant maps

$$K^\rho_\nu \doteq \left\{ \varphi : [a, b] \mapsto \Gamma^\nu; \quad \begin{aligned} &\frac{w_i(\varphi(x_k)) - w_i(\varphi(x_h))}{x_k - x_h} \leq \frac{5\rho}{x_h - a} \begin{cases} \text{for} & a < x_h < x_k < b, \\ & x_h, x_k \in \pi(w_i \circ \varphi), \\ \text{if } i \in \{p+1, \dots, n\} \end{cases} \\ &\frac{w_i(\varphi(x_k)) - w_i(\varphi(x_h))}{x_k - x_h} \leq \frac{5\rho}{b - x_k} \begin{cases} \text{for} & a < x_h < x_k < b, \\ & x_h, x_k \in \pi(w_i \circ \varphi), \\ \text{if } i \in \{1, \dots, p\} \end{cases} \end{aligned} \right\}. \tag{3.9}$$

The next lemma shows that we can approximate in $\mathbf{L}^1$ any map $\varphi \in K^\rho$ with a piecewise constant function $\varphi_\nu \in K^\rho_\nu$.

LEMMA 3.2. *For any given $\varphi \in K^\rho$, there exists a sequence of right continuous maps $\varphi_\nu \in K^\rho_\nu$, $\nu \geq 1$, such that*

(a) *for every $i \in \{1, \dots, n\}$, and for any $x_h \in \pi(w_i \circ \varphi_\nu)$, there holds*

$$w_i(\varphi_\nu(x_{h+1})) > w_i(\varphi_\nu(x_h)) \Longrightarrow w_i(\varphi(x_{h+1})) = w_i(\varphi(x_h)) + 2^{-\nu}; \tag{3.10}$$

(b) *there holds*

$$\varphi_\nu \to \varphi \qquad \text{in } \mathbf{L}^1([a,b]). \tag{3.11}$$

Proof. (1) First observe that, by Remark 2.5, any map $x \mapsto w_i(\varphi(x))$, $i \in \{1,\dots,n\}$, has finite total variation on $[a+\varepsilon, b-\varepsilon]$, $\varepsilon > 0$. Hence, we may assume that $w_i(\varphi(\cdot))$ admits left and right limits in any point $x \in]a,b[$ and that $w_i(\varphi(x)) = w_i(\varphi(x^+)) \doteq \lim_{\xi\to x^+} w_i(\varphi(\xi))$ for all $i \in \{1,\dots,n\}$. Let $\{y_{i,m};\ m \in \mathbb{N}\}$ be the countable set of discontinuities of $w_i(\varphi(\cdot))$, $i \in \{1,\dots,n\}$. Then, we can find a partition $\xi^1_{i,m} = y_{i,m} < \xi^2_{i,m} < \cdots < \xi^{\ell_{i,m}}_{i,m} = y_{i,m'}$ of each interval $[y_{i,m}, y_{i,m'}[$ where $x \mapsto w_i(\varphi(x))$ is continuous, so that

(i) for every $1 < \ell < \ell_{i,m}$ there holds

$$w_i(\varphi(\xi^\ell_{i,m})) \in 2^{-\nu}\mathbb{Z}; \tag{3.12}$$

(ii) for every $1 \le \ell < \ell_{i,m}$ one has

$$|w_i(\varphi(x)) - w_i(\varphi(\xi^\ell_{i,m}))| \le 2^{-\nu} \quad \forall x \in [\xi^\ell_{i,m}, \xi^{\ell+1}_{i,m}[. \tag{3.13}$$

Notice that the Oleinik-type conditions stated in the definition of K^ρ imply that, at any discontinuity point $y_{i,m}$ of $w_i(\varphi(\cdot))$, one has

$$\lim_{\xi\to y^-_{i,m}} w_i(\varphi(\xi)) > w_i(\varphi(y_{i,m})). \tag{3.14}$$

(2) Let $\varphi_\nu : [a,b] \mapsto \Gamma^\nu$ be the piecewise constant, right continuous map defined by setting, for every $i \in \{1,\dots,n\}$ and for any interval $[y_{i,m}, y_{i,m'}[$, where $w_i(\varphi(\cdot))$ is continuous,

$$w_i(\varphi_\nu(x)) \doteq \begin{cases} 2^{-\nu}\lfloor 2^\nu w_i(\varphi(\xi^1_{i,m}))\rfloor & \text{if } \begin{cases} x \in [\xi^1_{i,m}, \xi^2_{i,m}[, \text{ and} \\ w_i(\varphi_\nu(\xi^1_{i,m})) \\ \le 2^{-\nu}(\lfloor 2^\nu w_i(\varphi(\xi^1_{i,m}))\rfloor + 2^{-1}), \end{cases} \\ 2^{-\nu}(\lfloor 2^\nu w_i(\varphi(\xi^1_{i,m}))\rfloor + 1) & \text{if } \begin{cases} x \in [\xi^1_{i,m}, \xi^2_{i,m}[, \text{ and} \\ w_i(\varphi_\nu(\xi^1_{i,m})) \\ > 2^{-\nu}(\lfloor 2^\nu w_i(\varphi(\xi^1_{i,m}))\rfloor + 2^{-1}), \end{cases} \\ w_i(\varphi(\xi^\ell_{i,m})) & \text{if} \quad x \in [\xi^\ell_{i,m}, \xi^{\ell+1}_{i,m}[, \quad 1 < \ell < \ell_{i,m}, \end{cases} \tag{3.15}$$

where $\lfloor\cdot\rfloor$ denotes the integer part. Notice that, by construction and because of (3.12)–(3.14), the map $\varphi_\nu : [a,b] \mapsto \Gamma^\nu$ enjoys the following property:

$$\left.\begin{array}{l} w_i(\varphi_\nu(x_k)) > w_i(\varphi_\nu(x_h)) \\ x_h < x_k \in \pi(w_i \circ \varphi_\nu) \end{array}\right\} \Longrightarrow w_i(\varphi(x_k)) > w_i(\varphi(x_h)) + 2^{-(\nu+1)}. \tag{3.16}$$

Therefore, since $\varphi \in K^\rho$, relying on (3.13), (3.16), we deduce that for every $w_i(\varphi_\nu(\cdot))$, $i \in \{p+1,\dots,n\}$, and for any $x_h < x_k \in \pi(w_i \circ \varphi_\nu)$ such that $w_i(\varphi_\nu(x_k)) > w_i(\varphi_\nu(x_h))$, there holds

$$\begin{aligned} \frac{w_i(\varphi_\nu(x_k)) - w_i(\varphi_\nu(x_h))}{x_k - x_h} &\le \frac{w_i(\varphi(x_k)) - w_i(\varphi(x_h)) + 2^{-(\nu-1)}}{x_k - x_h} \\ &\le \frac{5(w_i(\varphi(x_k)) - w_i(\varphi(x_h)))}{x_k - x_h} \\ &\le \frac{5\rho}{x_h - a}. \end{aligned} \tag{3.17}$$

Clearly, with the same computations we can show that, for every $w_i(\varphi_\nu(\cdot))$, $i \in \{1,\dots,p\}$, and for any $x_h < x_k \in \pi(w_i \circ \varphi_\nu)$, there holds

$$\frac{w_i(\varphi_\nu(x_k)) - w_i(\varphi_\nu(x_h))}{x_k - x_h} \le \frac{5\rho}{b - x_k}. \tag{3.18}$$

The estimates (3.17)–(3.18), together, imply that $\varphi_\nu \in K^\rho_\nu$, while (3.13) yields (3.11). On the other hand, observe that, by construction and because of (3.14), the map φ_ν satisfies condition (3.10), which completes the proof of the lemma. □

We now provide a further estimate on the distance between two rarefaction fronts of a front tracking solution that, at a fixed time τ, attains a profile belonging to the set (3.9).

LEMMA 3.3. *Consider a front tracking solution $u(t,x)$ with values in Γ^ν, $\nu \ge 1$, constructed by the algorithm of section* 2.4 *on the region $[\tau,\tau'] \times [a,b]$. Assume that $u(\tau',\cdot)$ is right-continuous, verifies condition* (a) *of Lemma* 3.2*, and satisfies*

$$u(\tau',\cdot) \in K^{\rho'}_\nu, \qquad \rho' \doteq \frac{\lambda^{\min}}{6C_1}, \tag{3.19}$$

where $\lambda^{\min}$, C_1 are the minimum speed in (2.3) *and the constant of Lemma* 3.1*. Then, given any two adjacent rarefaction fronts of u located at $x(t) \le y(t)$, $t \in [\tau,\tau']$, and belonging to the same family, there holds*

$$x(\tau) < y(\tau). \tag{3.20}$$

Proof. To fix the ideas, assume that $x(t) \le y(t)$ are the locations of two adjacent rarefaction fronts of the $k \in \{p+1,\dots,n\}$th family and hence, by (2.2), have positive speeds. Observe that, by condition (a) of Lemma 3.2, one has

$$w_k(u(\tau', y(\tau'))) - w_k(u(\tau', x(\tau'))) = 2^{-\nu}. \tag{3.21}$$

Moreover, since u is a front tracking solution constructed by the algorithm of section 2.4 on the region $[\tau,\tau'] \times [a,b]$, we can apply Lemma 3.1. Thus, using (2.3), (3.1), and (3.21), and recalling the definition (3.9) of $K^{\rho'}_\nu$, we deduce

$$\begin{aligned} y(\tau') - x(\tau') &\le y(\tau) - x(\tau) + C_1(\tau' - \tau)2^{-\nu} \\ &\le y(\tau) - x(\tau) + C_1 \frac{x(\tau') - x(\tau)}{\lambda^{\min}} \cdot (w_k(\varphi_\nu(y(\tau'))) - w_k(\varphi_\nu(x(\tau')))) \\ &\le y(\tau) - x(\tau) + C_1 \frac{5\rho'}{\lambda^{\min}} \cdot (y(\tau') - x(\tau')) \end{aligned}$$

which, because of (3.19), implies

$$y(\tau) - x(\tau) \ge \left(1 - C_1 \frac{5\rho'}{\lambda^{\min}}\right) \cdot (y(\tau') - x(\tau')) > 0,$$

proving (3.20). □

We next derive a regularity property enjoyed by general solutions of Temple systems with boundary variation defined as a limit of front tracking approximations, which allows us to establish Proposition 2.13. This is an extension of the regularity results obtained in [18, 23, 13] for the solution with small total variation of general systems. The arguments of the proof are quite similar to the corresponding result

in [13], but we will repeat some of them for completeness, referring to [13] (see also [9, Theorem 10.4]) for further details.

LEMMA 3.4. *Let* (1.1) *be a system of Temple class satisfying the same assumptions as in Theorem* 2.9. *Consider a sequence* $u^\nu(t,\cdot) : [c,d] \mapsto \Gamma^\nu$, $t \in [r,s]$, *of front tracking approximate solutions of the mixed problem for* (1.1) *(constructed by the algorithm of section* 2.4*) that converges in* $\mathbf{L}^1$, *as* $\nu \to \infty$, *to some function* $u(t,\cdot) : [c,d] \mapsto \Gamma$, *for every* $t \in [r,s] \subset \mathbb{R}^+$. *Assume that*

$$Tot.Var.(u^\nu(t,\cdot)) \leq M, \quad Tot.Var.(u^\nu(\cdot,x)) \leq M \quad \forall t,x,\nu, \tag{3.22}$$

for some constant $M > 0$, *here and throughout the following,* $Tot.Var.(w)$ *denotes the total variation of the function* w. *Then, there exist a countable set of interaction points* $\Theta \doteq \{(\tau_l, x_l);\ l \in \mathbb{N}\} \subset D \doteq [r,s] \times [c,d]$, *and a countable family of Lipschitz continuous shock curves* $\Upsilon \doteq \{x = y_m(t);\ t \in]r_m, s_m[,\ m \in \mathbb{N}\}$, *such that the following hold.*

(i) *For each* $m \in \mathbb{N}$, *and for any* $\tau \in]r_m, s_m[$ *with* $(\tau, y_m(\tau)) \notin \Theta$, *there exist the left and right limits* (2.42) *of* u *at* $(\tau, y_m(\tau))$ *and the shock speed* $\dot{y}_m(\tau)$. *Moreover, these limits satisfy the Rankine–Hugoniot relations* (2.43) *and the Lax entropy inequality* (2.44), *for some* $i \in \{1,\dots,n\}$.

(ii) *The map* u *is continuous outside the set* $\Theta \cup \Upsilon$.

Proof. 1. To establish (i) we need to recall some technical tools introduced in [13] (see also [9, Theorem 10.4]). For every front tracking solution u^ν, we define the *interaction and cancellation measure* μ^{IC}_ν that is a positive, purely atomic measure on D, concentrated on the set of points P where two or more wave-fronts of u^ν interact. Namely, if the incoming fronts at P have size $\sigma_1,\dots,\sigma_\ell$ (w.r.t. the Riemann coordinates) and belong to the families $i_1,\dots,i_\ell$, respectively, we set

$$\mu^{IC}_\nu(P) \doteq \sum_{\alpha,\beta} |\sigma_\alpha \sigma_\beta| + \sum_i \left(\sum_{\{i_\alpha;\, i_\alpha = i\}} |\sigma_\alpha| - \left| \sum_{\{i_\alpha;\, i_\alpha = i\}} \sigma_\alpha \right| \right). \tag{3.23}$$

Since μ^{IC}_ν have a uniformly bounded total mass, by possibly taking a subsequence we can assume the weak convergence

$$\mu^{IC}_\nu \rightharpoonup \mu^{IC} \tag{3.24}$$

for some positive, purely atomic measure μ^{IC} on D. Call Θ the countable set of atoms of μ^{IC}, i.e., set

$$\Theta \doteq \left\{P \in D;\ \mu^{IC}(P) > 0\right\}.$$

For every approximate solution u^ν taking values in Γ^ν, $\nu \geq 1$, and for any fixed $\varepsilon \geq 2^{-\nu}$, by an *$\varepsilon$-shock front of the ith family* in u^ν we mean a polygonal line in D, with nodes $(\tau_0, x_0), \dots, (\tau_N, x_N)$, having the following properties.

(I) The nodes (τ_h, x_h) are interaction points or lie on the boundary of D, and the sequence of times is increasing $\tau_0 < \tau_1 < \cdots < \tau_N$.

(II) Along each segment joining (τ_{h-1}, x_{h-1}) with (τ_h, x_h), the function u^ν has an i-shock with strength $|\sigma_h| \geq \varepsilon$.

(III) For $h < N$, if two (or more) incoming i-shocks of strength $\geq \varepsilon$ interact at the node (τ_h, x_h), then the shock coming from (τ_{h-1}, x_{h-1}) has the larger speed, i.e., is the one coming from the left.

An ε-shock front, which is maximal with respect to the set theoretical inclusion, will be called a *maximal ε-shock front*. Observe that, because of (III), two maximal ε-shock fronts of the same family either are disjoint or coincide. Moreover, by (3.22), the number of maximal ε-shock fronts that start at the boundary of D is uniformly bounded by $3M/\varepsilon$. On the other hand, the special geometric features of Temple class systems guarantee that no new shock front can arise in the interior of D. Indeed, the coinciding shock and rarefaction assumption together with the existence of Riemann invariants prevents the creation of shocks of other families than those of the incoming fronts at any interaction point. Therefore, for fixed $\varepsilon > 0$, and $i \in \{1, \dots, n\}$, the number of maximal ε-shock fronts of the ith family remains uniformly bounded by $M_\varepsilon \doteq 3M/\varepsilon$ in all u^ν, $\nu \geq 1$. Denote such curves by

$$y^\varepsilon_{\nu,m} : [t^{\varepsilon,-}_{\nu,m}, t^{\varepsilon,+}_{\nu,m}] \mapsto \mathbb{R}, \qquad m = 1, \dots, M_\varepsilon.$$

By possibly extracting a further subsequence, we can assume the convergence

$$y^\varepsilon_{\nu,m}(\cdot) \longrightarrow y^\varepsilon_m(\cdot), \qquad t^{\varepsilon,\pm}_{\nu,m} \longrightarrow t^{\varepsilon,\pm}_m, \qquad m = 1, \dots, M_\varepsilon,$$

for some Lipschitz continuous paths $y^\varepsilon_m : [t^{\varepsilon,-}_m, t^{\varepsilon,+}_m] \mapsto \mathbb{R}$, $m = 1, \dots, M_\varepsilon$. Repeating this construction in connection with a sequence $\varepsilon_k \to 0$, and taking the union of all the paths thus obtained, we find, for each characteristic family $i \in \{1, \dots, n\}$, a countable family of Lipschitz continuous curves $y_m : [t^-_m, t^+_m] \mapsto \mathbb{R}$, $m \in \mathbb{N}$. Call Υ the union of all such curves.

2. Consider now a point $P = (\tau, y_m(\tau)) \notin \Theta$ along a curve $y_m \in \Upsilon$ of a family $i \in \{1, \dots, n\}$. Notice that, by construction and because of (3.24), no curve in Υ can cross y_m at P. Moreover, by (3.22), the function $u(\tau, \cdot)$ has bounded variation, and hence there exist the limits

$$\lim_{x \to y_m(\tau)-} u(\tau, x) \doteq u^-, \qquad \lim_{x \to y_m(\tau)+} u(\tau, x) \doteq u^+. \tag{3.25}$$

We claim also that the limits (2.42) exist and thus coincide with those in (3.25). To this end observe that, by construction, there exists a sequence of shocks curves $y_{\nu,m}$ of the ith family converging to y_m, along which each approximate solution u^ν has a jump of strength $\geq \varepsilon^*$, for some $\varepsilon^* > 0$. Then, relying on the assumption

$$\mu^{IC}(\{P\}) = 0 \tag{3.26}$$

and letting $B(P, r)$ denote the ball centered at P with radius r, one can establish the limits

$$\lim_{r \to 0+} \limsup_{\nu \to +\infty} \left(\sup_{(t,x) \in B(P,r) x < y_{\nu,m}(t)} |u^\nu(t, x) - u^-| \right) = 0, \tag{3.27}$$

$$\lim_{r \to 0+} \limsup_{\nu \to +\infty} \left(\sup_{(t,x) \in B(P,r) x > y_{\nu,m}(t)} |u^\nu(t, x) - u^-| \right) = 0, \tag{3.28}$$

which clearly yield (2.42). Indeed, if, for example, (3.27) do not hold, by possibly taking a subsequence we would find $\varepsilon > 0$ and points $P_\nu \doteq (t_\nu, \xi_\nu) \to P$ on the left of $y_{\nu,m}$ such that

$$|u^\nu(t_\nu, \xi_\nu) - u^-| \geq \varepsilon \quad \forall \nu.$$

On the other hand, by the first limit in (3.25), and since $u^\nu(\tau,x)\to u(\tau,x)$ for a.e. $x\in[\alpha,\beta]$, we could also find points $Q_\nu\doteq(\tau,\xi'_\nu)\to P$ on the left of $y_{\nu,m}$ such that

$$u^\nu(\tau,\xi'_\nu)\to u^-,\quad \frac{|\xi_\nu-\xi'_\nu|}{|t_\nu-\tau|}>\lambda^{\max}\quad\forall\nu,$$

where $\lambda^{\max}$ denotes the maximum speed at (2.3). But then, for each solution u^ν, the segment $\overline{P_\nu Q_\nu}$ would be crossed by an amount of waves of strength $\geq\varepsilon$. Hence, by strict hyperbolicity and genuine nonlinearity, this would generate a uniformly positive amount of interaction and cancellation within an arbitrary small neighborhood of P (see [9, Theorem 10.4, Step 5]) which by the definition (3.23) and because of (3.24) contradicts the assumption (3.26).

To complete the proof of (i), observe that, by construction, the states $u^-_{\nu,m}(\tau)$, $u^+_{\nu,m}(\tau)$ to the left and right of the jump in u^ν at $y_{\nu,m}(\tau)$ satisfy the Rankine–Hugoniot conditions. Thus, relying on (3.27)–(3.28) and on the convergence $y_{\nu,m}\to y_\nu$, one deduces (2.43). The proof of (ii) can be established with the same type of arguments (see [9, Theorem 10.4, Step 8]). □

As an immediate consequence of Lemma 3.4, we derive Proposition 2.13, stated in section 2.4.

Proof of Proposition 2.13. Consider a sequence $u^\nu(t,\cdot):[a,b]\mapsto\Gamma^\nu$ of front tracking approximate solutions of the mixed problem for (1.1) on the region $\Omega_T\doteq[0,T]\times[a,b]$ that converges in $\mathbf{L}^1$, as $\nu\to\infty$, to some function $u(t,\cdot):[a,b]\mapsto\Gamma$ for every $t\in[0,T]$. Observe that by Theorem 2.9 one can find another sequence $\{v^\nu\}_{\nu\geq1}$ of approximate solutions of (1.1) on the region Ω_T, whose initial and boundary data have a number of shocks $N_\nu\leq\nu$ for each characteristic family, and such that

$$\|u^\nu(t,\cdot)-v^\nu(t,\cdot)\|_{\mathbf{L}^1([a,b])}\leq1/\nu\quad\forall t\in[1/\nu,T].$$

Then, thanks to the Oleinik estimates (2.31)–(2.32) and because all v^ν take values in the compact set (2.13), there will be, for every fixed $\varepsilon>0$, some constant $M_\varepsilon>0$ such that

$$\begin{aligned}\text{Tot.Var.}\{v^\nu(t,\cdot);\ [a+\varepsilon,b-\varepsilon]\}&\leq M_\varepsilon\quad\forall t\in[\varepsilon,T],\\ \text{Tot.Var.}\{v^\nu(\cdot,x);\ [\varepsilon,T]\}&\leq M_\varepsilon\quad\forall x\in[a+\varepsilon,b-\varepsilon]\end{aligned}\tag{3.29}$$

$\forall\nu\in\mathbb{N}$. Thus, writing Ω_T as the countable union

$$\Omega_T=\cup_k D_k,\quad D_k\doteq[1/k,T]\times[a+(1/k),b-(1/k)],$$

and applying Lemma 3.4 to each sequence of maps $v^\nu_k\doteq v^\nu\restriction_{D_k},\nu\geq1$, defined as the restriction of v^ν to the domain D_k, we clearly reach the conclusion of Proposition 2.13. □

We are now in position to establish Theorem 2.11, relying on Proposition 2.13 and Theorem 2.10.

Proof of Theorem 2.11. Let $u^\nu(t,\cdot):[a,b]\mapsto\Gamma^\nu$ be a sequence of front tracking approximate solutions of the mixed problem for (1.1) on the region $\Omega_T\doteq[0,T]\times[a,b]$ that converges in $\mathbf{L}^1$, as $\nu\to\infty$, to some function $u(t,\cdot):[a,b]\mapsto\Gamma$ for every $t\in[0,T]$. Since, by construction, each u^ν is a weak solution of (1.1) and because $u^\nu(0,\cdot)\to u(0,\cdot)=\overline{u}$, the limit function u also is a weak solution of the Cauchy problem (1.1)–(1.2) on the region Ω_T. Moreover, applying Proposition 2.13, we deduce that u admits at $t=0$ and at $x=a$, $x=b$ the left and right limits stated in Theorem 2.11. On

the other hand, by the same arguments used in the proof of Proposition 2.13, we may assume that the initial and boundary data of each approximate solution u^ν have at most $N_\nu \leq \nu$ shocks for every characteristic family. Then, letting $\nu \to \infty$ in (2.31)–(2.32), by the lower semicontinuity of the total variation we find that u satisfies the entropy conditions (2.7)–(2.10) on the decay of positive waves. It follows that u is an entropy weak solution of the mixed problem (1.1)–(1.4) according to Definition 2.2. Hence, observing that by construction the map $(t,x) \to (u(t,\cdot), u(\cdot,x))$ takes values within the domain $\mathcal{D}_T$ defined in (2.37), and applying Theorem 2.10, we deduce that (2.41) is verified. □

4. Proof of Theorems 2.4–2.7.

Proof of Theorem 2.4. We shall first prove that, for every fixed $\overline{\tau} > 0$, there exists some constant $\rho = \rho(\overline{\tau}) > 0$ so that (2.17) holds. Given $\widetilde{u}_a \in \mathcal{U}_\tau^\infty$, $\widetilde{u}_b \in \mathcal{U}_\tau^\infty$, $\tau \geq \overline{\tau}$, let $u = u(t,x)$ be an entropy weak solution of (1.1)–(1.4) on the region $[0,\tau] \times [a,b]$ according to Definition 2.2. Then, the Oleinik-type estimates (2.8) on the decay of positive waves imply that, for $i \in \{p+1,\dots,n\}$, $\tau \geq \overline{\tau}$, and for a.e. $a < x < y < b$, there holds

$$\begin{aligned} \frac{w_i(\tau,y) - w_i(\tau,x)}{y-x} &\leq C \cdot \left\{ \frac{y-x}{\tau} + \log\left(\frac{y-a}{x-a}\right) \right\} \\ &\leq (b-a)C \cdot \left\{ \frac{1}{\overline{\tau}} + \frac{1}{x-a} \right\} \\ &\leq \frac{C(b-a)((b-a)+\overline{\tau})}{\overline{\tau}} \cdot \frac{1}{x-a}. \end{aligned} \tag{4.1}$$

Clearly, with the same computations, relying on the Oleinik-type estimates (2.7), we deduce that, for $i \in \{1,\dots,p\}$, $\tau \geq \overline{\tau}$, and for a.e. $a < x < y < b$, there holds

$$\frac{w_i(\tau,y) - w_i(\tau,x)}{y-x} \leq \frac{C(b-a)((b-a)+\overline{\tau})}{\overline{\tau}} \cdot \frac{1}{b-y}. \tag{4.2}$$

Hence, taking

$$\rho \geq \frac{C(b-a)((b-a)+\overline{\tau})}{\overline{\tau}} \tag{4.3}$$

from (4.1)–(4.2), we derive $u(\tau,\cdot) \in K^\rho$, which proves (2.17).

Concerning the second statement of the theorem, we will show that, letting $\lambda^{\min}$ and ρ' be the minimum speed in (2.3) and the constant (3.19) of Lemma 3.1 and taking

$$T \doteq \frac{4(b-a)}{\lambda^{\min}}, \tag{4.4}$$

the relation (2.18) is verified, i.e., given $\varphi \in K^{\rho'}$ and $\tau > T$, there exist $\widetilde{u}_a \in \mathcal{U}_\tau^\infty$, $\widetilde{u}_b \in \mathcal{U}_\tau^\infty$, and a solution $u(t,x)$ of (1.1)–(1.4) on $[0,\tau] \times [a,b]$ (according to Definition 2.2), such that $u(\tau,\cdot) \equiv \varphi$. Notice that, by Remark 2.5, we may assume that $w_i(\varphi(x))$ admits left and right limits in any point $x \in]a,b[$ and that $w_i(\varphi(x)) = w_i(\varphi(x^+)) \doteq \lim_{\xi \to x^+} w_i(\varphi(\xi))$ for all $i \in \{1,\dots,n\}$. The proof is divided into two steps.

Step 1. *Backward construction of front tracking approximations.* Letting $\rho' > 0$ be the constant in (3.19), consider a sequence $\{\varphi_\nu\}_{\nu \geq 1}$ of (right continuous) piecewise constant maps in $K_\nu^{\rho'}$, satisfying the conditions (a) and (b) of Lemma 3.2, and take

a piecewise constant approximation $\overline{u}^\nu : [a,b] \mapsto \Gamma^\nu$ of the initial data $\overline{u}$, so that $\overline{u}^\nu \to \overline{u}$ in $\mathbf{L}^1$. Given $\tau > T$ (T being the time defined in (4.4)), for each $\nu \geq 1$, we will construct here a front tracking solution $u^\nu(t,x)$ of (1.1) on the region $[0,\tau]\times[a,b]$, with initial data $u^\nu(0,\cdot) = \overline{u}^\nu$, so that

$$u^\nu(\tau,\cdot) = \varphi_\nu. \tag{4.5}$$

This goal is accomplished by proving the following two lemmas.

LEMMA 4.1. *Let $T, \rho' > 0$ be the constants in* (4.4) *and* (3.19). *Then, for every (right continuous) $\varphi_\nu \in K_\nu^{\rho'}$, $\nu \geq 1$, satisfying condition* (a) *of Lemma* 3.2 *and for any $\tau > T$, there exists a front tracking solution $u^\nu(t,x)$ of* (1.1) *on the region $[(3/4)T,\tau]\times[a,b]$, with boundary data $\widetilde{u}_a^\nu \doteq u^\nu(\cdot,a)$, $\widetilde{u}_b^\nu \doteq u^\nu(\cdot,b) \in \mathbf{L}^\infty([(3/4)T,\tau],\Gamma^\nu)$, so that*

$$u^\nu((3/4)T,x) \equiv \omega, \quad u^\nu(\tau,x) = \varphi_\nu(x) \quad \forall x \in [a,b] \tag{4.6}$$

for some constant state $\omega \in \Gamma^\nu$.

Proof. Given $\tau > T$ and $\varphi_\nu \in K_\nu^{\rho'}$, $\nu \geq 1$, satisfying condition (a) of Lemma 3.2, we will use the algorithm described in section 2.4 to construct backward in time a front tracking solution that takes value φ_ν at time τ. To this end, we first observe that according to the algorithm of section 2.4, we can always construct the backward solution of a Riemann problem with terminal data

$$u(t,x) = \begin{cases} u^L & \text{if } x < \xi, \\ u^R & \text{if } x > \xi \end{cases} \tag{4.7}$$

if the terminal states $u^L, u^R \in \Gamma^\nu$ have Riemann coordinates

$$w(u^L) \doteq w^L = (w_1^L, \ldots, w_n^L), \qquad w(u^R) \doteq w^R = (w_1^R, \ldots, w_n^R)$$

that satisfy

$$w_i^L < w_i^R \Longrightarrow w_i^R = w_i^L + 2^{-\nu} \quad \forall i. \tag{4.8}$$

Indeed, if we consider the intermediate states

$$z^i = \begin{cases} u^L & \text{if } i = 0, \\ u(w_1^L, \ldots, w_{n-i}^L, w_{n-i+1}^R, \ldots, w_n^R) & \text{if } 0 < i < n, \\ u^R & \text{if } i = n, \end{cases} \tag{4.9}$$

we realize that, because of (4.8), the solution of every Riemann problem with initial states (z^{i-1}, z^i) (defined as in section 2.4) contains only a single front. Thus, we can construct the solution to the Riemann problem with terminal data (4.7) in a backward neighborhood of (t,ξ) by piecing together the solutions to the simple Riemann problems (z^{i-1}, z^i), $i = 1, \ldots, n$.

A front tracking solution u^ν can now be constructed backward in time starting at $t = \tau$ and piecing together the backward solutions of the Riemann problems determined by the jumps in φ_ν. The resulting piecewise constant function $u^\nu(\tau-,\cdot)$ is then prolonged for $t < \tau$ tracing backward the incoming fronts at $t = \tau$, up to the first time $\tau_1 < \tau$ at which two or more discontinuities cross in the interior of Ω. Observe that, since u^ν is a front tracking solution constructed by the algorithm of section 2.4 on the region $[\tau_1,\tau]\times[a,b]$, we can apply Lemma 3.3. Hence, it follows that the left and

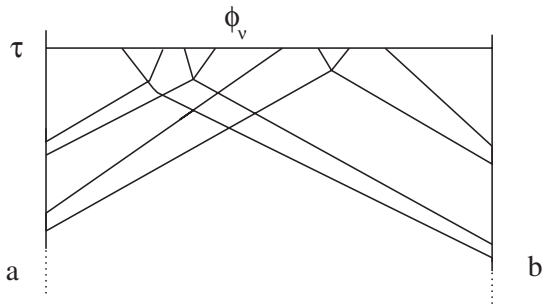


FIG. 1.

right states of the jumps occurring in $u^\nu(\tau_1,\cdot)$ satisfy condition (4.8), because (3.20) guarantees that two (or more) adjacent rarefaction fronts of the same family cannot cross at time τ_1. We then solve backward the resulting Riemann problems applying again the above procedure. This determines the solution $u^\nu(t,\cdot)$ until the time $\tau_2 < \tau_1$ at which another intersection between its fronts takes place in the interior of Ω, and so on (see Figure 1).

With this construction we define a front tracking solution $u^\nu(t,x)$ on the whole region $[(3/4)T,\tau]\times[a,b]$ that verifies the first equality in (4.6) and corresponds to the boundary data $\widetilde{u}^\nu_a \doteq u^\nu(\cdot,a)$, $\widetilde{u}^\nu_b \doteq u^\nu(\cdot,b) \in \mathbf{L}^\infty([(3/4)T,\tau],\Gamma^\nu)$. Clearly, the total number of wave-fronts in $u^\nu(t,\cdot)$ decreases, as $t \downarrow (3/4)T$, whenever a (backward) front crosses the boundary points $x=a$, $x=b$. Since (2.3) implies that the maximum time taken by fronts of u^ν to cross the interval $[a,b]$ is $(b-a)/\lambda^{\min}$, the definition (4.4) of T guarantees that all the (backward) fronts of u^ν will hit the boundaries $x=a$, $x=b$ within some time $\tau' \in](3/4)T,\tau[$, which shows also that the second equality in (4.6) is verified, thus completing the proof. □

LEMMA 4.2. *Let $T>0$ be the constant in* (4.4). *Then, for any piecewise constant function $\overline{u}^\nu \in \mathbf{L}^\infty([a,b],\Gamma^\nu)$ and for every state $\omega\in\Gamma^\nu$, there exists a front tracking solution $u^\nu(t,x)$ of* (1.1) *on the region $[0,(3/4)T]\times[a,b]$, corresponding to some boundary data $\widetilde{u}^\nu_a,\widetilde{u}^\nu_b \in \mathbf{L}^\infty([0,(3/4)T],\Gamma^\nu)$, so that*

$$u^\nu(0,x)=\overline{u}^\nu(x), \quad u^\nu((3/4)T,x)\equiv\omega \quad \forall x\in[a,b]. \tag{4.10}$$

Proof. The approximate solution u^ν is constructed as follows. By Remark 2.8, for $t\in[0,T/4]$, we can define $u^\nu(t,x)$ as the restriction to the region $[0,T/4]\times[a,b]$ of the front tracking solution to the Cauchy problem for (1.1), with initial data

$$\overline{u}(x)=\begin{cases}\overline{u}^\nu(a+) & \text{if } x<a,\\ \overline{u}^\nu(x) & \text{if } a\le x\le b,\\ \overline{u}^\nu(b-) & \text{if } x>b\end{cases}$$

(constructed as in [12] with the same type of algorithm described in section 2.4). Observe that, since u^ν contains only fronts originated at the points of the segment $\{(0,x);\ x\in[a,b]\}$, because of (2.3), (4.4), these wave-fronts cross the whole interval $[a,b]$ and exit from the boundaries $x=a$, $x=b$ before time $T/4$ (see Figure 2). Hence, there will be some state $\omega'\in\Gamma^\nu$ such that

$$u^\nu(T/4,x)\equiv\omega' \quad \forall x\in[a,b]. \tag{4.11}$$

Thus, introducing the intermediate state

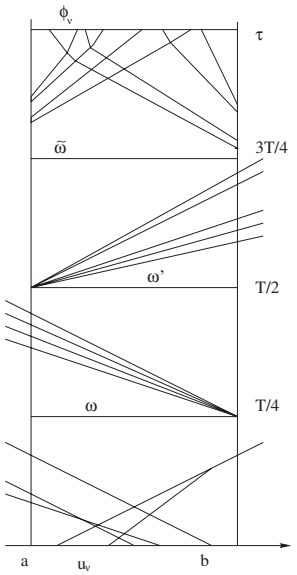


FIG. 2.

$$\widetilde{\omega} \doteq (\omega_1, \dots, \omega_p, \omega'_{p+1}, \dots, \omega'_n)$$

between ω' and ω, we will define $u^\nu(t,x)$, for $t \in [T/4, T/2]$, as the restriction to the region $[T/4, T/2] \times [a,b]$ of the approximate solution to the Riemann problem for (1.1), with initial data

$$u^\nu(T/4, x) = \begin{cases} u(\omega') & \text{if } x < b, \\ u(\widetilde{\omega}) & \text{if } x > b, \end{cases} \tag{4.12}$$

while, for $t \in [T/2, (3/4)T]$, we will let $u^\nu(t,x)$ be the restriction to the region $[T/2, (3/4)T] \times [a,b]$ of the approximate solution to the Riemann problem for (1.1), with initial data

$$u^\nu(T/2, x) = \begin{cases} u(\omega) & \text{if } x < a, \\ u(\widetilde{\omega}) & \text{if } x > a. \end{cases} \tag{4.13}$$

By the definition of $\widetilde{\omega}$, and because of (2.3), (4.4), on $[T/4, T/2]$ the solution of the Riemann problems with initial data (4.12) contains only wave-fronts originated at the point $(T/4, b)$ that cross the whole interval $[a,b]$ and exit from the boundary $x = a$ before time $T/2$. Similarly, still by (2.3), (4.4), for $t \in [T/2, (3/4)T]$ the solution of the Riemann problem with initial data (4.13) contains only wave-fronts originated at $(T/2, a)$ that cross the whole interval $[a,b]$ and exit from the boundary $x = b$ before time $(3/4)T$ (see Figure 2). Hence, $u^\nu(t,x)$ is a front tracking solution defined on the whole region $[0, (3/4)T] \times [a,b]$ that corresponds to the boundary data $\widetilde{u}^\nu_a \doteq u^\nu(\cdot, a)$, $\widetilde{u}^\nu_b \doteq u^\nu(\cdot, b) \in \mathbf{L}^\infty([0, (3/4)T], \Gamma^\nu)$, and verifies the conditions (4.10). □

Step 2. *Convergence of the approximate solutions.* By Step 1, for a given $\varphi \in K^{\rho'}$ (with ρ' as in (3.19)), we have found a sequence of initial data $\overline{u}^\nu$ and boundary data $\widetilde{u}^\nu_a, \widetilde{u}^\nu_b \in \mathcal{U}^\infty_\tau$, so that, letting $u^\nu(\tau, \cdot) \doteq E^\nu_\tau(\overline{u}^\nu, \widetilde{u}^\nu_a, \widetilde{u}^\nu_b)$ be the corresponding front tracking solution, there holds

$$\overline{u}^\nu \to \overline{u}, \qquad u^\nu(\tau, \cdot) \to \varphi \quad \text{in } \mathbf{L}^1([a,b]). \tag{4.14}$$

By the same arguments used in the proof of Proposition 2.13, we may assume that the initial and boundary data of each approximate solution u^ν have at most $N_\nu \leq \nu$ shocks for every characteristic family. Then, thanks to the Oleinik-type estimates (2.31) and because u^ν are uniformly bounded since they take values in the compact set (2.13), for every fixed $\varepsilon > 0$, there will be some constant $C_\varepsilon > 0$ such that

$$\begin{aligned} &\text{Tot.Var.}\{u^\nu(t,\cdot);\ [a+\varepsilon, b-\varepsilon]\} \leq C_\varepsilon \quad \forall t \in [\varepsilon, \tau], \\ &\int_{a+\varepsilon}^{b-\varepsilon} |u^\nu(t,x) - u^\nu(s,x)|\, dx \leq C_\varepsilon |t-s| \quad \forall t, s \in [\varepsilon, \tau] \end{aligned} \tag{4.15}$$

$\forall \nu \in \mathbb{N}$. Hence, applying Helly's theorem, we deduce that there exists a subsequence $\{u^{\nu_j}\}_{j\geq 0}$ that converges in $\mathbf{L}^1([a,b],\Gamma)$ to some function $u_\varepsilon(t,\cdot)$, for any $t \in [\varepsilon,\tau]$. Therefore, repeating the same construction in connection with a sequence $\varepsilon_k \to 0+$ and using a diagonal procedure, we obtain a subsequence $\{u^{\nu'}(t,\cdot)\}_{\nu'\geq 0}$ that converges in $\mathbf{L}^1([a,b],\Gamma)$ to some function $u(t,\cdot)$ for any $t \in [0,\tau]$. Then, by Theorem 2.11, there holds (2.41), with $\widetilde{u}_a \doteq u(\cdot,a), \widetilde{u}_b \doteq u(\cdot,b) \in \mathcal{U}_\tau^\infty$, while (4.14) implies $u(\tau,\cdot) = \varphi$, which shows $\varphi \in \mathcal{A}(\tau)$. This completes the proof of Theorem 2.4. □

We next establish the compactness of the attainable set (2.14) stated in Theorem 2.7. The proof is quite similar to that of [3, Theorem 2.3]. We repeat it for completeness.

Proof of Theorem 2.7. Fix $T > 0$, and consider a sequence $\{u^\nu\}_{\nu\geq 0}$ of entropy weak solutions to the mixed problem for (1.1) on $\Omega_T \doteq [0,T] \times [a,b]$ (according to Definition 2.2), with fixed initial data $\overline{u} \in \mathbf{L}^\infty([a,b],\Gamma)$. Since all u^ν are uniformly bounded and because of the Oleinik-type estimates (2.7)–(2.8), one can find, for every $\varepsilon > 0$, some constant $C_\varepsilon > 0$ so that (4.15) holds. Thus, with the same arguments used in Step 2 of the previous proof, we can construct a subsequence $\{u^{\nu'}\}_{\nu'\geq 0}$ so that, for any $t \in [0,T]$, $u^{\nu'}(t,\cdot)$ converges in $\mathbf{L}^1$ to some function $u(t,\cdot)$, which is continuous as a map from $]0,T]$ into $\mathbf{L}^1([a,b],\Gamma)$ and satisfies the entropy conditions (2.7)–(2.10) on the decay of positive waves. On the other hand, the weak traces $\Psi_a^{\nu'}, \Psi_b^{\nu'}$ of the fluxes $f(u^{\nu'})$ at the boundaries $x = a$, $x = b$ are uniformly bounded, and hence are weak* relatively compact in $\mathbf{L}^\infty([0,T])$. Thus, by possibly taking a further subsequence, we have

$$\Psi_a^{\nu'} \overset{*}{\rightharpoonup} \Psi_a, \qquad \Psi_b^{\nu'} \overset{*}{\rightharpoonup} \Psi_b \quad \text{in } \mathbf{L}^\infty([0,T]) \tag{4.16}$$

for some maps $\Psi_a, \Psi_b \in \mathbf{L}^\infty([0,T])$. Notice that, by the properties of the Riemann invariants, the set $f(\Gamma)$ is closed and convex, and hence also the weak limits Ψ_a, Ψ_b take values in $f(\Gamma)$. Moreover, since each u^ν is a distributional solution of (1.1)–(1.2) on Ω_T, the limit function u also is a distributional solution of the Cauchy problem (1.1)–(1.2) on the region Ω_T. Then, setting $\widetilde{u}_a \doteq f^{-1} \circ \Psi_a, \widetilde{u}_b \doteq f^{-1} \circ \Psi_b$, it follows that u is an entropy weak solution of the mixed problem (1.1)–(1.4) (with boundary data in $\mathcal{U}_T^\infty$) according to Definition 2.2, which shows that $u(T,\cdot) \in \mathcal{A}(T)$. This completes the proof of Theorem 2.7. □

5. Conclusion. The results presented in this paper represent a contribution to the development of a general theory on boundary controllability for systems of nonlinear hyperbolic equations within the context of entropy weak solutions. As is shown in [11] there is no hope of establishing exact controllability results for general systems of conservation laws due to the wave-front structure of the weak solutions, which may present shock waves that can never be canceled by interactions with rarefaction waves of the same characteristic family and that at the same time give rise to new shock

fronts by interacting with shock waves of the other characteristic families. Here we have analyzed the exact boundary controllability for the simplest class of nonlinear hyperbolic systems: the class of Temple systems with genuinely nonlinear characteristic fields, whose study is motivated by applications to multicomponent chromatography.

A natural direction in which to pursue this analysis is to consider Temple systems with linearly degenerate characteristic fields (or with a general "nonconvex" flux) which appear in several traffic flow models [6, 16] where one is usually interested in controlling the inflow of cars at the entry of a given road. Another relevant direction worthy of investigation is the controllability of systems of balance laws, i.e., of systems of conservation laws with the presence of source terms. Systems of balance laws belonging to Temple class arise, for example, in modeling chromatography reactors where chemical reactions take place allowing the different solutes (dissolved in the liquid) to transform into each other (see [25, 26]).

All of this type of analysis refers to the case of boundary control problems where total control on the boundary values is available. Of course one may consider more general controllability problems where the control acts only on some of the boundary conditions. For example, we may consider the system of isentropic gas dynamics describing a gas confined in a cylinder within two pistons. In this case it is reasonable to expect that, by controlling only the speed of one piston, it is possible to asymptotically stabilize the system at any constant state. To this purpose it is natural to study first the boundary controllability of the linearized system. A generic condition that guarantees the exact boundary controllability in finite time of a linear hyperbolic system with constant coefficients is obtained in [2].

Acknowledgment. The authors would like to thank Prof. Alberto Bressan for suggesting the problem.

REFERENCES

[1] D. Amadori, *Initial-boundary value problems for nonlinear systems of conservation laws*, NoDEA Nonlinear Differential Equations Appl., 4 (1997), pp. 1–42.

[2] F. Ancona and G. M. Coclite, *Exact boundary controllability of linear hyperbolic systems*, in preparation.

[3] F. Ancona and P. Goatin, *Uniqueness and stability of* $\mathbf{L}^\infty$ *solutions for Temple class systems with boundary and properties of the attainable sets*, SIAM J. Math. Anal., 34 (2002), pp. 28–63.

[4] F. Ancona and A. Marson, *On the attainable set for scalar non-linear conservation laws with boundary control*, SIAM J. Control Optim., 36 (1998), pp. 290–312.

[5] F. Ancona and A. Marson, *Scalar non-linear conservation laws with integrable boundary data*, Nonlinear Anal., 35 (1999), pp. 687–710.

[6] A. Aw and M. Rascle, *Resurrection of "second order" model of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[7] C. Bardos, A. Y. Leroux, and J. C. Nedelec, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[8] S. Bianchini, *Stability of* $\mathbf{L}^\infty$ *solutions for hyperbolic systems with coinciding shocks and rarefactions*, SIAM J. Math. Anal., 33 (2001), pp. 959–981.

[9] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One-Dimensional Cauchy Problem*, Oxford Univ. Press, Oxford, UK, 2000.

[10] A. Bressan, *Unique solutions for a class of discontinuous differential equations*, Proc. Amer. Math. Soc., 104 (1988), pp. 772–778.

[11] A. Bressan and G. M. Coclite, *On the boundary control of systems of conservation laws*, SIAM J. Control Optim., 41 (2002), pp. 607–622.

[12] A. Bressan and P. Goatin, *Stability of* $\mathbf{L}^\infty$ *solutions of Temple class systems*, Differential Integral Equations, 13 (2000), pp. 1503–1528.

[13] A. Bressan and P. G. LeFloch, *Structural stability and regularity of entropy solutions to hyperbolic systems of conservation laws*, Indiana Univ. Math. J., 48 (1999), pp. 43–84.
[14] G.-Q. Chen and H. Frid, *Divergence-measure fields and hyperbolic conservation laws*, Arch. Ration. Mech. Anal., 147 (1999), pp. 89–118.
[15] G.-Q. Chen and H. Frid, *Vanishing viscosity limit for initial-boundary value problems for conservation laws*, in Nonlinear Partial Differential Equations, G.-Q.Chen and E. DiBenedetto, eds., Contemp. Math. 238, 1999, pp. 35–51.
[16] R. M. Colombo, *A 2×2 hyperbolic traffic flow model. Traffic flow—modelling and simulation*, Math. Comput. Modelling, 35 (2002), pp. 683–688.
[17] C. M. Dafermos, *Generalized characteristic and the structure of solutions of hyperbolic conservation laws*, Indiana Univ. Math. J., 26 (1977), pp. 1097–1119.
[18] R. J. DiPerna, *Singularities of solutions of nonlinear hyperbolic systems of conservation laws*, Arch. Ration. Mech. Anal., 60 (1976), pp. 75–100.
[19] F. Dubois and P. G. LeFloch, *Boundary conditions for non-linear hyperbolic systems of conservation laws*, J. Differential Equations, 71 (1988), pp. 93–122.
[20] K. T. Joseph and P. G. LeFloch, *Boundary layers in weak solutions of hyperbolic conservation laws*, Arch. Ration. Mech. Anal., 147 (1999), pp. 47–88.
[21] T. Horsin, *On the controllability of the Burgers equation*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 83–95.
[22] P. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.
[23] T.-P. Liu, *Admissible solutions of hyperbolic conservation laws*, Mem. Amer. Math. Soc., 240 (1981), pp. 1–78.
[24] H. K. Rhee, R. Aris, and N. R. Amundson, *On the theory of multicomponent chromatography*, Philos. Trans. Roy. Soc. London Ser. A, 267 (1970), pp. 419–455.
[25] H. K. Rhee, R. Aris, and N. R. Amundson, *First-Order Partial Differential Equations:* Vol. I*, Theory and Application of Single Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1986.
[26] H. K. Rhee, R. Aris, and N. R. Amundson, *First-Order Partial Differential Equations:* Vol. II*, Theory and Application of Hyperbolic Systems of Quasilinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[27] M. Sablé-Tougeron, *Méthode de Glimm et probléme mixte*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 423–443.
[28] D. Serre, *Systemes de Lois de Conservation II*, Diderot Editeur, Paris, 1996.
[29] B. Temple, *Systems of conservation laws with invariant submanifolds*, Trans. Amer. Math. Soc., 280 (1983), pp. 781–795.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2191–2232

OPTIMAL BOUNDARY CONTROL FOR THE EVOLUTIONARY NAVIER–STOKES SYSTEM: THE THREE-DIMENSIONAL CASE*

A. V. FURSIKOV†, M. D. GUNZBURGER‡, AND L. S. HOU§

Dedicated to the memory of Olga Aleksandrovna Ladyzhenskaya

Abstract. Optimal boundary control problems for the three-dimensional, evolutionary Navier–Stokes equations in the exterior of a bounded domain are studied. Control is effected through the Dirichlet boundary condition and is sought in a subset of the trace space of velocity fields with almost minimal possible regularity. The control objective is to minimize the drag functional. The existence of an optimal solution is proved. A strong form of an optimality system of equations is derived on the basis of regularity results established in this work for the adjoint Oseen equations with regular initial data which do not satisfy the compatibility conditions.

Key words. optimal control, Navier–Stokes equations, boundary value problem, drag reduction

AMS subject classifications. 76D05, 49J20, 49K20, 35K50

DOI. 10.1137/S0363012904400805

1. Introduction.

This paper is devoted to the investigation of using boundary controls to minimize the drag about a three-dimensional body B moving at a constant velocity $\mathbf{v}_\infty$ in a fluid. The fluid flow is assumed to be governed by the time-dependent, viscous, incompressible Navier–Stokes system. The drag minimization problem is formulated as an optimal boundary control problem by introducing an appropriate drag functional and by choosing a suitable control space. The two-dimensional analogue of this problem was studied in [8], and this paper represents a continuation of that work in the three-dimensional case. Our aims are to prove the existence of an optimal solution and derive an optimality system of equations. In order to achieve these aims, we will need to define the correct mathematical formulation of the optimal control problem in question.

The type of optimal control problem we study has the following form:

$$\mathcal{J}(\mathbf{v}) \to \inf, \tag{1.1}$$

$$\mathrm{NS}(\mathbf{v}, p_1) = \mathbf{0}, \quad \mathbf{v}\big|_{t=0} = \mathbf{v}_0, \quad \mathbf{v}\big|_{|\mathbf{x}|\to\infty} = \mathbf{v}_\infty, \quad \mathbf{v}\big|_{(0,T)\times\partial B} = \mathbf{b}, \tag{1.2}$$

$$R(\mathbf{b}) \le M. \tag{1.3}$$

Here, $\mathbf{v}(t,\mathbf{x})$ and $p_1(t,\mathbf{x})$ for $t \in [0,T]$ and $\mathbf{x} \in \Omega \equiv \mathbb{R}^3 \setminus B$ are the vector-valued velocity field and scalar-valued pressure field, respectively, of the fluid flow surrounding the body B; we attach the coordinates to B, i.e., we treat B as fixed so that the fluid is in motion relative to B. Also, $\mathcal{J}(\mathbf{v})$ is the drag functional, $\mathrm{NS}(\mathbf{v}, p_1) = \mathbf{0}$ is

*Received by the editors January 2, 2002; accepted for publication (in revised form) October 4, 2004; published electronically May 27, 2005.
http://www.siam.org/journals/sicon/43-6/40080.html

†Department of Mechanics and Mathematics, Moscow State University, Moscow 119899, Russia (fursikov@mtu-net.ru). Part of this work was completed while the author was in residence at the Department of Mathematics, Iowa State University, Ames, IA 50011-2064, and at the School of Computational Science, Florida State University, Tallahassee, FL 32306-4120.

‡School of Computational Science, Florida State University, Tallahassee, FL 32306-4120 (gunzburg@csit.fsu.edu).

§Department of Mathematics, Iowa State University, Ames, IA 50011-2064 (hou@math.iastate.edu).

the Navier–Stokes system, $\mathbf{v}_0$ is the initial condition, and $\mathbf{b}$ is the Dirichlet boundary condition which acts as the control in our problem.

It is clear that the correct physical setting of the optimal drag reduction problem must contain the constraint (1.3), where $R(\mathbf{b})$ is a norm-like functional for functions defined on the boundary $\Sigma = (0,T)\times\partial\Omega \equiv (0,T)\times\partial B$. Indeed, if (1.3) is not imposed or the constant M in that condition is too large, then, instead of drag reduction, the boundary control $\mathbf{b}$ can actually push the body B in the direction opposite to $\mathbf{v}_\infty$ (see the relevant discussions in [8]).

Note that choosing the appropriate form for $R(\mathbf{b})$ is not a trivial task, since choosing R to be a simple functional such as the $\mathbf{L}^2(\Sigma)$-norm is unsuccessful even in two-dimensional case (see [8]); of course, in three dimensions, it is impossible as well. The correct choice of $R(\mathbf{b})$ is strictly connected to choosing a proper boundary control space. That is, they must be chosen in such a way that the validity of the use of the Lagrange multiplier principle is guaranteed (see [2, 6]). This amounts to the requirement that the state space $\mathbf{W}$ must be sufficiently regular so that the derivative operator $\mathrm{NS}'(\widetilde{\mathbf{v}}) : \mathbf{W} \to \mathbf{F}$ is an epimorphism for some suitably chosen function space $\mathbf{F}$. It turns out that this requirement can be met if the solution to the Navier–Stokes equations is unique within the space $\mathbf{W}$. It is well known (see [5]) that if we restrict ourselves to Hilbert space settings, then $\mathbf{W}$ should have the following form to ensure the uniqueness property:

$$\mathcal{V}^{(s)}(Q) \equiv \{\mathbf{v} \in L^2(0,T;\mathbf{H}^s(\Omega)) : \partial_t \mathbf{v} \in L^2(0,T;\mathbf{H}^{s-2}(\Omega)),\ \operatorname{div}\mathbf{v} = 0\} \tag{1.4}$$

for $s \geq 3/2$, where $Q \equiv (0,T)\times\Omega$ and the function spaces used will be precisely defined in subsection 2.4. In other words, if the smoothness of functions from $\mathbf{W}$ is less than $\mathcal{V}^{(3/2)}(Q)$, then it is not clear how to prove surjectivity of the operator $\mathrm{NS}'(\widetilde{\mathbf{v}}) : \mathbf{W} \to \mathbf{F}$ or the solvability of the corresponding Oseen equations with such nonsmooth coefficients. To find a suitable norm R, we could simply take the norm of space of restrictions to Σ of a function belonging to the space $\mathcal{V}^{(3/2)}(Q)$, whose trace space was already characterized in [9]. However, to simplify the definition of R, we choose the weakest norm R which avoids fractional derivatives and whose natural domain of definition is a subspace of the trace space of $\mathcal{V}^{(3/2)}(Q)$ (see [10]). Precisely, we choose

$$R(\mathbf{b}) \equiv \|\mathbf{b}\|^2_{\mathbf{H}^1(\Sigma)} = \int_\Sigma \left(|\partial_t \mathbf{b}|^2 + |\nabla_\tau \mathbf{b}|^2 + |\mathbf{b}|^2\right) ds\, dt, \tag{1.5}$$

where $\mathbf{H}^1(\Sigma)$ is the Sobolev space of vector-valued functions defined on Σ possessing square integrable first derivatives. (The definition of surface gradient ∇_τ will be given below; see (2.11).)

Since the norm (1.5) we choose for the boundary data $\mathbf{b}$ is stronger than the norm for the space generated by restricting the space (1.4) to the boundary, we should choose the solution space $\mathbf{W}$ for the Navier–Stokes equations to be the one that corresponds to the boundary norm (1.5) instead of choosing $\mathbf{W}$ to be simply (1.4). The problem of characterizing such a space $\mathbf{W}$ was solved in [10].

Note that our aim is to investigate the case for which there are no restrictions on the magnitude $|\mathbf{v}_\infty|$ of the velocity $\mathbf{v}_\infty$ of the body B. It is precisely the case of large $|\mathbf{v}_\infty|$ that is most interesting in applications. To fulfill such investigation under general assumptions is not possible because of the absence of theorems regarding the existence of smooth solutions for the three-dimensional evolutionary Navier–Stokes equations with arbitrary data; only when the corresponding norms of the data are sufficiently small has the existence of a smooth solution been proved.

To circumvent this difficulty we consider the following concrete and physically reasonable situation. Let a body B move in a steady-state regime with velocity $\mathbf{v}_\infty$. Then, at the instant t_0 (say $t_0 = 0$), we switch on the control $\mathbf{b}$ on ∂B and solve the optimal drag reduction problem over the time interval $(0, T)$, where $T > 0$ is a given arbitrary number. In contrast to the evolutionary case, the existence of a smooth solution $\mathbf{v}_0$ for the steady-state three-dimensional Navier–Stokes equations with the adhesion condition on ∂B and an arbitrary data $\mathbf{v}_\infty$ at infinity has been proved. So, we can take this steady-state solution $\mathbf{v}_0$ as the initial condition in (1.2); this in turn will allow us to investigate optimal control problem (1.1)–(1.3) on the basis of local existence theorems of smooth solutions for the three-dimensional evolutionary Navier–Stokes equations in a neighborhood of $\mathbf{v}_0$. In addition, we will fulfill the locality condition by choosing a sufficiently small parameter M in (1.3). Of course, in such an approach M depends on $|\mathbf{v}_\infty|$:

$$M(|\mathbf{v}_\infty|) \to 0 \qquad \text{as} \qquad |\mathbf{v}_\infty| \to \infty. \tag{1.6}$$

The plan described above will be realized mathematically in this paper. We point out one difficulty arising in this realization that is connected with the property $\mathbf{v}_0(x) - \mathbf{v}_\infty \notin \mathbf{L}^2(\Omega)$ for the steady-state solution. This property does not complicate very much our proof of the existence theorem for the control problem (1.1)–(1.3). However, the problem of constructing a weak solution for the optimality system becomes quite difficult. We will invoke a special form of the abstract Lagrange multiplier principle (see [6, Chap. 2, Thm. 1.6, Thm. 1.8]); to apply this result we are forced to introduce some special Orlicz spaces which are connected with the properties of a vector field $\mathbf{v}_0(x) \to \mathbf{v}_\infty$.

In section 2, we give a precise statement of the optimal control problems we consider. In section 3 and Appendix A, we establish regularity results for the Navier–Stokes system as well as for linearized Navier–Stokes systems and the adjoint linearized Navier–Stokes systems. Although several of these results are of interest in their own right, they are used in this paper as auxiliary results to help us prove the main results of this paper: the existence of optimal solutions and the derivation of weak and strong forms of an optimality system. Section 4 is devoted to the proof of the existence theorem for the control problems defined in section 2. In sections 5–7, we derive corresponding weak and strong forms of the optimality systems.

Finally, we emphasize that we consider this investigation to be a major, but not final, step towards a complete solution of the drag reduction problem. Even after this paper and [8, 9, 10], there remain a number of unresolved problems. For example, by virtue of (1.6) for sufficiently large $|\mathbf{v}_\infty|$, the size of the bound M in (1.3) may become too small to be of practical use in applications. The issue of how to increase M in this situation goes beyond the scope of this paper. Nevertheless, we consider the resolution of this issue quite realistic; indeed, as was shown in [7, 15], it is quite possible to solve nonlocal control problems in the space of smooth solutions for the three-dimensional evolutionary Navier–Stokes equations if the control function is supported on the whole boundary.

2. Formulation of the problem. In this section, we provide a precise statement of the optimal control problems treated in this paper.

2.1. The state system and the cost functional. As discussed in section 1, we consider the problem of using boundary controls to minimize the drag about a three-dimensional body B moving at a constant velocity $\mathbf{v}_\infty$ in a viscous fluid. In

coordinates attached to the body B, this problem transforms into the drag reduction problem for a fixed body B surrounded by a fluid flow having velocity $\mathbf{v}_\infty$ at infinity. Mathematically, the fluid flow is described as follows. Let $B \subset \mathbb{R}^3$ be a bounded domain and let $\Omega \equiv \mathbb{R}^3 \setminus B$. Suppose that the boundary $\partial\Omega$ of Ω is of class C^∞ and is a connected surface. (We impose the last assumption only because it is reasonable from the physical point of view; the generalization of our results to the case of unconnected $\partial\Omega$ is a simple matter.) In the flow domain Ω, we consider the Navier–Stokes system

$$\left\{\begin{array}{l} \partial_t \mathbf{v} - \Delta \mathbf{v} + (\mathbf{v}\cdot\nabla)\mathbf{v} + \nabla p_1 = \mathbf{0} \quad \text{in } Q, \\ \operatorname{div} \mathbf{v} = 0 \quad \text{in } Q, \\ \mathbf{v}|_{t=0} = \mathbf{v}_0 \quad \text{for } \mathbf{x} \in \Omega, \\ \mathbf{v}|_\Sigma = \mathbf{b} \quad \text{with} \quad \displaystyle\int_{\partial\Omega} \mathbf{b}\cdot\mathbf{n}\, ds = 0 \quad \text{for } t \in (0,T), \\ \mathbf{v} \to \mathbf{v}_\infty \quad \text{as } |\mathbf{x}| \to \infty, \end{array}\right. \tag{2.1}$$

where $Q = (0,T)\times\Omega$, $\Sigma = (0,T)\times\partial\Omega$, $\mathbf{v}(t,\mathbf{x}) = \big(v_1(t,\mathbf{x}), v_2(t,\mathbf{x}), v_3(t,\mathbf{x})\big)$ for $t \in [0,T]$, $\mathbf{x} \in \Omega$ is the velocity field, and $\nabla p_1(t,\mathbf{x})$ is a pressure gradient.

The vector field $\mathbf{b}$ is defined on Σ and is the control available to effect optimization. We wish to minimize the work due to drag through a proper choice of $\mathbf{b}$. The work due to drag, or the drag functional, is defined by the formula

$$\mathcal{W} = \int_0^T \int_{\partial\Omega} (\mathbf{v} - \mathbf{v}_\infty)\cdot \mathcal{T}\mathbf{n}\, ds\, dt,$$

where $\mathcal{T} = -p_1 I + 2\mathcal{D}$ is the stress tensor, $\mathcal{D} = \mathcal{D}(\mathbf{v}) = (\nabla\mathbf{v} + \nabla\mathbf{v}^T)/2$ is the rate of deformation tensor, and $\mathbf{n}$ is the unit, outward-pointing normal along the boundary $\partial\Omega$. We can derive, just as in the two-dimensional case (see [8]), the following equivalent expression for $\mathcal{W}$:

$$\begin{aligned} \mathcal{J}(\mathbf{v}) = \int_0^T \int_\Omega \mathcal{D}(\mathbf{v}) : \mathcal{D}(\mathbf{v})\, d\mathbf{x}\, dt + \frac{1}{2}\int_0^T \int_{\partial\Omega} |\mathbf{v} - \mathbf{v}_\infty|^2 \mathbf{v}\cdot\mathbf{n}\, ds\, dt \\ + \frac{1}{2}\int_\Omega |\mathbf{v}(T,\mathbf{x}) - \mathbf{v}_\infty|^2\, d\mathbf{x} - \frac{1}{2}\int_\Omega |\mathbf{v}_0 - \mathbf{v}_\infty|^2\, d\mathbf{x}. \end{aligned} \tag{2.2}$$

The functional (2.2) is precisely the functional to be minimized through a proper choice of the boundary control $\mathbf{b}$ on Σ.

2.2. The initial condition. The correct choice for the initial data $\mathbf{v}_0(\mathbf{x})$ is an important issue since it is related to the physical context in which we formulate the optimal control problem and affects the mathematical proof of existence of optimal solutions. As was mentioned in section 1, we suppose that $\mathbf{v}_0(\mathbf{x})$ is the solution of the steady-state Navier–Stokes problem:

$$\left\{\begin{array}{l} -\Delta\mathbf{v}_0 + (\mathbf{v}_0\cdot\nabla)\mathbf{v}_0 + \nabla p_0 = \mathbf{0} \quad \text{in } \Omega, \\ \operatorname{div}\mathbf{v}_0 = 0 \quad \text{in } \Omega, \\ \mathbf{v}_0\big|_{\partial\Omega} = \mathbf{0}, \\ \mathbf{v}_0 \to \mathbf{v}_\infty \quad \text{as } |\mathbf{x}| \to \infty. \end{array}\right. \tag{2.3}$$

The no-slip condition $\mathbf{v}_0\big|_{\partial\Omega} = \mathbf{0}$ is imposed purely for simplicity and is quite reasonable from a physical point of view. There is no difficulty in treating the case for which $\mathbf{v}_0$ satisfies an inhomogeneous boundary condition.

The following proposition asserts the existence of a solution of (2.3).

PROPOSITION 2.1. *There exists a solution* $(\mathbf{v}_0, p_0) \in [C^\infty(\overline{\Omega})]^4$ *to* (2.3) *satisfying*

$$\|\mathbf{v}_0 - \mathbf{v}_\infty\|_{\mathbf{L}^6(\Omega)} + \|\nabla \mathbf{v}_0\|_{\mathbf{L}^2(\Omega)} + \|\nabla \mathbf{v}_0\|_{\mathbf{C}^2(\overline{\Omega})} \le C|\mathbf{v}_\infty|. \tag{2.4}$$

A proof of this result can be found in, e.g., [11, 13]. Thus, throughout, we assume that the data $\mathbf{v}_0$ in the initial condition in (2.1) is a $\mathbf{C}^\infty(\overline{\Omega})$ solution of (2.3) satisfying (2.4).

2.3. Change of variables. We introduce the change of variables

$$\mathbf{w}(t,\mathbf{x}) = \mathbf{v}(t,\mathbf{x}) - \mathbf{v}_0(\mathbf{x}) \qquad \text{and} \qquad p(t,\mathbf{x}) = p_1(t,\mathbf{x}) - p_0(\mathbf{x}). \tag{2.5}$$

The unknown vector field $\mathbf{w}$ is more convenient to work with than $\mathbf{v}$ due to its zero initial condition and its vanishing values at infinity. After substitution of (2.5) into (2.2) we see that the last two integrals in (2.2) contain two terms $\mathbf{v}_0 - \mathbf{v}_\infty$ that do not belong to $\mathbf{L}^2(\Omega)$. But these terms annul each other. So the problem of minimizing the functional (2.2) subject to (2.1) and (1.3) is then recast as an optimization problem for $\mathbf{w}$ as follows: minimize the functional

$$\begin{aligned} \mathcal{J}(\mathbf{w}) = {} & \int_0^T \int_\Omega \mathcal{D}(\mathbf{w}+\mathbf{v}_0) : \mathcal{D}(\mathbf{w}+\mathbf{v}_0)\, d\mathbf{x}\, dt + \frac{1}{2}\int_0^T \int_{\partial\Omega} |\mathbf{w}-\mathbf{v}_\infty|^2 \mathbf{w}\cdot\mathbf{n}\, ds\, dt \\ & + \frac{1}{2}\int_\Omega (|\mathbf{w}(T,\mathbf{x})|^2 + 2[(\mathbf{w}(T,x)\cdot(\mathbf{v}_0(\mathbf{x}) - \mathbf{v}_\infty)])\, d\mathbf{x} \end{aligned} \tag{2.6}$$

subject to the constraints

$$\partial_t \mathbf{w} - \Delta \mathbf{w} + [(\mathbf{w}+\mathbf{v}_0)\cdot\nabla]\mathbf{w} + (\mathbf{w}\cdot\nabla)\mathbf{v}_0 + \nabla p = \mathbf{0} \quad \text{in } Q, \tag{2.7}$$

$$\operatorname{div} \mathbf{w} = 0 \quad \text{in } Q, \tag{2.8}$$

$$\mathbf{w}|_{t=0} = \mathbf{0} \quad \text{in } \Omega, \tag{2.9}$$

$$\mathbf{w} \to \mathbf{0} \quad \text{as } |\mathbf{x}| \to \infty, \tag{2.10}$$

$$R(\mathbf{w}) = \int_\Sigma \left(|\partial_t \mathbf{w}|^2 + |\nabla_\tau \mathbf{w}|^2 + |\mathbf{w}|^2 \right) ds\, dt \le M. \tag{2.11}$$

Recall that the surface gradient $\nabla_\tau \mathbf{w}\big|_\Sigma = (\nabla \mathbf{w})\big|_\Sigma - (\partial_n \mathbf{w})\big|_\Sigma$, where $\nabla \mathbf{w}$ is the usual gradient of $\mathbf{w}$ in $\mathbb{R}^3$ and $(\partial_n \mathbf{w})$ is the derivative of $\mathbf{w}$ with respect to the outward-pointing unit normal $\mathbf{n}$ on $\partial\Omega$.

In addition, we suppose that $\mathbf{w}$ satisfies the compatibility condition

$$\big(\mathbf{w}(t,\mathbf{x})|_\Sigma\big)|_{t=0} = \mathbf{0} \quad \text{in } \partial\Omega. \tag{2.12}$$

We omitted, from the system (2.7)–(2.11), the boundary condition and eliminated the unknown Dirichlet boundary control $\mathbf{b}$ in (2.1); instead, the Dirichlet control is expressed by $\mathbf{w}|_\Sigma$.

By virtue of (2.5) and $\mathbf{v}_0|_{\partial\Omega} = 0$, the functional $R(\mathbf{w})$ and the parameter M in (2.11) coincide with R and M in (1.3) and (1.5).

Note that functional (2.6) is well defined on $\mathbf{w}$ satisfying (2.7)–(2.11): this is shown in subsection 3.2.

2.4. Function spaces. In this subsection, we define function spaces for the state variables and the Dirichlet controls. The structure of the space where we look for the solution of the optimal control problem is quite complicated because it should be the space where solutions of the Navier–Stokes equations are unique, and besides it has to contain all solutions of the boundary value problem for the Navier–Stokes equations with arbitrary Dirichlet boundary conditions from $\mathbf{H}^1(\Sigma)$ (i.e., with not so smooth conditions) but with the right side belonging to the Lebeque space. Besides, we recall definitions of well-known Sobolev spaces.

The Sobolev spaces $H^k(\Omega)$ with k a nonnegative integer are defined by

$$H^k(\Omega) = \big\{u \in L^2(\Omega) \ : \ \|u\|^2_{H^k(\Omega)} \equiv \sum_{|\alpha|\le k} \int_\Omega |D^\alpha u(\mathbf{x})|^2 \, d\mathbf{x} < \infty\big\},$$

where $\mathbf{x} \in \Omega \subset \mathbb{R}^3$, $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ is a multi-index (α_j being a nonnegative integer), $|\alpha| = \alpha_1 + \alpha_2 + \alpha_3$, and $D^\alpha = \partial^{|\alpha|}/(\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \partial x_3^{\alpha_3})$. The Sobolev space $H^s(\Omega)$ for an arbitrary $s > 0$ is defined through the interpolation of the spaces $H^k(\Omega)$ for integer k; see [14]. By definition, $H_0^s(\Omega)$, $s > 0$, is the closure of $C_0^\infty(\Omega)$ in $H^s(\Omega)$. The space $H^{-s}(\Omega)$, $s > 0$, is defined as the dual space of $H_0^s(\Omega)$, i.e., $H^{-s}(\Omega) = (H_0^s(\Omega))'$, with the norm

$$\|f\|_{H^{-s}(\Omega)} = \sup_{\phi \in H_0^s(\Omega),\, \phi \ne 0} \frac{\langle f, \phi \rangle}{\|\phi\|_{H_0^s(\Omega)}},$$

where $\langle \cdot, \cdot \rangle$ denotes the duality between $H^{-s}(\Omega)$ and $H_0^s(\Omega)$ generated by the scalar product in $L^2(\Omega)$. Sobolev spaces on $\partial\Omega$ are denoted by $H^r(\partial\Omega)$ and are defined with the help of partition of unity techniques; for details, see, e.g., [14]. Vector-valued spaces (including vector-valued Sobolev spaces) are denoted by boldface letters, e.g., $\mathbf{H}^s(\Omega) = [H^s(\Omega)]^3$ for all $s \in \mathbb{R}$, $\mathbf{H}_0^s(\Omega) = [H_0^s(\Omega)]^3$, and $\mathbf{H}^r(\partial\Omega) = [H^r(\partial\Omega)]^3$. For $s \ge -1$, we define the divergence-free spaces

$$\mathbf{V}^s(\Omega) \equiv \big\{\mathbf{v} \in \mathbf{H}^s(\Omega) \ : \ \operatorname{div} \mathbf{v} = 0\big\}. \tag{2.13}$$

Also, we define

$$\mathbf{V}_0^0(\Omega) = \big\{\mathbf{v} \in \mathbf{L}^2(\Omega) \ : \ \operatorname{div} \mathbf{v} = 0, \ (\mathbf{v} \cdot \mathbf{n})|_{\partial\Omega} = 0\big\} \tag{2.14}$$

equipped with the $\mathbf{L}^2(\Omega)$ norm where both equalities are understood in the sense of distributions.

For $s \ge 0$, we introduce the spaces of functions depending on both spatial and temporal variables:

$$H^{1,s}(Q) = \big\{y(t, \mathbf{x}) \in L^2(0, T; H^{s+1}(\Omega)) \ : \ \partial_t y \in L^2(0, T; H^s(\Omega))\big\}$$

where $Q = (0, T) \times \Omega$ and

$$H^{1,s}(\Sigma) = \big\{y(t, \mathbf{x}) \in L^2(0, T; H^{s+1}(\partial\Omega)) \ : \ \partial_t y \in L^2(0, T; H^s(\partial\Omega))\big\},$$

where $\Sigma = (0, T) \times \partial\Omega$. Analogously, we introduce the following spaces of solenoidal vector fields defined on Q:

$$\mathbf{V}^{1,s}(Q) = L^2(0, T; \mathbf{V}^{s+1}(\Omega)) \cap H^1(0, T; \mathbf{V}^s(\Omega)). \tag{2.15}$$

Recall that Ω is the exterior of a bounded domain $B \subset \mathbb{R}^3$. Let $\rho > 0$ be a fixed number satisfying

$$\partial\Omega \subset \{\mathbf{x} \in \mathbb{R}^3 \ : \ |\mathbf{x}| < \rho\}. \tag{2.16}$$

For arbitrary $k \geq 0$, we set

$$\Omega_{\rho+k} = \Omega \cap \{\mathbf{x} \in \mathbb{R}^3 \ : \ |\mathbf{x}| < \rho + k\} \quad \text{and} \quad Q_{\rho+k} = (0,T) \times \Omega_{\rho+k}. \tag{2.17}$$

Let X_1 and X_2 be two Hilbert spaces. Then, the direct sum $X_1 + X_2 = \{x = x_1 + x_2 : x_1 \in X_1, x_2 \in X_2\}$ is also a Hilbert space with norm defined by

$$\|x\|^2_{X_1+X_2} = \inf_{x=x_1+x_2, x_1\in X_1, x_2 \in X_2} \Big(\|x_1\|^2_{X_1} + \|x_2\|^2_{X_2} \Big). \tag{2.18}$$

Evidently, (2.18) defines a Hilbert norm. We now define the space $\mathbf{V}^{1,1/2}_\rho(Q)$, which was introduced in [10]. Let $\rho > 0$ be fixed and satisfy (2.16). Then

$$\begin{aligned} \mathbf{V}^{1,1/2}_\rho(Q) = \big\{ \mathbf{v} \in \mathbf{V}^{1,1/2}(Q) \ : \ \operatorname{supp} \mathbf{v} \subset Q_{\rho+2}, \ \mathbf{v}\big|_{t=0} = \mathbf{0}, \\ \Delta\mathbf{v} \in \mathbf{L}^2(Q_{\rho+2}) + \mathbf{L}^2(0,T; \nabla H^{1/2}(\Omega_{\rho+2})) \big\} \end{aligned} \tag{2.19}$$

equipped with the norm

$$\|\mathbf{v}\|^2_{\mathbf{V}^{1,1/2}_\rho(Q)} = \|\mathbf{v}\|^2_{\mathbf{V}^{1,1/2}(Q)} + \|\Delta\mathbf{v}\|_{\mathbf{L}^2(Q_{\rho+2}) + \mathbf{L}^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2}))},$$

where the space $\nabla H^{1/2}(\Omega_{\rho+2}) = \{\nabla q(\mathbf{x}) : q \in H^{1/2}(\Omega_{\rho+2})\}$ is equipped with the norm

$$\|\nabla q\|_{\nabla H^{1/2}(\Omega_{\rho+2})} \equiv \|\nabla q\|_{(H^{1/2}_{00}(\Omega_{\rho+2}))'}.$$

We recall from [14] that $H^{1/2}_{00}(\Omega_{\rho+2})$ and $[H^{1/2}_{00}(\Omega_{\rho+2})]'$ are defined as follows. Let $r(\mathbf{x}) \in C^\infty(\overline{\Omega_{\rho+2}})$, $r(\mathbf{x}) > 0$ for $\mathbf{x} \in \Omega_{\rho+2}$, and $r(\mathbf{x}) = \operatorname{dist}(\mathbf{x}, \partial\Omega_{\rho+2})$ in a sufficiently small neighborhood of $\partial\Omega_{\rho+2}$. Then,

$$H^{1/2}_{00}(\Omega_{\rho+2}) = \big\{ u \ : \ u \in H^{1/2}(\Omega_{\rho+2}), \ r^{-1/2}u \in L^2(\Omega_{\rho+2}) \big\}$$

with the norm

$$\|u\|^2_{H^{1/2}_{00}(\Omega_{\rho+2})} = \|u\|^2_{H^{1/2}(\Omega_{\rho+2})} + \|r^{-1/2}u\|^2_{L^2(\Omega_{\rho+2})}.$$

By $(H^{1/2}_{00}(\Omega_{\rho+2}))'$ we denote the dual space of $H^{1/2}_{00}(\Omega_{\rho+2})$ with the norm

$$\|f\|_{(H^{1/2}_{00}(\Omega_{\rho+2}))'} = \inf_{\phi \in H^{1/2}_{00}(\Omega_{\rho+2}),\, \phi \neq 0} \frac{\langle f, \phi \rangle}{\|\phi\|_{H^{1/2}_{00}(\Omega_{\rho+2})}},$$

where $\langle \cdot, \cdot \rangle$ denotes the duality generated by the scalar product in $L^2(\Omega_{\rho+2})$.

Following the notation $\mathcal{V}^{(2)}(Q)$ introduced in [8], we define the spaces

$$\begin{aligned} \mathcal{V}^{(2)}_0(Q) = \big\{ \mathbf{v} \in L^2(0,T; \mathbf{V}^2(\Omega)) \ : \\ \partial_t \mathbf{v} \in L^2(0,T; \mathbf{V}^0(\Omega)), \ \mathbf{v}\big|_{t=0} = \mathbf{0}, \ \mathbf{v}\big|_\Sigma = \mathbf{0} \big\} \end{aligned} \tag{2.20}$$

The former is a subspace of the latter with the homogeneous boundary value. Analogous to (2.20), we introduce

$$\mathcal{V}_T^{(2)}(Q) = \big\{\mathbf{v} \in L^2(0,T;\mathbf{V}^2(\Omega)) \ : \ \partial_t \mathbf{v} \in L^2(0,T;\mathbf{V}^0(\Omega)), \mathbf{v}|_{t=T} = \mathbf{0}, \mathbf{v}|_\Sigma = \mathbf{0}\big\}. \tag{2.21}$$

Now, we can define the space in which the solution of the boundary value problem (2.7)–(2.10) is sought:

$$\mathbf{W} \equiv \mathbf{W}(Q) = \big\{\mathbf{v} \in \mathbf{V}_\rho^{1,1/2}(Q) + \mathcal{V}_0^{(2)}(Q) \ : \ \mathbf{v}\big|_\Sigma \in \mathbf{H}^1(\Sigma)\big\}. \tag{2.22}$$

The Sobolev space $\mathbf{H}^1(\Sigma)$ is defined with the help of partition of unity techniques; see [14]. The norm of the space $\mathbf{W}$ is defined by

$$\|\mathbf{v}\|_{\mathbf{W}}^2 = \|\mathbf{v}\|_{\mathbf{V}_\rho^{1,1/2}(Q)+\mathcal{V}_0^{(2)}(Q)}^2 + \|\mathbf{v}\|_{\mathbf{H}^1(\Sigma)}^2.$$

The boundary control will be sought in the following subspace of $\mathbf{H}^1(\Sigma)$ satisfying a compatibility condition at $t = 0$ (see (2.12)):

$$\widehat{\mathbf{H}}^1(\Sigma) = \Big\{\mathbf{v} \in \mathbf{H}^1(\Sigma) \ : \ \mathbf{v}|_{t=0} = \mathbf{0}, \ \int_{\partial\Omega} \mathbf{v}(t,\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\, ds = 0 \text{ a.e. } t \in [0,T]\Big\}. \tag{2.23}$$

We establish a compact embedding result that is used to prove, in Theorem 4.1, the existence of solutions of the optimization problems posed in section 2.5. Let ρ satisfying (2.16) be fixed, let $k > 0$ be arbitrary, and let $Q_{\rho+k}$ be defined as in (2.17). Let the function space $\mathbf{W}(Q_{\rho+k})$ be defined in exactly the same way as $\mathbf{W}(Q) \equiv \mathbf{W}$ (see (2.22)): we simply replace Q and Ω by $Q_{\rho+k}$ and $\Omega_{\rho+k}$, respectively.

LEMMA 2.2. *For each $k \in (0,\infty)$, the embedding $\mathbf{W}(Q_{\rho+k}) \hookrightarrow \mathbf{L}^2(Q_{\rho+k})$ is compact.*

Proof. From definitions (2.19), (2.20), and (2.22) for $\mathbf{W}(Q_{\rho+k})$ and definition (2.15) for $\mathbf{V}^{1,s}(Q_{\rho+k})$ (with Q replaced by $Q_{\rho+k}$ and Ω by $\Omega_{\rho+k}$ in all these relations), we easily conclude that the embedding $\mathbf{W}(Q_{\rho+k}) \hookrightarrow \mathbf{V}^{1,1/2}(Q_{\rho+k})$ is continuous. As is well known (see, e.g., [17, Chap. 4, sect. 3]), the embedding $\mathbf{V}^{1,1/2}(Q_{\rho+k}) \hookrightarrow \mathbf{L}^2(Q_{\rho+k})$ is compact. Combining these two results implies the desired assertion. ∎

2.5. Precise statement of the control problems. We now state precisely the optimal control problems to be studied.

PROBLEM I. *Let $\mathbf{v}_\infty \in \mathbb{R}^3$ and $M > 0$ be given, and suppose that $\mathbf{v}_0$ is constructed as in Proposition* 2.1. *Seek a $(\mathbf{w}, \nabla p) \in \mathbf{W} \times [L^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2})) + \mathbf{L}^2(Q)]$ that minimizes the functional* (2.6) *subject to the constraints* (2.7)–(2.12).

As in [8], we may replace the constraint (2.11) by adding a corresponding penalty term in the functional and consider the following penalized variant of the above optimal control problem.

PROBLEM II. *Let $\mathbf{v}_\infty \in \mathbb{R}^3$ and $N > 0$ be given, and suppose that $\mathbf{v}_0$ is constructed as in Proposition* 2.1. *Seek a $(\mathbf{w}, \nabla p) \in \mathbf{W} \times [L^2(0,T;\nabla H^{1/2}(\Omega_{N+2})) + \mathbf{L}^2(Q)]$ that*

minimizes the functional

$$\begin{aligned}\mathcal{J}_N(\mathbf{w}) &= \int_0^T \int_\Omega \mathcal{D}(\mathbf{w}+\mathbf{v}_0) : \mathcal{D}(\mathbf{w}+\mathbf{v}_0)\, d\mathbf{x}\, dt \\ &+\frac{1}{2}\int_\Omega (|\mathbf{w}(T,\mathbf{x})|^2 + 2[(\mathbf{w}(T,x)\cdot(\mathbf{v}_0(\mathbf{x})-\mathbf{v}_\infty)])\, d\mathbf{x} \\ &+\frac{1}{2}\int_0^T \int_{\partial\Omega} |\mathbf{w}-\mathbf{v}_\infty|^2 \mathbf{w}\cdot\mathbf{n}\, ds\, dt + \frac{N}{2}\int_\Sigma \left(|\partial_t \mathbf{w}|^2 + |\nabla_\tau \mathbf{w}|^2 + |\mathbf{w}|^2\right) ds\, dt\end{aligned} \tag{2.24}$$

subject to the constraints (2.7)–(2.10) *and* (2.12).

DEFINITION 2.3. *An element* $\mathbf{w} \in W$ *is called admissible if it satisfies* (2.7)–(2.12) *in the case of Problem* I *and satisfies* (2.7)–(2.10) *and* (2.12) *in the case of Problem* II. *The set of admissible elements is denoted by* $\mathcal{U}_{\rm ad}$.

DEFINITION 2.4. *An element* $\widehat{\mathbf{w}} \in \mathcal{U}_{\rm ad}$ *is called a solution of Problem* I *if*

$$\mathcal{J}(\widehat{\mathbf{w}}) = \inf_{\mathbf{w}\in\mathcal{U}_{ad}} \mathcal{J}(\mathbf{w}),$$

where $\mathcal{J}$ *is defined by* (2.6). *An element* $\widehat{\mathbf{w}} \in \mathcal{U}_{\rm ad}$ *is called a solution of Problem* II *if*

$$\mathcal{J}_N(\widehat{\mathbf{w}}) = \inf_{\mathbf{w}\in\mathcal{U}_{\rm ad}} \mathcal{J}_N(\mathbf{w}),$$

where $\mathcal{J}_N$ *is defined by* (2.24).

3. Preliminary results. In this section, we collect results that will be needed for the analysis of the optimal control problems defined in section 2.5.

3.1. Boundary value problems with inhomogeneous boundary conditions. To prove the existence of solutions (see section 4) of the optimal control problems stated in section 2.5, we need results regarding the existence and uniqueness of a solution of the boundary value problem (2.7)–(2.10) with an inhomogeneous Dirichlet boundary condition on Σ satisfying the compatibility condition (2.12). We rewrite that boundary value problem in terms of $\mathbf{w}$ as follows:

$$\begin{cases} \partial_t \mathbf{w} - \Delta\mathbf{w} + [(\mathbf{w}+\mathbf{v}_0)\cdot\nabla]\mathbf{w} + (\mathbf{w}\cdot\nabla)\mathbf{v}_0 + \nabla p = \mathbf{0} & \text{in } Q, \\ \operatorname{div}\mathbf{w} = 0 & \text{in } Q, \\ \mathbf{w}|_{t=0} = \mathbf{0} & \text{in } \Omega, \\ \mathbf{w}|_\Sigma = \mathbf{b}, \\ \mathbf{w} \to \mathbf{0} & \text{as } |\mathbf{x}| \to \infty. \end{cases} \tag{3.1}$$

We assume that the Dirichlet boundary data $\mathbf{b} \in \widehat{\mathbf{H}}^1(\Sigma)$.

The solution $(\mathbf{w}, \nabla p)$ of (3.1) is sought in $\mathbf{W} \times [L^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2})) + \mathbf{L}^2(Q)]$. This boundary value problem was analyzed in [10]. The key step was to prove the following result concerning the extension of the Dirichlet boundary condition $\mathbf{b}$ from Σ into Q.

PROPOSITION 3.1. *There exists a continuous extension operator*

$$\mathcal{E} : \widehat{\mathbf{H}}^1(\Sigma) \to \mathbf{V}_\rho^{1,1/2}(Q). \tag{3.2}$$

With the help of Proposition 3.1, the following result was also established in [10].

PROPOSITION 3.2. *Assume that* $\mathbf{b} \in \widehat{\mathbf{H}}^1(\Sigma)$ *and*

$$\|\mathbf{b}\|^2_{\mathbf{H}^1(\Sigma)} \le \epsilon, \tag{3.3}$$

where $\epsilon > 0$ *is sufficiently small. Assume further that* $\mathbf{v}_0 \in \mathbf{C}^\infty(\overline{\Omega})$ *and* $\mathbf{v}_0$ *satisfies* (2.4). *Then, there exists a unique solution* $(\mathbf{w}, \nabla p)$ *of problem* (3.1) *belonging to* $\mathbf{W} \times [L^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2})) + \mathbf{L}^2(Q)]$ *and satisfying the estimate*

$$\|\mathbf{w}\|^2_{\mathbf{W}} + \|\nabla p\|^2_{L^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2}))+\mathbf{L}^2(Q)} \le C(\epsilon), \tag{3.4}$$

where $C(\epsilon)$ *is a positive continuous function defined for all sufficiently small* ϵ.

3.2. The correctness of the functionals (2.6) and (2.24). Here we show that all integrals from (2.6) and (2.24) converge for each $\mathbf{w} \in \mathbf{W}$ satisfying (2.7)–(2.10). Indeed, the convergence of all terms except $\mathbf{w}(T,x) \cdot (\mathbf{v}_0(x) - \mathbf{v}_\infty)$ follow directly from the inclusion $\mathbf{w} \in \mathbf{W}$. Since by (2.4) we have $\mathbf{v}_0 - \mathbf{v}_\infty \in \mathbf{L}^6(\Omega)$, to prove the convergence of this aforementioned term we need to check that $\mathbf{w}(T,x) \in \mathbf{L}^{6/5}(\Omega)$.

LEMMA 3.3. *Let* $\mathbf{w} \in \mathbf{W}$ *satisfy* (2.7)–(2.10). *Then,* $\mathbf{w}(T,x) \in \mathbf{L}^{6/5}(\Omega)$.

Proof. Evidently, it is enough to prove the inclusion $\mathbf{w}(T,x) \in \mathbf{L}^{6/5}(\mathbb{R}^3 \setminus \Omega_{\rho+2})$ where ρ is the number from definitions (2.19) and (2.22). Using the techniques of [10] we extend a solution $\mathbf{w} \in \mathbf{W}$ of (2.7)–(2.10) from $[0,T]\times\mathbb{R}^3\setminus Q$ into a vector field $\widetilde{\mathbf{w}}$ on $[0,T]\times\mathbb{R}^3$ satisfying $\widetilde{\mathbf{w}} \in L^2(0,T;\mathbf{V}^2(\mathbb{R}^3)) \cap H^1(0,T;\mathbf{V}^0(\mathbb{R}^3))$ and $\widetilde{\mathbf{w}}|_{t=0} = 0$. We also extend the gradient ∇p in (2.7)–(2.10) from $[0,T]\times\mathbb{R}^3\setminus Q$ into $\nabla\widetilde{p} \in L^2((0,T)\times\mathbb{R}^3)$, and extend $\mathbf{v}_0(x)$ into a C^∞-vector field on $\mathbb{R}^3$. Substituting $(\widetilde{\mathbf{w}}, \nabla\widetilde{p})$ into the left-hand side of (2.7), we obtain

$$\partial_t\widetilde{\mathbf{w}} - \Delta\widetilde{\mathbf{w}} = -[(\widetilde{\mathbf{w}} + \mathbf{v}_0)\cdot\nabla]\widetilde{\mathbf{w}} - (\widetilde{\mathbf{w}}\cdot\nabla)\mathbf{v}_0 + \nabla\widetilde{p} + \mathbf{g}(t,\mathbf{x}) \quad \text{in } (0,T)\times\mathbb{R}^3, \tag{3.5}$$

$$\operatorname{div}\widetilde{\mathbf{w}} = 0 \quad \text{in } (0,T)\times\mathbb{R}^3, \qquad \widetilde{\mathbf{w}}|_{t=0} = \mathbf{0} \quad \text{in } \mathbb{R}^3, \tag{3.6}$$

where $\mathbf{g}(t,\mathbf{x}) \in L^2((0,T)\times\mathbb{R}^3)$ and $\operatorname{supp}\mathbf{g} \subset Q_{\rho+2}$. Evidently, $\mathbf{g} \in L^{6/5}((0,T)\times\mathbb{R}^3)$.

Recall that if $\Omega \subset \mathbb{R}^3$ is a domain and $Q = [0,T]\times\Omega$, then the Sobolev space $W^{1,2}_q(Q)$ with $1 \le q < \infty$ is defined as follows:

$$W^{1,2}_q(Q) = \Big\{ u(t,x) \in L^q(Q) \,:\, \|u\|^q_{W^{1,2}_q(Q)} \equiv \int_Q |u|^q + |\partial_t u|^q + |\nabla u|^q + \sum_{i,j=1}^{3} \left|\frac{\partial u}{\partial x_i \partial x_j}\right|^q dx < \infty \Big\}. \tag{3.7}$$

By virtue of Sobolev embedding theorem (see [3]), if $1 < p < q < \infty$ then the following embeddings are continuous:

$$W^{1,2}_p(Q) \subset L^q(Q) \quad \text{for } \frac{1}{p} - \frac{1}{q} \le \frac{2}{5} \tag{3.8}$$

and

$$\nabla W^{1,2}_p(Q) := \{\nabla f : f \in W^{1,2}_p(Q)\} \subset L^q(Q) \quad \text{for } \frac{1}{p} - \frac{1}{q} \le \frac{2}{5}. \tag{3.9}$$

Using Holder's inequality, (3.8), (3.9), and (2.4), we obtain

$$\|(\widetilde{\mathbf{w}}\cdot\nabla)\widetilde{\mathbf{w}}\|_{L^{6/5}(Q)}\le\|\widetilde{\mathbf{w}}\|_{\mathbf{L}^3(Q)}\|\nabla\widetilde{\mathbf{w}}\|_{\mathbf{L}^2(Q)}\le\|\widetilde{\mathbf{w}}\|^2_{\mathbf{W}_2^{1,2}(Q)},$$

$$\|(\widetilde{\mathbf{w}}\cdot\nabla)\mathbf{v}_0\|_{L^{6/5}(Q)}\le\|\widetilde{\mathbf{w}}\|_{\mathbf{L}^3(Q)}\|\nabla\mathbf{v}_0\|_{\mathbf{L}^2(Q)}\le c\|\widetilde{\mathbf{w}}\|_{\mathbf{W}_2^{1,2}(Q)},$$

and

$$\|(\mathbf{v}_0\cdot\nabla)\widetilde{\mathbf{w}}\|_{L^{3/2}(Q)}\le\|\mathbf{v}_0\|_{\mathbf{L}^6(Q)}\|\nabla\widetilde{\mathbf{w}}\|_{\mathbf{L}^2(Q)}\le C\|\widetilde{\mathbf{w}}\|_{\mathbf{W}_2^{1,2}(Q)}.$$

Thus, $(\widetilde{\mathbf{w}}\cdot\nabla)\widetilde{\mathbf{w}}+(\widetilde{\mathbf{w}}\cdot\nabla)\mathbf{v}_0\in\mathbf{L}^{6/5}(Q)$ and $(\mathbf{v}_0\cdot\nabla)\widetilde{\mathbf{w}}\in\mathbf{L}^{3/2}(Q)$ so that the right-hand side of (3.5) belongs to $\mathbf{L}^{3/2}([0,T]\times\mathbb{R}^3)$. By virtue of the well-known estimates for solutions of the Cauchy problem for the Stokes equations (see [13, Chap. 4, sect. 6]) we obtain from (3.5) and (3.6) that $\widetilde{\mathbf{w}}\in\mathbf{W}^{1,2}_{3/2}([0,T]\times\mathbb{R}^3)$. Using this inclusion we deduce

$$\|(\mathbf{v}_0\cdot\nabla)\widetilde{\mathbf{w}}\|_{L^{6/5}(Q)}\le\|\mathbf{v}_0\|_{\mathbf{L}^6(Q)}\|\nabla\widetilde{\mathbf{w}}\|_{\mathbf{L}^{3/2}(Q)}\le C\|\widetilde{\mathbf{w}}\|_{\mathbf{W}^{1,2}_{3/2}(Q)}.$$

Therefore, the right-hand side of (3.5) belongs to $\mathbf{L}^{6/5}(Q)$, which implies that $\widetilde{\mathbf{w}}\in\mathbf{W}^{1,2}_{6/5}(Q)$ so that $\widetilde{\mathbf{w}}(T,\cdot)\in\mathbf{L}^{6/5}(\mathbb{R}^3)$. □

3.3. Linearized boundary value problems. To derive and analyze (see section 5) weak formulations of the optimality systems for the control problems, we will need the following theorem concerning the solvability of the (homogeneous) boundary value problem for the Oseen equations

$$\begin{cases}\partial_t\mathbf{h}-\Delta\mathbf{h}+[(\widehat{\mathbf{w}}+\mathbf{v}_0)\cdot\nabla]\mathbf{h}+[\mathbf{h}\cdot\nabla](\widehat{\mathbf{w}}+\mathbf{v}_0)+\nabla q=\mathbf{g} & \text{in } Q,\\ \operatorname{div}\mathbf{h}=0 & \text{in } Q,\\ \mathbf{h}|_{t=0}=\mathbf{0} & \text{in } \Omega,\\ \mathbf{h}|_{\Sigma}=\mathbf{0},\\ \mathbf{h}\to\mathbf{0} & \text{as } |\mathbf{x}|\to\infty\end{cases}\tag{3.10}$$

that are the linearization of (3.1) about a given vector field $\widehat{\mathbf{w}}\in\mathbf{W}$.

PROPOSITION 3.4. *Assume that* $\mathbf{v}_0\in\mathbf{C}^\infty(\overline{\Omega})$ *satisfies* (2.4), $\widehat{\mathbf{w}}\in\mathbf{W}$, *and* $\mathbf{g}\in\mathbf{L}^2(Q)$. *Then, there exists a unique solution* $(\mathbf{h},\nabla q)\in\mathcal{V}_0^{(2)}(Q)\times\mathbf{L}^2(Q)$ *of the problem* (3.10). *Moreover,*

$$\|\mathbf{h}\|^2_{\mathcal{V}_0^{(2)}(Q)}+\|\nabla q\|^2_{\mathbf{L}^2(Q)}\le C\|\mathbf{g}\|^2_{\mathbf{L}^2(Q)}.\tag{3.11}$$

The proof is well known; see, e.g., [5, 10, 13, 16].

3.4. Adjoint boundary value problems. To derive strong forms (see sections 6 and 7) of the optimality systems, we will need results concerning the solvability and regularity of the adjoint boundary value problem for (3.10):

$$\begin{cases}\partial_t\mathbf{q}+\Delta\mathbf{q}+[(\widehat{\mathbf{w}}+\mathbf{v}_0)\cdot\nabla]\mathbf{q}-[\nabla(\widehat{\mathbf{w}}+\mathbf{v}_0)]^*\,\mathbf{q}+\nabla r=\mathbf{k} & \text{in } Q,\\ \operatorname{div}\mathbf{q}=0 & \text{in } Q,\\ \mathbf{q}|_{t=T}=\mathbf{q}_0(\mathbf{x}) & \text{in } \Omega,\\ \mathbf{q}|_{\Sigma}=\mathbf{0},\\ \mathbf{q}\to\mathbf{0} & \text{as } |\mathbf{x}|\to\infty,\end{cases}\tag{3.12}$$

where $[\nabla\widehat{\mathbf{w}}]^*\,\mathbf{q} = (\sum_{i=1}^3 \frac{\partial \widehat{w}_i}{\partial x_j} q_i,\ j = 1,2,3)$ for $\widehat{\mathbf{w}} = (\widehat{w}_1, \widehat{w}_2, \widehat{w}_3)$ and $\mathbf{q} = (q_1, q_2, q_3)$. The following assertion concerning the unique solvability of (3.12) in the case of $\mathbf{q}_0 = \mathbf{0}$ is completely analogous to Proposition 3.4; the proof is also identical to that of the proposition.

PROPOSITION 3.5. *Assume that* $\mathbf{v}_0 \in \mathbf{C}^\infty(\overline{\Omega})$ *satisfies* (2.4), $\widehat{\mathbf{w}} \in \mathbf{W}$, $\mathbf{k} \in \mathbf{L}^2(Q)$, *and* $\mathbf{q}_0 = \mathbf{0}$. *Then, there exists a unique solution* $(\mathbf{q}, \nabla r) \in \mathcal{V}_T^{(2)}(Q) \times \mathbf{L}^2(Q)$ *of the problem* (3.12) (*with* $\mathbf{q}_0 = \mathbf{0}$) *satisfying the estimate*

$$\|\mathbf{q}\|^2_{\mathcal{V}_T^{(2)}(Q)} + \|\nabla r\|^2_{\mathbf{L}^2(Q)} \le C\|\mathbf{k}\|^2_{\mathbf{L}^2(Q)}. \tag{3.13}$$

We will also need results concerning problem (3.12) with $\mathbf{k} = \mathbf{0}$ and $\mathbf{q}_0 \neq \mathbf{0}$. In this case, we reduce (3.12) to the following system through the change of variable $\tau = T - t$ and redenoting τ by t:

$$\begin{cases} \partial_t \mathbf{q} - \Delta\mathbf{q} - [(\widehat{\mathbf{w}} + \mathbf{v}_0)\cdot\nabla]\mathbf{q} + [\nabla(\widehat{\mathbf{w}} + \mathbf{v}_0)]^*\,\mathbf{q} - \nabla r = \mathbf{0} \quad \text{in } Q, \\ \operatorname{div}\mathbf{q} = 0 \quad \text{in } Q, \\ \mathbf{q}|_{t=0} = \mathbf{q}_0(\mathbf{x}) \qquad \text{in } \Omega, \\ \mathbf{q}|_\Sigma = \mathbf{0}, \\ \mathbf{q} \to \mathbf{0} \quad \text{as } |\mathbf{x}| \to \infty. \end{cases} \tag{3.14}$$

Using well-known energy methods (see [13, 16]), we can prove the following result.

LEMMA 3.6. *Assume that* $\mathbf{v}_0 \in \mathbf{C}^\infty(\overline{\Omega})$ *satisfies* (2.4), $\widehat{\mathbf{w}} \in \mathbf{W}$, *and* $\mathbf{q}_0 \in \mathbf{V}_0^0(\Omega)$. *Then there exists a solution* $(\mathbf{q}, \nabla r)$ *of the problem* (3.14) *satisfying the energy estimate*

$$\begin{aligned} &\|\mathbf{q}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\nabla\mathbf{q}\|^2_{\mathbf{L}^2(Q)} \\ &\qquad \le C\big(\|\widehat{\mathbf{w}}\|_{\mathbf{V}^{1,1/2(\Omega)}} + \|\nabla\mathbf{v}_0\|_{\mathbf{C}(\overline{\Omega})}\big)\|\mathbf{q}_0\|^2_{\mathbf{L}^2(\Omega)}, \end{aligned} \tag{3.15}$$

where $C(\gamma)$ *is a positive function increasing in* γ.

Improved estimates for $\mathbf{q}$ when $\mathbf{q}_0$ is smoother will also be needed in order to derive strong forms of the optimality systems; these estimates are obtained in Appendix A.3.

4. Existence of solutions for the optimal control problems.

4.1. The solvability of Problem I. We consider the solvability of Problem I formulated in section 2.5.

THEOREM 4.1. *Suppose that the constant* M *in* (2.11) *is sufficiently small. Then there exists a solution* $(\widehat{\mathbf{w}}, \nabla\widehat{p}) \in \mathbf{W} \times (\mathbf{L}^2(Q) + L^2(0,T;\nabla H^{1/2}(\Omega)))$ *for Problem* I.

Proof. Recall that a pair $(\mathbf{w}, p) \in \mathbf{W} \times \mathbf{L}^2(Q)$ is called admissible if it satisfies (2.7)–(2.12) and the functional (2.6) evaluated at $(\mathbf{w}, p)$ is finite. Evidently, the admissible set $\mathcal{U}_{\rm ad} \neq \emptyset$, as $(\mathbf{w}, p) = (\mathbf{0}, 0) \in \mathcal{U}_{\rm ad}$. Let $\{(\mathbf{w}_n, \nabla p_n)\} \in \mathcal{U}_{\rm ad}$ be a minimizing sequence for the functional $\mathcal{J}(\mathbf{w})$:

$$\lim_{n\to\infty} \mathcal{J}(\mathbf{w}_n) = J_{\min} \equiv \inf_{(\mathbf{w},\nabla p)\in\mathcal{U}_{\rm ad}} \mathcal{J}(\mathbf{w}).$$

By virtue of (2.11) we have $\|\mathbf{w}_n|_\Sigma\|^2_{\mathbf{H}^1(\Sigma)} \le M$. Let $\mathbf{b}_n \equiv \mathbf{w}_n|_\Sigma$; note that $\mathbf{b}_n \in \widehat{\mathbf{H}}^1(\Sigma)$. Consider the boundary value problem (3.1) with $\mathbf{b} = \mathbf{b}_n$. Let $\epsilon > 0$ be a sufficiently

small number determined by Proposition 3.2 and suppose $M < \epsilon$. Proposition 3.2 then implies that

$$\|\mathbf{w}_n\|_{\mathbf{W}}^2 + \|\nabla p_n\|_{L^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2}))+\mathbf{L}^2(Q)} \le C(M), \tag{4.1}$$

where $C(M)$ is a positive constant depending on M. The estimate (4.1) allows us to choose a subsequence of $\{\mathbf{w}_n\}$ (denoted by the same) such that

$$\mathbf{w}_n \rightharpoonup \widehat{\mathbf{w}} \quad \text{weakly in } \mathbf{W}.$$

The definition of $\mathbf{W}$ (see (2.22)) then implies that

$$\mathbf{w}_n|_\Sigma = \mathbf{b}_n \rightharpoonup \widehat{\mathbf{b}} \equiv \widehat{\mathbf{w}}|_\Sigma \quad \text{weakly in } \widehat{\mathbf{H}}^1(\Sigma). \tag{4.2}$$

Since $\mathbf{b}_n$ satisfies (2.11) and the set $\{\mathbf{w} \in \widehat{\mathbf{H}}^1(\Sigma) : \mathbf{w} \text{ satisfies (2.11)}\}$ is convex and closed (hence sequentially weakly closed), we see that $\widehat{\mathbf{b}} \in \widehat{\mathbf{H}}^1(\Sigma)$ and $\mathbf{b}$ satisfies (2.11). $\widehat{\mathbf{w}}$ obviously satisfies (2.8)–(2.10). To prove that $\widehat{\mathbf{w}}$ satisfies (2.7) with some $\nabla\widehat{p} \in L^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2})) + \mathbf{L}^2(Q)$, we proceed by noting that

$$\begin{aligned}&\int_Q (\partial_t \mathbf{w}_n - \Delta \mathbf{w}_n)\cdot\boldsymbol{\phi}\,d\mathbf{x}\,dt\\ &\qquad\to \int_Q (\partial_t \widehat{\mathbf{w}} - \Delta\widehat{\mathbf{w}})\cdot\boldsymbol{\phi}\,d\mathbf{x}\,dt \qquad \forall\,\boldsymbol{\phi} \in L^2(0,T;\mathbf{V}_0^1(\Omega)),\end{aligned} \tag{4.3}$$

where $\mathbf{V}_0^1(\Omega)$ is defined by (A.2). Using Lemma 2.2 with $k = 1,2,3,\dots$, we may choose subsequences $\{\mathbf{w}_{n,k}\}$ converging to $\widehat{\mathbf{w}}$ in $\mathbf{L}^2(Q_{\rho+k})$. Then, by choosing the diagonal subsequence $\{\mathbf{w}_j\}$, we infer that

$$\mathbf{w}_j \to \widehat{\mathbf{w}} \qquad \text{strongly in } \mathbf{L}^2(Q_{\rho+k}) \text{ for each } k = 1,2,3,\dots. \tag{4.4}$$

We now take the $\mathbf{L}^2(Q)$ inner product between an arbitrary $\boldsymbol{\phi} \in L^2(0,T;\mathbf{V}_0^1(\Omega))$ and (2.7) for $(\mathbf{w}_j, \nabla p_j)$. The term involving ∇p_j obviously vanishes. Integrating by parts and passing to the limit with the help of (4.3) and (4.4), we obtain the following equation for $\widehat{\mathbf{w}}$:

$$\begin{aligned}\int_Q \Big(\boldsymbol{\phi}\cdot\partial_t\widehat{\mathbf{w}} + (\nabla\widehat{\mathbf{w}}) : (\nabla\boldsymbol{\phi}) + (\widehat{\mathbf{w}}\cdot\nabla)\mathbf{v}_0\cdot\boldsymbol{\phi} - [(\widehat{\mathbf{w}}+\mathbf{v}_0)\cdot\nabla]\boldsymbol{\phi}\cdot\widehat{\mathbf{w}}\Big)\,d\mathbf{x}\,dt\\ = 0 \qquad \forall\,\boldsymbol{\phi} \in L^2(0,T;\mathbf{V}_0^1(\Omega)).\end{aligned} \tag{4.5}$$

Since $\widehat{\mathbf{w}} \in \mathbf{W}$, we see from definition (2.22) that

$$\widehat{\mathbf{w}} = \mathcal{E}\widehat{\mathbf{b}} + \mathbf{w}, \tag{4.6}$$

where $\mathcal{E}$ is the extension operator of Proposition 3.1, $\mathcal{E}\widehat{\mathbf{b}} \in \mathbf{V}_\rho^{1,1/2}(Q)$, and $\mathbf{w} \in \mathcal{V}_0^{(2)}(Q)$. Substitution of (4.6) into (4.5) yields

$$\begin{aligned}&\int_Q \Big(\partial_t\mathbf{w} - \Delta\mathbf{w} + (\mathbf{w}\cdot\nabla)(\mathbf{v}_0 + \mathcal{E}\widehat{\mathbf{b}})\\ &\qquad\qquad + [(\mathcal{E}\widehat{\mathbf{b}} + \mathbf{v}_0 + \mathbf{w})\cdot\nabla]\mathbf{w} - \mathbf{f}_1 + \nabla p_1\Big)\cdot\boldsymbol{\phi}\,d\mathbf{x}\,dt\\ &\quad= \int_Q \Big(\partial_t\mathbf{w} - \Delta\mathbf{w} + (\mathbf{w}\cdot\nabla)(\mathbf{v}_0 + \mathcal{E}\widehat{\mathbf{b}})\\ &\qquad\qquad + [(\mathcal{E}\widehat{\mathbf{b}} + \mathbf{v}_0 + \mathbf{w})\cdot\nabla]\mathbf{w} - \mathbf{f}_1\Big)\cdot\boldsymbol{\phi}\,d\mathbf{x}\,dt\\ &\quad= 0 \qquad \forall\,\boldsymbol{\phi} \in L^2(0,T;\mathbf{V}_0^1(\Omega)),\end{aligned} \tag{4.7}$$

where $-\mathbf{f}_1 = \partial_t \mathcal{E}\widehat{\mathbf{b}} + \mathbf{g} + (\mathcal{E}\widehat{\mathbf{b}} \cdot \nabla)\mathbf{v}_0 + [(\mathcal{E}\widehat{\mathbf{b}} + \mathbf{v}_0) \cdot \nabla]\mathcal{E}\widehat{\mathbf{b}}$ and $-\Delta\mathcal{E}\widehat{\mathbf{b}} = \mathbf{g} + \nabla p_1$ with $\mathbf{g} \in \mathbf{L}^2(Q)$ and $\nabla p_1 \in L^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2}))$. The inclusions $\mathbf{w} \in \mathcal{V}_0^{(2)}(Q)$ and $\mathcal{E}\widehat{\mathbf{b}} \in \mathbf{V}_\rho^{1,1/2}(Q) \subset \mathbf{V}^{1,1/2}(Q)$ and Sobolev embedding theorems imply that $\mathbf{f}_1 \in \mathbf{L}^2(Q)$ and

$$\partial_t \mathbf{w} - \Delta\mathbf{w} + (\mathbf{w} \cdot \nabla)(\mathbf{v}_0 + \mathcal{E}\widehat{\mathbf{b}}) + [(\mathcal{E}\widehat{\mathbf{b}} + \mathbf{v}_0 + \mathbf{w}) \cdot \nabla]\mathbf{w} \in \mathbf{L}^2(Q).$$

Recalling the Weyl decomposition

$$\mathbf{L}^2(Q) = \mathbf{V}_0^0(\Omega) \oplus \nabla H^1(\Omega),$$

where $\mathbf{V}_0^0(\Omega)$ is defined by (2.14) and

$$\nabla H^1(\Omega) = \{\mathbf{v} \in \mathbf{L}^2(\Omega) \ : \ \mathbf{v} = \nabla p, \ p \in H^1_{\mathrm{loc}}(\Omega)\},$$

we obtain from (4.7) that there exists a $\nabla p \in L^2(0,T;\nabla H^1(\Omega))$ such that

$$\begin{aligned} \partial_t \mathbf{w} - \Delta\mathbf{w} + (\mathbf{w} \cdot \nabla)(\mathbf{v}_0 + \mathcal{E}\widehat{\mathbf{b}}) \\ +[(\mathcal{E}\widehat{\mathbf{b}} + \mathbf{v}_0 + \mathbf{w}) \cdot \nabla]\mathbf{w} + \nabla p - \mathbf{f}_1 = \mathbf{0} \quad \text{in } Q, \end{aligned} \tag{4.8}$$

where (4.8) is understood as an equality in $\mathbf{L}^2(Q)$. Substituting (4.6) into (4.8) yields the equality

$$\partial_t \widehat{\mathbf{w}} - \Delta\widehat{\mathbf{w}} + (\widehat{\mathbf{w}} \cdot \nabla)\mathbf{v}_0 + [(\widehat{\mathbf{w}} + \mathbf{v}_0) \cdot \nabla]\widehat{\mathbf{w}} + \nabla\widehat{p} = 0 \qquad \text{in } Q,$$

where $\nabla\widehat{p} = \nabla p_1 + \nabla p \in L^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2})) + L^2(0,T;\nabla H^1_{\mathrm{loc}}(\Omega))$. Thus, we have proved that $(\widehat{\mathbf{w}}, \nabla\widehat{p})$ satisfies (2.7).

Since $\mathbf{w}_n \rightharpoonup \widehat{\mathbf{w}}$ in $\mathbf{W}$ and the functional

$$\begin{aligned} \mathcal{J}_1(\mathbf{w}) = \int_0^T \int_\Omega \mathcal{D}(\mathbf{w} + \mathbf{v}_0) : \mathcal{D}(\mathbf{w} + \mathbf{v}_0)\, d\mathbf{x}\, dt \\ + \frac{1}{2} \int_\Omega (|\mathbf{w}(T,\mathbf{x})|^2 + 2[(\mathbf{w}(T,x) \cdot (\mathbf{v}_0(\mathbf{x}) - \mathbf{v}_\infty))])\, d\mathbf{x} \end{aligned}$$

is convex (and therefore it is lower semicontinuous with respect to weak convergence), we deduce that

$$\mathcal{J}_1(\widehat{\mathbf{w}}) \le \liminf_{n\to\infty} \mathcal{J}_1(\mathbf{w}_n). \tag{4.9}$$

The facts that Σ is compact and $\dim\Sigma = 3$ allow us to use embedding theorems to deduce that the embedding $\widehat{\mathbf{H}}^1(\Sigma) \hookrightarrow \mathbf{L}^3(\Sigma)$ is compact. Then (4.2) implies $\mathbf{w}|_\Sigma = \mathbf{b}_n \to \widehat{\mathbf{b}} = \widehat{\mathbf{w}}|_\Sigma$ strongly in $\mathbf{L}^3(\Sigma)$. Thus, by defining the functional

$$\mathcal{J}_2(\mathbf{w}) = \frac{1}{2} \int_0^T \int_{\partial\Omega} |\mathbf{w} - \mathbf{v}_\infty|^2 \mathbf{w} \cdot \mathbf{n}\, ds\, dt,$$

we have

$$\mathcal{J}_2(\widehat{\mathbf{w}}) \le \lim_{n\to\infty} \mathcal{J}_2(\mathbf{w}_n). \tag{4.10}$$

The relations (4.9) and (4.10) and the equality $\mathcal{J}(\mathbf{w}) = \mathcal{J}_1(\mathbf{w}) + \mathcal{J}_2(\mathbf{w})$ yield

$$\mathcal{J}(\widehat{\mathbf{w}}) \le \liminf_{n\to\infty} \mathcal{J}(\mathbf{w}_n) = J_{\inf}.$$

Therefore, the pair $(\widehat{\mathbf{w}}, \widehat{p})$ is a solution of Problem I. □

4.2. The solvability of Problem II. We next prove that there exists a solution to Problem II for sufficiently large N, where N is the parameter appearing in the definition of $\mathcal{J}_N(\mathbf{w})$; see (2.24).

THEOREM 4.2. *Suppose that the parameter N of the functional $\mathcal{J}_N(\mathbf{w})$ satisfies $N \geq N_0$ for a sufficiently large, fixed constant $N_0 > 0$. Then there exists a solution $(\widehat{\mathbf{w}}, \widehat{p}) \in \mathbf{W} \times [\mathbf{L}^2(Q) + L^2(0,T; \nabla H^{1/2}(\Omega_{\rho+2}))]$ for Problem* II.

Proof. Recall that a pair $(\mathbf{w}, \nabla p) \in \mathbf{W} \times \mathbf{L}^2(Q)$ is called admissible for Problem II if it satisfies (2.7)–(2.10) and (2.12). We denote by $\mathcal{U}_{\rm ad}^N$ the set of all admissible pairs for Problem II.

Let $\epsilon > 0$ be a sufficiently small number such that for any boundary data $\mathbf{b} \in \widehat{\mathbf{H}}^1(\Sigma)$ satisfying (3.3), the assertions of Proposition 3.2 are true. We fix the constant M in (2.11) as

$$M \in (0, \epsilon). \tag{4.11}$$

Let $(\widehat{\mathbf{w}}, \widehat{p})$ be a solution of Problem I (the existence of such a pair is guaranteed by Theorem 4.1). We define $N_0 > 0$ by the relation

$$\frac{2}{N_0}\mathcal{J}_{N_0}(\widehat{\mathbf{w}}) \equiv \frac{2}{N_0}\mathcal{J}(\widehat{\mathbf{w}}) + R(\widehat{\mathbf{w}}) = \epsilon, \tag{4.12}$$

where the functionals $\mathcal{J}_N$, $\mathcal{J}$, and R are defined by (2.24), (2.6), and (2.11), respectively. The number N_0 satisfying (4.12) is well defined thanks to (4.11) and the estimate $R(\widehat{\mathbf{w}}) \leq M$ for every solution $\widehat{\mathbf{w}}$ of Problem I. Thus, for each $N > N_0$ we have

$$\frac{2}{N}\mathcal{J}_N(\widehat{\mathbf{w}}) \leq \frac{2}{N_0}\mathcal{J}_{N_0}(\widehat{\mathbf{w}}) = \epsilon. \tag{4.13}$$

Set

$$\mathcal{U}_{\rm ad}^{N,\epsilon} = \Big\{(\mathbf{w}, \nabla p) \in \mathcal{U}_{\rm ad}^N \ : \ \frac{1}{N}\mathcal{J}_N(\widehat{\mathbf{w}}) \leq \epsilon\Big\}.$$

By virtue of (4.13), $\mathcal{U}_{\rm ad}^{N,\epsilon}$ is not empty. We now choose a minimizing sequence $\{(\mathbf{w}_n, \nabla p_n)\} \subset \mathcal{U}_{\rm ad}^{N,\epsilon}$ for Problem II:

$$\lim_{n\to\infty} \frac{2}{N}\mathcal{J}_N(\mathbf{w}_n) = J_{N,\inf} \equiv \inf_{(\mathbf{w},\nabla p)\in\mathcal{U}_{\rm ad}^{N,\epsilon}} \frac{2}{N}\mathcal{J}_N(\mathbf{w}).$$

Since $\mathbf{w}_n \in \mathcal{U}_{\rm ad}^{N,\epsilon}$, it follows from (4.13) and (2.6) that for every $N > N_0$,

$$R(\mathbf{w}_n|_\Sigma) \leq \epsilon. \tag{4.14}$$

Denoting $\mathbf{b}_n \equiv \mathbf{w}_n|_\Sigma$ and using (4.14) and (2.11) (the definition of functional R), we see that $\|\mathbf{b}_n\|_{\mathbf{H}^1(\Sigma)} < \epsilon$, i.e., the boundary condition $\mathbf{b}_n$ satisfies the assumptions of Proposition 3.1. Thus, $(\mathbf{w}_n, \nabla p_n)$, being the solution of (3.1) with $\mathbf{b}$ replaced by $\mathbf{b}_n$, satisfies the estimate (3.11) in which $\mathbf{w}$ and p are replaced by $\mathbf{w}_n$ and p_n, respectively. Then, by repeating the relevant segment of the proof of Theorem 4.1, we prove the existence of a solution $(\widehat{\widehat{\mathbf{w}}}, \widehat{\widehat{p}})$ for ProblemII. ∎

5. A weak formulation of an optimality system for Problem II and regularity of the adjoint velocity.

5.1. Abstract Lagrange multiplier principles. We consider an abstract minimization problem. Let X_1 and X_2 be two Banach spaces. Let $f : X_1 \to \mathbb{R}$ and $g : X_1 \to \mathbb{R}$ be functionals and $F : X_1 \to X_2$ be a mapping. We seek a $z \in X_1$ such that

$$f(z) = \inf_{u \in \mathcal{U}_{\rm ad}} f(u), \tag{5.1}$$

where

$$\mathcal{U}_{\rm ad} = \big\{u \in X_1 : F(u) = 0 \text{ and } g(u) \le 0\big\}.$$

The Lagrange functional for the minimization problem (5.1) is defined by

$$\mathcal{L}(z, \lambda_0, \lambda, q) = \lambda_0 f(z) + \langle F(z), q\rangle + \lambda g(z) \tag{5.2}$$

for all $z \in X_1$, $\lambda_0 \in \mathbb{R}$, $\lambda \in \mathbb{R}$, and $q \in X_2^*$, where X_2^* is the dual space of X_2 and $\langle\cdot,\cdot\rangle$ denotes the duality pairing between X_2 and X_2^*. We quote a standard abstract Lagrange principle in the following particular form (see [2]).

THEOREM 5.1. *Let z be a solution of* (5.1). *Assume that the mappings f, g, and F are continuously differentiable and that the image of the operator $F'(z) : X_1 \to X_2$ is closed. Then there exists a $q \in X_2^*$, $\lambda_0 \in \mathbb{R}$, and $\lambda \in \mathbb{R}$ such that the triplet $(q, \lambda_0, \lambda) \neq (0, 0, 0)$ (i.e., the quantities in the triplet do not all vanish simultaneously),*

$$\langle \mathcal{L}_z(z, \lambda_0, \lambda, q), h\rangle = 0 \qquad \forall\, h \in X_1, \tag{5.3}$$

$$\lambda_0 \ge 0, \quad \lambda \ge 0, \quad \textit{and} \quad \lambda g(z) = 0, \tag{5.4}$$

where $\mathcal{L}_z(\cdot,\cdot,\cdot,\cdot)$ denotes the Fréchet derivative of $\mathcal{L}$ with respect to the first argument. Furthermore, if $F'(z) : X_1 \to X_2$ is an epimorphism and the constraint $g(z) \le 0$ is absent in problem (5.1), *then $\lambda_0 \neq 0$ and λ_0 can be taken as* 1.

5.2. The weak formulation of an optimality system. In this subsection we apply Theorem 5.1 to derive a weak form of an optimality system of equations for Problem II by applying a trick employed in [6, Chap. 1, Thm. 1.8] that consists of using the space of variations which does not contain the solution of the considered extreme problem.

In order to apply Theorem 5.1, we first have to define the space $\mathbf{V}_{A^*}(Q)$ in which we search for the adjoint vector field for the optimality system. This space is determined in Appendix B; see (B.14).

THEOREM 5.2. *Assume that $(\widehat{\mathbf{w}}, \nabla\widehat{p}) \in \mathbf{W} \times [\mathbf{L}^2(Q) + L^2(0, T; \nabla H^{1/2}(\Omega_{\rho+2}))]$ is a solution for Problem* II. *Then there exists a $\widehat{\mathbf{q}} \in \mathbf{V}_{A^*}(Q)$ such that*

$$\begin{aligned}
&\int_Q \big\{\partial_t \mathbf{h} - \Delta\mathbf{h} + [(\mathbf{v}_0 + \widehat{\mathbf{w}})\cdot\nabla]\mathbf{h} + (\mathbf{h}\cdot\nabla)(\mathbf{v}_0 + \widehat{\mathbf{w}})\big\}\cdot\widehat{\mathbf{q}}\, d\mathbf{x}\, dt \\
&+2\int_Q \mathcal{D}(\widehat{\mathbf{w}} + \mathbf{v}_0) : \mathcal{D}(\mathbf{h})\, d\mathbf{x}\, dt + \int_\Omega (\widehat{\mathbf{w}}(T, \mathbf{x}) + \mathbf{v}_0(\mathbf{x}) - \mathbf{v}_\infty)\cdot\mathbf{h}\, d\mathbf{x} \\
&+\int_\Sigma \Big(\mathbf{h}\cdot(\widehat{\mathbf{w}} - \mathbf{v}_\infty)\widehat{\mathbf{w}}\cdot\mathbf{n} + \frac{1}{2}|\widehat{\mathbf{w}} - \mathbf{v}_\infty|^2\mathbf{h}\cdot\mathbf{n} \\
&\qquad + N[\partial_t\widehat{\mathbf{w}}\cdot\partial_t\mathbf{h} + \nabla_\tau\widehat{\mathbf{w}} : \nabla_\tau\mathbf{h} + \widehat{\mathbf{w}}\cdot\mathbf{h}]\Big)\, ds\, dt = 0 \qquad \forall\, \mathbf{h} \in \mathbf{V}_A^{1,2}(Q)
\end{aligned} \tag{5.5}$$

Proof. First we convert Problem II into an equivalent optimal control problem through the change of variables

$$\mathbf{w} = \widehat{\mathbf{w}} + \mathbf{z} \quad \text{and} \quad \nabla p = \nabla \widehat{p} + \nabla p_1. \tag{5.6}$$

The optimal control problem for $(\mathbf{z}, \nabla p_1)$ is

$$\mathcal{J}_N(\widehat{\mathbf{w}} + \mathbf{z}) \to \inf \tag{5.7}$$

subject to the constraints

$$\partial_t \mathbf{z} - \Delta \mathbf{z} + [(\mathbf{v}_0 + \widehat{\mathbf{w}} + \mathbf{z}) \cdot \nabla]\mathbf{z} + (\mathbf{z} \cdot \nabla)(\widehat{\mathbf{w}} + \mathbf{v}_0) = -\nabla p_1, \qquad \operatorname{div} \mathbf{z} = 0, \qquad \text{in } Q, \tag{5.8}$$

$$\mathbf{z}|_{t=0} = \mathbf{0} \quad \text{and} \quad \mathbf{z} \to \mathbf{0} \text{ as } |\mathbf{x}| \to \infty. \tag{5.9}$$

We consider the problem (5.7)–(5.9) for $\mathbf{z}$ running through the space $\mathbf{V}^{1,2}_A(Q)$ defined in Appendix B; see (B.21). Evidently, for each $\mathbf{z} \in \mathcal{W}^{(2)}_0(Q)$ (this space is defined in (B.21)), the left-hand side of the first equation in (5.8) belongs to $\mathbf{L}^2(Q)$. Let

$$P : \mathbf{L}^2(Q) \to L^2(0, T; \mathbf{V}^0_0(\Omega))$$

be the projection operator. Then Weyl's decomposition allows us to transform (5.8) into

$$P\Big(\partial_t \mathbf{z} - \Delta \mathbf{z} + [(\mathbf{v}_0 + \widehat{\mathbf{w}} + \mathbf{z}) \cdot \nabla]\mathbf{z} + (\mathbf{z} \cdot \nabla)(\widehat{\mathbf{w}} + \mathbf{v}_0)\Big) = 0 \quad \text{and} \quad \operatorname{div} \mathbf{z} = 0. \tag{5.10}$$

The embedding $\mathcal{W}^{(2)}_0(Q) \hookrightarrow \mathbf{W}$ and the assumption

$$(\widehat{\mathbf{w}}, \widehat{p}) \in \mathbf{W} \times [\mathbf{L}^2(Q) + L^2(0, T; \nabla H^{1/2}(\Omega_{\rho+2}))]$$

being a solution for Problem II imply that $\widehat{\mathbf{z}} \equiv \mathbf{0}$ is a solution of optimal control problem (5.7) and (5.9)–(5.10). We now apply Theorem 5.1 to this control problem. We set $X_1 = \mathbf{V}^{1,2}_A(Q)$ and $X_2 = \mathbf{V}_A(Q)$ (see (B.21) and (B.12)). We define the mappings $f : X_1 \to \mathbb{R}$ and $F : X_1 \to X_2$ as follows:

$$\begin{aligned} &f(\mathbf{z}) = \mathcal{J}_N(\mathbf{z} + \widehat{\mathbf{w}}), \\ &F(\mathbf{z}) = P\Big(\partial_t \mathbf{z} - \Delta \mathbf{z} + [(\mathbf{v}_0 + \widehat{\mathbf{w}} + \mathbf{z}) \cdot \nabla]\mathbf{z} + (\mathbf{z} \cdot \nabla)(\widehat{\mathbf{w}} + \mathbf{v}_0)\Big). \end{aligned} \tag{5.11}$$

Note that the constraint (5.9) is built into the space X_1 and the inequality constraint $g \le 0$ is absent in Problem II. We have to show that f and F defined in (5.11) are continuously differentiable. Here, we will only prove the continuous differentiability of F since this is more difficult than the corresponding property of f. The proof that the operator defined by the left-hand side of (5.8) acts continuously from X_1 to $\mathbf{L}^2(Q)$ is evident. To prove its continuity from X_1 to $\mathbf{L}^{6/5}(Q)$ we have to repeat the proof of Lemma 3.3 for all terms except for $(\mathbf{v}_0 \cdot \nabla)\mathbf{z}$. Using Holder's inequality and interpolation bound, we have

$$\begin{aligned} \|(\mathbf{v}_0 \cdot \nabla)\mathbf{z}\|_{L^{6/5}(Q)} &\le \|\mathbf{v}_0\|_{\mathbf{L}^6(Q)} \|\nabla \mathbf{z}\|_{\mathbf{L}^{3/2}(Q)} \\ &\le \|\mathbf{v}_0\|_{\mathbf{L}^6(Q)} \|\nabla \mathbf{z}\|^{1/2}_{\mathbf{L}^{6/5}(Q)} \|\nabla \mathbf{z}\|^{1/2}_{\mathbf{L}^2(Q)} \le C \|\nabla \mathbf{z}\|_{\mathbf{V}^{1/2}_A(Q)}. \end{aligned}$$

Therefore, the continuity of $F(z)$ from (5.11) is reduced to the proof of the continuity of the projector $P : \mathbf{L}_A(Q) \to \mathbf{V}_A(Q)$ that is actually contained in the proof of decomposition (B.16). Hence we have proved the continuity of $P : \mathbf{L}_A \to \mathbf{V}_A$. Consequently, we have proved the continuity of $F : \mathbf{V}_A^{1,2}(Q) \to \mathbf{V}_A(Q)$.

The derivative of F at the solution $\mathbf{0}$ for the problem (5.7) and (5.9)–(5.10) is given by

$$F'(\mathbf{0})\mathbf{h} = P\Big(\partial_t \mathbf{h} - \Delta \mathbf{h} + [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\mathbf{h} + (\mathbf{h} \cdot \nabla)(\widehat{\mathbf{w}} + \mathbf{v}_0)\Big)$$

and the operator $F'(0) : X_1 \to X_2$ is continuous, which can be proved in a way analogous to the proof of the continuity of operator F. To show $F'(0)$ is surjective, it suffices to prove that for each $\mathbf{f} \in \mathbf{V}_A(Q)$ there exists a solution $\mathbf{h} \in \mathbf{V}_A^{1,2}(Q)$ for the problem

$$\partial_t \mathbf{h} - \Delta \mathbf{h} + [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\mathbf{h} + (\mathbf{h} \cdot \nabla)(\widehat{\mathbf{w}} + \mathbf{v}_0) + \nabla p = \mathbf{f} \tag{5.12}$$

$$\operatorname{div} \mathbf{h} = 0, \quad \mathbf{h}|_{t=0} = \mathbf{0}, \quad \mathbf{h}|_{\Sigma} = \mathbf{0}, \quad \text{and} \quad \mathbf{h} \to \mathbf{0} \text{ as } |\mathbf{x}| \to \infty \tag{5.13}$$

with some $\nabla p \in \mathbf{L}^2(Q) \cap \mathbf{L}^{6/5}(Q)$.

Since $\mathbf{V}_A(Q) \subset \mathbf{L}^2(Q)$, by virtue of Proposition 3.4 there exists a unique solution $\mathbf{h} \in \mathcal{W}_0^{(2)}(Q)$ of (5.12) and (5.13). Moving the last three terms in the left-hand side of (5.12) to the right-hand side and using arguments of the proof of Lemma 3.3, we see that this new right-hand side belongs to $\mathbf{L}^{6/5}(Q)$. Extending $\mathbf{h}$ in (5.12) from $(0,T) \times \Omega$ into $\widetilde{\mathbf{h}} \in \mathcal{W}_0^{(2)}((0,T) \times \mathbb{R}^3)$ and using estimates of solutions of the Cauchy problem for the Stokes equations, we obtain, as in the proof of Lemma 3.3, that $\mathbf{h} \in \mathcal{W}_0^{(2)}((0,T) \times \mathbb{R}^3) \cap \mathbf{W}_{6/5}^{1,2}$. Hence $\widetilde{\mathbf{h}} = \widetilde{\mathbf{h}}|_Q \in \mathbf{V}_A^{1,2}(Q)$.

Hence, we have verified all assumptions of Theorem 5.1 and that theorem implies that there exists a $\hat{\mathbf{q}} \in \mathbf{V}_{A^*}(Q)$ such that (5.3) holds with $\lambda_0 = 1$, λ absent, and

$$\begin{aligned} \mathcal{L}(\mathbf{z}, \lambda_0, \mathbf{q}) &= \mathcal{J}_N(\mathbf{z} + \widehat{\mathbf{w}}) \\ &+ \int_Q \Big(\partial_t \mathbf{z} - \Delta \mathbf{z} + [(\mathbf{v}_0 + \widehat{\mathbf{w}} + \mathbf{z}) \cdot \nabla]\mathbf{z} + (\mathbf{z} \cdot \nabla)(\widehat{\mathbf{w}} + \mathbf{v}_0)\Big) \cdot \hat{\mathbf{q}}\, d\mathbf{x}\, dt. \end{aligned} \tag{5.14}$$

Equation (5.3) with $\mathcal{L}$ defined by (5.14) takes on the form of (5.5). ∎

We express $\widehat{\mathbf{q}}(t, \mathbf{x})$ in the form

$$\widehat{\mathbf{q}}(t, \mathbf{x}) = \mathbf{q}(t, \mathbf{x}) - (\mathbf{v}_0(\mathbf{x}) - \mathbf{v}_\infty), \tag{5.15}$$

where $\mathbf{v}_0(\mathbf{x})$ is the steady state solution from Proposition 2.1 and $\mathbf{v}_\infty \in \mathbb{R}^3$ is the vector from (2.4). By virtue of definitions (B.3), (B.4), and (B.14), the inclusion $\mathbf{v}_0(x) - \mathbf{v}_\infty \in \mathbf{L}^6(Q) \cap \mathbf{C}^\infty(Q)$ implies $\mathbf{v}_0(x) - \mathbf{v}_\infty \in \mathbf{V}_{A^*}(Q)$. Since $\hat{\mathbf{q}} \in \mathbf{V}_{A^*}$, we have $\mathbf{q} \in \mathbf{V}_{A^*}$.

5.3. Regularity of the adjoint velocity in the optimality system. In this subsection we will derive some regularity estimates for the adjoint variable $\mathbf{q}$ defined by Theorem 5.2 and (5.15).

We substitute (5.15) in (5.5) and restrict, in (5.5), $\mathbf{h}$ to $\mathbf{V}_A^{1,2}(Q) \cap \mathcal{V}_0^{(2)}(Q)$ (see (2.20)). Then we obtain

$$\begin{aligned}
&\int_Q \Big[\partial_t \mathbf{h} - \Delta \mathbf{h} + [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\mathbf{h} + (\mathbf{h} \cdot \nabla)(\mathbf{v}_0 + \widehat{\mathbf{w}})\Big] \cdot \Big[\mathbf{q} - (\mathbf{v}_0 - \mathbf{v}_\infty)\Big]\, d\mathbf{x}\, dt \\
&\quad + \int_Q 2\mathcal{D}(\mathbf{h}) : \mathcal{D}(\mathbf{v}_0 + \widehat{\mathbf{w}})\, d\mathbf{x}\, dt \\
&\quad + \int_\Omega \mathbf{h}(T,x) \cdot \big[\widehat{\mathbf{w}}(T,x) + (\mathbf{v}_0(x) - \mathbf{v}_\infty)\big]\, d\mathbf{x} = 0 \quad \forall\, \mathbf{h} \in \mathbf{V}_A^{1,2} \cap \mathcal{V}_0^{(2)}(Q).
\end{aligned} \tag{5.16}$$

The relation (5.16) implies that $\mathbf{q} \in L^2(0,T;\mathbf{V}_0^0(\Omega))$ is the generalized solution of the boundary value problem

$$P\Big(\partial_t \mathbf{q} + \Delta \mathbf{q} + [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\mathbf{q} - [\nabla(\mathbf{v}_0 + \widehat{\mathbf{w}})]^* \mathbf{q}\Big) = P\Phi, \tag{5.17}$$

$$\operatorname{div} \mathbf{q} = 0, \qquad \mathbf{q}\big|_\Sigma = \mathbf{0}, \qquad \mathbf{q} \to \mathbf{0} \text{ as } |\mathbf{x}| \to \infty \tag{5.18}$$

and

$$\mathbf{q}\big|_{t=T} = -P\widehat{\mathbf{w}}(T,\cdot). \tag{5.19}$$

Here

$$\Phi(t,x) = \Delta \mathbf{v}_0 + [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\mathbf{v}_0 - [\nabla(\mathbf{v}_0 + \widehat{\mathbf{w}})]^*(\mathbf{v}_0 - \mathbf{v}_\infty) - \Delta(\mathbf{v}_0 + \widehat{\mathbf{w}}). \tag{5.20}$$

We emphasize that the "initial" condition for $\mathbf{q}$, i.e., the right-hand side of (5.19), does not contain $\mathbf{v}_0 - \mathbf{v}_\infty$ although this term is present in integral over Ω in (5.16). We have the following regularity result for $\mathbf{q}(T,\cdot)$.

LEMMA 5.3. *Let* $\widehat{\mathbf{w}} \in \mathbf{W}$ *be a solution of Problem* I *or* II. *Then* $P\widehat{\mathbf{w}}(T,\cdot) \in \mathbf{H}^{3/4}(\Omega) \cap \mathbf{V}_0^0(\Omega)$, *where* P *is the projection operator defined by* (A.16).

Proof. We consider the operator $P = P_0 : \mathbf{L}^2(\Omega) \to \mathbf{V}_0^0(\Omega)$, which was defined as the Weyl orthogonal projection operator. We recall that for $s \in [0,2]$, $P\mathbf{H}^s(\Omega) \subset \mathbf{H}^s(\Omega)$ and the operator $P : \mathbf{H}^s(\Omega) \to \mathbf{H}^s(\Omega)$ is bounded. Indeed, for each $\mathbf{u} \in \mathbf{L}^2(\Omega)$, $P\mathbf{u} = \mathbf{u} - \nabla p$, where $\nabla p \in G_0 \equiv \{\nabla\phi \in \mathbf{L}^2(\Omega) : \phi \in H^1_{\mathrm{loc}}(\Omega)\}$ is the solution of the variational problem

$$\int_\Omega \nabla p(\mathbf{x}) \cdot \nabla\phi(\mathbf{x})\, d\mathbf{x} = \int_\Omega \mathbf{u} \cdot \nabla\phi(\mathbf{x})\, d\mathbf{x} \quad \forall\, \nabla\phi \in G_0. \tag{5.21}$$

The existence and uniqueness of a solution for this problem is well known (see [13]). Let $\mathbf{u} \in \mathbf{H}^2(\Omega)$. Integration by parts in (5.21) yields that ∇p is the solution of the boundary value problem: $\nabla p \in \mathbf{L}^2(\Omega)$,

$$-\Delta p = \operatorname{div} \mathbf{u} \quad \text{in } \Omega \qquad \text{and} \qquad \frac{\partial p}{\partial n}\Big|_{\partial\Omega} = (\mathbf{u} \cdot \mathbf{n})\big|_{\partial\Omega}.$$

By elliptic regularity and the regularity for div-curl problems (see [16]), we obtain that $p \in H^2_{\mathrm{loc}}(\Omega)$ and

$$\|\nabla p\|_{\mathbf{H}^2(\Omega)} \le C\Big(\|\operatorname{div} \mathbf{u}\|_{H^1(\Omega)} + \|(\mathbf{u} \cdot \mathbf{n})\big|_{\partial\Omega}\|_{H^{3/2}(\partial\Omega)}\Big) \le C\|\mathbf{u}\|_{\mathbf{H}^2(\Omega)}.$$

Thus, we have that the operators $I - P : \mathbf{L}^2(\Omega) \to \mathbf{L}^2(\Omega)$ and $I - P : \mathbf{H}^2(\Omega) \to \mathbf{H}^2(\Omega)$ are bounded. By interpolation theorems, the operator $P : \mathbf{H}^s(\Omega) \to \mathbf{H}^s(\Omega)$ and $I - P : \mathbf{H}^s(\Omega) \to \mathbf{H}^s(\Omega)$ are bounded for each $s \in [0,2]$.

Since $\widehat{\mathbf{w}} \in \mathbf{W} \subset L^2(0,T;\mathbf{H}^{3/2}(\Omega))$, we deduce that

$$P\widehat{\mathbf{w}} \in L^2(0,T;\mathbf{H}^{3/2}(\Omega)). \tag{5.22}$$

Because $\widehat{\mathbf{w}}$ is a solution of Problem I or II, it satisfies (2.7)–(2.10). Integrating (2.7) over $t \in [0,\tau]$ and then applying the operator P, we obtain

$$P\widehat{\mathbf{w}}(\tau,\cdot) = \int_0^\tau \Big(P\Delta\widehat{\mathbf{w}}(t,\cdot) - P[(\widehat{\mathbf{w}}+\mathbf{v}_0)\cdot\nabla]\widehat{\mathbf{w}} - P(\widehat{\mathbf{w}}\cdot\nabla)\mathbf{v}_0 \Big)\, dt. \tag{5.23}$$

Since $\widehat{\mathbf{w}} \in \mathbf{W}$ and therefore $\Delta\widehat{\mathbf{w}} = g + \nabla p$ with $g \in \mathbf{L}^2(Q)$, $\nabla p \in L^2(0,T;\nabla H^{1/2}(\Omega))$, we easily see that

$$P\widehat{\mathbf{w}}(t,\cdot) \in L^2(0,T;\mathbf{V}_0^0(\Omega)) \text{ and } \widehat{\mathbf{w}} \in L^2(0,T;\mathbf{H}^{3/2}(\Omega)) \cap L^\infty(0,T;\mathbf{H}^1(\Omega)).$$

From the last inclusion, the inclusion $\mathbf{v}_0 \in \mathbf{C}^2(\overline{\Omega})$, and Sobolev embedding theorems, we conclude that the integrand from the right-hand side of (5.23) belongs to $L^2(0,T;\mathbf{V}_0^0(\Omega))$. Hence, by differentiating (5.23) with respect to τ we obtain

$$\partial_t P\widehat{\mathbf{w}}(t,\cdot) \in L^2(0,T;\mathbf{V}_0^0(\Omega)) \subset L^2(0,T;\mathbf{L}^2(\Omega)). \tag{5.24}$$

Then (5.22), (5.24), and the trace theorems of [14] imply that

$$P\widehat{\mathbf{w}}(t,\cdot) \in \mathbf{C}(0,T;\mathbf{H}^{3/4}(\Omega)). \qquad \square$$

The spaces $\mathbf{V}_\sigma^s$ used below are defined and studied in Appendix A.1.

THEOREM 5.4. *Assume that* $\mathbf{q} \in \mathbf{V}_{A^*}(Q)$ *satisfies* (5.16). *Then for each* $\delta > 0$, $\mathbf{q} \in L^2(0,T-\delta;\mathbf{V}_\sigma^2(\Omega)) \cap H^1(0,T-\delta;\mathbf{V}_\sigma^0(\Omega))$. *Furthermore, there exists a constant* $C > 0$ *such that for each* $\delta \in [0,T]$ *and each* $\epsilon \in (0,1/2)$, $\mathbf{q}$ *satisfies the estimate*

$$\begin{aligned} &\int_0^{T-\delta} \Big(\|\mathbf{q}(t,\cdot)\|^2_{\mathbf{V}_\sigma^2(\Omega)} + \|\partial_t\mathbf{q}(t,\cdot)\|^2_{\mathbf{V}_\sigma^0(\Omega)} \Big)\, dt \\ &\le C\Big(\delta^{-\epsilon-1/2}\|P\widehat{\mathbf{w}}(T,\cdot)\|^2_{\mathbf{V}_\sigma^{-\epsilon+1/2}(\Omega)} + \|P\Phi\|^2_{L^2(0,T;\mathbf{V}_\sigma^0(\Omega))} \Big). \end{aligned} \tag{5.25}$$

In particular, $\mathbf{q}$ *satisfies* (5.17) *in* $L^2(0,T-\delta;\mathbf{V}_\sigma^0(\Omega))$ *and* (5.18) *for every* $\delta \in (0,T)$ *and satisfies* (5.19) *in the space* $\mathbf{V}_\sigma^{-\epsilon+1/2}(\Omega)$ *for every* $\epsilon \in (0,1/2)$.

Proof. Since $\widehat{\mathbf{w}} \in \mathbf{W}$ and $\mathbf{v}_0 \in C^\infty(\overline{\Omega})$ satisfies (2.4), the vector field Φ defined by (5.20) belongs to $L^2(0,T;\mathbf{V}_0^0(\Omega))$. Therefore problem (5.17)–(5.19) has a solution $\widetilde{\mathbf{q}} \in L^2(0,T;\mathbf{V}_\sigma^1(\Omega))$. This follows directly from Proposition 3.5 and Lemma 3.6. Moreover, since $\Phi \in \mathbf{L}^2(Q)$, Proposition 3.5 and Theorem A.8 imply that $\widetilde{\mathbf{q}} \in L^2(0,T-\delta;\mathbf{V}_\sigma^2(\Omega)) \cap H^1(0,T-\delta;\mathbf{V}_\sigma^0(\Omega))$ for all $\delta \in (0,T)$ and $\widetilde{\mathbf{q}}$ satisfies (5.25).

Multiplying (5.17) by $\mathbf{h} \in \mathcal{V}_0^{(2)}(Q)$, integrating over Q, and performing integration by parts, we see that $\widetilde{\mathbf{q}}$ is a generalized solution of (5.17)–(5.19), i.e., $\widetilde{\mathbf{q}}$ satisfies (5.16).

Now we prove the uniqueness of the generalized solution in the space $\mathbf{V}_{A^*}(Q)$ for (5.17)–(5.19) (recall that the space $\mathbf{V}_{A^*}(\Omega)$ contains $L^2(0,T;\mathbf{V}_0^0(\Omega))$.) Let $\mathbf{q}$ and $\widetilde{\mathbf{q}}$ both belong to $\mathbf{V}_{A^*}(Q)$ and satisfy (5.16). Denote $\mathbf{g} = \mathbf{q} - \widetilde{\mathbf{q}}$. Substituting (5.16) for $\widetilde{\mathbf{q}}$ from (5.16) for $\mathbf{q}$ we obtain

$$\int_Q \Big([\partial_t\mathbf{h} - \Delta\mathbf{h} + [(\mathbf{v}_0+\widehat{\mathbf{w}})\cdot\nabla]\mathbf{h} + (\mathbf{h}\cdot\nabla)(\widehat{\mathbf{w}}+\mathbf{v}_0) \Big) \cdot \mathbf{g}\, d\mathbf{x}\, dt = 0 \tag{5.26}$$

$$\forall\, \mathbf{h} \in \mathbf{V}_A^{1,2}(Q) \cap \mathcal{V}_0^{(2)}(Q).$$

Let us consider the boundary value problem

$$P\big(\partial_t \mathbf{h} - \Delta \mathbf{h} + [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\mathbf{h} + (\mathbf{h} \cdot \nabla)(\widehat{\mathbf{w}} + \mathbf{v}_0\big) = \mathbf{g}_1, \tag{5.27}$$

$$\operatorname{div} \mathbf{h} = 0, \qquad \mathbf{h}\big|_{\Sigma} = \mathbf{0}, \qquad \mathbf{h} \to 0 \text{ as } |\mathbf{x}| \to \infty, \qquad \text{and} \qquad \mathbf{h}\big|_{t=0} = \mathbf{0}, \tag{5.28}$$

where $\mathbf{g}_1 \in \mathbf{V}_A(Q)$. The problem (5.27)–(5.28) is equivalent to the problem (5.12)–(5.13), whose solvability has been established in the proof of Theorem 5.2. Thus, for each $\mathbf{g}_1 \in \mathbf{V}_A(Q)$ there exists the unique solution $\mathbf{h} \in \mathbf{V}_A^{1,2}(Q)$ of (5.27)–(5.28). The spaces $\mathbf{V}_A(Q)$ and $\mathbf{V}_{A^*}(Q)$ are dual, and therefore by a well-known corollary of the Hanh–Banach theorem, for a given $\mathbf{g} \in \mathbf{V}_{A^*}$ there exists a $\mathbf{g}_1 \in \mathbf{V}_A(Q)$ (which we consider here as a functional on $\mathbf{V}_{A^*}(Q)$) such that $\|\mathbf{g}_1\|_{\mathbf{V}_A(Q)} = 1$ and

$$\int_Q \mathbf{g}_1 \cdot \mathbf{g}\, d\mathbf{x}\, dt = \|\mathbf{g}\|_{\mathbf{L}_{A^*}(Q)}. \tag{5.29}$$

If we substitute into (5.26) the solution $\mathbf{h}$ of (5.27)–(5.28), we obtain that the left-hand side of (5.26) is equal to (5.29). Hence, $\mathbf{g} \equiv \mathbf{0}$ and uniqueness is proved. Equality (5.19) is true in the space $\mathbf{V}_\sigma^{-\epsilon+1/2}(\Omega)$ by virtue of Lemmas 5.3, A.4, A.6, and A.7. To summarize, we have proved all the assertions of Theorem 5.4. □

6. The strong form of the optimality system for Problem II. Using the regularity results for the adjoint velocity field established in Theorem 5.4, we now proceed to derive (see Theorem 6.4 below) the optimality system of partial differential equations and boundary, initial, and terminal conditions for Problem II.

6.1. The adjoint pressure. We first establish the existence of an adjoint pressure variable.

LEMMA 6.1. *Let* $\mathbf{q} \in L^2(0.T; \mathbf{V}_0^0(\Omega))$ *be the adjoint variable defined in* (5.15) *by* $\widehat{\mathbf{q}}$ *found in Theorem* 5.2. *Then there exists a distribution* $\widetilde{r}(t, \mathbf{x})$ *on* Q *such that the pair* $(\mathbf{q}, \nabla\widetilde{r})$ *is a solution of the problem* (5.18) *and*

$$\partial_t \mathbf{q} + \Delta \mathbf{q} + [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\mathbf{q} - [\nabla(\mathbf{v}_0 + \widehat{\mathbf{w}})]^* \mathbf{q} + \nabla\widetilde{r} = \Phi, \tag{6.1}$$

where Φ *is defined in* (5.20). *Furthermore,* $\nabla\widetilde{r}$ *has the decomposition*

$$\nabla\widetilde{r} = \nabla r_1 + \nabla r_2 + \nabla r_3, \tag{6.2}$$

with ∇r_i, $i = 1, 2, 3$, *satisfying the estimates*

$$\int_0^{T-\delta} \|\nabla r_1(t, \cdot)\|^2_{\mathbf{L}^2(\Omega)}\, dt \le C\Big(\delta^{-\epsilon-1/2} \|P\widehat{\mathbf{w}}(T, \cdot)\|^2_{\mathbf{V}^{-\epsilon+1/2}(\Omega)} + \|P\Phi\|^2_{L^2(0,T;\mathbf{V}^0_\sigma(\Omega))}\Big), \tag{6.3}$$

$$\operatorname{supp}(\nabla r_2) \in Q_{\rho+2}, \qquad \|\nabla r_2\|_{L^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2}))} \le C\|\widehat{\mathbf{w}}\|_{\mathbf{V}_\rho^{1,1/2}(Q)}, \tag{6.4}$$

and

$$\|\nabla r_3\|_{\mathbf{L}^2(Q)} \le C\|(\Phi + \Delta\widehat{\mathbf{w}}) - \mathbf{w}_2\|_{\mathbf{L}^2(Q)} \le C_1\Big(\|\widehat{\mathbf{w}}\|_{\mathbf{V}_\rho^{1,1/2}(Q)} + \|\Delta\mathbf{v}_0\|_{\mathbf{L}^2(Q)}\Big), \tag{6.5}$$

where C in (6.3) *depends on* $\|\mathbf{v}_0\|_{\mathbf{C}^2(\overline{\Omega})}+\|\widehat{\mathbf{w}}\|_{\mathbf{V}^{1,1/2}(Q)}$, C_1 *in* (6.5) *depends on* $\|\mathbf{v}_0\|_{\mathbf{C}^2(\Omega)}+\|\nabla\mathbf{v}_0\|_{\mathbf{L}^2(\Omega)}+|\mathbf{v}_\infty|$, *and* $\mathbf{w}$ *is defined below in* (6.7).

Proof. First, we claim that (5.17) can be rewritten as (6.1). Indeed, since $\partial_t\mathbf{q}+\Delta\mathbf{q}+[(\mathbf{v}_0+\widehat{\mathbf{w}})\cdot\nabla]\mathbf{q}-[\nabla(\mathbf{v}_0+\widehat{\mathbf{w}})]^*\mathbf{q}\in\mathbf{L}^2((0,T-\delta)\times\Omega)$ for each $\delta\in(0,T)$, we obtain from Weyl's decomposition that there exists a $\nabla r_1\in\mathbf{L}^2((0,T-\delta)\times\Omega)$ for each $\delta\in(0,T)$ such that

$$\partial_t\mathbf{q}+\Delta\mathbf{q}+[(\mathbf{v}_0+\widehat{\mathbf{w}})\cdot\nabla]\mathbf{q}-[\nabla(\mathbf{v}_0+\widehat{\mathbf{w}})]^*\mathbf{q}+\nabla r_1=P\Phi. \tag{6.6}$$

Expressing ∇r_1 through (6.6) and using (5.25) and Sobolev embedding theorems, we deduce estimate (6.3). On the other hand, it follows from the definition of $\mathbf{W}$ and $\mathbf{V}_\rho^{1,1/2}(Q)$ (see (2.22) and (2.19)) and the inclusion $\widehat{\mathbf{w}}\in\mathbf{W}$, that

$$\widehat{\mathbf{w}}=\mathbf{v}_1+\mathbf{v}_2\quad\text{where}\quad\Delta\mathbf{v}_2=\mathbf{w}_2+\nabla r_2 \tag{6.7}$$

and where $\mathbf{v}_1\in\mathcal{V}_0^{(2)}(Q),\mathbf{v}_2\in\mathbf{V}_\rho^{1,1/2}(Q),\mathbf{w}_2\in\mathbf{L}^2(Q_{\rho+2})$, and $\nabla r_2\in L^2(0,T;\nabla H^{1/2}(\Omega_{\rho+2}))$. Evidently, the estimate (6.4) holds. Moreover, $P(\Phi+\Delta\widehat{\mathbf{w}}-\mathbf{w}_2)=\Phi+\Delta\widehat{\mathbf{w}}-\mathbf{w}_2-\nabla r_3$ so that (6.1) and the estimate (6.5) hold. ∎

We next show that the traces of $\partial\widehat{\mathbf{w}}/\partial n$ and r_2 with respect to Σ are well defined.

LEMMA 6.2. *Let r_2 be defined in* (6.7). *The restrictions of r_2 and $\partial\widehat{\mathbf{w}}/\partial n$ on Σ are well defined and*

$$\frac{\partial\widehat{\mathbf{w}}}{\partial n}\Big|_\Sigma\in\mathbf{L}^2(\Sigma)\qquad\text{and}\qquad r_2|_\Sigma\in L^2(\Sigma). \tag{6.8}$$

Proof. To define r_2 from (6.7) uniquely, we assume that $\int_{\Omega_{\rho+2}}r_2(t,\mathbf{x})\,d\mathbf{x}=0$ a.e. $t\in[0,T]$. Since the inclusion $\mathbf{v}_1\in\mathcal{V}_0^{(2)}(Q)$ implies that $\frac{\partial\mathbf{v}_1}{\partial n}|_\Sigma\in L^2(\Sigma)$, the inclusion $\frac{\partial\widehat{\mathbf{w}}}{\partial n}\in L^2(\Sigma)$ follows from (6.7) and $\frac{\partial\mathbf{v}_2}{\partial n}|_\Sigma\in L^2(\Sigma)$.

Using Lemma 4.4 from [9], we introduce coordinates (y_1,y_2,y_3) in $\Omega(\varepsilon)$ (an ε-neighborhood of $\partial\Omega$), where $y_3(\mathbf{x})=\operatorname{dist}(\mathbf{x},\partial\Omega)$ for $\mathbf{x}\in\Omega(\varepsilon)$ and $(y_1,y_2)=(y_1^k(\mathbf{x}),y_2^k(\mathbf{x}))$ are local coordinates in $\mathcal{U}^k$ and $\cup_k\mathcal{U}^k$ is a finite covering of $\partial\Omega$. For each k, the coordinates $(y_1,y_2,y_3)=(y_1^k,y_2^k,y_3)$ are orthogonal, i.e.,

$$\nabla y_i(\mathbf{x})\cdot\nabla y_j(\mathbf{x})=\delta_{ij},\quad i,j=1,2,3.$$

This property implies that if $\mathbf{v}(\mathbf{y})$ is a rewriting of a vector field $\mathbf{w}(\mathbf{x})$ in terms of the coordinates (y_1,y_2,y_3) and $\operatorname{div}\mathbf{w}(\mathbf{x})=\sum_{i=1}^3\partial_{x_i}w_i=0$, then $\operatorname{div}\mathbf{v}(\mathbf{y})=\sum_{i=1}^3\partial_{y_i}v_i=0$; see [4].

Let $\widehat{\mathbf{v}}(t,\mathbf{y})$, $\mathbf{z}(t,\mathbf{y})$, and $\nabla_\mathbf{y}s(t,\mathbf{y})$ be rewritings in terms of the coordinates (y_1,y_2,y_3) of the corresponding vector fields $\mathbf{v}_2(t,\mathbf{x})$, $\mathbf{w}_2(t,\mathbf{x})$, and $\nabla_\mathbf{x}r_2(t,\mathbf{x})$. Then, the decomposition $\Delta\mathbf{v}_2=\mathbf{w}_2+\nabla r_2$ can be rewritten as follows:

$$\Delta_y\widehat{\mathbf{v}}=\mathbf{z}+\sum_{i=1}^3C_i(\mathbf{y})\frac{\partial\widehat{\mathbf{v}}}{\partial y_i}+\nabla_\mathbf{y}s, \tag{6.9}$$

where $C_i(\mathbf{y})$ are certain infinitely smooth vector fields. We denote $Q(\varepsilon)=(0,T)\times\Omega(\varepsilon)$. Taking into account the fact that $\widehat{\mathbf{v}}\in\mathbf{H}^{1,1/2}(Q(\varepsilon))$, which in turn implies

$$\partial\widehat{\mathbf{v}}/\partial y_3\in L^2(0,T;\mathbf{H}^{1/2}(\Omega(\varepsilon)))\cap H^1(0,T;\mathbf{H}^{-1/2}(\Omega(\varepsilon))),$$

we obtain that, for $j=1,2$,

$$\frac{\partial^2\widehat{\mathbf{v}}}{\partial y_3\partial y_j}\in L^2(0,T;\mathbf{L}^2(0,\varepsilon;\mathbf{H}^{-1/2}(\partial\Omega_{(\cdot)}))), \tag{6.10}$$

where $(0,\varepsilon)$ is the interval for the local variables y_3 and $\partial\Omega_{(z)} = \{\mathbf{x} \in \Omega(\varepsilon) : y_3(\mathbf{x}) = z\}$, $z \in (0,\varepsilon)$. Relation (6.10) means that $\partial^2_{y_3y_j}\widehat{\mathbf{v}}(\cdot,y_3) \in L^2(0,T;\mathbf{H}^{-1/2}(\partial\Omega_{(y_3)}))$ for almost all y_3 and the function $y_3 \mapsto \|\partial^2\widehat{\mathbf{v}}/\partial y_3\partial y_j(\cdot,y_3)\|_{L^2(0,T;\mathbf{H}^{-1/2}(\partial\Omega_{(y_3)}))} \in L^2(0,\varepsilon)$. Differentiating the equality $\operatorname{div}\widehat{\mathbf{v}} = 0$ with respect to y_3 and taking into account (6.10), we obtain

$$\frac{\partial^2\widehat{v}_3}{\partial y_3^2} = -\frac{\partial^2\widehat{v}_1}{\partial y_3\partial y_1} - \frac{\partial^2\widehat{v}_2}{\partial y_3\partial y_2} \in L^2(0,\varepsilon;L^2(0,T;\mathbf{H}^{-1/2}(\partial\Omega_{(\cdot)}))). \tag{6.11}$$

Since $\widehat{\mathbf{v}} \in \mathbf{H}^{1,1/2}(Q(\varepsilon)) \subset L^2(0,T;\mathbf{H}^{3/2}(\Omega(\varepsilon)))$, we obtain

$$\widehat{v}_j \in L^2(0,\varepsilon;L^2(0,T;H^{3/2}(\partial\Omega_{(\cdot)}))), \quad j = 1,2,3. \tag{6.12}$$

Define the diffeomorphism $\kappa : \Omega(\varepsilon) \to (0,\varepsilon)\times\partial\Omega$ which transforms each point $\mathbf{x} \in \partial\Omega(\varepsilon)$ possessing coordinates $(y_1^k(\mathbf{x}), y_2^k(\mathbf{x}), y_3(\mathbf{x}))$ into the pair $(y_3(\mathbf{x}),\mathbf{z})$, where $y_3 \in (0,\epsilon)$ and $\mathbf{z} \in \partial\Omega$ is a point possessing coordinates $(y_1(\mathbf{x}), y_2(\mathbf{x}))$. Evidently, under this diffeomorphysm the inclusions (6.11) and (6.12) transform into the inclusions

$$\frac{\partial^2\widehat{v}_3}{\partial y_3^2} \in L^2(0,\varepsilon;L^2(0,T;H^{-1/2}(\partial\Omega))), \quad \widehat{v}_3 \in L^2(0,\varepsilon;L^2(0,T;H^{3/2}(\partial\Omega))). \tag{6.13}$$

The inclusions in (6.13) and trace theorems (see [14]) yield that

$$\frac{\partial\widehat{v}_3}{\partial y_3} \in C(0,\varepsilon;L^2(0,T;L^2(\partial\Omega))) \qquad \text{and} \qquad \frac{\partial\widehat{v}_3}{\partial y_3}\Big|_\Sigma \in L^2(\Sigma). \tag{6.14}$$

Since $\widehat{v}_3 \in H^{1,1/2}(Q(\varepsilon))$ and $z \in \mathbf{L}^2(Q(\varepsilon))$, (6.9) implies that

$$\begin{aligned}\partial_{y_3}\big(\partial_{y_3}\widehat{v}_3 - s\big) &= -\big(\partial^2_{y_1y_1} + \partial^2_{y_2y_2}\big)\widehat{v}_3 - z_3 + \sum_{j=1}^3 C_j^3(\mathbf{y})\frac{\partial\widehat{v}^3}{\partial y_j}\\ &\in L^2(0,\varepsilon;L^2(0,T;H^{-1/2}(\partial\Omega))).\end{aligned} \tag{6.15}$$

Moreover, since $\partial_{y_3}\widehat{v}_3 \in L^2(0,T;H^{1/2}(\Omega(\varepsilon)))$ and $s \in L^2(0,T;H^{1/2}(\Omega(\varepsilon)))$, we have that

$$\partial_{y_3}\widehat{v}_3 - s \in L^2(0,\varepsilon;L^2(0,T;H^{1/2}(\partial\Omega_{(\cdot)}))). \tag{6.16}$$

Applying the diffeomorphysm κ and after using trace theorems, we obtain from (6.15) and (6.16) that

$$\partial_{y_3}\widehat{v}_3 - s \in C(0,\varepsilon;L^2(0,T;L^2(\partial\Omega_{(\cdot)}))) \quad \text{and} \quad (\partial_{y_3}\widehat{v}_3 - s)\big|_\Sigma \in L^2(\Sigma). \tag{6.17}$$

Then (6.14) and (6.17) imply that

$$s|_\Sigma \in L^2(\Sigma). \tag{6.18}$$

Finally, (6.9) implies, for $k = 1,2$,

$$\frac{\partial^2\widehat{v}_k}{\partial y_3^2} = -\left(\frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial y_2^2}\right)\widehat{v}_k - z_3 + \sum_{j=1}^3 C_j^k\frac{\partial\widehat{v}}{\partial y_j} - \frac{\partial s}{\partial y_k} \in L^2(0,\varepsilon;L^2(0,T;H^{-1/2}(\partial\Omega_{(\cdot)}))).$$

This relation together with (6.12) implies, by virtue of trace theorems, that

$$\frac{\partial \widehat{v}_k}{\partial y_3} \in C(0,\varepsilon; L^2(0,T; L^2(\partial\Omega_{(\cdot)}))) \quad \text{and} \quad \frac{\partial \widehat{v}_k}{\partial y_3}\Big|_{\Sigma} \in L^2(\Sigma), \quad k=1,2. \tag{6.19}$$

The relations (6.7), (6.14), (6.18), and (6.19) imply (6.8). □

Note that (6.3), (6.4), and (6.5) imply that $\widetilde{r}(t,\mathbf{x})$ determined in (6.2) has a well-defined trace on Σ and

$$\widetilde{r}|_{\Sigma} \in L^2(\Sigma). \tag{6.20}$$

6.2. An additional condition on the boundary. We now derive an additional condition that holds on the boundary. This is done, roughly speaking, by integration by parts in (5.5).

Let $\widehat{p}$ be the pressure in (2.7), where $\mathbf{w} = \widehat{\mathbf{w}}$. Using (6.7) and Lemma 6.2 and repeating relevant arguments from (6.1) to (6.20), we can derive from (2.7) that $\nabla\widehat{p}$ has a well-defined trace on Σ and

$$\widehat{p}|_{\Sigma} \in L^2(\Sigma). \tag{6.21}$$

Note that through the change of variable

$$\nabla\widetilde{r} = -\nabla r - \nabla\widehat{p}, \tag{6.22}$$

where $\widehat{p}$ is the pressure in the first equation in (2.7), we may convert (6.1) into

$$\partial_t \mathbf{q} + \Delta\mathbf{q} + [(\mathbf{v}_0 + \widehat{\mathbf{w}})\cdot\nabla]\mathbf{q} - [\nabla(\mathbf{v}_0 + \widehat{\mathbf{w}})]^*\mathbf{q} - \nabla r = \Phi + \nabla\widehat{p}. \tag{6.23}$$

By (6.20), (6.21), and (6.22), $r|_{\Sigma}$ is well defined and

$$r|_{\Sigma} \in L^2(\Sigma). \tag{6.24}$$

In order to define r and $\widehat{p}$ uniquely we assume that

$$\int_{\partial\Omega} \widehat{p}(t,\mathbf{x})\, ds = 0 \qquad \text{and} \qquad \int_{\partial\Omega} r(t,\mathbf{x})\, ds = 0 \qquad \text{a.e. } t \in [0,T]. \tag{6.25}$$

LEMMA 6.3. *Let* $\mathbf{q} \in L^2(0.T; \mathbf{V}_0^0(\Omega))$ *be the adjoint variable defined in* (5.15) *by* $\widehat{\mathbf{q}}$ *found in Theorem* 5.2. *Then*

$$\begin{aligned} &\int_{\Sigma} \mathbf{h}\cdot\Big(\mathcal{T}(\widehat{\mathbf{w}},\widehat{p})\mathbf{n} + \mathcal{T}(\mathbf{q},r)\mathbf{n} + 2\mathcal{D}(\mathbf{v}_0)\mathbf{n} + (\widehat{\mathbf{w}} - \mathbf{v}_\infty)\widehat{\mathbf{w}}\cdot\mathbf{n} \\ &+ \frac{1}{2}|\widehat{\mathbf{w}} - \mathbf{v}_\infty|^2\mathbf{n} + N(-\partial_{tt}\widehat{\mathbf{w}} + \widehat{\mathbf{w}})\Big)\, ds\, dt \\ &+ N\int_{\Sigma} \nabla_\tau\widehat{\mathbf{w}} : \nabla_\tau\mathbf{h}\, ds\, dt = 0 \quad \forall\, \mathbf{h} \in \widetilde{\mathcal{V}}_0^{(2)}(Q) \quad \textit{with} \quad \mathbf{h}(T,\cdot) = 0, \end{aligned} \tag{6.26}$$

where

$$\mathcal{T}(\widehat{\mathbf{w}},\widehat{p}) = -\widehat{p}I + 2\mathcal{D}(\widehat{\mathbf{w}}) \qquad \textit{and} \qquad \mathcal{T}(\mathbf{q},r) = -rI + 2\mathcal{D}(\mathbf{q}). \tag{6.27}$$

Proof. We substitute (5.15) into (5.5) and manipulate the resulting expression with $\mathbf{q}$ as follows. We subdivide Q into two disjoint cylinders $Q_{T-\delta} = (0, T-\delta)\times\Omega$ and $Q_\delta = (T-\delta, T)\times\Omega$. Denote $\Sigma_{T-\delta} = (0,T-\delta)\times\partial\Omega$. We express the integral

over Q in (5.5) as the sum of integrals over $Q_{T-\delta}$ and Q_δ and perform integration by parts over $Q_{T-\delta}$. This leads us to

$$
\begin{aligned}
&\int_{Q_{T-\delta}} \Big(- \big[\partial_t \mathbf{q} + \Delta \mathbf{q} + [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\mathbf{q} \\
&\qquad - [\nabla(\mathbf{v}_0 + \widehat{\mathbf{w}})]^* \mathbf{q}\big] \cdot \mathbf{h} \Big)\, d\mathbf{x}\, dt + \int_{Q_{T-\delta}} \mathbf{h} \cdot \Phi \, dx\, dt \\
&\quad + \int_{Q_\delta} \Big(2\mathcal{D}(\widehat{\mathbf{w}} + \mathbf{v}_0) : \mathcal{D}(\mathbf{h}) + \big[\partial_t \mathbf{h} - \Delta \mathbf{h} + [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\mathbf{h} \\
&\qquad + (\mathbf{h} \cdot \nabla)(\mathbf{v}_0 + \widehat{\mathbf{w}})\big] \cdot (\mathbf{q} - \mathbf{v}_0 + \mathbf{v}_\infty) \Big)\, d\mathbf{x}\, dt \\
&\quad + \int_{\Sigma_{T-\delta}} \Big(\mathbf{h} \cdot 2\mathcal{D}(\mathbf{v}_0 + \widehat{\mathbf{w}})\mathbf{n} + \mathbf{h} \cdot \partial_n \mathbf{q} \Big)\, ds\, dt \\
&\quad + \int_\Omega (\mathbf{q}(T - \delta, \mathbf{x}) - \mathbf{v}_0 + \mathbf{v}_\infty) \cdot \mathbf{h}(T - \delta, \mathbf{x})\, d\mathbf{x} \\
&\quad + \int_\Omega (\widehat{\mathbf{w}}(T, \mathbf{x}) + \mathbf{v}_0(\mathbf{x}) - \mathbf{v}_\infty) \cdot \mathbf{h}(T, \mathbf{x})\, d\mathbf{x} \\
&\quad + \int_\Sigma \Big(\mathbf{h} \cdot (\widehat{\mathbf{w}} - \mathbf{v}_\infty)\widehat{\mathbf{w}} \cdot \mathbf{n} + \frac{1}{2}|\widehat{\mathbf{w}} - \mathbf{v}_\infty|^2 \mathbf{h} \cdot \mathbf{n} \\
&\qquad + N[\partial_t \widehat{\mathbf{w}} \cdot \partial_t \mathbf{h} + \nabla_\tau \widehat{\mathbf{w}} : \nabla_\tau \mathbf{h} + \widehat{\mathbf{w}} \cdot \mathbf{h}] \Big)\, ds\, dt = 0 \quad \forall\, \mathbf{h} \in \mathbf{V}_A^{1,2}(Q).
\end{aligned} \tag{6.28}
$$

Note that (6.22) and (6.25) imply

$$
\int_{\partial\Omega} \widetilde{r}(t, \cdot)\, ds = 0 \quad \text{a.e. } t \in (0, T). \tag{6.29}
$$

Expressing $\nabla \widetilde{r}$ by (6.1), we see that the integrals over $Q_{T-\delta}$ in (6.28) are equal to

$$
\int_{Q_{T-\delta}} \mathbf{h} \cdot \nabla \widetilde{r}\, d\mathbf{x}\, dt = \int_{\Sigma_{T-\delta}} \mathbf{h} \cdot \widetilde{r}\, \mathbf{n}\, ds,
$$

where the integration by parts is valid because of (6.8). Thus, (6.28) and the relation $\mathbf{v}_0|_{\partial\Omega} = 0$ reduce to

$$
\begin{aligned}
&\int_{Q_\delta} \Big\{ 2\mathcal{D}(\mathbf{h}) : \mathcal{D}(\mathbf{v}_0 + \widehat{\mathbf{w}}) + [\partial_t \mathbf{h} - \Delta \mathbf{h} + ((\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla)\mathbf{h} \\
&\qquad + (\mathbf{h} \cdot \nabla)(\mathbf{v}_0 + \widehat{\mathbf{w}})] \cdot (\mathbf{q} - \mathbf{v}_0 + \mathbf{v}_\infty) \Big\}\, d\mathbf{x}\, dt \\
&\quad + \int_{\Sigma_{T-\delta}} \mathbf{h} \cdot 2\mathcal{D}(\widehat{\mathbf{w}})\mathbf{n}\, ds\, dt + \int_{\Sigma_{T-\delta}} \mathbf{h} \cdot (\partial_n \mathbf{q} + \widetilde{r}\mathbf{n})\, ds\, dt \\
&\quad + \int_\Omega \Big(\mathbf{h}(T, \mathbf{x}) \cdot (\widehat{\mathbf{w}}(T, \mathbf{x}) + \mathbf{v}_0(\mathbf{x}) - \mathbf{v}_\infty) \\
&\qquad + \mathbf{h}(T - \delta, \mathbf{x}) \cdot (\mathbf{q}(T - \delta, \mathbf{x}) - \mathbf{v}_0 + \mathbf{v}_\infty) \Big)\, d\mathbf{x} \\
&\quad + \int_\Sigma \Big(\mathbf{h} \cdot (\widehat{\mathbf{w}} - \mathbf{v}_\infty)\widehat{\mathbf{w}} \cdot \mathbf{n} + \frac{1}{2}|\widehat{\mathbf{w}} - \mathbf{v}_\infty|^2 \mathbf{h} \cdot \mathbf{n} \Big)\, ds\, dt \\
&\quad + \int_\Sigma \Big(N(\partial_t \mathbf{w} \cdot \partial_t \mathbf{h} - \nabla_\tau \mathbf{w} : \nabla_\tau \mathbf{h} + \widehat{\mathbf{w}} \cdot \mathbf{h}) \Big)\, ds\, dt = 0 \quad \forall\, \mathbf{h} \in \mathbf{V}_A^{1,2}(Q).
\end{aligned} \tag{6.30}
$$

Now we examine passage to the limit in (6.30) as $\delta \to 0$. The integral over Q_δ tends to zero as $\delta \to 0$ since $\mathbf{h} \in \mathbf{V}_A^{1,2}(Q)$, $(q - \mathbf{v}_0 + \mathbf{v}_\infty) \in L_{A^*}(Q)$, and $\text{meas}(Q_\delta) \to 0$. By virtue of Lemma 6.2,

$$\int_{\Sigma_{T-\delta}} \mathbf{h} \cdot 2\mathcal{D}(\widehat{\mathbf{w}})\mathbf{n}\, ds\, dt \to \int_{\Sigma_T} \mathbf{h} \cdot 2\mathcal{D}(\widehat{\mathbf{w}})\mathbf{n}\, ds\, dt \quad \text{as } \delta \to 0.$$

Since $\mathbf{q}$ is a solution of (3.12), by Proposition 3.5 and Lemma 3.6 it satisfies the energy estimate so that $\mathbf{q} \in C([0,T]; \mathbf{L}_w^2(\Omega))$, where $\mathbf{L}_w^2(\Omega)$ is $\mathbf{L}^2(\Omega)$ endowed with the weak topology. Thus, the integral over Ω in (6.30) tends to zero as $\delta \to 0$ thanks to Theorem 5.4 and the fact that $\mathbf{h} \in \mathbf{V}_A^{1,2}(Q) \subset C([0,T]; \mathbf{L}_A(\Omega))$. We also have

$$\int_{\Sigma_{T-\delta}} \mathbf{h} \cdot (\partial_n \mathbf{q} + \widetilde{r}\mathbf{n})\, ds\, dt \to \int_{\Sigma_T} \mathbf{h} \cdot (\partial_n \mathbf{q} + \widetilde{r}\mathbf{n})\, ds\, dt \quad \text{as } \delta \to 0. \tag{6.31}$$

Indeed, (6.30) is valid for all $\delta > 0$ and each term in equality (6.30), except for the term $\int_{\Sigma_{T-\delta}} \mathbf{h} \cdot (\partial_n \mathbf{q} + \widetilde{r}\mathbf{n})\, ds\, dt$, has a limit as $\delta \to 0$. These facts imply that the integral $\int_{\Sigma_{T-\delta}} \mathbf{h} \cdot (\partial_n \mathbf{q} + \widetilde{r}\mathbf{n})\, ds\, dt$ has a limit which, by the definition of improper integrals, equals the right-hand side of (6.31). Hence, passing to limit in (6.30) as $\delta \to 0$ yields

$$\begin{aligned} &\int_\Sigma \Big(\mathbf{h} \cdot 2\mathcal{D}(\widehat{\mathbf{w}} + \mathbf{v}_0)\mathbf{n} + \mathbf{h} \cdot (\partial_n \mathbf{q} + \widetilde{r}\mathbf{n}) + \mathbf{h} \cdot (\widehat{\mathbf{w}} - \mathbf{v}_\infty)\widehat{\mathbf{w}} \cdot \mathbf{n} \\ &\qquad + \frac{1}{2}|\widehat{\mathbf{w}} - \mathbf{v}_\infty|^2 \mathbf{h} \cdot \mathbf{n} + N(\partial_t \widehat{\mathbf{w}} \cdot \partial_t \mathbf{h} + \nabla_\tau \widehat{\mathbf{w}} : \nabla_\tau \mathbf{h} + \widehat{\mathbf{w}} \cdot \mathbf{n}) \Big)\, ds\, dt = 0 \\ &\quad \forall\, \mathbf{h} \in \mathbf{V}_A^{1,2}(Q). \end{aligned} \tag{6.32}$$

We now show that

$$\int_{\partial\Omega} \partial_n \mathbf{q} \cdot \mathbf{h}\, ds = 2 \int_{\partial\Omega} \mathbf{h} \cdot \mathcal{D}(\mathbf{q})\mathbf{n}\, ds. \tag{6.33}$$

For almost every $t \in [0,T]$, we have Green's identity

$$\int_\Omega \nabla \mathbf{q} : \nabla \mathbf{h}\, d\mathbf{x} = \int_{\partial\Omega} \partial_n \mathbf{q} \cdot \mathbf{h}\, ds - \int_\Omega \mathbf{h} \cdot \Delta \mathbf{q}\, d\mathbf{x}. \tag{6.34}$$

Since $\mathbf{q}\big|_{\partial\Omega} = \mathbf{0}$ and $\operatorname{div} \mathbf{h} = 0$, we obtain

$$\int_\Omega \nabla \mathbf{q} : \nabla \mathbf{h}\, d\mathbf{x} = \int_\Omega [\nabla \mathbf{q} + (\nabla \mathbf{q})^T] : \nabla \mathbf{h}\, d\mathbf{x} = 2 \int_\Omega \nabla \mathbf{h} \cdot \mathcal{D}(\mathbf{q})\, d\mathbf{x}$$

so that by applying Green's identity to the last term we have

$$\int_\Omega \nabla \mathbf{q} : \nabla \mathbf{h}\, d\mathbf{x} = 2 \int_{\partial\Omega} \mathbf{h} \cdot \mathcal{D}(\mathbf{q})\mathbf{n}\, ds - \int_\Omega \mathbf{h} \cdot \Delta \mathbf{q}\, d\mathbf{x}. \tag{6.35}$$

Equalities (6.34) and (6.35) imply (6.33).

Taking into account (6.21) and (6.24) and using (6.22) and (6.33), we can rewrite (6.32) as (6.26). ∎

6.3. The optimality system. We are now in a position to derive an optimality system for Problem II in the form of a boundary value problem. The surface Laplacian $\Delta_\tau \mathbf{u}$ defined on $\partial\Omega$ (and on Σ) can be determined by

$$\int_{\partial\Omega} \nabla_\tau \mathbf{u} : \nabla_\tau \mathbf{v}\, ds = -\int_{\partial\Omega} \mathbf{v} \cdot \Delta_\tau \mathbf{u}\, ds, \tag{6.36}$$

where $\mathbf{u}$ and $\mathbf{v}$ are the restrictions of $\mathbf{w}_i \in \mathbf{V}^2(\Omega)$, $i = 1, 2$, onto $\partial\Omega$, i.e., $\mathbf{u} = \mathbf{w}_1\big|_{\partial\Omega}$ and $\mathbf{v} = \mathbf{w}_2\big|_{\partial\Omega}$.

THEOREM 6.4. *Assume* $(\widehat{\mathbf{w}}, \widehat{p}) \in \mathbf{W} \times \mathbf{L}^2(Q)$ *is a solution for Problem* II *and* $\mathbf{q} \in L^2(0, T; \mathbf{V}^1_\sigma(\Omega))$ *is the Lagrange multiplier defined in* (5.15) *by* $\widehat{\mathbf{q}}$ *introduced in Theorem* 5.2. *Then there exists an* r *defined in the form of* (6.22), *with* $\widetilde{r}$ *defined in Lemma* 6.1 *such that the quadruple* $(\widehat{\mathbf{w}}, \widehat{p}, \mathbf{q}, r)$ *satisfies the partial differential equations*

$$\partial_t \widehat{\mathbf{w}} - \Delta \widehat{\mathbf{w}} + [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\widehat{\mathbf{w}} + (\widehat{\mathbf{w}} \cdot \nabla)\mathbf{v}_0 + \nabla \widehat{p} = \mathbf{0} \qquad \text{in } Q, \tag{6.37}$$

$$\operatorname{div} \widehat{\mathbf{w}} = 0 \qquad \text{in } Q, \tag{6.38}$$

$$\begin{aligned} -\partial_t \mathbf{q} - \Delta \mathbf{q} - [(\mathbf{v}_0 + \widehat{\mathbf{w}}) \cdot \nabla]\mathbf{q} + [\nabla(\mathbf{v}_0 + \widehat{\mathbf{w}})]^T \mathbf{q} + \nabla r \\ = \Phi - \nabla \widehat{p} \qquad \text{in } Q, \end{aligned} \tag{6.39}$$

where Φ *is defined by* (5.20) *and*

$$\operatorname{div} \mathbf{q} = 0 \qquad \text{in } Q, \tag{6.40}$$

where $\widehat{p}$ *and* r *satisfy* (6.25). *Furthermore,* $(\widehat{\mathbf{w}}, \mathbf{q})$ *satisfy the initial and terminal conditions*

$$\widehat{\mathbf{w}}(0, \mathbf{x}) = \mathbf{0} \quad \text{in } \Omega \tag{6.41}$$

and

$$\mathbf{q}(T, \mathbf{x}) + \widehat{\mathbf{w}}_\sigma(T, \mathbf{x}) = 0 \quad \text{in } \Omega, \tag{6.42}$$

where, by definition, $\mathbf{v}_\sigma = P\mathbf{v}$ *is the* $\mathbf{V}^0_0(\Omega)$*-projection of* $\mathbf{v}$, *and the lateral boundary conditions*

$$\mathbf{q}\big|_\Sigma = \mathbf{0} \tag{6.43}$$

and

$$\begin{aligned} \big[N(-\partial_{tt}\widehat{\mathbf{w}} - \Delta_\tau \widehat{\mathbf{w}}) + \mathcal{A}(\widehat{\mathbf{w}}) + \mathcal{T}(\mathbf{q}, \tau)\mathbf{n} \\ + \mathcal{T}(\widehat{\mathbf{w}}, \widehat{p})\mathbf{n} + 2\mathcal{D}(\mathbf{v}_0)\mathbf{n}\big]\big|_\Sigma = -\boldsymbol{\eta}(t)\mathbf{n}\big|_\Sigma, \end{aligned} \tag{6.44}$$

where $\mathcal{T}(\widehat{\mathbf{w}}, \widehat{p})$ *and* $\mathcal{T}(\mathbf{q}, \tau)$ *are defined by* (6.27),

$$\mathcal{A}(\widehat{\mathbf{w}}) = N\widehat{\mathbf{w}} + \frac{1}{2}|\widehat{\mathbf{w}} - \mathbf{v}_\infty|^2 \mathbf{n} + (\widehat{\mathbf{w}} - \mathbf{v}_\infty)\widehat{\mathbf{w}} \cdot \mathbf{n}, \tag{6.45}$$

and

$$\boldsymbol{\eta}(t) = \frac{\displaystyle\int_{\partial\Omega} \left[N\Delta_\tau \widehat{\mathbf{w}} - \mathcal{A}(\widehat{\mathbf{w}})\right] \cdot \mathbf{n}\, ds}{\displaystyle\int_{\partial\Omega} ds}. \tag{6.46}$$

Furthermore, the following compatibility conditions hold:

$$\begin{aligned} &\left(\widehat{\mathbf{w}}\big|_\Sigma\right)\Big|_{t=0} = \mathbf{0}, \quad \partial_t\big(\widehat{\mathbf{w}}\big|_\Sigma\big)_\tau\Big|_{t=T} = \mathbf{0}, \\ &\left(\big|(\widehat{\mathbf{w}} - \mathbf{v}_\infty)_\pi\big|_\Sigma + 2N\partial_t\big(\widehat{\mathbf{w}}\big|_\Sigma\big) \cdot \mathbf{n}\right)\Big|_{t=T} = 0, \end{aligned} \tag{6.47}$$

where $(\widehat{\mathbf{w}}|_\Sigma)_\tau$ *is the tangential projection of* $\widehat{\mathbf{w}}|_\Sigma$ *onto* Σ *and* $\mathbf{w}_\pi$ *is the projection of* $\mathbf{w}$ *in the following Weyl decomposition of* $\mathbf{V}^0(\Omega)$*:*

$$\mathbf{w} = \mathbf{w}_\sigma + \nabla w_\pi, \tag{6.48}$$

where $\mathbf{w}_\sigma \in \mathbf{V}_0^0(\Omega)$ *and* $\nabla w_\pi \in \nabla H_\pi(\Omega) \equiv \{\nabla\tau \in \mathbf{L}^2(\Omega) : \tau \in H^1_{\mathrm{loc}}(\Omega), \Delta\tau = 0\}$. *The primitive* w_π *of* ∇w_π *is determined by the equality*

$$\int_{\partial\Omega} w_\pi\, ds = 0. \tag{6.49}$$

Proof. The equations in (6.25) are simply conditions for pinning down $\widehat{p}$ and r uniquely and can always be satisfied by redefining $\widehat{p}$ and r if necessary.

The relations (6.37)–(6.38) and (6.41) were built into the formulation of Problem II and are automatically satisfied. Equalities (6.39)–(6.40) and (6.43) were proved in Theorem 5.4.

The equalities (6.26) and (6.36) yield

$$\begin{aligned} \int_\Sigma \mathbf{h} \cdot \Big(&\mathcal{T}(\widehat{\mathbf{w}}, \widehat{p})\mathbf{n} + \mathcal{T}(\mathbf{q}, r)\mathbf{n} + 2\mathcal{D}(\mathbf{v}_0)\mathbf{n} + \mathcal{A}(\widehat{\mathbf{w}}) \\ &+ \left[N(-\partial_{tt}\widehat{\mathbf{w}} - \Delta_\tau\widehat{\mathbf{w}})\right]\Big)\, ds\, dt = 0 \end{aligned} \tag{6.50}$$

for all $\mathbf{h} \in \mathbf{V}_A^{1,2}(Q)$, $\mathbf{h}(T, \cdot) = 0$, where $\mathcal{T}(\mathbf{w}, \widehat{p})$, $\mathcal{T}(\mathbf{q}, r)$, and $\mathcal{A}(\widehat{\mathbf{w}})$ are defined in (6.27) and (6.45). In [8], it was proved that

$$\int_\Sigma \mathbf{n} \cdot \left(\mathcal{T}(\widehat{\mathbf{w}}, \widehat{p})\mathbf{n} + \mathcal{T}(\mathbf{q}, r)\mathbf{n} + 2\mathcal{D}(\mathbf{v}_0)\mathbf{n} - N\partial_{tt}\widehat{\mathbf{w}}\right) ds\, dt = 0. \tag{6.51}$$

Recall that, by assumption, $\partial\Omega$ is connected set. (In the case of unconnected sets, the proof would be the same, but in the formulas (6.44) and (6.46) we would have to change $\eta(t)$ to $\eta_j(t)$ and $\partial\Omega$, Σ to $\partial\Omega_j$, $\Sigma_j = (0, T) \times \partial\Omega_j$, where $\partial\Omega_j$ is a connected component of $\partial\Omega$.) As is well known (see [9]), a vector field $\mathbf{g}(x'), x' \in \partial\Omega$, is a restriction on $\partial\Omega$ of a solenoidal vector field defined on Ω if and only if $\int_{\partial\Omega} \mathbf{g}\cdot\mathbf{n}\, ds = 0$. Hence, (6.50) and (6.51) imply (6.44) with $\eta(t)$ defined by (6.46).

The relations (6.47) can be proved in the same way as in [8]. ∎

7. The optimality system for Problem I. We now derive the optimality system for Problem I.

THEOREM 7.1. *Assume* $(\widehat{\mathbf{w}},\widehat{p}) \in \mathbf{W}\times\mathbf{L}^2(Q)$ *is a solution of Problem* I. *Then there exists a triple* $(\mathbf{q},\nabla\tau,\lambda)\in L^2(0,T;\mathbf{V}^1_\sigma)(\Omega)\times\mathbf{L}^2(Q)\times\mathbb{R}_+$ *such that* $(\mathbf{q},\nabla\tau,\lambda)\neq(\mathbf{0},\mathbf{0},0)$ *and* $(\widehat{\mathbf{w}},\widehat{p},\mathbf{q},\tau,\lambda)$ *satisfies* (6.37)–(6.43), (6.25), *and the boundary conditions*

$$\lambda(-\partial_{tt}\widehat{\mathbf{w}}-\Delta_\tau\widehat{\mathbf{w}})+\widetilde{\mathcal{A}}(\widehat{\mathbf{w}})+\mathcal{T}(\mathbf{q},\tau)\mathbf{n}+\mathcal{T}(\widehat{\mathbf{w}},\widehat{p})\mathbf{n}+\mathcal{D}(\mathbf{v}_0)\mathbf{n}=-\widetilde{\boldsymbol{\eta}}(t)\mathbf{n}, \tag{7.1}$$

where $\mathcal{T}(\widehat{\mathbf{w}},\widehat{p})$ *and* $\mathcal{T}(\mathbf{q},\tau)$ *are defined by* (6.30),

$$\widetilde{\mathcal{A}}(\widehat{\mathbf{w}})=\lambda\widehat{\mathbf{w}}+\frac{1}{2}|\widehat{\mathbf{w}}-\mathbf{v}_\infty|^2\mathbf{n}+(\widehat{\mathbf{w}}-\mathbf{v}_\infty)\widehat{\mathbf{w}}\cdot\mathbf{n}, \tag{7.2}$$

and

$$\widetilde{\boldsymbol{\eta}}(t)=\frac{\displaystyle\int_{\partial\Omega}\Big(\lambda\Delta_\tau\widehat{\mathbf{w}}-\widetilde{\mathcal{A}}(\widehat{\mathbf{w}})\mathbf{n}\Big)\,ds}{\displaystyle\int_{\partial\Omega}ds}. \tag{7.3}$$

In addition, the following compatibility conditions hold:

$$\lambda\partial_t\big(\widehat{\mathbf{w}}\big|_\Sigma\big)_\tau\Big|_{t=T}=\mathbf{0}\quad\text{and}\quad\Big(\widehat{w}_\pi\big|_\Sigma+2\lambda\partial_t\big(\widehat{\mathbf{w}}\big|_\Sigma\big)\cdot\mathbf{n}\Big)\Big|_{t=T}=0, \tag{7.4}$$

where $\widehat{\mathbf{w}}_\tau$ *is the tangential projection of* $\widehat{\mathbf{w}}$ *onto* Σ *and* $\widehat{w}_\pi$ *is defined by* (6.48) *and* (6.49). *Moreover, the nonnegativity and complementary slackness conditions hold, i.e.,*

$$\lambda\ge 0 \tag{7.5}$$

and

$$\lambda\Big[\int_\Sigma\Big(|\partial_t\widehat{\mathbf{w}}|^2+|\nabla_\tau\widehat{\mathbf{w}}|^2+|\widehat{\mathbf{w}}|^2\Big)\,ds\,dt-M\Big]=0. \tag{7.6}$$

Proof. First, we derive a weak form of optimality system for Problem I in exactly the same manner as was done in the appropriate part of Theorem 6.9 in [8] for the two-dimensional case:

$$\begin{aligned}
&\lambda_0\int_Q\mathcal{D}(\widehat{\mathbf{w}}+\mathbf{v}_0):\mathcal{D}(\mathbf{h})\,d\mathbf{x}\,dt\\
&\quad+\int_Q\big\{\partial_t\mathbf{h}-\Delta\mathbf{h}+[(\mathbf{v}_0+\widehat{\mathbf{w}})\cdot\nabla]\mathbf{h}+(\mathbf{h}\cdot\nabla)(\mathbf{v}_0+\widehat{\mathbf{w}})\big\}\cdot\widehat{\mathbf{q}}\,d\mathbf{x}\,dt\\
&\quad+\int_\Omega(\widehat{\mathbf{w}}(T,\mathbf{x})+\mathbf{v}_0(\mathbf{x})-\mathbf{v}_\infty)\cdot\mathbf{h}\,d\mathbf{x}\\
&\quad+\int_\Sigma\Big(\mathbf{h}\cdot(\widehat{\mathbf{w}}-\mathbf{v}_\infty)\widehat{\mathbf{w}}\cdot\mathbf{n}+\frac{1}{2}|\widehat{\mathbf{w}}-\mathbf{v}_\infty|^2\mathbf{h}\cdot\mathbf{n}\\
&\qquad+\lambda[\partial_t\widehat{\mathbf{w}}\cdot\partial_t\mathbf{h}+\nabla_\tau\widehat{\mathbf{w}}:\nabla_\tau\mathbf{h}+\widehat{\mathbf{w}}\cdot\mathbf{h}]\Big)\,ds\,dt=0\quad\forall\,\mathbf{h}\in\mathbf{V}^{1,2}_A(Q).
\end{aligned} \tag{7.7}$$

The difference between this integral equality and (5.5) is that, in (7.7), the parameter N is renamed to λ and λ_0 multiplies the first term. Therefore, if we repeat all the

arguments that lead to the assertions of Theorem 6.4, we obtain the optimality system (6.37)–(6.49), where N is renamed as λ and all terms generated by the first term in (5.5) are multiplied by λ_0. Hence, to prove Theorem 7.1, we have to show that $\lambda_0 \neq 0$. We do this analogously to the proof of the appropriate part of Theorem 6.9 in [8]. As in that theorem, the assumption that $\lambda_0 = 0$ leads to the inequality $\lambda > 0$. As a result, the proof is reduced to establishing the following assertion:

$$\begin{aligned} &\text{there exists a } \mathbf{y} \in \gamma_\Sigma \mathbf{V}_A^{1,2}(Q) \text{ such that} \\ &\int_\Sigma \Big(\partial_t \widehat{\mathbf{w}} \cdot \partial_t \mathbf{y} + \nabla_\tau \widehat{\mathbf{w}} : \nabla_\tau \mathbf{y} + \widehat{\mathbf{w}} \cdot \mathbf{y} \Big)\, ds\, dt \neq 0. \end{aligned} \tag{7.8}$$

Here, $\gamma_\Sigma \mathbf{V}_A^{1,2}(Q)$ is the set of restrictions of vector fields belonging to the space $\mathbf{V}_A^{1,2}(Q)$ defined in (B.22).

We prove (7.8) by contradiction. By virtue of (2.11), the solution $\widehat{\mathbf{w}}$ of Problem I satisfies $\widehat{\mathbf{w}} \in \mathbf{H}^1(\Sigma)$ and $\int_{\partial\Omega} \widehat{\mathbf{w}} \cdot \mathbf{n}\, ds = 0$ for almost every $t \in (0,T)$. As is well known, there exists a sequence $\mathbf{y}_n(t,x) \in \mathbf{C}^\infty(Q)$, $\int_{\partial\Omega} \mathbf{y}_n \cdot \mathbf{n}\, ds = 0$ a.e. $t \in (0,T)$, such that $\|\widehat{\mathbf{w}} - \mathbf{y}_n\|_{\mathbf{H}^1(\Sigma)} \to 0$ as $n \to 0$. It is well known (and the proof can be easily reproduced, e.g., from [9]) that there exists $\mathbf{z}_n \in \mathbf{V}_A^{1,2}(Q)$ such that $\mathbf{z}_n\big|_\Sigma = \mathbf{y}_n$. Hence, (7.6) and (7.8) imply that

$$\begin{aligned} 0 &= \int_\Sigma \Big(\partial_t \widehat{\mathbf{w}} \cdot \partial_t \mathbf{y}_n + \nabla_\tau \widehat{\mathbf{w}} : \nabla_\tau \mathbf{y}_n + \widehat{\mathbf{w}} \cdot \mathbf{y}_n \Big)\, ds\, dt \\ &\to \int_\Sigma \Big(|\partial_t \widehat{\mathbf{w}}|^2 + |\nabla_\tau \widehat{\mathbf{w}}|^2 + |\widehat{\mathbf{w}}|^2 \Big)\, ds\, dt = M. \end{aligned}$$

This contradiction completes the proof. ∎

Appendix A. Regularity results for some auxiliary boundary value problems. We now derive some nonstandard regularity results for the Stokes problem involving the function space $\mathbf{V}_\sigma^s(\Omega)$. We will also derive similar regularity results for the adjoint boundary value problem in the form (3.14). First, we introduce an appropriate functions spaces.

A.1. Spectral function spaces. In this subsection, we will define the function space $\mathbf{V}_\sigma^s(\Omega)$ and study its properties. The tools we will need are direct integrals and intermediate spaces. These spaces will be used in the study of the regularity for the adjoint velocity in the optimality system.

Consider the operator

$$A = P(-\Delta + I) : \mathbf{V}_0^0(\Omega) \to \mathbf{V}_0^0(\Omega), \tag{A.1}$$

where $P : \mathbf{L}^2(\Omega) \to \mathbf{V}_0^0(\Omega)$ is the orthogonal projection operator onto $\mathbf{V}_0^0(\Omega)$. The domain $\mathcal{D}(A)$ of the operator A is defined by $\mathcal{D}(A) = \mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega)$, where

$$\mathbf{V}_0^1(\Omega) = \{\mathbf{v} \in \mathbf{V}^1(\Omega) \ : \ \mathbf{v}|_{\partial\Omega} = \mathbf{0}\}. \tag{A.2}$$

Applying standard arguments for proving the solvability of the steady-state Stokes equations (see [13]), we can conclude that the operator A is a positive, self-adjoint operator whose spectrum lies in $[1,\infty)$. We wish to use a spectral decomposition theorem for self-adjoint operators. To this end we first recall the concept of a direct integral of Hilbert spaces (see [12] and [14, Chap. 1, sect. 2.3])

$$\Upsilon = \int_1^\infty \oplus H(\lambda)\, d\mu(\lambda). \tag{A.3}$$

Here, $d\mu(\lambda)$ is a nonnegative Borel measure supported in $[1,\infty)$, $H(\lambda)$ is a family of Hilbert spaces with norm and inner product denoted by $\|\cdot\|_{H(\lambda)}$ and $(\cdot,\cdot)_{H(\lambda)}$, respectively, and $H(\lambda)$ is assumed μ-measurable. The μ-measurability of $H(\lambda)$ is defined as follows: there exists a family $\mathcal{M}$ of functions $f\colon [1,\infty)\ni\lambda\mapsto f(\lambda)\in H(\lambda)$ satisfying

1. for all $f\in\mathcal{M}$, the function $\lambda\mapsto\|f(\lambda)\|_{H(\lambda)}$ is μ-measurable;
2. if a function $g:[1,\infty)\ni\lambda\mapsto g(\lambda)\in H(\lambda)$ is such that $\lambda\mapsto\big(f(\lambda),g(\lambda)\big)_{H(\lambda)}$ is μ-measurable for each $f\in\mathcal{M}$, then $g\in\mathcal{M}$; and
3. there exists a sequence $\{f_i\}_{i=1}^{\infty}\subset\mathcal{M}$ such that for every $\lambda\in[1,\infty)$ the set $\{f_i(\lambda)\}_{i=1}^{\infty}$ is dense in $H(\lambda)$.

In other words, $\mathcal{M}$ is the set of μ-measurable functions taking values in $H(\lambda)$. The space (A.3) is the set of functions $f\in\mathcal{M}$ for which

$$\|f\|_{\Upsilon}^2=\int_1^{\infty}\|f(\lambda)\|_{H(\lambda)}^2\,d\mu(\lambda).$$

The scalar product in Υ is defined by the formula

$$(f,g)_{\Upsilon}=\int_1^{\infty}(f(\lambda),g(\lambda))_{H(\lambda)}^2 d\mu(\lambda).$$

In [12], it was proved that Υ is a Hilbert space.

By virtue of the spectral decomposition theorem (see [12]) for the operator (A.1), there exists a direct integral of Hilbert spaces (A.3) and a unitary operator $U:\mathbf{V}_0^0(\Omega)\to\Upsilon$ which maps $\mathcal{D}(A)$ into

$$\Upsilon_A=\{f\in\Upsilon:\lambda f\in\Upsilon\},\qquad \|f\|_{\Upsilon_A}^2=\int_1^{\infty}\lambda^2\|f(\lambda)\|_{H(\lambda)}^2 d\mu(\lambda)$$

and satisfies $U(A\mathbf{v})=\lambda(U\mathbf{v})$. For $s\in[0,2]$, we introduce the spaces

$$\mathbf{V}_\sigma^s(\Omega)=\Big\{\mathbf{v}\in\mathbf{V}_0^0(\Omega)\ :\ U\mathbf{v}=\int_1^{\infty}\oplus\widehat{\mathbf{v}}(\lambda)\,d\mu(\lambda),\ \int_1^{\infty}\lambda^s\|\widehat{\mathbf{v}}(\lambda)\|_{H(\lambda)}^2 d\mu(\lambda)<\infty\Big\}\tag{A.4}$$

with norms

$$\|\mathbf{v}\|_{\mathbf{V}_\sigma^s(\Omega)}^2=\int_1^{\infty}\lambda^s\|\widehat{\mathbf{v}}(\lambda)\|_{H(\lambda)}^2 d\mu(\lambda).\tag{A.5}$$

The definition of the spectral decomposition implies that $\mathbf{V}_\sigma^0(\Omega)=\mathbf{V}_0^0(\Omega)$, $\mathbf{V}_\sigma^2(\Omega)=\mathcal{D}(A)=\mathbf{V}^2(\Omega)\cap\mathbf{V}_0^1(\Omega)$, and, in the spaces $\mathbf{V}_\sigma^s(\Omega)$, the norms (A.5) with $s=0,2$ are equivalent to the norms of $\mathbf{L}^2(\Omega)$ and $\mathbf{H}^2(\Omega)$, respectively. Furthermore, the following equality holds:

$$\mathbf{V}_\sigma^1(\Omega)=\mathbf{V}_0^1(\Omega)\tag{A.6}$$

and, in this space, $\|\cdot\|_{\mathbf{V}_\sigma^1(\Omega)}$ is equivalent to the $\mathbf{H}^1(\Omega)$ norm. Indeed, for each $\mathbf{v}\in\mathcal{D}(A)=\mathbf{V}_\sigma^2(\Omega)$, we have

$$\begin{aligned}\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}^2&=-\int_\Omega\mathbf{v}\cdot\Delta\mathbf{v}\,d\mathbf{x}+\|\mathbf{v}\|_{\mathbf{L}^2(\Omega)}^2\\&=\int_1^{\infty}(\lambda+1)\|\widehat{\mathbf{v}}(\lambda)\|_{H(\lambda)}^2\,d\mu(\lambda)\le 2\|\mathbf{v}\|_{\mathbf{V}_\sigma^1(\Omega)}^2.\end{aligned}$$

Also, we obviously have $\|\mathbf{v}\|_{\mathbf{V}^1_\sigma(\Omega)} \le \|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}$. Hence, the completions of $\mathbf{C}^\infty_0(\Omega) \cap \mathbf{V}^0_0(\Omega)$ under $\|\cdot\|_{\mathbf{H}^1(\Omega)}$ and $\|\cdot\|_{\mathbf{V}^1_\sigma(\Omega)}$ yield the same space, i.e., (A.6) holds.

Let X and Y be Hilbert spaces satisfying

(A.7) $\quad X \subset Y \quad$ and $\quad X$ is dense in Y and is continuously embedded into Y.

Assume that $\Lambda : Y \to Y$ is a positive, self-adjoint linear operator with domain $\mathcal{D}(\Lambda) = X$ and that $\|\cdot\|_X$ is equivalent to the norm $X \ni u \mapsto (\|u\|^2_Y + \|\Lambda u\|^2_Y)^{1/2}$. Recall that the intermediate space $[X, Y]_\theta$, $\theta \in [0, 1]$, is by definition the space $\mathcal{D}(\Lambda^{1-\theta})$ endowed with the norm

$$\|u\|^2_{[X,Y]_\theta} = \|u\|^2_Y + \|\Lambda^{1-\theta} u\|^2_Y \qquad \forall\, u \in [X, Y]_\theta\,;$$

see [14] for details.

LEMMA A.1. *Suppose that X is a closed subspace of $\mathbf{H}^k(\Omega)$ with k being a positive integer, Y is a closed subspace of $\mathbf{L}^2(\Omega)$, $\|\cdot\|_X$ is equivalent to $\|\cdot\|_{\mathbf{H}^k(\Omega)}$, $\|\cdot\|_Y$ is equivalent to $\|\cdot\|_{\mathbf{L}^2(\Omega)}$, and X and Y satisfy* (A.7). *Then, for each $\theta \in [0, 1]$, $[X, Y]_\theta$ is a closed subspace of $\mathbf{H}^{k(1-\theta)}(\Omega)$ and the norm of $[X, Y]_\theta$ is equivalent to the norm of $\mathbf{H}^{k(1-\theta)}(\Omega)$.*

Proof. Recall that, by definition, $\mathbf{H}^{k(1-\theta)}(\Omega) = [\mathbf{H}^k(\Omega), \mathbf{L}^2(\Omega)]_\theta$. Denoting by I the embedding operator we see that $I : X \to \mathbf{H}^k(\Omega)$ and $I : Y \to \mathbf{L}^2(\Omega)$ are continuous. Hence, by interpolation theorems of [14, Chap. 1, sect. 5], we deduce that $I : [X, Y]_\theta \to \mathbf{H}^{k(1-\theta)}(\Omega)$ is also embedding and

$$\|u\|_{\mathbf{H}^{k(1-\theta)}(\Omega)} \le C\|u\|_{[X,Y]_\theta} \qquad \forall\, u \in [X, Y]_\theta. \tag{A.8}$$

We consider now a bounded extension operator $L : \mathbf{L}^2(\Omega) \to \mathbf{L}^2(\mathbb{R}^3)$ (i.e., $Lu(\mathbf{x}) = u(\mathbf{x})$ for every $\mathbf{x} \in \Omega$) such that its restriction on $\mathbf{H}^k(\Omega)$ is the bounded extension operator $L : \mathbf{H}^k(\Omega) \to \mathbf{H}^k(\mathbb{R}^3)$ (such an extension can be easily constructed by the well-known Whitney formula). Clearly, LX is a closed subspace of $\mathbf{H}^k(\mathbb{R}^3)$ and we equip LX with the $\mathbf{H}^k(\mathbb{R}^3)$ norm. Analogously, LY is a closed subspace of $\mathbf{L}^2(\mathbb{R}^3)$ and we equip LY with the $\mathbf{L}^2(\mathbb{R}^3)$ norm. Denote by r the restriction operator which maps any function $f(\mathbf{x})$ defined in $\mathbb{R}^3$ into the same function f restricted to Ω. Evidently, operators $r : LX \to X$ and $r : LY \to Y$ are continuous so that interpolation theorems imply that $r : [LX, LY]_\theta \to [X, Y]_\theta$ is also continuous and

$$\|\mathbf{u}\|_{[X,Y]_\theta} = \|rL\mathbf{u}\|_{[X,Y]_\theta} \le C\|L\mathbf{u}\|_{[LX,LY]_\theta} \qquad \forall\, \mathbf{u} \in [X, Y]_\theta. \tag{A.9}$$

The Sobolev norms on LX and LY can be expressed in terms of Fourier transforms:

$$\|L\mathbf{u}\|^2_{\mathbf{H}^k(\mathbb{R}^3)} = \int_{\mathbb{R}^3} (1 + |\boldsymbol{\xi}|^2)^k |\widehat{L\mathbf{u}}(\boldsymbol{\xi})|^2 \, d\boldsymbol{\xi} \quad \forall\, \mathbf{u} \in LX$$

and

$$\|L\mathbf{u}\|^2_{\mathbf{H}^k(\mathbb{R}^3)} = \int_{\mathbb{R}^3} |\widehat{L\mathbf{u}}(\boldsymbol{\xi})|^2 \, d\boldsymbol{\xi} \quad \forall\, \mathbf{u} \in LY.$$

Thus,

$$\|L\mathbf{u}\|_{[LX,LY]_\theta} = \|L\mathbf{u}\|_{\mathbf{H}^{k(1-\theta)}(\mathbb{R}^3)} \le C\|\mathbf{u}\|_{\mathbf{H}^{k(1-\theta)}(\Omega)} \qquad \forall\, \mathbf{u} \in [X, Y]_\theta, \tag{A.10}$$

where in the last step we used interpolation theorems on the extension operators $L : \mathbf{H}^k(\Omega) \to \mathbf{H}^k(\mathbb{R}^3)$ and $L : \mathbf{L}^2(\Omega) \to \mathbf{L}^2(\mathbb{R}^3)$. Combining (A.8), (A.9), and (A.10), we prove Lemma A.1. ∎

The following lemma is a direct consequence of Lemma A.1 by taking $k = 2$, $X = \mathbf{V}^2_\sigma(\Omega) = \mathbf{V}^2(\Omega) \cap \mathbf{V}^1_0(\Omega)$, and $Y = \mathbf{V}^0_\sigma(\Omega) = \mathbf{V}^0_0(\Omega)$.

LEMMA A.2. *For each $s \in [0,2]$, the norms $\|\cdot\|_{\mathbf{V}^s_\sigma(\Omega)}$ and $\|\cdot\|_{\mathbf{H}^s(\Omega)}$ are equivalent on the space $\mathbf{V}^s_\sigma(\Omega)$.*

Let $G \subset \mathbb{R}^3$ be a bounded domain with a C^∞ boundary ∂G. We consider the problem

$$\mathbf{curl}\,\mathbf{w} = \mathbf{v} \quad \text{in } G \qquad \text{and} \qquad \mathbf{w}\cdot\mathbf{n}\big|_{\partial G} = 0. \tag{A.11}$$

LEMMA A.3. *Assume that $\mathbf{v} \in \mathbf{H}^s(G) \cap \mathbf{V}^0(G)$, $s \geq 0$, and $\mathbf{v}\cdot\mathbf{n}\big|_{\partial\Omega} = 0$. Then there exists a solution $\mathbf{w} \in \mathbf{H}^{s+1}(G)$ for the problem* (A.11) *satisfying the condition $\mathbf{w}\big|_\Gamma = \mathbf{0}$.*

The proof of this lemma is entirely analogous to that of [9, Lem. 4.3] and is omitted here.

LEMMA A.4. *For each $s \in (0, 1/2)$, $\mathbf{V}^s_\sigma(\Omega) = \mathbf{H}^s(\Omega) \cap \mathbf{V}^0_0(\Omega)$.*

Proof. By virtue of Lemma A.2 and the definition of intermediate spaces, we have

$$\mathbf{V}^s_\sigma(\Omega) = \text{closure of } \mathbf{V}^2(\Omega) \cap \mathbf{V}^1_0 \text{ in } \mathbf{H}^s(\Omega). \tag{A.12}$$

Since the topology of $\mathbf{H}^s(\Omega)$ is stronger than the $\mathbf{L}^2(\Omega)$ topology, the right-hand side of (A.12) is a subset of $\mathbf{V}^0_0(\Omega)$, so that $\mathbf{V}^s_\sigma(\Omega) \subset \mathbf{V}^0_0(\Omega) \cap \mathbf{H}^s(\Omega)$. Next we proceed to prove the reverse embedding. Let an arbitrary $\mathbf{v} \in \mathbf{H}^s(\Omega) \cap \mathbf{V}^0_0(\Omega)$ be given, and we choose a sequence $\{\mathbf{v}_n\} \subset \mathbf{V}^2(\Omega) \cap \mathbf{V}^1_0$ such that $\|\mathbf{v}_n - \mathbf{v}\|_{\mathbf{H}^s(\Omega)} \to 0$ as $n \to \infty$. Let $\rho > 0$ be a fixed, sufficiently large number satisfying (2.16). We set

$$\Omega_\rho \equiv \{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{x}| < \rho\} \cap \Omega \qquad \text{and} \qquad \Omega^c_\rho \equiv \mathbb{R}^3 \setminus \Omega_\rho.$$

Firstly, we decompose v as follows:

$$\begin{aligned} v(x) = v_1(\mathbf{x}) + v_2(\mathbf{x}), \quad &\text{where} \quad \operatorname{supp} v_1 \subset \Omega_{\rho+3}, \\ &\operatorname{supp} v_2 \subset \Omega^c_{\rho+2}, \quad v_i \in \mathbf{V}^0_0 \cap \mathbf{H}^s(\Omega),\ i = 1,2. \end{aligned} \tag{A.13}$$

Consider $w(x) \in \mathbf{H}^{s+1}(\Omega_{\rho+3})$ satisfying (A.11) with $G = \Omega_{\rho+3}$. The existence of such a vector field is established, e.g., in [9, 16]. Let $\varphi_1(\mathbf{x}) \in C^\infty(\Omega)$, $\varphi_1(\mathbf{x}) = 1$, for $|\mathbf{x}| < \rho + 1$, $\varphi_1(x) = 0$ for $|\mathbf{x}| > \rho + 2$, and $\varphi_2(\mathbf{x}) = 1 - \varphi_1(\mathbf{x})$. Then the functions

$$\begin{aligned} &v_1(\mathbf{x}) = \mathbf{curl}\,(w(\mathbf{x})\varphi_1(\mathbf{x})), \\ &v_2(\mathbf{x}) = \mathbf{curl}\,(w(x)\varphi_2(\mathbf{x})) \ \text{ for } |\mathbf{x}| < \rho + 3, \qquad v_2(\mathbf{x}) = v(\mathbf{x}) \ \text{ for } |\mathbf{x}| > \rho + 2 \end{aligned}$$

satisfy conditions (A.13).

It is enough to find sequences

$$\{v_{in}\} \subset \mathbf{V}^2(\Omega) \cap \mathbf{V}^1_0(\Omega) \quad \text{such that} \quad \|v_{in} - v_i\|_{H^s(\Omega)} \to 0 \quad \text{as} \quad n \to \infty \tag{A.14}$$

for $i = 1, 2$.

In the case $i = 2$, we can take as v_{2n} the Friedrichs average

$$v_{2n}(\mathbf{x}) = n^3 \int_\Omega j(n(\mathbf{x} - \mathbf{y}))v_2(\mathbf{y})\,d\mathbf{y},$$

where $j(\mathbf{x}) \in C^\infty_0(\mathbb{R}^3)$, $j(\mathbf{x}) \geq 0$, $\operatorname{supp} j \subset \{|\mathbf{x}| \leq 1\}$, and $\int_\Omega j(\mathbf{x})\,d\mathbf{x} = 1$. Evidently, $v_{2n} \in \mathbf{V}^2 \cap \mathbf{V}^1_0(\Omega)$. Since the $v_{2n}(\mathbf{x})$ are well defined for $\mathbf{x} \in \mathbb{R}^3$, the relation (A.14) can be proved in this case with the help of the Fourier transforms.

To prove (A.14) for $i = 1$, we consider (A.11) with $G = \Omega_{\rho+3}$, and $v(\mathbf{x})$ changed to $v_1(\mathbf{x})$. By virtue of Lemma A.3, there exists a solution $w_1(\mathbf{x}) \in H^{s+1}(\Omega_{\rho+3})$ of this problem which satisfies $w_1(\mathbf{x})|_{\partial\Omega_{\rho+3}} = 0$. This equality and the condition $0 < s < 1/2$ yield that $w_1(\mathbf{x}) \in H_0^{s+1}(\Omega_{\rho+3})$. By the definition of the space $H_0^{s+1}(\Omega_{\rho+3})$, there exists a sequence $w_{1n}(x) \in C_0^\infty(\Omega_{\rho+3})$ such that $\|w_{1n} - w_1\|_{H^{1+s}(\Omega_{\rho+3})} \to 3$ as $n \to \infty$. This relation implies that sequence $v_{1n} = \mathbf{curl}\, w_{1n}$ satisfies the relation (A.14). ∎

For $s \in [0,1]$, we may define $\mathbf{V}_\sigma^{-s}(\Omega)$ as the closure of $\mathbf{V}_0^0(\Omega)$ with respect to the norm

$$\|\mathbf{v}\|_{\mathbf{V}_\sigma^{-s}(\Omega)} = \Big(\int_1^\infty \lambda^{-s}\|\widehat{\mathbf{v}}(\lambda)\|^2_{H(\lambda)}\, d\mu(\lambda)\Big)^{1/2} \quad \text{if} \quad U\mathbf{v} = \int_1^\infty \oplus\widehat{\mathbf{v}}(\lambda)\, d\mu(\lambda). \tag{A.15}$$

LEMMA A.5. *For $s \in (0, 1/2)$, $\mathbf{V}_\sigma^{-s}(\Omega) \subset \mathbf{H}^{-s}(\Omega)$. Furthermore, the norm $\|\cdot\|_{\mathbf{V}_\sigma^{-s}(\Omega)}$ is equivalent to $\|\cdot\|_{\mathbf{H}^{-s}(\Omega)}$ on $\mathbf{V}_\sigma^{-s}(\Omega)$.*

This lemma can be proved with the help of Lemmas A.2 and A.4 just as the corresponding assertions in [6, Chap. 3, Lem. 4.5] were proved. We omit the details here.

A.2. Regularity results for the Stokes problem. For $s \in [0,1]$, we may define the operator $P = P_s : \mathbf{H}^{-s}(\Omega) \to \mathbf{V}_\sigma^{-s}(\Omega)$ by the formula

$$\langle P\mathbf{u}, \mathbf{v}\rangle_{\mathbf{V}_0^0(\Omega)} = \langle \mathbf{u}, \mathbf{v}\rangle_{\mathbf{L}^2(\Omega)} \qquad \forall\, \mathbf{v} \in \mathbf{V}_\sigma^s(\Omega), \tag{A.16}$$

where $\langle\cdot,\cdot\rangle_{\mathbf{V}_0^0(\Omega)}$ and $\langle\cdot,\cdot\rangle_{\mathbf{L}^2(\Omega)}$ denote the duality generated by the scalar products on $\mathbf{V}_0^0(\Omega)$ and $\mathbf{L}^2(\Omega)$, respectively. Note that for $s = 0$, P coincides with the orthogonal projection operator from $\mathbf{L}^2(\Omega)$ to $\mathbf{V}_0^0(\Omega)$; see the statement immediately following (A.1).

We first consider the Stokes problem with a vanishing forcing term and an inhomogeneous initial value $\mathbf{q}_0$. Note that $P\nabla r = \mathbf{0}$ for $\nabla r \in \mathbf{H}^{-s}(\Omega)$ and $P\partial_t\mathbf{q} = \partial_t\mathbf{q}$ for $\partial_t\mathbf{q}$ in $L^2(0,T;\mathbf{V}_\sigma^{-s}(\Omega))$; thus, we may write the Stokes problem as

$$\partial_t\mathbf{q}(t,\mathbf{x}) + A\mathbf{q} = \mathbf{0} \quad \text{in } Q \qquad \text{and} \qquad \mathbf{q}|_{t=0} = \mathbf{q}_0 \quad \text{in } \Omega, \tag{A.17}$$

where $A = P(-\Delta + I)$. Since we look for $\mathbf{q}$ in the space $L^2(0,T;\mathbf{V}_\sigma^{-s+2}(\Omega))$ and $-s+2 \ge 1$, a solution $\mathbf{q}$ to (A.17) automatically satisfies

$$\operatorname{div}\mathbf{q} = 0 \qquad \text{and} \qquad \mathbf{q}|_\Sigma = \mathbf{0}. \tag{A.18}$$

As was mentioned in Lemma 3.6, the existence and uniqueness of a solution $\mathbf{q} \in L^2(0,T;\mathbf{V}_\sigma^1(\Omega)) \cap H^1(0,T;\mathbf{V}_\sigma^{-1}(\Omega))$ to (A.17) is well known. The following regularity result holds.

LEMMA A.6. *Assume that $\mathbf{q}_0 \in \mathbf{V}_\sigma^s(\Omega)$, $s \ge 0$. Then a solution $\mathbf{q}$ of the problem* (A.17) *satisfies the estimate*

$$\begin{aligned}\int_t^\infty \Big(\|\mathbf{q}(\tau,\cdot)\|^2_{\mathbf{V}_\sigma^{\nu+s+1}(\Omega)} + \|\partial_t\mathbf{q}(\tau,\cdot)\|^2_{\mathbf{V}_\sigma^{\nu+s-1}(\Omega)}\Big)\, d\tau \\ = \|\mathbf{q}(t,\cdot)\|^2_{\mathbf{V}_\sigma^{\nu+s}(\Omega)} \le e^{-\nu}\Big(\frac{\nu}{2}\Big)^\nu t^{-\nu}\|\mathbf{q}_0\|^2_{\mathbf{V}_\sigma^s(\Omega)}\end{aligned} \tag{A.19}$$

for every $\nu \ge 0$.

Proof. We apply to (A.17) the unitary operator U introduced in the definition of the spectral decomposition of A as direct integrals. Then (A.17) reduces to

$$\int_1^\infty \oplus\Big(\partial_t\widehat{\mathbf{q}}(t,\lambda) + \lambda\widehat{\mathbf{q}}(t,\lambda)\Big)\, d\mu(\lambda) = 0$$

and

$$\int_1^\infty \oplus\Big(\widehat{\mathbf{q}}(t,\lambda)|_{t=0} - \widehat{\mathbf{q}}_0(\lambda)\Big)\, d\mu(\lambda) = 0.$$

By the definition of direct integrals, the integrands belong to different Hilbert spaces $H(\lambda)$ for different λ. This allows us to deduce that for almost every λ (with respect to the μ-measure),

$$\partial_t \widehat{\mathbf{q}}(t,\lambda) + \lambda \widehat{\mathbf{q}}(t,\lambda) = \mathbf{0} \quad \text{in } Q \qquad \text{and} \qquad \widehat{\mathbf{q}}(t,\lambda)|_{t=0} = \widehat{\mathbf{q}}_0(\lambda) \quad \text{in } \Omega. \tag{A.20}$$

Equalities (A.20) are understood as equalities in the Hilbert spaces $H(\lambda)$. More precisely, the first equality is considered in $L^2(0,T;H(\lambda))$. Equalities (A.20) imply

$$\widehat{\mathbf{q}}(t,\lambda) = e^{-\lambda t}\widehat{\mathbf{q}}_0(\lambda). \tag{A.21}$$

Using definitions (A.5) and (A.15) of the norms for $\mathbf{V}^s_\sigma(\Omega)$, we obtain

$$\|\mathbf{q}(t,\cdot)\|^2_{\mathbf{V}^{\nu+s}_\sigma(\Omega)} = \int_1^\infty e^{-2\lambda t}\lambda^{\nu+s}\|\widehat{\mathbf{q}}_0(\lambda)\|^2_{H(\lambda)}\, d\mu(\lambda). \tag{A.22}$$

By solving a simple extremal problem we obtain that, for each $t > 0$,

$$\max_{\lambda\in[1,\infty)} (e^{-2\lambda t}\lambda^\nu) = e^{-\max\{2t,\nu\}}\big[\max\{1,\nu/(2t)\}\big]^\nu \le e^{-\nu}\Big(\frac{\nu}{2}\Big)^\nu t^{-\nu}. \tag{A.23}$$

From (A.22) and (A.23) we deduce that

$$\|\mathbf{q}(t,\cdot)\|^2_{\mathbf{V}^{\nu+s}_\sigma(\Omega)} \le e^{-\nu}\Big(\frac{\nu}{2}\Big)^\nu t^{-\nu}\|\mathbf{q}_0\|^2_{\mathbf{V}^s_\sigma(\Omega)}. \tag{A.24}$$

Integrating (A.22) in t and applying (A.22) to the result, we have

$$\begin{aligned}\int_t^\infty &\|\mathbf{q}(\tau,\cdot)\|^2_{\mathbf{V}^{\nu+s+1}_\sigma(\Omega)}\, d\tau \\ &= \frac{1}{2}\int_1^\infty e^{-2\lambda t}\lambda^{\nu+s}\|\widehat{\mathbf{q}}_0(\lambda)\|^2_{H(\lambda)}\, d\mu(\lambda) = \frac{1}{2}\|\mathbf{q}(t,\cdot)\|^2_{\mathbf{V}^{\nu+s}_\sigma(\Omega)}.\end{aligned} \tag{A.25}$$

Differentiating (A.21) with respect to t and repeating the arguments used in deriving (A.25), we are led to

$$\begin{aligned}\int_t^\infty &\|\partial_t\mathbf{q}(\tau,\cdot)\|^2_{\mathbf{V}^{\nu+s-1}_\sigma(\Omega)}\, d\tau \\ &= \int_t^\infty\int_1^\infty e^{-2\lambda\tau}\lambda^{\nu+s+1}\|\mathbf{q}_0\|^2_{\mathbf{V}^s_\sigma(\Omega)}\, d\mu(\lambda)\, d\tau = \frac{1}{2}\|\mathbf{q}(t,\cdot)\|^2_{\mathbf{V}^{\nu+s}_\sigma(\Omega)}.\end{aligned} \tag{A.26}$$

Then inequalities (A.24)–(A.26) imply (A.19). □

We next consider the Stokes problem with nonzero forcing and the zero initial value:

$$\partial_t\mathbf{q}(t,\mathbf{x}) + A\mathbf{q} = \mathbf{h}(t,\mathbf{x}) \quad \text{in } Q \qquad \text{and} \qquad \mathbf{q}|_{t=0} = \mathbf{0} \quad \text{in } \Omega. \tag{A.27}$$

LEMMA A.7. *Assume* $\mathbf{h} \in L^2(0,T;\mathbf{V}^{-s}_\sigma(\Omega))$, $s \in [0,1]$. *Then the solution* $\mathbf{q}$ *of the problem* (A.27) *satisfies the equality*

$$\int_0^T \Big(\|\partial_t\mathbf{q}(t,\cdot)\|^2_{\mathbf{V}^{-s}_\sigma(\Omega)} + \|\mathbf{q}(t,\cdot)\|^2_{\mathbf{V}^{2-s}_\sigma(\Omega)}\Big)\, dt = \int_0^T \|\mathbf{h}(t,\cdot)\|^2_{\mathbf{V}^{-s}_\sigma(\Omega)}\, dt. \tag{A.28}$$

Proof. As in the proof of Lemma A.6, we reduce the problem (A.27) to an ordinary differential equation problem for the spectral decomposition $\widehat{\mathbf{q}}(t,\lambda)$ of $\mathbf{q}(t,\mathbf{x})$:

$$\partial_t \widehat{\mathbf{q}}(t,\lambda) + \lambda \widehat{\mathbf{q}}(t,\lambda) = \widehat{\mathbf{h}}(t,\lambda) \quad \text{in } Q \tag{A.29}$$

and

$$\widehat{\mathbf{q}}(t,\lambda)|_{t=0} = \mathbf{0} \tag{A.30}$$

for μ-almost-every $\lambda \in [1,\infty)$, where $\widehat{\mathbf{h}}$ is the spectral decomposition of $\mathbf{h}$. We extend $\widehat{\mathbf{h}}$ in t by zero outside the interval $[0,T]$ and extend $\widehat{\mathbf{q}}$ by zero for $t<0$ and by $e^{-\lambda t}\int_0^T e^{\lambda\tau}\mathbf{h}(\tau,\lambda)\,d\tau$ for $t>T$. This integral defines the solution of (A.29) for $t>T$ if $\mathbf{h}=0$ for $t>T$. We still denote the extended functions by $\widehat{\mathbf{q}}$ and $\widehat{\mathbf{h}}$, respectively. By virtue of (A.30), the extended functions satisfy (A.29)–(A.30) for $t\in\mathbb{R}$. Applying the Fourier transform in t, i.e.,

$$\widetilde{\mathbf{q}}(\tau,\lambda) = \int_{\mathbb{R}} e^{-i\tau t}\widehat{\mathbf{q}}(t,\lambda)\,dt \qquad \text{and} \qquad \widetilde{\mathbf{h}}(\tau,\lambda) = \int_{\mathbb{R}} e^{-i\tau t}\widehat{\mathbf{h}}(t,\lambda)\,dt,$$

to (A.29) we obtain

$$\widetilde{\mathbf{q}}(\tau,\lambda) = \frac{\widetilde{\mathbf{h}}(\tau,\lambda)}{i\tau+\lambda} \quad \text{and} \quad i\tau\widetilde{\mathbf{q}}(\tau,\lambda) = \frac{i\tau\widetilde{\mathbf{h}}(\tau,\lambda)}{i\tau+\lambda}.$$

Taking the $\|\cdot\|_{\mathbf{V}_\sigma^{-s}(\Omega)}$ norms yields

$$\begin{aligned} &\|\widetilde{\mathbf{q}}(\tau,\cdot)\|^2_{\mathbf{V}_\sigma^{2-s}(\Omega)} + \|i\tau\widetilde{\mathbf{q}}(\tau,\cdot)\|^2_{\mathbf{V}_\sigma^{-s}(\Omega)} \\ &\quad = \int_1^\infty \left(\frac{\lambda^{2-s}}{|i\tau+\lambda|^2} + \frac{\lambda^{-s}|i\tau|^2}{|i\tau+\lambda|^2}\right)\|\widetilde{\mathbf{h}}(\tau,\lambda)\|^2_{H(\lambda)}\,d\lambda \\ &\quad = \int_1^\infty \lambda^{-s}\|\widetilde{\mathbf{h}}(\tau,\lambda)\|^2_{H(\lambda)}\,d\lambda = \|\widetilde{\mathbf{h}}(\tau,\cdot)\|^2_{\mathbf{V}_\sigma^{s}(\Omega)}. \end{aligned} \tag{A.31}$$

Applying Parseval's equality to both ends of (A.31) we obtain (A.28). (Note that standard uniqueness arguments applied to (A.27) imply that we have obtained estimates precisely for the solution of (A.27).) ∎

A.3. Regularity estimate for the solution of the adjoint boundary value problem. We will apply the regularity results for the Stokes problem to derive regularity estimates for the solution $\mathbf{q}$ of the adjoint boundary value problem (3.14). (3.14) can be rewritten as

$$\partial_t\mathbf{q}(t,\mathbf{x}) + A\mathbf{q} = L\mathbf{q} \quad \text{in } Q \qquad \text{and} \qquad \mathbf{q}|_{t=0} = \mathbf{q}_0 \quad \text{in } \Omega, \tag{A.32}$$

where

$$L\mathbf{q} = P\big(\mathbf{q} + [(\mathbf{v}_0+\widehat{\mathbf{w}})\cdot\nabla]\mathbf{q} - (\nabla\widehat{\mathbf{w}})^*\mathbf{q}\big). \tag{A.33}$$

Indeed, we consider the first equation in (A.32) in the space $L^2(0,T;\mathbf{V}_\sigma^{-s}(\Omega))$, $s\in[0,1]$. Since the definition (A.16) implies that P is a projection and we look for $\partial_t\mathbf{q}$ in $L^2(0,T;\mathbf{V}_\sigma^{-s}(\Omega))$, we have that $P\partial_t\mathbf{q} = \partial_t\mathbf{q}$. The equality $A = P(-\Delta+I)$ follows from the definitions (A.1) and (A.16) of the operators A (on $\mathbf{V}_0^0(\Omega)$) and P and

the definition of the space $\mathbf{V}_\sigma^{-s}(\Omega)$ (see (A.15)). Relations (A.18) are built into the space $L^2(0,T;\mathbf{V}_\sigma^{-s+2}(\Omega))$, $-s+2\geq 1$. Note that the existence and uniqueness of the solution was established in Lemma 3.6. We have the following regularity results for the adjoint boundary value problem.

PROPOSITION A.8. *Assume that* $\mathbf{q}_0\in\mathbf{V}_\sigma^s(\Omega)$ *for an* $s\in[0,1/2)$. *Let* $\mathbf{q}$ *be the unique solution of the problem* (A.32)–(A.33) *satisfying the inequality* (3.15). *Then the following estimate holds:*

$$\int_t^\infty \left(\|\mathbf{q}(\tau,\cdot)\|^2_{\mathbf{V}^2_\sigma(\Omega)}+\|\partial_t\mathbf{q}(\tau,\cdot)\|^2_{\mathbf{V}^0_\sigma(\Omega)}\right)d\tau\leq C_0t^{s-1}\|\mathbf{q}_0\|^2_{\mathbf{V}^s_\sigma(\Omega)}, \tag{A.34}$$

where the constant C_0 *depends only on* $\|\widehat{\mathbf{w}}\|_{\mathbf{V}^{1,1/2}(Q)}$ *and* $\|\mathbf{v}_0\|_{\mathbf{C}^2(\overline{\Omega})}$.

Proof. By virtue of Theorem 2.1, we have that

$$\sup_{\mathbf{x}\in\overline{\Omega}}|\mathbf{v}_0(\mathbf{x})-\mathbf{w}_\infty|+\sum_{i=1}^3\sup_{\mathbf{x}\in\overline{\Omega}}|\partial_{x_i}\mathbf{v}_0(\mathbf{x})|+\sum_{i,j=1}^3\sup_{\mathbf{x}\in\overline{\Omega}}|\partial_{x_i}\partial_{x_j}\mathbf{v}_0(\mathbf{x})|\leq C|\mathbf{v}_\infty|.$$

As indicated in Theorems 4.1 and 4.2 and by (2.22), (2.19), and (2.15), the function $\widehat{\mathbf{w}}$ satisfies

$$\widehat{\mathbf{w}}\in\mathbf{V}^{1,1/2}(Q)\subset\mathbf{C}([0,T];\mathbf{H}^1(\Omega)). \tag{A.35}$$

Therefore, for $\mathbf{q}\in L^2(0,T;\mathbf{V}^1_\sigma(\Omega))$, we have $(\mathbf{v}_0\cdot\nabla)\mathbf{q}\in L^2(0,T;\mathbf{L}^2(\Omega))$, and we obtain by Sobolev embedding theorems and the Holder inequality that

$$\begin{aligned}\Big|\int_\Omega(\widehat{\mathbf{w}}\cdot\nabla)\mathbf{q}\cdot\boldsymbol{\phi}\,dx\Big|&\leq\|\widehat{\mathbf{w}}\|_{\mathbf{L}^6(\Omega)}\|\nabla\mathbf{q}\|_{\mathbf{L}^2(\Omega)}\|\boldsymbol{\phi}\|_{\mathbf{L}^3(\Omega)}\\&\leq C\|\widehat{\mathbf{w}}\|_{\mathbf{H}^1(\Omega)}\|\nabla\mathbf{q}\|_{\mathbf{L}^2(\Omega)}\|\boldsymbol{\phi}\|_{\mathbf{H}^{1/2}(\Omega)},\end{aligned}$$

which, upon using (A.35), yields

$$\|(\widehat{\mathbf{w}}\cdot\nabla)\mathbf{q}\|_{L^2(0,T;\mathbf{H}^{-1/2}(\Omega))}\leq C\|\mathbf{q}\|_{L^2(0,T;\mathbf{V}^1_\sigma(\Omega))}\|\widehat{\mathbf{w}}\|_{\mathbf{V}^{1,1/2}(Q)}. \tag{A.36}$$

Similarly, we obtain

$$\begin{aligned}\Big|\int_\Omega(\nabla\widehat{\mathbf{w}})^*\mathbf{q}\cdot\boldsymbol{\phi}\,dx\Big|&\leq\|\nabla\widehat{\mathbf{w}}\|_{\mathbf{L}^2(\Omega)}\|\mathbf{q}\|_{\mathbf{L}^6(\Omega)}\|\boldsymbol{\phi}\|_{\mathbf{L}^3(\Omega)}\\&\leq C\|\nabla\widehat{\mathbf{w}}\|_{\mathbf{L}^2(\Omega)}\|\mathbf{q}\|_{\mathbf{H}^1(\Omega)}\|\boldsymbol{\phi}\|_{\mathbf{H}^{1/2}(\Omega)}\end{aligned} \tag{A.37}$$

and

$$\|(\nabla\widehat{\mathbf{w}})^*\mathbf{q}\|_{L^2(0,T;\mathbf{H}^{-1/2}(\Omega))}\leq C\|\mathbf{q}\|_{L^2(0,T;\mathbf{V}^1_\sigma(\Omega))}\|\widehat{\mathbf{w}}\|_{\mathbf{V}^{1,1/2}(Q)}. \tag{A.38}$$

From (A.33), (A.36), and (A.38), we deduce that

$$\|L\mathbf{q}\|_{L^2(t,T;\mathbf{V}_\sigma^{-1/2}(\Omega))}\leq C\|\mathbf{q}\|_{L^2(t,T;\mathbf{V}^1_\sigma(\Omega))},\qquad t\in(0,T). \tag{A.39}$$

We decompose the solution $\mathbf{q}$ of (A.32) into

$$\mathbf{q}=\mathbf{q}_1+\mathbf{q}_2, \tag{A.40}$$

where $\mathbf{q}_1$ is the solution of (A.17) and $\mathbf{q}_2$ is the solution of (A.27) with $\mathbf{h}=L\mathbf{q}$. Then, by Lemma A.7, we have

$$\|\mathbf{q}_2\|^2_{L^2(0,T;\mathbf{V}^{3/2}_\sigma(\Omega))}+\|\partial_t\mathbf{q}_2\|^2_{L^2(0,T;\mathbf{V}_\sigma^{-1/2}(\Omega))}\leq C\|L\mathbf{q}\|^2_{L^2(0,T;\mathbf{V}_\sigma^{-1/2}(\Omega))}. \tag{A.41}$$

The relations (A.40)–(A.41) and Lemma A.6 with $s+\nu = 1/2$ and (A.39) imply that, for each $t \in (0,T)$,

$$\int_t^T \Big(\|\mathbf{q}(\tau,\cdot)\|^2_{\mathbf{V}^{3/2}_\sigma(\Omega)} + \|\partial_t \mathbf{q}(\tau,\cdot)\|^2_{\mathbf{V}^{-1/2}_\sigma(\Omega)} \Big)\, d\tau \le C\Big(\|\mathbf{q}_1(t,\cdot)\|^2_{\mathbf{V}^{1/2}_\sigma(\Omega)} + \|\mathbf{q}\|^2_{L^2(t,T;\mathbf{V}^1_\sigma(\Omega))} \Big). \tag{A.42}$$

Repeating arguments similar to those used in the derivation of (A.36)–(A.39), we obtain

$$\|L\mathbf{q}\|^2_{L^2(t,T;\mathbf{V}^{-\varepsilon}_\sigma(\Omega))} \le C \|\mathbf{q}\|^2_{L^2(t,T;\mathbf{V}^{3/2}_\sigma(\Omega))}, \tag{A.43}$$

where $\varepsilon > 0$ is an arbitrary small number. Also, analogous to (A.42), we have that

$$\|\mathbf{q}\|^2_{L^2(t,T;\mathbf{V}^{2-\varepsilon}_\sigma(\Omega))} + \|\partial_t \mathbf{q}\|^2_{L^2(t,T;\mathbf{V}^{-\varepsilon}_\sigma(\Omega))} \le C\Big(\|\mathbf{q}_1(t,\cdot)\|^2_{\mathbf{V}^1_\sigma(\Omega)} + \|L\mathbf{q}\|^2_{L^2(t,T;\mathbf{V}^{-\varepsilon}_\sigma(\Omega))} \Big). \tag{A.44}$$

We substitute (A.43) into (A.44) and use the obtained estimate as well as the arguments which led us to (A.43) and (A.44) to deduce that (A.43) and (A.44) hold with $\varepsilon = 0$. Combining the estimates (A.42)–(A.44) (with $\varepsilon > 0$ and with $\varepsilon = 0$), we are led to

$$\int_t^T \Big(\|\mathbf{q}(\tau,\cdot)\|^2_{\mathbf{V}^2_\sigma(\Omega)} + \|\partial_t \mathbf{q}(\tau,\cdot)\|^2_{\mathbf{V}^0_\sigma(\Omega)} \Big)\, d\tau \le C\Big(\|\mathbf{q}_1(t,\cdot)\|^2_{\mathbf{V}^1_\sigma(\Omega)} + \|\mathbf{q}\|^2_{L^2(t,T;\mathbf{V}^1_\sigma(\Omega))} \Big). \tag{A.45}$$

Applying Lemma A.6 with $\nu = 1 - s$ and the estimate (3.15) to the right-hand side of (A.45), we arrive at (A.34). ∎

Appendix B. Orlicz spaces. In section 5.2, we needed to determine the space in which to search for the adjoint vector field for the optimality system. For this purpose, we first calculate the dual space for the function space $L^2(Q) \cap L^{6/5}(Q)$. We can consider $L^2(Q) \cap L^{6/5}(Q)$ as the Orlicz space with the N-function

$$A(t) = \max\,(t^2, t^{6/5}), \qquad t \ge 0.$$

(For the N-function as well as other notations and assertions connected with Orlicz spaces, see [1, Chap. 8].) In other words, the Orlicz space $L^2(Q) \cap L^{6/5}(Q)$ can be defined as follows:

$$\begin{aligned} L^2(Q) \cap L^{6/5}(Q) &= L_A(Q) \\ &\equiv \Big\{ f(t,\mathbf{x}),\ (t,\mathbf{x}) \in Q : \int_Q A(|f(t,\mathbf{x})|)\, d\mathbf{x}\, dt < \infty \Big\}. \end{aligned} \tag{B.1}$$

The norm in a general Orlicz space $L_A(Q)$ is defined as follows:

$$\|f\|_{L_A(Q)} = \inf \Big\{ k > 0 : \int_Q A\Big(\frac{|f(t,\mathbf{x})|}{k} \Big)\, d\mathbf{x}\, dt \le 1 \Big\}. \tag{B.2}$$

Define the Legendre transform

$$A^*(s) = \max_{t \ge 0} (st - A(t))$$

for $A(t) = \max(t^2, t^{6/5})$. Then $A^*(s)$ is an N-function and the Orlicz space

$$L_{A^*}(Q) = \left\{ f(t,\mathbf{x}),\ (t,\mathbf{x}) \in Q : \int_Q A^*(|f(t,\mathbf{x})|)\, d\mathbf{x}\, dt < \infty \right\} \tag{B.3}$$

is the function space dual to space (B.1) (see [1, Chap. 8]). Straightforward calculations show that

$$A^*(s) = \begin{cases} 5^5 s^6/6^6, & s \in [0, 6/5], \\ s - 1, & s \in [6/5, 2], \\ s^2/4, & s \geq 2. \end{cases} \tag{B.4}$$

The norm $\|\cdot\|_{L_{A^*}(Q)}$ of the space $L_{A^*}(Q)$ is defined similarly to (B.2) with the help of the N-function $A^*(t)$.

Let us consider now the spaces $\mathbf{L}_A(Q)$ and $\mathbf{L}_{A^*}(Q)$ of vector functions that are defined by (B.1) and (B.3), respectively, where f is the vector function $f = (f_1(t,x), f_2(t,x), f_3(t.x))$.

LEMMA B.1. *The spaces $\mathbf{L}_A(Q)$ and $\mathbf{L}_{A^*}(Q)$ are dual.*

Proof. We will prove that $(\mathbf{L}_A(Q))^* = \mathbf{L}_{A^*}(Q)$, where $(\mathbf{L}_A(Q))^*$ is the adjoint space of $\mathbf{L}_A(Q)$. Each $\mathbf{v} \in \mathbf{L}_{A^*}(Q)$ forms a linear bounded functional on $\mathbf{L}_A(Q)$ by the formula

$$l(u) = \int_Q \mathbf{v}(t,x) \cdot \mathbf{u}(t,\mathbf{x})\, d\mathbf{x}\, dt \qquad \forall \mathbf{u} \in \mathbf{L}_A(Q) \tag{B.5}$$

because of the generalized Holder inequality (see [1, Chap. 8]):

$$\int_Q \mathbf{v}(t,x) \cdot \mathbf{u}(t,\mathbf{x})\, d\mathbf{x}\, dt \leq 2\|\mathbf{v}\|_{\mathbf{L}_{A^*}(Q)} \|\mathbf{u}\|_{\mathbf{L}_A(Q)}. \tag{B.6}$$

Therefore $\mathbf{L}_{A^*}(Q) \subset (\mathbf{L}_A(Q))^*$. To prove the inverse inclusion we establish first that $\mathbf{L}_A(Q)$ is isomorphic to $L_A(Q) \times L_A(Q) \times L_A(Q)$. For this we have to prove the estimates

$$\|\mathbf{v}\|_{\mathbf{L}_A(Q)} \leq \sum_{j=1}^3 \|v_j\|_{L_A(Q)} \leq 3\|\mathbf{v}\|_{\mathbf{L}_A(Q)} \quad \forall\, \mathbf{v} = (v_1, v_2, v_3) \in \mathbf{L}_A(Q). \tag{B.7}$$

Suppose first that for each j, $v_j(t,\mathbf{x}) \neq 0$ on a set of positive Lebesque measure. Then, using the definition (B.2), the inequality $|\mathbf{v}(\mathbf{x})| \leq \sum_{j=1}^3 |v_j(\mathbf{x})|$, and the convexity of $A(\lambda)$, we obtain

$$\begin{aligned}
\|\mathbf{v}\|_{\mathbf{L}_A(Q)} &= \inf\left\{ k > 0 : \int_\Omega A\left(\frac{|\mathbf{v}(\mathbf{x})|}{k}\right) d\mathbf{x} \leq 1 \right\} \\
&\leq \inf\left\{ k > 0 : \int_\Omega A\left(\frac{\sum_{j=1}^3 |v_j(\mathbf{x})|}{k}\right) d\mathbf{x} \leq 1 \right\} \\
&= \inf\left\{ \sum_{i=1}^3 k_i : k_i > 0,\ i = 1,2,3, \int_\Omega A\left(\sum_{j=1}^3 \frac{k_j}{k_1 + k_2 + k_3} \frac{|v_j(\mathbf{x})|}{k_j} \right) d\mathbf{x} \leq 1 \right\} \\
&\leq \inf\left\{ \sum_{i=1}^3 k_i : k_i > 0,\ i = 1,2,3, \sum_{j=1}^3 \frac{k_j}{k_1 + k_2 + k_3} \int_\Omega A\left(\frac{|v_j(\mathbf{x})|}{k_j} \right) d\mathbf{x} \leq 1 \right\} \\
&\leq \inf\left\{ \sum_{i=1}^3 k_i : k_i > 0, \int_\Omega A\left(\frac{|v_j(\mathbf{x})|}{k_j} \right) d\mathbf{x} \leq 1\ i = 1,2,3 \right\} \\
&\leq \sum_{i=1}^3 \inf\left\{ k > 0 : \int_\Omega A\left(\frac{|v_i(\mathbf{x})|}{k} \right) d\mathbf{x} \leq 1 \right\} = \sum_{i=1}^3 \|v_i\|_{L_A(Q)}.
\end{aligned}$$

This inequality and the homogeniety property of norms imply the first estimate in (B.7)

Denote $\hat{v}(t,\mathbf{x}) = \max_{j=1,2,3} |v_j(t,\mathbf{x})|$. For each $(t,\mathbf{x})$ we have

$$\frac{1}{3}|\mathbf{v}(t,\mathbf{x})| \le \hat{v}(t,\mathbf{x}) \le |\mathbf{v}(t,\mathbf{x})|$$

so that $A(\hat{v}(t,\mathbf{x})/k) \le A(|\mathbf{v}(t,\mathbf{x})|/k)$ for all $k > 0$. Thus,

$$\begin{aligned}\sum_{j=1}^{3} \|v_j\|_{L_A(Q)} \le 3\|\hat{v}\|_{L_A(Q)} &= 3\inf\left\{k>0:\ \int_Q A(\frac{|\hat{v}(t,\mathbf{x})|}{k})\,d\mathbf{x}\,dt \le 1\right\} \\ &\le 3\inf\left\{k>0:\ \int_Q A\left(\frac{|\mathbf{v}(t,\mathbf{x})|}{k}\right)d\mathbf{x}\,dt \le 1\right\} = 3\|\mathbf{v}\|_{\mathbf{L}_A(Q)}.\end{aligned} \tag{B.8}$$

The bound (B.8) implies the second estimate in (B.7).

The estimate (B.7) in the case when $v_j(t,\mathbf{x}) = 0$ almost everywhere for one or two j can be considered similarly. The case $v_j(t,\mathbf{x}) = 0$ a.e. for $j = 1,2,3$ is trivial.

By the lemma on the form of a functional on direct product of spaces (see [2, sect. 2.1.2]), each functional $\Lambda \in (L_A(Q) \times L_A(Q) \times L_A(Q))^*$ has the form

$$\Lambda(\mathbf{u}) = l_1(u_1) + l_2(u_2) + l_3(u_3) \quad \forall\ \mathbf{u} = (u_1,u_2,u_3) \in [L_A(Q)]^3, \tag{B.9}$$

where $l_j \in (L_A(Q))^*$, $j = 1,2,3$. By [1, Chap. 8], there exist $v_j \in L_{A^*}(Q)$, $j = 1,2,3$, such that

$$l_j(u) = \int_Q v_j(t,\mathbf{x})u(t,\mathbf{x})\,d\mathbf{x}\,dt, \quad j = 1,2,3 \quad \forall\ u \in L_A(Q). \tag{B.10}$$

Relations (B.9), (B.10), and the isomorphism $\mathbf{L}_A(Q) \cong L_A(Q) \times L_A(Q) \times L_A(Q)$ proved by (B.7) yield the embedding $\mathbf{L}_{A^*}(Q) \supset (\mathbf{L}_A(Q))^*$.

The proof of the relation $(\mathbf{L}_{A^*}(Q))^* = \mathbf{L}_A(Q)$ is similar. □

We will need the following assertion.

LEMMA B.2. *Let Y be a closed subspace of a reflexive space X. Assume that the norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ are equivalent. Then Y is a reflexive space.*

Proof. By virtue of the Eberlein–Shmulian theorem [18, Appendix], the reflexivity of Y is equivalent to the following property: each bounded sequence $y_n \in Y$ has a subsequence $\{y_{n_k}\}$ converging weakly to a $\widehat{y} \in Y$. Let a bounded sequence $y_n \in Y$ be given. Since the sequence $\{y_n\}$ is bounded in X and X is reflexive, there exists a subsequence $\{y_{n_k}\} \subset \{y_n\}$ that converges weakly in X to a $\widehat{y} \in X$. Note that Y being closed and convex, it is sequentially weakly closed. By the Hahn–Banach theorem, the weak convergence in X of $\{y_{n_k}\} \subset Y$ implies the weak convergence in Y. Hence, Y is a reflexive space. □

B.1. Orlicz spaces of solenoidal fields. Now we consider solenoidal vector fields. Using the space of vector fields

$$\mathcal{V}(Q) = \{\mathbf{v}(t,\mathbf{x}) \in \mathbf{C}_0^\infty(Q):\ \operatorname{div}\mathbf{v} = 0, \quad \operatorname{supp}\mathbf{v} \subset\subset Q\}$$

we introduce

$$L^2(0,T;\mathbf{V}_0^0(\Omega)) = \text{closure of } \mathcal{V}(Q) \text{ in } \mathbf{L}^2(Q). \tag{B.11}$$

It is well known that $\mathbf{V}_0^0(\Omega)$ in (B.11) is the space (2.14). Define also

$$\mathbf{V}_A(Q) = \text{closure of } \mathcal{V}(Q) \text{ in } \mathbf{L}_A(Q). \tag{B.12}$$

Evidently $\mathbf{V}_A(Q) = L^2(0,T;\mathbf{V}_0^0(\Omega)) \cap \mathbf{L}^{6/5}(Q)$. We set

$$\mathbf{V}_{A^*}(Q) = \text{closure of } \mathcal{V}(Q) \text{ in } \mathbf{L}_{A^*}(Q). \tag{B.13}$$

All elements of the spaces (B.11), (B.12), and (B.13) possess the properties $\operatorname{div} \mathbf{u} = 0$ and $(\mathbf{u}\cdot\mathbf{n})|_{\partial\Omega} = 0$ (these equalities are understood in the sense of distributions).

Denote

$$\mathbf{G}_A(Q) = \{\nabla p \in \mathbf{L}_A(Q) :\ p \in L^2(0,T;H^1_{\mathrm{loc}}(\Omega))\}, \tag{B.14}$$

LEMMA B.3. *The relations* $(\mathbf{V}_{A^*}(Q))^* = \mathbf{V}_A(Q)$ *and* $(\mathbf{V}_A(Q))^* = \mathbf{V}_{A^*}(Q)$ *hold.*

Proof. We begin from the first identity. Each $\mathbf{u} \in \mathbf{V}_A(Q)$ defines a functional on $\mathbf{V}_{A^*}(Q)$ by (B.5) and (B.6). Hence $\mathbf{V}_A(Q) \subset (\mathbf{V}_{A^*}(Q))^*$. To prove the reverse inclusion let a functional $l \in (\mathbf{V}_{A^*}(Q))^*$ be given, and we extend it by the Hahn–Banach theorem into an $\widehat{l} \in (\mathbf{L}_{A^*}(Q))^*$. By Lemma B.1, there exists a $\mathbf{u} \in \mathbf{L}_A(Q)$ such that $\widehat{l}(\mathbf{v}) = \int_Q \mathbf{u}\cdot\mathbf{v}\,d\mathbf{x}\,dt$. Thus,

$$l(\mathbf{v}) = \int_Q \mathbf{u}\cdot\mathbf{v}\,d\mathbf{x}\,dt \qquad \forall\, \mathbf{v} \in \mathbf{V}_{A^*}(Q). \tag{B.15}$$

Note that the following decomposition is true:

$$\mathbf{L}_A(Q) = \mathbf{V}_A(Q) + \mathbf{G}_A(Q). \tag{B.16}$$

Indeed, since $\mathbf{L}_A(Q) \subset \mathbf{L}^2(Q)$, by virtue of Weyl's decomposition for each $\mathbf{u} \in \mathbf{L}_A(Q)$, there exist $\nabla p \in \mathbf{G}(Q) = \{\nabla p \in \mathbf{L}^2(Q) :\ p \in L^2(0,T;H^1_{\mathrm{loc}}(\Omega))\}$ and $\mathbf{w} \in L^2(0,T;\mathbf{V}_0^0(\Omega))$ such that

$$\mathbf{u} = \mathbf{w} + \nabla p. \tag{B.17}$$

To prove Weyl's decomposition, one defines $\nabla p(t,\mathbf{x})$ to be the solution of the problem

$$\int_Q \nabla p(t,\mathbf{x})\cdot\nabla q(t,\mathbf{x})\,d\mathbf{x}\,dt = \int_Q \mathbf{u}(t,\mathbf{x})\cdot\nabla q(t,\mathbf{x})\,d\mathbf{x}\,dt \qquad \forall\, \nabla q. \tag{B.18}$$

If $\nabla q \in \mathbf{L}^2(Q)$, then (B.18) defines $\nabla p \in \mathbf{L}^2(Q)$. But since $\mathbf{u} \in \mathbf{V}_A(Q) = \mathbf{L}^2(Q) \cap \mathbf{L}^{6/5}(Q)$, the right-hand side of (B.18) is a bounded functional with respect to $\nabla q \in \mathbf{L}^6(Q)$, so that $\nabla p \in \mathbf{L}^{6/5}(Q)$. Hence $\nabla p \in \mathbf{V}_A(Q)$ and, by (B.17), $\mathbf{w} \in \mathbf{V}_A(Q)$, which proves (B.16). Note that the relation $\mathbf{V}_A(Q) \cap \mathbf{G}_A(Q) = 0$ follows from the orthogonality between $\mathbf{w}$ and ∇p in $\mathbf{L}^2(Q)$.

Now we substitute (B.17) into (B.15) and take into account that $\int_Q \nabla p\cdot\mathbf{v}\,d\mathbf{x}\,dt = 0$ for every $\nabla p \in \mathbf{G}_A(Q)$ and every $\mathbf{v} \in \mathbf{V}_{A^*}(Q)$. As a result we deduce

$$l(\mathbf{v}) = \int_Q \mathbf{w}\cdot\mathbf{v}\,d\mathbf{x}\,dt \qquad \forall\, \mathbf{v} \in \mathbf{V}_{A^*}(Q) \tag{B.19}$$

with $\mathbf{w} \in \mathbf{V}_A(Q)$. Hence, we have proved the equality $(\mathbf{V}_{A^*}(Q))^* = \mathbf{V}_A(Q)$. This equality yields $(\mathbf{V}_A(Q))^* = \mathbf{V}_{A^*}(Q)$, since by Lemma B.2, the spaces $\mathbf{V}_A(Q))$ and $\mathbf{V}_{A^*}(Q)$ are reflexive. ∎

We set

$$\mathcal{W}_0^{(2)}(Q) = \{\mathbf{v} \in L^2(0,T;\mathbf{V}^2(\Omega)) : \partial_t \mathbf{v} \in L^2(0,T;\mathbf{V}^0(\Omega)),\ \mathbf{v}\big|_{t=0} = \mathbf{0}\} \tag{B.20}$$

and

$$\mathbf{V}_A^{1,2}(Q) = \mathcal{W}_0^{(2)}(Q)) \cap \mathbf{W}_{6/5}^{1,2}(Q), \tag{B.21}$$

where $\mathbf{W}_{6/5}^{1,2}(Q)$ is Sobolev space defined by (3.7).

REFERENCES

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

[2] V. Alekseev, V. Tikhomirov, and S. Fomin, *Optimal Control*, Consultants Bureau, New York, 1987.

[3] O. Besov, V. Il'in, and S. Nikol'skiy, *Integral Representations of Functions and Embedding Theorems*, Nauka, Moscow, 1975 (in Russian).

[4] B. Dubrovin, A. Fomenko, and S. Novikov, *Modern Geometry: Methods and Applications. Part* 1. *The Geometry of Surfaces, Transformation Groups and Fields,* 2nd ed., Springer, New York, 1992.

[5] A. Fursikov, *Control problems and theorems concerning the unique solvability of mixed boundary value problems for the three-dimensional Navier–Stokes and Euler equations,* Math USSR Sb., 43 (1982), pp. 251–273.

[6] A.Fursikov, *Optimal Control of Distributed Systems. Theory and Applications,* Transl. Math. Monogr. 187, American Mathematical Society, Providence, RI, 2000.

[7] A. Fursikov and O. Emanuilov, *Exact controllability of Navier–Stokes and Boussinesq equations,* Russian Math. Surveys, 54 (1999), pp. 565–618.

[8] A. Fursikov, M. Gunzburger, and L. Hou, *Boundary value problems and boundary optimal control for the Navier–Stokes systems: The two dimensional case,* SIAM J. Control Optim., 36 (1998), pp. 852–894.

[9] A. Fursikov, M. Gunzburger, and L. Hou, *Trace theorems for three-dimensional time-dependent solenoidal vector fields and their applications,* Trans. Amer. Math. Soc., 354 (2002), pp. 1079–1116.

[10] A. Fursikov, M. Gunzburger, and L. Hou, *Boundary value problems for three-dimensional evolutionary Navier–Stokes equations,* J. Math. Fluid Mech., 4 (2002), pp. 45–75.

[11] G. Galdi, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations*, Vol. 2, Springer, New York, 1994.

[12] I. Gelfand and N. Vilenkin, *Generalized Functions: Applications of Harmonic Analysis*, Vol. 4, Academic, New York, 1964.

[13] O. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1969.

[14] J.-L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer, Berlin, 1972.

[15] P. Shorygin, *Approximate controllability of the Navier–Stokes system in unbounded domains*, Sbornik: Math., 194 (2003), pp. 1725–1745.

[16] R. Temam, *Navier–Stokes Equations—Theory and Numerical Analysis*, AMS Chelsea Publishing, Providence, RI, 2001.

[17] M. Vishik and A. Fursikov, *Mathematical Problems of Statistical Hydrodynamics*, Kluwer, Dordrecht, 1988.

[18] K. Yosida, *Functional Analysis*, Springer, Berlin, 1965.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2233–2244

RAPID EXPONENTIAL FEEDBACK STABILIZATION WITH UNBOUNDED CONTROL OPERATORS*

JOSE MANUEL URQUIZA†

Abstract. We study the wellposedness and the main features of a class of feedback control systems. The involved control system is composed of the generator of a strongly continuous group for the free part and of an unbounded control operator, so that the results can be applied to boundary or point control problems for partial differential equations of hyperbolic or Petrowski type. The feedback operator is explicit and one can achieve an arbitrary large decay rate for the closed-loop system. These results are proved under a controllability assumption and the proofs rely on general results about the algebraic Riccati equation associated with the linear quadratic regulator problem.

Key words. strongly continuous group, unbounded control operator, stabilization, controllability, algebraic Riccati equation, wave equation

AMS subject classifications. 93B52, 93D15, 93C20, 93B05, 35L90, 35B37, 49J20

DOI. 10.1137/S0363012901388452

1. Introduction.

The purpose of this paper is to depict a class of full state stabilizing feedback operators for some infinite-dimensional linear systems with unbounded control operators and to prove the wellposedness of the resulting closed-loop feedback systems. The class of infinite-dimensional equations that is considered contains time-reversible partial differential equations such as first-order hyperbolic, wave, and plate-like equations. The assumptions on the control operator allow us to consider boundary or point control problems for these equations.

One of the main features of the class of feedback operators which is presented in this paper is that they are built in a straightforward and explicit way. Moreover, the exponential decay rate of the resulting closed-loop feedback system can be made as large as desired by freely tuning a parameter on which the feedback operator depends. The controllability of the system is the main assumption for the results to hold.

This class of feedback operators was already known for finite-dimensional systems. Moreover, it constitutes a limiting but unexplored class of feedback operators studied recently by Komornik [12] for boundary control problems of partial differential equations. In [12], a lower bound of the exponential decay rate can also be chosen arbitrarily. For the class of feedback operators presented here, we can give the optimal decay rate explicitly as a function of the tuning parameter and of the growth bound of the backward in time semigroup associated with the system.

In order to state the problem treated in this paper an introductory exposition is given next. It begins with, in addition to well-known results of the control theory of linear finite-dimensional systems, some other results that seem to be scarcely present in the literature. These results are given in Russell's book [24] to which we refer the reader for details. They deal with easy ways to stabilize linear constant finite-dimensional systems that are controllable.

*Received by the editors April 23, 2001; accepted for publication (in revised form) August 30, 2004; published electronically May 27, 2005.
http://www.siam.org/journals/sicon/43-6/38845.html
†Centre de Recherches Mathématiques, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal H3C 3J7, QC, Canada (Urquiza@CRM.UMontreal.CA).

Consider the linear constant finite-dimensional system

$$\begin{cases} \dot{y}(t) = Ay(t) + Bu(t), \\ y(0) = y_0, \qquad y_0 \in \mathbb{R}^n, \end{cases} \tag{1.1}$$

where $A \in \mathbb{R}^{n,n}$, $B \in \mathbb{R}^{n,p}$ for some $n, p \in \mathbb{N}^*$. The control system (A, B) is said to be controllable in time $T > 0$ if for every $y_0, y_T \in \mathbb{R}^n$, there is a control u such that the solution y of (1.1) satisfies $y(T) = y_T$. A necessary and sufficient condition is that the symmetric matrix

$$\Lambda_{0,T} \equiv \int_0^T e^{-tA} B B^* e^{-tA^*} dt$$

is positive definite for some $T > 0$ (and thus for every $T > 0$). $\Lambda_{0,T}$ is called a *controllability gramian* and it can be used to derive a control candidate. For instance, with

$$u(t) = -B^* e^{-tA^*} \Lambda_{0,T}^{-1} y_0, \quad 0 \le t \le T,$$

the solution of (1.1) satisfies $y(T) = 0$.

Gramians can thus be used for the controllability problem, but can also be used for the stabilization problem, that is, to find a matrix operator $F \in \mathbb{R}^{p,n}$ such that all solutions of

$$\dot{y}(t) = Ay(t) + BFy(t) = (A + BF)y(t), \quad t \ge 0, \tag{1.2}$$

decay asymptotically to the null state.

If $\Lambda_{0,T}$ is positive definite, the same conclusion holds for the *modified gramians* $\Lambda_{\omega,T}$ defined for every $0 < T < \infty$ and every $\omega \ge 0$ by

$$\Lambda_{\omega,T} \equiv \int_0^T e^{-2\omega t} e^{-tA} B B^* e^{-tA^*} dt$$

and also for the *extended controllability gramians* $\Lambda_{\omega,\infty}$when ω is sufficiently large:

$$\Lambda_{\omega,\infty} \equiv \int_0^\infty e^{-2\omega t} e^{-tA} B B^* e^{-tA^*} dt \ > \ 0.$$

It is well known that, at least for linear constant finite-dimensional systems, controllability implies stabilizability. A common feature of the gramians described above is their ability to be used to build explicit asymptotically stabilizing feedback operators. Indeed, let $0 \le \omega < \infty$ and $0 < T \le \infty$ be such that $\Lambda_{\omega,T}$ is well defined and regular, and set

$$F_{\omega,T} = -B^* \Lambda_{\omega,T}^{-1}.$$

Then $A + BF_{\omega,T}$ is a stable matrix (all its eigenvalues have negative real parts), so that the solutions of (1.2) with $F = F_{\omega,T}$ decay uniformly and exponentially to zero: there is $M, \lambda > 0$ such that

$$\left| e^{t(A+BF_{\omega,T})} y_0 \right| \le M e^{-\lambda t} |y_0|, \quad t \ge 0, \tag{1.3}$$

for every $y_0 \in \mathbb{R}^n$.

For $\omega = 0$ and $0 < T < \infty$ this result was presented by Kleinman [10] and Lukes [21] in a finite-dimensional setting. The case $\omega > 0$, $0 < T < \infty$ appeared in Slemrod [26] in a broader setting including infinite-dimensional systems with bounded control operators. The result for this latter case was extended by Komornik [12] to unbounded control operators, with applications to some boundary or point control problems for partial differential equations.

The interest to take $\omega > 0$ lies in the fact that these feedbacks lead to an exponential stability of the system with an exponential decay rate greater than ω. As ω may be taken freely, the control system can be uniformly stabilized with an exponential decay rate as large as desired. This property has been called *complete stabilizability* [26]. Of course, it is well known that for controllable linear constant finite-dimensional systems, poles can be placed at will, but it seems that such a result is difficult to generalize to infinite-dimensional systems. Nevertheless, the uniform estimate (1.3) with $\lambda = \omega$ also holds for some infinite-dimensional systems to which applications of these feedback operators (with $\omega > 0$ and $T < \infty$) were done, although, for both finite-dimensional systems and infinite-dimensonal systems, we cannot say more about the *growth bound* (see the definition in the next section) of the closed-loop system in general when T is finite.

Also, based on Komornik's result [12], some numerical simulations have been conducted (see Briffaut [4]) on the spatial semidiscretizations of boundary control problems for some canonical partial differential equations of dynamical elasticity which have the particularity of being conservative systems. It appeared that the decay rate for the approximated systems was in fact approximately twice better than the theorical bound ω. From these observations, and by first considering some simple small finite-dimensional systems for which we can easily make the calculus explicit, we were led to consider the feedback operators built with the extended gramians, that is, to take $T = \infty$ (and ω sufficiently large). As a particular case of the main result to be proved here, we obtain that for the class of skew-adjoint operators ($A^* = -A$) the decay rate is exactly 2ω.

The easy way of using $\Lambda_{\omega,\infty}$ to stabilize a controllable system is mentioned in Russell's book [24, pp. 117–118] as Bass's method. According to [24], R. W. Bass introduced it in some *"Lecture notes on control and optimization presented at NASA Langley Research Center in August* 1961."

Assuming we are still in the finite-dimensional context, let us give some details and characteristics about this class of feedback operators and the resulting feedback systems. From the expression of $\Lambda_{\omega,\infty}$, we easily deduce that it solves the Liapunov equation

$$\Lambda_{\omega,\infty}(A + \omega I)^* + (A + \omega I)\Lambda_{\omega,\infty} = BB^*. \tag{1.4}$$

From the stability of $-(A + \omega I)$ and the controllability assumption, the theory says that it is its unique symmetric solution.

Multiplying this equation by $\Lambda_{\omega,\infty}^{-1}$, we obtain the relation

$$BB^*\Lambda_{\omega,\infty}^{-1} = \Lambda_{\omega,\infty}A^*\Lambda_{\omega,\infty}^{-1} + A + 2\omega I \tag{1.5}$$

from which we deduce that

$$A + BF_{\omega,\infty} = \Lambda_{\omega,\infty}\left(-A^* - 2\omega I\right)\Lambda_{\omega,\infty}^{-1}.$$

We thus have a simple expression for the feedback system and its spectral elements in terms of the free system characteristics. In particular, if $A^* = -A$, then the spectrum

of A is purely imaginary and the feedback operator translates it to the axis of abscissa -2ω. In the general case, the spectral values of $A + BF_{\omega,\infty}$ are those of $-A^*$ after a translation of 2ω. Moreover, its spectral elements can be expressed in terms of those of $-A^*$ and of the matrix operator $\Lambda_{\omega,\infty}$.

If we now do the left multiplication of the relation (1.5) by $\Lambda^{-1}_{\omega,\infty}$, we obtain

$$A^*P + PA + 2\omega\Lambda^{-1}_{\omega,\infty} - PBB^*P = 0 \tag{1.6}$$

with $P = \Lambda^{-1}_{\omega,\infty}$. Equation (1.6) is nothing but the algebraic Riccati equation in P deduced from the linear quadratic regulator problem for the control system (A, B) associated with the cost functional

$$\int_0^\infty \left[2\omega \left(\Lambda^{-1}_{\omega,\infty} y(t), y(t)\right)_{\mathbb{R}^n} + \|u(t)\|^2_{\mathbb{R}^p}\right] dt.$$

The purpose of this paper is to extend the result to unbounded operators A and B as they appear in boundary control problems for partial differential equations. To this end we follow the indirect way of considering the optimal control problem over an infinite time interval that this feedback operator solves. By conveniently choosing the quadratic cost functional, and proving that we have the expected feedback operator, the theorical results on the linear quadratic regulator problem established by Flandoli, Lasiecka, and Triggiani [7] give the wellposedness of the resulting feedback system.

A more direct way would consist in defining the feedback operator and establishing its admissibility. But if A and (most of all) B are unbounded operators, it does not seem to be a trivial task (see, for instance, Weiss [31] for definitions of admissible control and feedback operators).

Moreover, if the control operator B is bounded, then the operator equations that correspond in the finite-dimensional case to the matrix equations (1.4)–(1.6) are posed in the domain of the unbounded operator A or of its adjoint. Whereas with a more general (unbounded) control operator, B has to be replaced in (1.6) by a nonunique extension in order to obtain a suitable algebraic Riccati equation posed in the domain of A. This was noticed by Staffans [28] and Weiss and Weiss [33] and illustrated by Weiss and Zwart [32] with a demonstrative example. In order to prove our results we use a result of [7] which links the solution of the algebraic Riccati equation to the solution of a dual Riccati equation with bounded control operator and that reduces to the Liapunov equation (1.4) posed in the domain of A^*.

The main results of this paper are stated in the next section. Then in section 3 we recall some results on the linear quadratic regulator problem and prove the main results: the wellposedness and the additional stability properties of the closed-loop feedback system.

2. Notations, assumptions, and main results. Let U and Y be two separable Hilbert spaces, the control space and the state space, with norms denoted by $\|\cdot\|_U$ and $\|\cdot\|_Y$, and with associated scalar products denoted by $(\cdot,\cdot)_U$ and $(\cdot,\cdot)_Y$, respectively. For two Banach spaces E and F, $\mathcal{L}(E, F)$ denotes the space of linear bounded operators from E to F, and $\mathcal{L}(E, E)$ is denoted by $\mathcal{L}(E)$. Moreover, we denote by $L^2(I; E)$ and $\mathcal{C}^k(I; E)$ the space of square Lebesgue integrable functions on an interval I of the real line and the space of k times continuously differentiable functions on I with values in the Banach space E, respectively.

Following Flandoli, Lasiecka, and Triggiani [7], we consider the following dynamics

on Y:

$$\begin{cases} y(t) = e^{tA}y_0 + (Lu)(t), & y_0 \in Y, \\ (Lu)(t) = \displaystyle\int_0^t e^{(t-\tau)A}Bu(\tau)\ d\tau \end{cases} \tag{2.1}$$

that formally corresponds to the initial value problem

$$\begin{cases} \dot{y}(t) = Ay(t) + Bu(t) & \text{on } [\mathcal{D}(A^*)]', \\ y(0) = y_0, & y_0 \in Y. \end{cases} \tag{2.2}$$

A is an unbounded linear operator on Y with dense domain $\mathcal{D}(A)$ that we endow with the usual graph norm. A^* denotes the Y-adjoint of A and $[\mathcal{D}(A^*)]'$ is the dual of $\mathcal{D}(A^*)$ with respect to the Y-topology.

We shall make the following assumptions on operators A and B.

(H1) *A is the infinitesimal generator of a strongly continuous group e^{tA}, $t \in \mathbb{R}$, on Y.*

(H2) *B is linear continuous $U \longrightarrow [\mathcal{D}(A^*)]'$.*

The group property in assumption (H1) always holds for the linear time reversible systems we have in mind: second-order linear equations of hyperbolic or Petrowski type, first-order linear hyperbolic systems.

B is said to be a bounded control operator if $B \in \mathcal{L}(U, Y)$ and an unbounded one otherwise. Assumption (H2) on the control operator B enables one to consider boundary and point control problems for the examples of systems mentioned above (see [7], Lasiecka and Triggiani [15], and Bensoussan et al. [2], [3] for precise examples).

Note that, by identifying $[\mathcal{D}(A^*)]''$ with $\mathcal{D}(A)$, we get

$$B^* \in \mathcal{L}\left(\mathcal{D}(A^*); U\right).$$

As Y is assumed to be a Hilbert space, A^* is the generator of a strongly continuous semigroup which is the dual of the semigroup generated by A. Because of assumption (H1), it is the generator of a strongly continuous group e^{tA^*}, $t \in \mathbb{R}$ on Y.

We shall furthermore assume the following.

(H3) *Regularity property. For every $0 < T < \infty$ there is $C_T > 0$ such that*

$$\int_0^T \left\| B^* e^{-tA^*} y \right\|_U^2 dt \le C_T \|y\|_Y^2, \quad y \in \mathcal{D}(A^*). \tag{2.3}$$

(H4) *Controllability property. There is $T > 0$ and $c_T > 0$ such that*

$$\int_0^T \left\| B^* e^{-tA^*} y \right\|_U^2 dt \ge c_T \|y\|_Y^2, \quad y \in \mathcal{D}(A^*). \tag{2.4}$$

Usually, assumptions (H3) and (H4) are stated with $-A^*$ replaced by A^*, but, when considering the time reversibility assumption (H1), these assumptions are equivalent to the more standard ones (see Lemma 7.0 in Flandoli, Lasiecka, and Triggiani [7]). Note that, by density and continuity, inequalities (2.3) and (2.4) hold for any y on Y too.

By assumption (H3), we can extend the operator $B^*e^{-tA^*}$ to a continuous one from Y to $L_2(0, T; U)$. When B stands for a boundary control operator for a partial differential equation this assumption is a sharp trace regularity result with respect to partial differential equations theories for it is not obtained directly by just applying

the usual trace results (see, for instance, Lasiecka, Lions, and Triggiani [14] or Lions [19] in the case of the wave equation with Dirichlet boundary control). Also, it is well known that under this assumption the following regularity property holds [7]:

$$L \text{ continuous from } L^2(0,T;U) \text{ to } \mathcal{C}(0,T;Y) \quad \forall T > 0.$$

Assumption (H4) is the observability property for the dual system. Under all foregoing assumptions, we know (see Dolecki and Russell [6]) that it is equivalent to the controllability of the pair (A, B) over the time interval $(0, T)$:

For every $y_0,\ y_1 \in Y$ *there is* $u \in L^2(0,T;U)$ *such that the solution* y *of* (2.1) *satisfies* $y(T) = y^1$.

Moreover, it may hold only for T sufficiently large (like in the case of hyperbolic equations because of the finite speed of propagation). We refer the reader to Lions [19] and Komornik [11] for examples of this sort.

Before stating the main results we recall some basic facts about the asymptotic behavior of semigroups of linear operators that the reader can find in Pazy [23] and van Neerven [22]. Let us introduce *the type* $g(A)$ of a strongly continuous semigroup e^{At}, $t \geq 0$, of bounded linear operators on a Hilbert space Y:

$$g(A) = \inf_{t>0} \frac{1}{t} \log \left\| e^{tA} \right\|_{\mathcal{L}(Y)}.$$

$g(A)$ is also called the *growth bound* of the generated semigroup.

If $g(A) < 0$, then A is the generator of an exponentially stable semigroup and is said to be an exponentially stable generator. Moreover, $-g(A)$ is said to be the *stability margin* of the semigroup. Let us recall that $g(A)$ is finite or equal to $-\infty$,

$$g(A) = \lim_{t \longrightarrow \infty} \frac{1}{t} \log \left\| e^{tA} \right\|_{\mathcal{L}(Y)},$$

and $g(A)$ is the infimum of those $\omega \in \mathbb{R}$ for which there is $M_\omega > 0$ such that

$$\left\| e^{tA} \right\|_{\mathcal{L}(Y)} \leq M_\omega e^{\omega t}, \qquad t \geq 0.$$

Moreover, for every $\lambda \in \mathbb{R}$, the operator $A + \lambda I$ of domain $\mathcal{D}(A)$ is the generator of the strongly continuous semigroup $e^{\lambda t} e^{tA}$, $t \geq 0$, and

$$g(A + \lambda I) = g(A) + \lambda.$$

Finally, as Y is assumed to be a Hilbert space, $\left\| e^{tA} \right\|_{\mathcal{L}(Y)} = \left\| e^{tA^*} \right\|_{\mathcal{L}(Y)}$, $t \geq 0$, thus we have

$$g(A^*) = g(A). \tag{2.5}$$

We now state our main result.

THEOREM 2.1. *Consider* A *and* B *under assumptions* (H1)–(H4). *Let* $\omega \in \mathbb{R}$ *such that*

$$\omega > \max(g(-A), 0). \tag{2.6}$$

Then we have the following.

(i) *Operator* Λ_ω. *The symmetric positive operator* Λ_ω *defined by*

$$(\Lambda_\omega x, z)_Y = \int_0^\infty \left(B^* e^{-t(A+\omega I)^*} x, B^* e^{-t(A+\omega I)^*} z \right)_U dt\ , \qquad x,\ z \in Y, \tag{2.7}$$

is coercive and an isomorphism on Y.

(ii) *Wellposedness. Let* $F_\omega = -B^*\Lambda_\omega^{-1}$. *The operator* $A + BF_\omega$ *with*

$$\mathcal{D}(A + BF_\omega) = \Lambda_\omega\,(\mathcal{D}(A^*))$$

is the infinitesimal generator of the strongly continuous semigroup

$$T_\omega(t) = e^{t(A+BF_\omega)}, \quad t \geq 0,$$

on Y *and is exponentially stable.*

(iii) *Exponential decay rate. Let* $S_\omega(t)$ *be the strongly continuous semigroup on* Y *generated by* $-A^* - 2\omega I$. *Then*

$$T_\omega(t) = \Lambda_\omega S_\omega(t)\Lambda_\omega^{-1}, \quad t \geq 0, \tag{2.8}$$

so that $T_\omega(t)$ *is an exponentially stable semigroup of the type*

$$g(A + BF_\omega) = g(-A) - 2\omega. \tag{2.9}$$

Remark 2.2. (1) Assumption $\omega > g(-A)$ from (2.6) allows us to define the linear operator Λ_ω by (2.7), and together with assumption $\omega > 0$ they ensure that we have an exponentially stable resulting closed-loop feedback system as shown by (2.9). Moreover, ω is an upper bound of the stability margin which can thus be made as large as desired by conveniently choosing ω.

(2) In the case of skew-adjoint operators ($A = -A^*$, $\mathcal{D}(A^*) = \mathcal{D}(A)$), as announced in the introduction, the achieved stability margin is exactly 2ω since we have $g(-A) = g(A) = 0$.

3. Proof of the main results. The following proposition gives an existence and uniqueness result for a Liapunov equation and also assertion (i) of Theorem 2.1. It is a direct consequence of a more general result for Liapunov equations demonstrated by Grabowski [8] (see also Hansen and G. Weiss [9]). We omit the proof.

PROPOSITION 3.1. *Under assumptions* (H1)–(H4) *and with*

$$\omega > \max(0, g(-A))$$

the operator Λ_ω *as given by* (2.7) *is a well-defined self-adjoint bounded linear operator on* Y *and is coercive. Moreover,* Λ_ω *is the unique solution in* $\{\Lambda \in \mathcal{L}(Y); \Lambda = \Lambda^*\}$ *of the following Liapunov equation:*

$$\begin{aligned}&\forall x, z \in \mathcal{D}(A^*),\\ &\qquad (\Lambda_w(A + \omega I)^* x, z)_Y + (\Lambda_\omega x, (A + \omega I)^* z)_Y = (B^* x, B^* z)_Y.\end{aligned}$$

In the remaining part of this section we often use some general results on the linear quadratic regulator problem as they are given in the relatively complete text [7] (see also [15] for an exposition of the results) in order to establish the wellposedness of the particular feedback system we are dealing with.

We consider the optimal control problem on the infinite horizon:

For any $y_0 \in Y$, *find* $u^\infty \in \mathcal{U} \equiv L_2(0, \infty; U)$ *such that the solution* y^∞ *of* (2.1) *with* $u = u^\infty$ *satisfies*

$$J(y_0, u^\infty) = \inf_{u \in \mathcal{U}} J(y_0, u), \tag{3.1}$$

where

$$J(y_0,u) \equiv \int_0^\infty \left[(Ry(t),y(t))_Y + \|u(t)\|_U^2\right] dt$$

and y is the solution of (2.1) *due to u.*

Here, R is assumed to satisfy

$$R \in \mathcal{L}(Y), \quad R = R^* \geq 0. \tag{3.2}$$

From the general results in [7, Theorem 2.2, Theorem 2.3] we can extract the following result.

LEMMA 3.2. *Under assumptions* (H1)–(H4) *and with R satisfying* (3.2)*, there is $P_\infty \in \mathcal{L}(Y)$, $P_\infty = P_\infty^* \geq 0$, such that $A_{P_\infty} = A - BB^*P_\infty$ is the infinitesimal generator of an exponentially stable strongly continuous semigroup*

$$T_\infty(t) = e^{tA_{P_\infty}} = e^{t(A-BB^*P_\infty)}, \quad t \geq 0,$$

and P_∞ is a solution of the algebraic Riccati equation

$$\begin{aligned} &\forall x,z \in \mathcal{D}(A_{P_\infty}),\\ &\quad (Px,Az)_Y + (PAx,z)_Y + (Rx,z)_Y = (B^*Px,B^*Pz)_Y\,. \end{aligned} \tag{3.3}$$

If moreover $R > 0$, then P_∞ is the unique solution within the class of all $P \in \mathcal{L}(Y)$ such that

$$P = P^* \geq 0, \quad A^*P \in \mathcal{L}\left(\mathcal{D}(A_{P_\infty});Y\right).$$

Remark 3.3. Some of the statements of Lemma 3.2 hold in fact under weaker assumptions, but this lemma is sufficient for our purpose and we refer the reader to [7] for stronger results. Let us just report that under the assumptions of Lemma 3.2, the optimal pair (y^∞,u^∞) of the optimization problem (3.1) is such that

$$y^\infty(t) = T_\infty(t)y_0 = e^{t(A-BB^*P_\infty)}y_0, \quad t \geq 0,$$

and we have

$$(P_\infty y_0,y_0)_Y = J(y_0,u_\infty) \equiv \int_0^\infty \left[(Ry_\infty(t),y_\infty(t))_Y + \|u_\infty(t)\|_U^2\right] dt.$$

Some comments are in order. In order to prove Theorem 2.1, the strategy consists in taking $R = 2\omega\Lambda_\omega^{-1}$ and then deducing from the Liapunov equation (Proposition 3.1) and from the Riccati equation (3.3) that $P_\infty = \Lambda_\omega^{-1}$.

Unfortunately, the algebraic Riccati equation (3.3) holds in $\mathcal{D}(A_{P_\infty})$. As $\mathcal{D}(A_{P_\infty})$ depends on the solution itself, (3.3) is not suitable in order to achieve our goal. If the algebraic Riccati equation (3.3) were also valid on $\mathcal{D}(A)$, as it would be if the control operator B were bounded (see, e.g., Curtain and Zwart [5]), we could prove Theorem 2.1 directly.

In fact, as the left-hand side of (3.3) is well defined on $\mathcal{D}(A)$, (3.3) induces an extension $[B^*P]_e$ (in the linear span of $\mathcal{D}(A_{P_\infty})$ and $\mathcal{D}(A^*)$) so that we have

$$\begin{aligned} &\forall x,z \in \mathcal{D}(A),\\ &\quad (Px,Az)_Y + (PAx,z)_Y + (Rx,z)_Y = \left([B^*P]_e\,x,[B^*P]_e\,z\right)_Y \end{aligned} \tag{3.4}$$

with $P = P_\infty$. This extension of B^*P on $\mathcal{D}(A)$ was implicitly done in [7] to derive the same Riccati equation on $\mathcal{D}(A_{P_\infty})$ and $\mathcal{D}(A)$. This was pointed out by Weiss and Zwart [32], who show that extensions $[B^*P]_e$ of B^*P such that (3.4) holds are generally not unique. Starting from a uniquely identified extension of B^* and using the spectral factorization approach to the linear quadratic optimal control problem with unbounded control operators developed by Staffans [27], [28] and Weiss and Weiss [33], they determine the optimal feedback operator and give an associated Riccati equation on $\mathcal{D}(A)$ that does not correspond to (3.3) exactly. The authors of [32] recall that an additional spectral factorization assumption leads to a suitable Riccati equation on $\mathcal{D}(A)$.

Later or at the same time, using a perturbation/regularization approach Triggiani [29] and Barbu, Lasiecka, and Triggiani [1] better identified a suitable extension of B^* on $P\mathcal{D}(A)$ so that a Riccati equation on $\mathcal{D}(A)$ of the type of (3.4) holds.

We stop these comments here and refer the reader to the relevant literature for details since none of the algebraic Riccati equations on $\mathcal{D}(A)$ provided by the two approaches cited above will be directly used in the following. Instead, we use further results of [7] relating the Riccati operator P_∞ of Lemma 3.2 to the solution to a *dual* algebraic Riccati equation.

Given a self-adjoint positive operator $R \in \mathcal{L}(Y)$, consider the dual dynamics to (2.1)

$$\begin{cases} \dot{z}(t) = -A^*z(t) + R^{1/2}v(t) & \text{on } [\mathcal{D}(A)]', \\ z(0) = z_0, & z_0 \in Y. \end{cases} \tag{3.5}$$

For every $v \in L^2(0,T;Y)$ the solution z of (3.5) belongs to $C(0,T;Y)$ and $B^*z \in L^2(0,T;U)$ [7]. As usual, the controllability of the pair $\left(-A^*, R^{1/2}\right)$ in some time T means that the linear map

$$v \in L^2(0,T;Y) \longrightarrow z(T) \in Y,$$

where z is the solution of (3.5) with $z_0 = 0$, is onto. Consider now the following optimal control problem for this dual dynamics.

For every $z_0 \in Y$, find $v^\infty \in \mathcal{V} \equiv L^2(0,\infty;Y)$ such that the solution z^∞ of (3.5) *with $v = v^\infty$ satisfies*

$$K(z_0, v^\infty) = \inf_{v \in \mathcal{V}} K(z_0, v),$$

where

$$K(z_0, v) \equiv \int_0^\infty \left[\|B^*z(t)\|_U^2 + \|v(t)\|_Y^2 \right] dt$$

and z is the solution of (3.5) *due to v.*

From the results of Flandoli, Lasiecka, and Triggiani [7, Theorems 2.4, 2.6, and 2.7] we obtain the following one as a particular case.

LEMMA 3.4. *Under the assumptions of Lemma* 3.2, *if $\left(-A^*, R^{1/2}\right)$ is controllable, then P_∞ is positive definite and is an isomorphism on Y. If we set $Q_\infty = P_\infty^{-1}$, then Q_∞ is the unique solution in the class $\{Q \in \mathcal{L}(Y);\ Q = Q^* > 0\}$ of the dual algebraic Riccati equation*

$$\begin{aligned} &\forall x, z \in \mathcal{D}(A^*), \\ &\qquad (Qx, A^*z)_Y + (QA^*x, z)_Y + (RQx, Qz)_Y = (B^*x, B^*z)_Y. \end{aligned} \tag{3.6}$$

Moreover, $-A^* - RQ_\infty$ is the infinitesimal generator of a strongly continuous semigroup

$$S_\infty(t) = e^{t(-A^* - RQ_\infty)}, \quad t \geq 0,$$

on Y and we have the following relation:

$$S_\infty(t) = P_\infty T_\infty(t) Q_\infty, \quad t \geq 0. \tag{3.7}$$

Remark 3.5. The operator $Q_\infty = P_\infty^{-1}$ synthesizes the optimal pair (z^∞, v^∞) [7]:

$$z^\infty(t) = S_\infty(t) z_0 = e^{t(-A^* - RQ_\infty)} z_0, \quad t \geq 0,$$
$$v^\infty(t) = -R^{1/2} Q_\infty z^\infty(t), \quad \text{a.e. } t \geq 0.$$

Remark 3.6. The cited results of [7] use a lemma (Lemma 6.2 in [7]) whose proof contains a misprint in page 354, line 3, and the following lines were accidentally omitted [18], leading to confusion. The missing step in the proof of Lemma 6.2 is the demonstration of the dense inclusion of $\mathcal{D} \equiv P_\infty \mathcal{D}(A_{P_\infty})$ into $\mathcal{D}(A^*)$ (with our proper notations). As explained by Lasiecka and Triggiani [18] an exactly similar argument to the use of Lemma 6.2 of [7] occurs in Lasiecka [13, Proposition 3.1 to Lemma 3.1, pp. 249–251] with all the lines in place and gives the perfect counterpart of Lemma 6.2 in [7]. More precisely, it occurs in the first part of Lemma 3.1 in $\mathcal{D}(A^*_{P_\infty})$, but the same scheme of proof and the same arguments lead to the desired result. Finally, according to [18] it seems that under the same assumptions the stronger result $\mathcal{D} = \mathcal{D}(A^*)$ holds. This is, for instance, stated in [1] without proof, and [18] refers to the third volume to be published as a sequel to [16], [17]. A complete proof is to appear in [30].

We can now end the proof of Theorem 2.1 with the use of Proposition 3.1 and Lemmas 3.2 and 3.4.

Proof of Theorem 2.1. The proof of part (i) is given in Proposition 3.1. Again by Proposition 3.1 Λ_ω is the unique solution of the Liapunov equation

$$\begin{aligned} &\forall x, z \in \mathcal{D}(A^*), \\ &\quad (\Lambda_w(A + \omega I)^* x, z)_Y + (\Lambda_\omega x, (A + \omega I)^* z)_Y = (B^* x, B^* z)_Y. \end{aligned} \tag{3.8}$$

Thus Λ_ω satisfies

$$\begin{aligned} &\forall x, z \in \mathcal{D}(A^*), \\ &\quad (\Lambda_\omega A^* x, z)_Y + (\Lambda_\omega x, A^* z)_Y + 2\omega \left(\Lambda_\omega^{-1} \Lambda_\omega x, \Lambda_\omega z\right)_Y = (B^* x, B^* z)_Y. \end{aligned} \tag{3.9}$$

Set $R = 2\omega \Lambda_\omega^{-1}$. Since $\omega > 0$, R is also a coercive self-adjoint bounded linear operator on Y. By Lemma 3.2 there exists a positive self-adjoint bounded linear operator P_∞ satisfying (3.3) (and it is its unique solution in the sense explained in Lemma 3.2), and $A - BB^* P_\infty$ is an exponentially stable generator of a semigroup on Y. Assume that $\left(-A^*, R^{1/2}\right)$ is controllable in some time T; then Lemma 3.4 gives that P_∞ is an isomorphism and $Q_\infty = P_\infty^{-1}$ is the unique solution of the algebraic Riccati equation

$$\begin{aligned} &\forall x, z \in \mathcal{D}(A^*), \\ &\quad (Qx, A^* z)_Y + (QA^* x, z)_Y + 2\omega \left(\Lambda_\omega^{-1} Qx, Qz\right)_Y = (B^* x, B^* z)_Y \end{aligned} \tag{3.10}$$

in the class $\{Q \in \mathcal{L}(Y);\ Q = Q^*\}$. As Q_∞ is the unique solution, from (3.9) and (3.10) we deduce that

$$Q_\infty = \Lambda_\omega, \quad P_\infty = \Lambda_\omega^{-1}.$$

The proof of part (ii) is thus almost complete if we prove that with $R = 2\omega\Lambda_\omega^{-1}$ the pair $(-A^*, R^{1/2})$ is controllable. From Louis and Wexler [20] we know that if A generates a group, then (A, I) is controllable. From the invertibility of $R^{1/2}$ we easily get the desired result.

By Lemma 3.4 we have

$$S_\infty(t) = e^{(-A-RQ_\infty)^* t} = e^{(-A^*-2\omega I)t}, \quad t \geq 0,$$

and relation (3.7) in Lemma 3.4 writes

$$e^{(-A^*-2\omega I)t} = \Lambda_\omega^{-1} e^{\left(A-BB^*\Lambda_\omega^{-1}\right)t}\Lambda_\omega \ , \quad t \geq 0, \tag{3.11}$$

which reduces to (2.8) in Theorem 2.1. Relation $\mathcal{D}(A + BF_\omega) = \Lambda_\omega\left(\mathcal{D}(A^*)\right)$ and equality

$$g(A + BF_\omega) = g(-A^*) - 2\omega$$

are consequences of (2.8) or (3.11) and of the definition of the growth bound of a semigroup (see Schumacher [25, Lemma 4.5]). This ends the proof of part (ii) and part (iii) in Theorem 2.1. □

Acknowledgments. The author is indebted to F. Bourquin for his encouragements, to the anonymous referees whose suggestions have led to a significant improvement of the paper, and, last but not least, to I. Lasiecka and R. Trigianni for their extremely useful comments and suggestions.

REFERENCES

[1] V. Barbu, I. Lasiecka, and R. Triggiani, *Extended algebraic Riccati equations in the abstract hyperbolic case*, Nonlinear Anal., 40 (2000), pp. 105–129.

[2] A. Bensoussan, G. Da Prato, M. Delfour, and S. Mitter, *Representation and Control of Infinite-Dimensional Systems,* Vol. 1, Birkhäuser, Boston, Basel, Berlin, 1993.

[3] A. Bensoussan, G. Da Prato, M. Delfour, and S. Mitter, *Representation and Control of Infinite-Dimensional Systems,* Vol. 2, Birkhäuser, Boston, Basel, Berlin, 1993.

[4] J. Briffaut, *Méthodes numériques pour le contrôle et la stabilisation de grandes structures élastiques*, thèse de l'Ecole Nat. des Ponts et Chaussées, 1999.

[5] R. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[6] S. Dolecki and D. L. Russell, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–220.

[7] F. Flandoli, I. Lasiecka, and R. Triggiani, *Algebraic Riccati equations with non-smoothing observation arising in hyperbolic and Euler–Bernouilli boundary control problems*, Ann. Mat. Pura Appl. (4), 153 (1988), pp. 307–382.

[8] P. Grabowski, *On the spectral-Lyapunov approach to parametric optimization of distributed parameter systems*, IMA J. Math. Control Inform., 7 (1990), pp. 317–338.

[9] S. Hansen and G. Weiss, *New results on the operator Carleson measure criterion*, IMA J. Math. Control Inform., 14 (1997), pp. 3–32.

[10] D. Kleinman, *An easy way to stabilize a linear constant system*, IEEE Trans. Automat. Control, 15 (1970), p. 692.

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

[12] V. Komornik, *Rapid boundary stabilization of linear distributed systems*, SIAM J. Control Optim., 35 (1997), pp. 1591–1613.

[13] I. Lasiecka, *Exponential stabilization of hyperbolic systems with nonlinear, unbounded perturbations—Riccati operator approach*, Appl. Anal., 42 (1991), pp. 243–261.

[14] I. Lasiecka, J.-L. Lions, and R. Triggiani, *Nonhomogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl. (9), 65 (1986), pp. 149–192.

[15] I. Lasiecka and R. Triggiani, *Differential and Algebraic Riccati Equations with Application to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, Lecture Notes in Control and Inform. Sci. 164, Springer-Verlag, Berlin, 1991.
[16] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations*: *Continuous and Approximation Theories* I, Encyclopedia Math. Appl. 74, Cambridge University Press, Cambridge, UK, 2000.
[17] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations*: *Continuous and Approximation Theories* II, Encyclopedia Math. Appl. 75, Cambridge University Press, Cambridge, UK, 2000.
[18] I. Lasiecka and R. Triggiani, *private communication*, 2002.
[19] J.-L. Lions, *Controlabilité Exacte, Perturbations et Stabilisation de Systèmes distribués*, Vol. 1, Masson, Paris, 1988.
[20] J.-C. Louis and D. Wexler, *On exact controllability in Hilbert spaces*, J. Differential Equations, 49 (1983), pp. 258–269.
[21] D. Lukes, *Stabilizability and optimal control*, Funkcial. Ekvac., 11 (1968), pp. 39–50.
[22] J. van Neerven, *The Asymptotic Behavior of Semigroups of Linear Operators*, Oper. Theory Adv. Appl. 88, Birkhäuser Verlag, Basel, 1996.
[23] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.
[24] D. Russell, *Mathematics of Finite Dimensional Control Systems*, Marcel Dekker, New York, Basel, 1979.
[25] J. M. Schumacher, *Dynamic Feedback in Finite- and Infinite-Dimensional Linear Systems*, Math. Centre Tracts 143, Mathematisch Centrum, Amsterdam, 1981.
[26] M. Slemrod, *A note on complete controllability and stabilizability for linear control systems*, SIAM J. Control Optim., 12 (1974), pp. 500–508.
[27] O. J. Staffans, *Coprime factorizations and well-posed linear systems*, SIAM J. Control Optim., 36 (1998), pp. 1268–1292.
[28] O. J. Staffans, *Quadratic optimal control of well-posed linear systems*, SIAM J. Control Optim., 37 (1998), pp. 131–164.
[29] R. Triggiani, *The algebraic Riccati equation with unbounded control operator: The abstract hyperbolic case revisited*, Contemp. Math., 209 (1997), pp. 315–338.
[30] R. Triggiani, *The dual algebraic Riccati equations: Additional results under isomorphism of the Riccati operator*, Appl. Math. Lett., to appear.
[31] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.
[32] G. Weiss and H. Zwart, *An example in linear quadratic optimal control*, Systems Control Lett., 33 (1998), pp. 339–349.
[33] M. Weiss and G. Weiss, *Optimal control of stable weakly regular linear systems*, Math. Control Signals Systems, 10 (1997), pp. 287–330.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2245–2277

STOCHASTIC CONTROL FOR LINEAR SYSTEMS DRIVEN BY FRACTIONAL NOISES*

YAOZHONG HU† AND XUN YU ZHOU‡

Abstract. This paper is concerned with optimal control of stochastic linear systems involving fractional Brownian motion (FBM). First, as a prerequisite for studying the underlying control problems, some new results on stochastic integrals and stochastic differential equations associated with FBM are established. Then, three control models are formulated and studied. In the first two models, the state is scalar-valued and the control is taken as Markovian. Either the problems are completely solved based on a Riccati equation (for model 1, where the cost is a quadratic functional on state and control variables) or optimality is characterized (for model 2, where the cost is a power functional). The last control model under investigation is a general one, where the system involves the Stratonovich integral with respect to FBM, the state is multidimensional, and the admissible controls are not limited to being Markovian. A new Riccati-type equation, which is a backward stochastic differential equation involving both FBM and normal Brownian motion, is introduced. Optimal control and optimal value of the model are explicitly obtained based on the solution to this Riccati-type equation.

Key words. fractional Brownian motion (FBM), stochastic linear-quadratic (LQ) control, Itô integral, Stratonovich integral, Hu–Meyer formula, multiple integral, Riccati equation, Malliavin derivative

AMS subject classifications. 60H05, 60H07, 93E20

DOI. 10.1137/S0363012903426045

1. Introduction.

Stochastic processes with long memory have been studied in the literature. One important class of such processes is the fractional Brownian motion (FBM). In recent years there has been considerable research interest in stochastic calculus for FBM. A central issue is to define a proper stochastic integral with respect to the FBM. Attempts were made in Lin [25] and Dai and Heyde [6]; however, the stochastic integrals defined in [25], [6] do not satisfy the familiar zero-mean property as opposed to the normal Brownian motion counterpart. More recently, Duncan, Hu, and Pasik-Duncan [9] employed the Wick calculus to define a fractional stochastic integral whose mean is indeed zero. This property is very convenient for both theoretical development and practical applications. A substantial stochastic calculus was established for this new integral in [9].

It is natural and interesting to study the optimal control of systems driven by fractional noises, namely, systems described by stochastic differential equations (SDEs) involving FBM. Some such control problems as well as their applications in mathematical finance have been studied in [3], [15], [20], [22] within the framework of fractional calculus of [9].

*Received by the editors April 10, 2003; accepted for publication (in revised form) August 29, 2004; published electronically May 27, 2005.

http://www.siam.org/journals/sicon/43-6/42604.html

†Department of Mathematics, University of Kansas, 405 Snow Hall, Lawrence, KS 66045-2142 (hu@math.ku.edu), and Wuhan Institute of Physics and Mathematics, The Chinese Academy of Science, Wuhan 430071, China. The work of this author was supported in part by National Science Foundation grants DMS 0204613 and EPS-9874732.

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk). The work of this author was supported in part by RGC Earmarked grants CUHK4435/99E, CUHK4175/00E, and CUHK4234/01E.

On the other hand, a classically important class of stochastic control problems is the so-called linear-quadratic (LQ) control class, where the system is linear with respect to the state and control variables and the cost functional is quadratic in the two variables. Such a problem can often be solved explicitly by solving a relevant Riccati equation. While stochastic LQ control for systems driven by the (normal) Brownian motion (i.e., the systems are perturbed by the usual "white noise") is a classical problem on which a vast amount of research has been done, there has recently been renewed interest in the problem, in the form of the so-called "indefinite stochastic LQ control." See [1], [4], [5], [23], [28], [29], and the references therein.

This paper represents the first attempt, to our knowledge, of systematically solving some optimal control problems on linear systems driven by FBM, including LQ control with fractional noises. The controlled systems under consideration include

$$dx_t = (A_t x_t + B_t u_t)dt + (C_t x_t + D_t u_t)dW_t^H, \quad x_0 = x \text{ is given} \tag{1.1}$$

and

$$dx_t = (A_t x_t + B_t u_t)dt + (C_t x_t + D_t u_t) \circ dW_t^H, \quad x_0 = x \text{ is given}, \tag{1.2}$$

where A_t, B_t, C_t, and D_t are given (matrix-valued) stochastic processes, $\{W_t^H, t \geq 0\}$ is an FBM of Hurst parameter $H \in [1/2, 1)$, and dW_t^H and $\circ dW_t^H$ are the Itô and Stratonovich integrals, respectively, as defined in [9]. Our problem is to minimize a cost functional under constraint (1.1) or (1.2).

The above problem, in its greatest generality, seems to be too difficult to solve at the moment owing to the long-range-dependence property of the FBM. In this paper we propose and study three special models. The first model is for system (1.1), where the state x_t is one-dimensional, D_t is absent, and the cost functional is quadratic. Moreover, we consider optimal control among the class of linear Markovian controls, namely, those of the form

$$u_t = K_t x_t, \tag{1.3}$$

where K_t is a deterministic (matrix-valued) function of t. (Notice that linear Markovian control constitutes an important control class due to its mathematical simplicity and ease of implementation. Also, it is well known that the optimal stochastic LQ control with normal Brownian motion turns out to be a linear Markovian control.) In this case, the optimal control (or, equivalently, the optimal K_t) is found based on the solution to a Riccati equation. This Riccati equation can be solved by using the classical theory. Thus, for this special yet interesting control model we obtain a complete solution to the LQ problem with fractional noises.

The second model is more general than the first in that $D_t \not\equiv 0$ and the cost functional is of a power form. The presence of D_t is a crucially different and difficult feature even in the normal stochastic LQ control (see [29] for detailed discussions). For this second model we use calculus of variation to obtain an integral equation that the optimal K_t must satisfy. We then demonstrate how to solve this integral equation in several special cases.

The third control model is an LQ problem, where the system is governed by (1.2), the state variable is multidimensional, and the admissible controls are not limited to being Markovian. However, D_t is assumed to be identically zero. We introduce a Riccati-type equation, which is a backward SDE involving integrals with respect to both FBM and normal Brownian motion. Assuming the solvability of this equation,

along with some technical conditions, we derive an optimal control along with the optimal value based on the completion-of-square technique. As a by-product the optimal control is in linear Markovian form. We also illustrate how we obtain such a Riccati-type equation heuristically using the method of approximation.

It is worth mentioning that the study of stochastic control problems with fractional noises will inevitably involve stochastic calculus on FBM, the associated stochastic integrals, and differential equations. However, a rich, complete theory on such calculus is still lacking because the very definition of the fractional stochastic integral we are employing was introduced only very recently [9]. Therefore, in this paper we make some effort to derive results on fractional stochastic calculus that are necessary for studying the subsequent control models. Some of the results are new and interesting in their own right.

The remainder of the paper is organized as follows. Section 2 is devoted to fractional stochastic calculus useful in what follows. Three different control models are treated in sections 3, 4, and 5, respectively. Finally, section 6 concludes the paper with some remarks.

2. Fractional stochastic calculus. This section presents stochastic calculus on FBM that is necessary for the control problems under investigation. Most of the results are new and interesting in their own right from the point of view of fractional calculus.

2.1. FBM. In this subsection we outline some of the notation and results of FBM. The primary reference in this regard is [9]. Let $\Omega = C_0(0, T; \mathbb{R})$ be the Banach space of a real-valued continuous function on $[0, T]$ with the initial value zero and the super norm. There is a probability measure P on $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is the Borel σ-algebra on Ω, such that on the probability space $(\Omega, \mathcal{F}, P)$ the process $\{W_t^H, 0 \le t \le T\}$ defined as

$$W_t^H(\omega) = \omega(t) \ \ \forall \omega \in \Omega$$

is a (one-dimensional) Gaussian process with mean

$$\mathbb{E}\ W_t^H = W_0^H = 0 \ \ \forall t \in [0, T]$$

and covariance

$$\mathbb{E}\ \left(W_t^H W_s^H\right) = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right) \ \ \forall t, s \in [0, T].$$

This canonical process $\{W_t^H, 0 \le t \le T\}$ is called a (standard) FBM of *Hurst parameter* H. In this paper $H \in [\frac{1}{2}, 1)$ is fixed. An FBM with $H = \frac{1}{2}$ reduces to the (normal) Brownian motion.

An FBM generates a filtration $\{\mathcal{F}_t,\ 0 \le t \le T\}$ with $\mathcal{F}_t = \sigma\{W_s^H, 0 \le s \le t\}$. Throughout this paper, only the canonical process and the associated filtered probability space $(\Omega, \mathcal{F}, P; \mathcal{F}_t)$, as described above, are used. Also notice that the FBM is assumed to be one-dimensional only for notational simplicity; there is no essential difficulty with a multidimensional FBM, where each component is a one-dimensional FBM with the *same* Hurst parameter H.

The following lemma is useful (see [7], [13], [26]).

LEMMA 2.1. *An FBM W_t^H with $H \in (\frac{1}{2}, 1)$ can be represented as*

$$W_t^H = \int_0^t Z(t, s) dW_s, \tag{2.1}$$

where $\{W_t\,, 0 \le t \le T\}$ *is a one-dimensional Brownian motion, and*

$$Z(t,s) := \left(H - \frac{1}{2}\right) c_H s^{(1/2)-H} \int_s^t r^{H-1/2}(r-s)^{H-(3/2)} dr,$$

with

$$c_H = \sqrt{\frac{H(2H-1)\Gamma\left(\frac{3}{2} - H\right)}{\Gamma\left(H - \frac{1}{2}\right)\Gamma(2-2H)}}$$

($\Gamma(\cdot)$ *is the gamma function*). *Moreover,* W_t *and* W_t^H *generate the same filtrations.*

Similar to the normal Brownian motion, the FBM is not differentiable a.s. (almost surely). As the "derivative" of a Brownian motion is usually referred to as white noise, that of an FBM is referred to as *fractional noise*.

For a fixed $H \in (\frac{1}{2}, 1)$, denote a function $\phi : [0,T] \times [0,T] \to \mathbb{R}_+$ by

$$\phi(s,t) := H(2H-1)|s-t|^{2H-2} \quad \forall s,t \in [0,T], \tag{2.2}$$

and let $L_\phi^2([0,T])$ be the Hilbert space of Borel measurable, scalar-valued functions f such that

$$|f|_\phi^2 := \int_0^T\!\!\int_0^T \phi(s,t) f_s f_t ds dt < \infty, \tag{2.3}$$

with the norm $|f|_\phi$; see [11], [12], [13] for a detailed discussion on this space. The inner product on $L_\phi^2([0,T])$ is denoted by $\langle \cdot, \cdot \rangle_\phi$. Define a mapping Φ on $L_\phi^2([0,T])$:

$$(\Phi g)_t := \int_0^T \phi(t,s) g_s ds \quad \forall g \in L_\phi^2([0,T]). \tag{2.4}$$

Then one can define the ϕ-derivative of a random variable $F \in L^p(\Omega, \mathcal{F}, P)$ $(p \ge 1)$ in the direction of Φg, where $g \in L_\phi^2([0,T])$, by

$$D_{\Phi g} F(\omega) := \lim_{\delta \to 0} \frac{1}{\delta} \left\{ F\left(\omega + \delta \int_0^{\cdot} (\Phi g)_t dt\right) - F(\omega) \right\} \tag{2.5}$$

if the limit exists in $L^p(\Omega, \mathcal{F}, P)$. Moreover, if there is a process $\{D_s^H F,\ 0 \le s \le T\}$ satisfying

$$D_{\Phi g} F = \int_0^T [D_s^H F] g_s ds \quad \text{a.s.} \quad \forall g \in L_\phi^2([0,T]), \tag{2.6}$$

then F is said to be ϕ-*differentiable* and $D_s^H F$ is called the *Malliavin derivative* of F. Refer to [9], [19] for more information about the Malliavin derivative.

Finally, we remark that if $H = 1/2$, then the function $\phi(\cdot,\cdot)$ should be replaced by the Delta function $\delta(\cdot,\cdot)$ satisfying

$$\int_0^T \int_0^T \delta(s,s') f(s,s') ds ds' = \int_0^T f(s,s) ds \quad \forall f(\cdot,\cdot) \in L^1([0,T] \times [0,T]).$$

2.2. Stratonovich integrals. Next, we discuss the Stratonovich stochastic integral, which will be dealt with later in this paper. First of all, let X_t be a semimartingale of the form

$$X_t = X_0 + \int_0^t f_s ds + \int_0^t g_s dW_s,$$

where $\{W_s, 0 \le s \le T\}$ is the Brownian motion given by Lemma 2.1 and X_0 is a deterministic constant. If $f, g \in L^2_{\mathcal{F}}(0,T)$, where $L^2_{\mathcal{F}}(0,T)$ denotes the set of scalar-valued, $\mathcal{F}_t$-adapted, square integrable processes on $[0,T]$, then it is a well-known result that (see, e.g., [21, Chapter III])

$$\int_0^t X_s \circ dW_s = \int_0^t X_s dW_s + \frac{1}{2}\int_0^t g_s ds \quad \forall t \in [0,T], \tag{2.7}$$

where $\int_0^t X_s \circ dW_s$ denotes the Stratonovich (symmetric) integral with respect to the Brownian motion W_s. As a result,

$$\mathbb{E}\left(\int_0^T X_s \circ dW_s\right) = \frac{1}{2}\mathbb{E}\int_0^T g_s ds. \tag{2.8}$$

However, if a process Y_t is given as

$$Y_t = Y_0 + \int_0^t f_s ds + \int_0^t g_s dW_s^H, \tag{2.9}$$

where the stochastic integral with respect to the FBM W_s^H is defined as in [9], then we have a quite different result.

THEOREM 2.1. *Let f_s and g_s satisfy the following conditions:*

$$\mathbb{E}\int_0^T |f_s|^2 ds < \infty, \quad \sup_{0\le s\le T}\mathbb{E}\,|g_s|^2 ds < \infty, \quad \text{and} \quad \mathbb{E}\int_0^T |D_s^H g_s|^2 ds < \infty, \tag{2.10}$$

and let Y_t be given by (2.9)*, where $H \in (\frac{1}{2}, 1)$. Then*

$$\int_0^t Y_s \circ dW_s = \int_0^t Y_s dW_s \quad \forall t \in [0,T]. \tag{2.11}$$

As a consequence, we have

$$\mathbb{E}\left[\int_0^t Y_s \circ dW_s\right] = 0. \tag{2.12}$$

Proof. Under condition (2.10) on f and g we see that Y_t is well defined (see [9]). Let $\pi : 0 = t_0 < t_1 < t_2 < \cdots < t_n = t$ be a partition of the interval $[0,t]$. By definition,

$$\begin{aligned}
\int_0^t Y_s \circ dW_s &= \lim_{|\pi|\to 0}\sum_{k=0}^{n-1}\frac{1}{2}\left(Y_{t_k} + Y_{t_{k+1}}\right)\left(W_{t_{k+1}} - W_{t_k}\right) \\
&= \lim_{|\pi|\to 0}\sum_{k=0}^{n-1} Y_{t_k}\left(W_{t_{k+1}} - W_{t_k}\right) + \lim_{|\pi|\to 0}\sum_{k=0}^{n-1}\frac{1}{2}\left(Y_{t_{k+1}} - Y_{t_k}\right)\left(W_{t_{k+1}} - W_{t_k}\right) \\
&:= \lim_{|\pi|\to 0} I_1 + \lim_{|\pi|\to 0} I_2,
\end{aligned}$$

where the limit is taken in the sense of "in probability" and $|\pi| := \max_{0\le k\le n-1}(t_{k+1}-t_k)$. It is easy to see by the definition of the Itô integral that the term I_1 converges to $\int_0^t Y_s dW_s$.

To consider the term I_2 we estimate $\mathbb{E}\,|Y_{t_{k+1}}-Y_{t_k}|^2$. From the definition of Y_t it follows that

$$\begin{aligned}\mathbb{E}\,\left|Y_{t_{k+1}}-Y_{t_k}\right|^2 &\le 2\mathbb{E}\,\left|\int_{t_k}^{t_{k+1}} f_s ds\right|^2 + 2\mathbb{E}\,\left|\int_{t_k}^{t_{k+1}} g_s dW_s^H\right|^2\\ &:= 2(I_3+I_4).\end{aligned}$$

It follows from (2.10) that

$$I_3 \le \mathbb{E}\int_{t_k}^{t_{k+1}} |f_s|^2 ds\cdot|t_{k+1}-t_k|. \tag{2.13}$$

Moreover, by virtue of the Itô isometry [9, Theorem 3.9], we have

$$\begin{aligned}I_4 &= \int_{t_k}^{t_{k+1}}\int_{t_k}^{t_{k+1}} \phi(t,s)\mathbb{E}\,(g_s g_t)\,dsdt + \mathbb{E}\,\left[\int_{t_k}^{t_{k+1}} D_s^H g_s ds\right]^2\\ &:= I_5+I_6.\end{aligned}$$

The assumption of the theorem and the definition of ϕ yield

$$\begin{aligned}I_5 &\le C\int_{t_k}^{t_{k+1}}\int_{t_k}^{t_{k+1}} \phi(t,s)dsdt\\ &\le C|t_{k+1}-t_k|^{2H},\end{aligned}$$

where here (and elsewhere) C is a generic positive constant that may vary from place to place. On the other hand,

$$I_6 \le \mathbb{E}\int_{t_k}^{t_{k+1}} |D_s^H g_s|^2 ds\cdot|t_{k+1}-t_k|.$$

Combining the above computations, we conclude that there is a constant C such that

$$\mathbb{E}\,|Y_{t_{k+1}}-Y_{t_k}|^2 \le C|t_{k+1}-t_k|^{2H} + U_k|t_{k+1}-t_k|, \tag{2.14}$$

where

$$U_k := \mathbb{E}\int_{t_k}^{t_{k+1}} (|f_s|^2+|D_s^H g_s|^2)ds.$$

Therefore,

$$\begin{aligned}2\mathbb{E}\,I_2 &\le \sum_{k=0}^{n-1}\mathbb{E}\,\left|Y_{t_{k+1}}-Y_{t_k}\right|\left|W_{t_{k+1}}-W_{t_k}\right|\\ &\le \sum_{k=0}^{n-1}\left[\mathbb{E}\,\left|Y_{t_{k+1}}-Y_{t_k}\right|^2\right]^{1/2}\left[\mathbb{E}\,\left|W_{t_{k+1}}-W_{t_k}\right|^2\right]^{1/2}\\ &\le C\sum_{k=0}^{n-1}(t_{k+1}-t_k)^{1/2+H} + C\sum_{k=0}^{n-1}[U_k(t_{k+1}-t_k)]^{1/2}\left[\mathbb{E}\,\left|W_{t_{k+1}}-W_{t_k}\right|^2\right]^{1/2}\\ &:= I_7+I_8.\end{aligned}$$

It is easy to check that

$$I_7 \le C \max_{0\le k\le n-1}(t_{k+1}-t_k)^{H-1/2}\sum_{k=0}^{n-1}(t_{k+1}-t_k) \to 0$$

and

$$\begin{aligned} I_8 &\le C\sum_{k=0}^{n-1} U_k^{1/2}(t_{k+1}-t_k) \\ &\le C\left(\sum_{k=0}^{n-1} U_k\right)^{1/2}\left(\sum_{k=0}^{n-1}(t_{k+1}-t_k)^2\right)^{1/2} \\ &\le C\left[\mathbb{E}\int_0^T (|f_s|^2+|D_s^H g_s|^2)ds\right]^{1/2}|\pi|^{1/2} \to 0. \end{aligned}$$

This implies that I_2 converges to 0 in $L^1(\Omega,\mathcal{F},P)$ and, consequently, in probability. This completes the proof. □

Remark 2.1. Whereas Theorem 2.1 is useful in section 5, it is also interesting in its own right. In particular, we see that (2.11) is different from the case when the integrand is a semimartingale; see (2.7). On the other hand, one has a different formula than (2.11) if the integration there is replaced by one with respect to the FBM W_s^H (under different conditions on the integrand, however); for details see [9, Theorem 3.12].

Remark 2.2. Since the Stratonovich-type integral can be represented by using the Itô-type integral [9], (2.9) and (2.10) are still true if Y_t is represented as a Stratonovich integral

$$Y_t = Y_0 + \int_0^t f_s ds + \int_0^t g_s \circ dW_s^H$$

under some mild conditions on f and g. Details are left to the interested readers.

2.3. Multiple integrals. In this section we recall some results from [9] on multiple integrals. These results are needed in this paper and we refer to [9] for more detail.

A function $f:[0,T]^n \to \mathbb{R}$ is called symmetric on $[0,T]^n$ if

$$f(s_{i_1},\dots,s_{i_n}) = f(s_1,\dots,s_n),\quad (s_1,\dots,s_n)\in[0,T]^n$$

for any permutation $(i_1,i_2,\dots,i_n)$ of $(1,2,\dots,n)$. Denote

$$\begin{aligned} L_\phi^2([0,T]^n) := \{&f:[0,T]^n\to\mathbb{R} \text{ is measurable and symmetric in its arguments,} \\ &|f|_\phi^2 := \langle f,f\rangle_\phi < \infty\}, \end{aligned}$$

where

$$\begin{aligned} \langle f,g\rangle_\phi := \int_{[0,T]^{2n}} &\phi(u_1,v_1)\phi(u_2,v_2)\cdots\phi(u_n,v_n)f(u_1,u_2,\dots,u_n) \\ &\cdot g(v_1,v_2,\dots,v_n)du_1du_2\cdots du_n dv_1dv_2\cdots dv_n. \end{aligned}$$

If $f \in L^2_\phi([0,T]^n)$, then the multiple Itô integral

$$\begin{aligned} I_n(f) \equiv I_{n,T}(f) &= \int_{0\le t_1,\dots,t_n\le T} f(t_1,\dots,t_n) dW^H_{t_1}\cdots dW^H_{t_n} \\ &= n! \int_{0\le t_1<\cdots<t_n\le T} f(t_1,\dots,t_n) dW^H_{t_1}\cdots dW^H_{t_n} \end{aligned}$$

is well defined, and (see [9, Lemma 6.6][1])

$$\mathbb{E}\ (I_n(f)I_m(g)) = \begin{cases} n!\langle f,g\rangle_\phi & \text{if } n=m, \\ 0 & \text{if } n\ne m. \end{cases} \tag{2.15}$$

The following result concerns the Malliavin derivative of a multiple integral.

THEOREM 2.2. *We have*

$$D^H_s I_n(f_n) = nI_{n-1}\left(\int_0^t \phi(s,r) f_{n-1}(r)dr\right) \quad \forall f_n \in L^2_\phi([0,T]^n), \tag{2.16}$$

where for any $0\le r\le T$, $f_{n-1}(r)$ denotes a function of $n-1$ variables given by

$$f_{n-1}(r)(s_1,\dots,s_{n-1}) := f_n(s_1,\dots,s_{n-1},r).$$

Proof. For all $\delta\in\mathbb{R}$ and $h\in L^2_\phi([0,T])$, one has

$$\begin{aligned} &I_n(f_n)(W^H + \delta\int_0^\cdot h(s)ds) \\ &\quad = I_n(f_n)(W^H) + \delta\sum_{k=1}^n \int_{0<s_1,\dots,s_n<T} \\ &\qquad\qquad \cdot f_n(s_1,\dots,s_n) dW^H_{s_1}\cdots \widehat{dW^H_{s_k}}\cdots dW^H_{s_n} h(s_k)ds_k + o(\delta) \\ &\quad = I_n(f_n)(W^H) + \delta n\int_0^T I_{n-1}(f_{n-1}(r))h(r)dr + o(\delta), \end{aligned}$$

where $\widehat{dW^H_{s_k}}$ means that the $dW^H_{s_k}$ term is taken off the product. Thus

$$\begin{aligned} D_{\Phi g}F &= n\int_0^T I_{n-1}(f_{n-1}(r))\Phi g(r)dr \\ &= n\int_0^T I_{n-1}(f_{n-1}(r))\int_0^T \phi(s,r)g(s)dsdr. \end{aligned}$$

By the definition of D^H_s (see (2.6)) we have

$$\begin{aligned} D^H_s I_n(f_n) &= n\int_0^T \phi(s,r) I_{n-1}(f_{n-1}(r))dr \\ &= nI_{n-1}\left(\int_0^T \phi(s,r) f_{n-1}(r)dr\right). \end{aligned} \tag{2.17}$$

□

[1]There is a typo in [9, Lemma 6.6]: the factorial $n!$ is missing from the term $\langle f,g\rangle_\phi$ on the right-hand side of (6.5) there.

We also need to employ the multiple Stratonovich integrals. For this we first recall the definition of the k-trace Tr_ϕ^k. Let $f \in L^2_\phi([0,T]^n)$. For any positive integer $k \le n/2$, the k-trace of f is defined as a function of $n-2k$ variables:

$$\mathrm{Tr}_\phi^k f(t_1,\dots,t_{n-2k}) := \int_{[0,T]^{2k}} f(s_1,s_2,\dots,s_{2k-1},s_{2k},t_1,\dots,t_{n-2k}) \cdot\phi(s_1,s_2)\phi(s_3,s_4)\cdots\phi(s_{2k-1},s_{2k})ds_1\cdots ds_{2k}$$

if the right-hand side is integrable.

The multiple Stratonovich integral

$$S_n(f) \equiv S_{n,T}(f) = \int_{0\le t_1,\dots,t_n\le T} f(t_1,\dots,t_n)\circ dW_{t_1}^H\cdots\circ dW_{t_n}^H = n!\int_{0\le t_1<\cdots<t_n\le T} f(t_1,\dots,t_n)\circ dW_{t_1}^H\cdots\circ dW_{t_n}^H$$

is well defined, and one has the following Hu–Meyer formula (see [16], [17], [18], and [9, equation (6.10)]):

$$S_n(f) = \sum_{k\le[n/2]} \frac{n!}{2^k k!(n-2k)!} I_{n-2k}(\mathrm{Tr}_\phi^k f). \tag{2.18}$$

2.4. SDEs with FBM. As the dynamics of the control problems under consideration in this paper is described by SDEs with FBM, we need first to study linear fractional SDEs in this section.

We introduce the two linear equations

$$\begin{cases} dx_t = A_t x_t dt + C_t x_t dW_t^H, \\ x_0 \in \mathbb{R}^n \qquad \text{given and deterministic,} \end{cases} \tag{2.19}$$

and

$$\begin{cases} dx_t = A_t x_t dt + C_t x_t \circ dW_t^H, \\ x_0 \in \mathbb{R}^n \qquad \text{given and deterministic,} \end{cases} \tag{2.20}$$

where dW_t^H and $\circ dW_t^H$ denote, respectively, the Itô-type and Stratonovich-type differentials in the sense of [9].

THEOREM 2.3. *If A_t and C_t are measurable and essentially bounded deterministic functions in t, then* (2.19) *admits a unique solution. Moreover, the solution satisfies*

$$\sup_{0\le t\le T} \mathbb{E}\,|x_t|^p < \infty \quad \forall\, p \ge 1. \tag{2.21}$$

Furthermore, the solution $x.$ has a continuous modification.

Proof. The results in which $H = \frac{1}{2}$ are clearly true by virtue of the classical SDE theory for normal Brownian motion. So we assume $H > \frac{1}{2}$. Let $\Psi(t,s)$ denote the fundamental solution associated with the deterministic part of (2.19), namely, $\Psi(t,s)$ satisfies

$$\frac{d}{dt}\Psi(t,s) = A_t\Psi(t,s)\,, \quad 0 \le s < t\,; \quad \Psi(s,s) = I,$$

where I denotes the identity matrix. Since A_t is uniformly bounded in t, it is routine to prove, via Gronwall's inequality, that $\Psi(t,s)$ is uniformly bounded in (t,s). This in turn implies, via the above equation, that $\Psi(t,s)$ is Lipschitz in t with a Lipschitz constant independent of (t,s).

Equation (2.19) is thus equivalent to

$$x_t = \Psi(t,0)x_0 + \int_0^t \Psi(t,s)C_s x_s dW_s^H. \tag{2.22}$$

Replacing x_s in the integral by this equation we have

$$\begin{aligned} x_t = \Psi(t,0)x_0 &+ \int_0^t \Psi(t,s)C_s\Psi(s,0)x_0 dW_s^H \\ &+ \int_0^t \Psi(t,s_2)C_{s_2}\int_0^{s_2}\Psi(s_2,s_1)C_{s_1}x_{s_1}dW_{s_1}^H dW_{s_2}^H. \end{aligned}$$

Repeatedly applying this procedure, we obtain

$$\begin{aligned} x_t = \Psi(t,0)x_0 &+ \int_0^t \Psi(t,s)C_s\Psi(s,0)x_0 dW_s^H \\ &+ \sum_{n=2}^{\infty}\int_{0<s_1<\cdots<s_n<t}\Psi(t,s_n)C_{s_n}\Psi(s_n,s_{n-1})C_{s_{n-1}}\cdots \\ &\cdot\Psi(s_2,s_1)C_{s_1}\Psi(s_1,0)x_0 dW_{s_1}^H\cdots dW_{s_n}^H. \end{aligned} \tag{2.23}$$

Now we show that (2.23) is convergent in $L^2(\Omega,\mathcal{F},P)$ for any fixed t. Denote

$$f_{n,t}(s_1,\ldots,s_n) := \Psi(t,s_n)C_{s_n}\Psi(s_n,s_{n-1})C_{s_{n-1}}\cdots C_{s_2}\Psi(s_2,s_1)C_{s_1}\Psi(s_1,0)x_0, \tag{2.24}$$

where $0<s_1<\cdots<s_n<t$. We also use the notation $f_n^t \equiv f_n^t(\cdot,\ldots,\cdot)$ to denote the symmetric extension of $f_{n,t}$ to the domain $[0,t]^n$, namely,

$$f_n^t(s_1,\ldots,s_n) := \Psi(t,s_{i_n})C_{s_{i_n}}\Psi(s_{i_n},s_{i_{n-1}})C_{s_{i_{n-1}}}\cdots C_{s_{i_2}}\Psi(s_{i_2},s_{i_1})C_{s_{i_1}}\Psi(s_{i_1},0)x_0$$

if $s_{i_1}\le s_{i_2}\le\cdots\le s_{i_n}$, where $i_1,\ldots,i_n$ is a permutation of $1,2,\ldots,n$.

By virtue of the underlying assumption it is easy to see that there is a positive constant K_t, depending on t, such that

$$\sup_{0\le s_1,\ldots,s_n\le t}|f_n^t(s_1,\ldots,s_n)|\le K_t^n.$$

The right-hand side of (2.23) can be rewritten as

$$Z_t := \sum_{n=0}^{\infty}\frac{1}{n!}I_{n,t}(f_n^t),$$

where

$$I_{n,t}(f_n^t) := \int_{0<s_1,\ldots,s_n<t} f_n^t(s_1,\ldots,s_n)dW_{s_1}^H\cdots dW_{s_n}^H$$

is the multiple Itô integral defined in section 2.3. It follows from (2.15) that

$$\begin{aligned}\mathbb{E}\,|Z_t|^2 &= \sum_{n=0}^{\infty} \frac{1}{(n!)^2}\mathbb{E}\,\left[I_{n,t}(f_n^t)\right]^2 \\ &= \sum_{n=0}^{\infty} \frac{1}{n!}\langle f_n^t, f_n^t\rangle_\phi.\end{aligned}$$

By its definition (see section 2.3), we have

$$\begin{aligned}\langle f_n^t, f_n^t\rangle_\phi &= \int_{[0,t]^{2n}} \phi(s_1,t_1)\cdots\phi(s_n,t_n) f_n^t(s_1,\ldots,s_n) f_n^t(t_1,\ldots,t_n) ds_1\cdots ds_n dt_1\cdots dt_n \\ &\le K_t^{2n}\int_{[0,t]^{2n}} \phi(s_1,t_1)\cdots\phi(s_n,t_n) ds_1\cdots ds_n dt_1\cdots dt_n \\ &\le K_t^{2n},\end{aligned}$$

where K_t is a generic constant (depending on t) whose values may vary. This leads to

$$\mathbb{E}\,|Z_t|^2 \le \sum_{n=0}^{\infty}\frac{1}{n!}K_t^{2n} < \infty.$$

Hence, the right-hand side of (2.23) converges in $L^2(\Omega,\mathcal{F},P)$, and x_t is well defined.

Next we show that $\int_0^t C_s x_s dW_s^H$ is well defined for each $t\in[0,T]$, where $x.$ is defined by (2.23). Let us prove this for $t=T$, the other case being similar. From [9, p. 591] there are several conditions for defining $\int_0^T C_t x_t dW_t^H$. Let us check

$$\mathbb{E}\int_0^T |D_s^H(C_s x_s)|^2 ds < \infty.$$

Other conditions are routine and easier to check. Since C_t is a bounded deterministic function of t, it suffices to verify

$$\mathbb{E}\int_0^T |D_s^H x_s|^2 dt < \infty. \tag{2.25}$$

To this end, recall that f_n^t is the symmetric extension of $f_n(t;\cdot,\ldots,\cdot)$ defined by (2.20). With this notation we have

$$x_t = \sum_{n=0}^{\infty}\frac{1}{n!}I_{n,t}(f_n^t).$$

Its Malliavin derivative is given by, appealing to Theorem 2.2,

$$D_s^H x_t = \sum_{n=1}^{\infty}\frac{1}{(n-1)!}I_{n-1,t}\left(\int_0^t \phi(s,r) f_{n-1}^t(r)dr\right).$$

Therefore we have

$$\mathbb{E}\,|D_s^H x_t|^2 = \sum_{n=1}^{\infty}\frac{1}{(n-1)!}\left\langle \int_0^t \phi(s,r) f_{n-1}^t(r)dr\,,\int_0^t \phi(s,r) f_{n-1}^t(r)dr\right\rangle_\phi.$$

Since f_n^t is bounded on $[0,T]$, i.e.,

$$|f_n^t| \le K_T^n,$$

we have

$$\begin{aligned}
&\left|\left\langle \int_0^t \phi(s,r) f_{n-1}^t(r)dr\,, \int_0^t \phi(s,r) f_{n-1}^t(r)dr \right\rangle_\phi\right| \\
&\le \left| \int_0^t \int_0^t \int_{\substack{0<s_1,\dots,s_{n-1}<t \\ 0<t_1,\dots,t_{n-1}<t}} f_n^t(s_1,\dots,s_{n-1},r_1) f_n^t(t_1,\dots,t_{n-1},r_2) \right. \\
&\qquad \left. \cdot \phi(s_1,t_1)\cdots\phi(s_{n-1},t_{n-1}) ds_1 dt_1 \cdots ds_{n-1}dt_{n-1}dr_1 dr_2 \right| \\
&\le K_T^n,
\end{aligned}$$

where here and in what follows we use K_T to denote a generic constant, which depends only on T and may be different in different places. Consequently, we have

$$\sup_{0\le s,t\le T} \mathbb{E}\,|D_s^H x_t|^2 \le K_T. \tag{2.26}$$

This easily implies (2.25).

Next, to obtain estimate (2.21) we use a hypercontractivity inequality. Define the number operator $\Gamma(\alpha)$, where $\alpha \in \mathbb{R}$, as

$$\Gamma(\alpha)F := \sum_{n=0}^\infty \alpha^n I_{n,t}(g_n) \qquad \text{if} \quad F = \sum_{n=0}^\infty I_{n,t}(g_n),$$

where $g_n \in L_\phi^2([0,t]^n)$ with $\sum_{n=0}^\infty I_{n,t}(g_n)$ convergent in $L^2(\Omega,\mathcal{F},P)$. Nelson's hypercontractivity theorem (see [8], [14]) yields

$$\left(\mathbb{E}\,\left|\Gamma\left(\frac{1}{p-1}\right)F\right|^p\right)^{1/p} \le \left(\mathbb{E}\,|F|^2\right)^{1/2} \quad \forall p \ge 2.$$

Now for f_n^t defined earlier, we set

$$Z_{t,p} := \sum_{n=0}^\infty \frac{(p-1)^n}{n!} I_{n,t}(f_n^t).$$

By the same estimate for Z_t, we obtain

$$\mathbb{E}\,|Z_{t,p}|^2 < \infty.$$

However, noting that $\Gamma(\frac{1}{p-1})Z_{t,p} = Z_t$ and applying the hypercontractivity inequality to $F = Z_{t,p}$, we have

$$(\mathbb{E}\,|x_t|^p)^{1/p} \equiv (\mathbb{E}\,|Z_t|^p)^{1/p} \le \left(\mathbb{E}\,|Z_{t,p}|^2\right)^{1/2} < \infty \ \ \forall p \ge 2.$$

When $1 \le p < 2$, inequality (2.21) follows from the Cauchy–Schwarz inequality.

Finally, to see that $x_\cdot$ has a continuous version, we estimate $\mathbb{E}\,|x_t - x_s|^2$. We shall need the uniform Lipschitz condition of $\Psi(t,s)$ in t which was proved earlier. From expression (2.23) it follows that

$$x_t - x_s = \Psi(t,0)x_0 - \Psi(s,0)x_0 + \sum_{n=1}^{\infty} \frac{1}{n!}\left[I_{n,t}(f_n^t) - I_{n,s}(f_n^s)\right].$$

By the orthogonality of different chaos, we have

$$\begin{aligned}
\mathbb{E}\,|x_t - x_s|^2 &= \left[\Psi(t,0)x_0 - \Psi(s,0)x_0\right]^2 + \sum_{n=1}^{\infty} \frac{1}{(n!)^2}\mathbb{E}\,\left[I_{n,t}(f_n^t) - I_{n,s}(f_n^s)\right]^2 \\
&\le \left[\Psi(t,0)x_0 - \Psi(s,0)x_0\right]^2 + \sum_{n=1}^{\infty} \frac{1}{(n!)^2}\mathbb{E}\,\left[I_{n,s}(f_n^t) - I_{n,s}(f_n^s)\right]^2 \\
&+ \sum_{n=1}^{\infty} \frac{1}{(n!)^2}\mathbb{E}\,\left[I_{n,t}(f_n^t) - I_{n,s}(f_n^t)\right]^2 \\
&=: I_1 + \sum_{n=1}^{\infty} \frac{1}{(n!)^2} I_{2,n} + \sum_{n=1}^{\infty} \frac{1}{(n!)^2} I_{3,n},
\end{aligned}$$

where

$$\begin{aligned}
I_{2,n} &= \mathbb{E}\,\left[I_{n,s}(f_n^t) - I_{n,s}(f_n^s)\right]^2 \\
&= \mathbb{E}\,\left[I_{n,s}(f_n^t - f_n^s)\right]^2 \\
&= \langle f_n^t - f_n^s\,, f_n^t - f_n^s\rangle_\phi.
\end{aligned}$$

Now it is straightforward to see that

$$I_{2,n} \le n! K_T^n |t-s|^2.$$

Denote by $T_n(s,t)$ the region

$$\begin{aligned}
T_n(s,t) &:= \{0 < s_1 < \cdots < s_n < t\} \setminus \{0 < s_1 < \cdots < s_n < s\} \\
&= \cup_{k=0}^{n-1} \{0 < s_1 < \cdots < s_k < s, s_{k+1} < \cdots < s_n < t\} \\
&=: \cup_{k=0}^{n-1} T_{n,k}(s,t).
\end{aligned}$$

Then

$$\begin{aligned}
&\frac{1}{(n!)^2} I_{3,n} = \frac{1}{(n!)^2}\mathbb{E}\,\left[I_{n,t}(f_n^t) - I_{n,s}(f_n^t)\right]^2 \\
&= \mathbb{E}\,\left[\int_{T_n(s,t)} f_n(t,s_1,\ldots,s_n) dW_{s_1}^H \cdots dW_{s_n}^H\right]^2 \\
&= \int_{T_n(s,t)}\int_{T_n(s,t)} f_n(t,s_1,\ldots,s_n) f_n(t,t_1,\ldots,t_n)\phi(s_1,t_1)\cdots\phi(s_n,t_n) ds_1 dt_1 \cdots ds_n dt_n \\
&\le \sum_{k,j=0}^{n-1} \int_{T_{n,k}(s,t)}\int_{T_{n,j}(s,t)} f_n(t,s_1,\ldots,s_n) f_n(t,t_1,\ldots,t_n)\phi(s_1,t_1)\cdots\phi(s_n,t_n) ds_1 dt_1 \cdots ds_n dt_n \\
&\le \sum_{k,j=0}^{n-1} \sqrt{\int_{T_{n,k}(s,t)}\int_{T_{n,k}(s,t)} f_n(t,s_1,\ldots,s_n) f_n(t,t_1,\ldots,t_n)\phi(s_1,t_1)\cdots\phi(s_n,t_n) ds_1 dt_1 \cdots ds_n dt_n} \\
&\cdot \sqrt{\int_{T_{n,j}(s,t)}\int_{T_{n,j}(s,t)} f_n(t,s_1,\ldots,s_n) f_n(t,t_1,\ldots,t_n)\phi(s_1,t_1)\cdots\phi(s_n,t_n) ds_1 dt_1 \cdots ds_n dt_n}.
\end{aligned}$$

It is easy to check that for all $0 \le k \le n-1$, we have

$$\begin{aligned}
&\int_{T_{n,k}(s,t)} \int_{T_{n,k}(s,t)} f_n(t,s_1,\dots,s_n) f_n(t,t_1,\dots,t_n)\phi(s_1,t_1)\cdots\phi(s_n,t_n)ds_1dt_1\cdots ds_ndt_n \\
&\quad \le \int_{\substack{0<s_1<\cdots<s_k<T \\ 0<t_1<\cdots<t_k<T}} \int_{\substack{s<s_{k+1}<\cdots<s_n<t \\ s<t_{k+1}<\cdots<t_n<t}} f_n(t,s_1,\dots,s_n) f_n(t,t_1,\dots,t_n)\phi(s_1,t_1)\cdots \\
&\qquad \cdot\phi(s_n,t_n)ds_1dt_1\cdots ds_ndt_n \\
&\quad \le \frac{1}{k!(n-k)!} K_T^n |t-s|^{2H} \\
&\quad \le |t-s|^{2H} K_T^n/n!.
\end{aligned}$$

This concludes that

$$\frac{1}{(n!)^2} I_{3,n} \le \frac{1}{n!} K_T^n |t-s|^{2H}.$$

Observing that

$$I_1 \le K_T |t-s|^2,$$

we obtain

$$\mathbb{E}\ |x_t - x_s|^2 \le K_T |t-s|^{2H}.$$

Since $2H > 1$ we see that x_t has a continuous version by the Kolmogorov lemma. This completes the proof of the theorem. □

Remark 2.3. There has been some study on fractional SDEs (see, for example, [27], [24], and the references therein). However, we could not find any result on the type of equation (2.19) in the literature. When the equation is one-dimensional, explicit representation of the solution was obtained in [3], [19]; see also Theorem 2.5 below. Here we have used the Wiener chaos expansion approach to handle the multidimensional case. Note that the L^p estimate of solution (2.21) is also new.

THEOREM 2.4. *If A_t and C_t are measurable and essentially bounded deterministic functions in t, then* (2.20) *admits a unique solution.*

Proof. Similar to the argument proving Theorem 2.3, a solution candidate to (2.20) is

(2.27)

$$\begin{aligned}
Y_t := \sum_{n=0}^{\infty} \int_{0<s_1<\cdots<s_n<t} &\Psi(t,s_n)C_{s_n}\Psi(s_n,s_{n-1})C_{s_{n-1}}\cdots\Psi(s_2,s_1)C_{s_1}\Psi(s_1,0)x_0 \circ dW_{s_1}^H \\
&\cdots \circ dW_{s_n}^H.
\end{aligned}$$

Fix t. Since

$$\begin{aligned}
S_n := \int_{0<s_1<\cdots<s_n<t} &\Psi(t,s_n)C_{s_n}\Psi(s_n,s_{n-1})C_{s_{n-1}}\cdots\Psi(s_2,s_1)C_{s_1}\Psi(s_1,0)x_0 \circ dW_{s_1}^H \\
&\cdots \circ dW_{s_n}^H,
\end{aligned}$$

$n = 0, 1, 2, \dots$, are not orthogonal, we consider the convergence in $L_1(\Omega, \mathcal{F}, P)$. We have

$$\mathbb{E}\ |Y_t| \le \sum_{n=0}^{\infty} \mathbb{E}\ |S_n| \le \sum_{n=0}^{\infty} \left(\mathbb{E}\ |S_n|^2\right)^{1/2}.$$

Defining $f_{n,t}$ as in (2.24), we can write

$$S_n = \int_{0<s_1<\cdots<s_n<t} f_{n,t}(s_1,\dots,s_n) \circ dW_{s_1}^H \cdots \circ dW_{s_n}^H.$$

As before, let f_n^t be the symmetric extension of $f_{n,t}$ to $[0,t]^n$. Since A and C are bounded, we have $|f_{n,t}(s_1,\dots,s_n)| \le K_T^n$ for some constant K_T depending only on T.

With this notation we have

$$S_n = \frac{1}{n!} S_{n,t}(f_n^t),$$

where

$$S_{n,t}(f_n^t) := \int_{0<s_1,\dots,s_n<t} f_n^t(s_1,\dots,s_n) \circ dW_{s_1}^H \cdots \circ dW_{s_n}^H.$$

By the Hu–Meyer formula (see (2.18)) we have

$$S_{n,t}(f_n^t) = \sum_{k\le n/2} \frac{n!}{2^k k!(n-2k)!} I_{n-2k,t}\left(\mathrm{Tr}_\phi^k f_n^t\right).$$

Therefore

$$\mathbb{E}\,|S_{n,t}(f_n^t)|^2 = \sum_{k\le n/2} \frac{(n!)^2}{2^{2k}(k!)^2(n-2k)!} |\mathrm{Tr}_\phi^k f_n^t|_\phi^2. \tag{2.28}$$

But

$$\begin{aligned}
|\,\mathrm{Tr}_\phi^k f_n^t|_\phi^2 &= \int_{\substack{0<s_1,\dots,s_{2k}<t\\ 0<t_1,\dots,t_{2k}<t}} \int_{\substack{0<t_1,\dots,t_{n-2k}<t\\ 0<t_1',\dots,t_{n-2k}'<t}} f_n^t(s_1,s_2,\dots,s_{2k},t_1,\dots,t_{n-2k}) \\
&\qquad \cdot f_n^t(r_1,r_2,\dots,r_{2k},t_1',\dots,t_{n-2k}')\phi(s_1,s_2)\cdots\phi(s_{2k-1},s_{2k}) \\
&\qquad \cdot \phi(r_1,r_2)\cdots\phi(r_{2k-1},r_{2k})\phi(t_1,t_1')\cdots \\
&\qquad \cdot \phi(t_{n-2k},t_{n-2k}')ds_1dr_1\cdots ds_{2k}dr_{2k}dt_1dt_1'\cdots dt_{n-2k}dt_{n-2k}' \\
&\le K_T^n \int_{\substack{0<s_1,\dots,s_{2k}<t\\ 0<t_1,\dots,t_{2k}<t}} \int_{\substack{0<t_1,\dots,t_{n-2k}<t\\ 0<t_1',\dots,t_{n-2k}'<t}} \phi(s_1,s_2)\cdots\phi(s_{2k-1},s_{2k}) \\
&\qquad \cdot \phi(r_1,r_2)\cdots\phi(r_{2k-1},r_{2k})\phi(t_1,t_1')\cdots \\
&\qquad \cdot \phi(t_{n-2k},t_{n-2k}')ds_1dr_1\cdots ds_{2k}dr_{2k}dt_1dt_1'\cdots dt_{n-2k}dt_{n-2k}' \\
&\le K_T^n.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}\,|S_n|^2 &\le \sum_{k\le n/2} \frac{K_T^n}{(k!)^2(n-2k)!} \\
&\le \frac{K_T^n}{n!}.
\end{aligned}$$

Thus Y_t is well defined. It is now routine to verify that Y_t is the solution to (2.20). □

Remark 2.4. Unlike with the case of Itô integral (see Theorem 2.3), it remains an open problem whether the solution in Theorem 2.4 has a continuous modification.

Remark 2.5. The nonhomogeneous versions of (2.19) and (2.20) are

$$\begin{cases} dx_t = (A_t x_t + f_t)dt + (C_t x_t + g_t)dW_t^H, \\ x_0 \in \mathbb{R}^n \qquad \text{given and deterministic} \end{cases} \tag{2.29}$$

and

$$\begin{cases} dx_t = (A_t x_t + f_t)dt + (C_t x_t + g_t) \circ dW_t^H, \\ x_0 \in \mathbb{R}^n \qquad \text{given and deterministic,} \end{cases} \tag{2.30}$$

respectively, where f_t and g_t are generally random processes. One may also use the Wiener chaos expansion technique to study the solvability of these equations, taking the "fundamental matrix" associated with the homogeneous versions (2.19) and (2.20) as the starting point. However, much more involved technicalities will have to be gone through. Details are left to the interested readers. Note that for studying the first and second control models in what follows (2.19) and (2.20) suffice since only the Markovian type of controls is considered.

The following theorem gives an explicit expression of the solution to (2.19) in the one-dimensional case. Note that the result was proved in [3] using the Wick product. Here we give a different yet simple proof.

THEOREM 2.5. *Let $n = 1$. If A_t and C_t are measurable and essentially bounded deterministic functions in t, then the solution of* (2.19) *can be represented as*

$$x_t = x_0 \exp\left[\int_0^t C_s dW_s^H + \int_0^t A_s ds - \frac{1}{2}\int_0^t \int_0^t \phi(s,s')C_s C_{s'} ds ds'\right]. \tag{2.31}$$

Proof. Define

$$\begin{aligned} y_t &:= \exp\left(\int_0^t C_s dW_s^H\right), \\ z_t &:= x_0 \exp\left\{\int_0^t A_s ds - \frac{1}{2}\int_0^t \int_0^t \phi(s,s')C_s C_{s'} ds ds'\right\} \\ &\equiv x_0 \exp\left\{\int_0^t A_s ds - \int_0^t \int_0^s \phi(s,s')C_s C_{s'} ds' ds\right\}, \\ \bar{x}_t &:= y_t z_t. \end{aligned}$$

It suffices to prove that $\bar{x}_t$ satisfies (2.19). To this end, noting that C_t is deterministic, we can apply a simplified Itô's formula ([9, Corollary 4.4][2]) to get

$$dy_t = C_t y_t dW_t^H + C_t y_t \int_0^t \phi(t,s)C_s ds\, dt. \tag{2.32}$$

[2]There is a typo in [9, Corollary 4.4]. The last expression on page 599 of [9] should be $\int_0^t \frac{\partial^2 f}{\partial x^2}(s,\eta_s)a_s \int_0^s \phi(s,v)a_v dv ds$.

Hence,

$$\begin{aligned} d\bar{x}_t = d(y_t z_t) &= y_t z_t \left[A_t dt - \int_0^t \phi(t,s') C_t C_{s'} ds' \, dt \right] \\ &\quad + y_t z_t \left[C_t dW_t^H + C_t \int_0^t \phi(t,s) C_s ds \, dt \right] \\ &= A_t \bar{x}_t dt + C_t \bar{x}_t dW_t^H. \end{aligned}$$

This completes the proof. □

The following two lemmas are useful in what follows. In particular, the first one represents the Malliavin derivative of the solution (2.31).

LEMMA 2.2. *Let x_t be the solution of* (2.19) *under the assumptions of Theorem* 2.5. *Then*

$$D_t^H x_t = x_t \int_0^t \phi(t,s) C_s ds \quad \forall t \in [0,T] \quad a.s. \tag{2.33}$$

Proof. Equation (2.19) is understood as

$$x_t = x_0 + \int_0^t A_s x_s ds + \int_0^t C_s x_s dW_s^H.$$

From the explicit expression (2.31) of the solution we could compute its Malliavin derivative. However, here we supply a simpler method. By [9, Theorem 4.2], we have

$$D_r^H x_t = \int_0^t A_s D_r^H x_s ds + \int_0^t C_s D_r^H x_s dW_s^H + \int_0^t \phi(r,s) C_s x_s ds \quad \forall r,t \in [0,T] \text{ a.s.}$$

Fix r. Denote $z_t := D_r^H x_t$. Then the above equation can be written as

$$dz_t = A_t z_t dt + C_t z_t dW_t^H + \phi(r,t) C_t x_t dt \tag{2.34}$$

with $z_0 = 0$. We want to solve this equation to find $D_r^H x_t$. Try the solution of the form $\tilde{z}_t := \rho_t x_t$, where ρ_t is deterministic and differentiable with $\rho_0 = 0$. Then

$$\begin{aligned} d\tilde{z}_t &= \dot{\rho}_t x_t dt + \rho_t A_t x_t dt + \rho_t C_t x_t dW_t^H \\ &= \dot{\rho}_t x_t dt + A_t \tilde{z}_t dt + C_t \tilde{z}_t dW_t^H. \end{aligned} \tag{2.35}$$

Comparing (2.35) with (2.34), we conclude that if $\dot{\rho}_t = \phi(r,t) C_t$, then z_t would be equal to $\rho_t x_t$ owing to the uniqueness of the solution to (2.34). This implies that

$$\rho_t = \int_0^t \phi(r,s) C_s ds.$$

The proof is completed. □

LEMMA 2.3. *Let x_t be the solution of* (2.19) *under the assumptions of Theorem* 2.5, *and let p_t and C_t be continuously differentiable deterministic functions in t. Then*

$$d(p_t x_t^2) = x_t^2 \left[\dot{p}_t dt + 2A_t p_t dt + 2p_t C_t \int_0^t \phi(t,s) C_s ds dt + 2C_t p_t dW_t^H \right]. \tag{2.36}$$

Proof. Again it suffices to consider the case when $H > \frac{1}{2}$. We apply Itô's formula [9, Theorem 4.5] to $p_t x_t^2$. To do this we need to check the following conditions:

(1) $\mathbb{E}\ \sup_{0\le s\le T}|A_t x_t|<\infty$;
(2) $\mathbb{E}\ \int_0^T |C_t x_t D_t^H x_t|^2 dt<\infty$;
(3) there is $\alpha>1-H$ such that $\mathbb{E}\ |C_t x_t - C_s x_s|^2 \le K|t-s|^{2\alpha}$, where $|t-s|\le\delta$ for some $\delta>0$ and $\lim_{|t-s|\to 0}\mathbb{E}\ |D_t^H(C_t x_t - C_s x_s)|^2=0$;
(4) the function $f(t,x):=p_t x^2$ has bounded derivatives.

Condition (4) can be replaced by the following (see [13]):

(4′) the function $f(t,x)$ has polynomial growth in x.

To prove (1), we use (2.31) to conclude that

$$\sup_{0\le t\le T}|A_t x_t| \le K_T \exp\left\{\sup_{0\le t\le T}\left|\int_0^t C_s dW_s^H\right|\right\}.$$

Condition (1) thus follows easily [10]. Next, combining (2.33) and (2.21) we derive condition (2).

To prove (3), since C_t is continuously differentiable, it suffices to show that

$$\mathbb{E}\ |x_t - x_s|^2 \le K|t-s|^{2\alpha}$$

and

$$\lim_{|t-s|\to 0}\mathbb{E}\ |D_t^H(x_t - x_s)|^2 = 0.$$

The first inequality (with $\alpha = H > 1-H$) was proved in the proof of Theorem 2.3, and the second equality is seen immediately by the fact that $D_r^H x_t = \int_0^t \phi(r,s)C_s ds$, which was proved in the proof of Lemma 2.2.

Finally, condition (4′) is obvious. Now applying [9, Theorem 4.5] we conclude that

$$\begin{aligned} d(p_t x_t^2) &= \dot p_t x_t^2 dt + 2p_t x_t dx_t + 2p_t C_t x_t D_t^H x_t dt \\ &= \dot p_t x_t^2 dt + 2A_t p_t x_t^2 dt + 2C_t p_t x_t^2 dW_t^H + 2p_t C_t x_t^2 \int_0^t \phi(t,s)C_s ds dt \\ &= x_t^2\left[\dot p_t dt + 2A_t p_t dt + 2p_t C_t \int_0^t \phi(t,s)C_s ds dt + 2C_t p_t dW_t^H\right], \end{aligned} \tag{2.37}$$

where the second equality is due to Lemma 2.2. This proves the lemma. □

3. Control model 1: Scalar state, quadratic cost, and Markovian control. We start with our first control model in this section. The controlled dynamics is given by the following Itô-type SDE, where the state is scalar-valued:

$$\begin{cases} dx_t = (A_t x_t + B_t u_t)dt + (C_t x_t + D_t u_t)dW_t^H, \\ x_0 \in \mathbb{R} \qquad \text{given and deterministic,} \end{cases} \tag{3.1}$$

where A_t, $B_t := (B_t^1,\dots,B_t^m)$, C_t, and $D_t := (D_t^1,\dots,D_t^m)$, $0\le t\le T$, are given essentially bounded deterministic functions of t. A control $u_t = (u_t^1,\dots,u_t^m)^*$, $0\le t\le T$, where * denotes the matrix transpose, is taken to be of the Markovian linear feedback type, namely,

$$u_t = K_t x_t, \tag{3.2}$$

where $K_t := (K_t^1, \dots, K_t^m)^*$ is an essentially bounded deterministic function of t. Such a control is also referred to as an *admissible (Markovian linear feedback) control* in this section.

Under each admissible control $u_t = K_t x_t$, system (3.1) reduces to the following linear SDE:

$$\begin{cases} dx_t = (A_t + B_t K_t)x_t dt + (C_t + D_t K_t)x_t dW_t^H, \\ \\ x_0 \in \mathbb{R} \qquad \text{given and deterministic.} \end{cases} \tag{3.3}$$

Hence $K_\cdot$ itself, also termed the *feedback gain*, can be regarded as a control.

For every initial state x_0 and admissible control $u_t = K_t x_t$, there is an associated cost

$$J(x_0, u_\cdot) \equiv J(x_0, K_\cdot) := \mathbb{E}\left[\int_0^T (Q_t x_t^2 + u_t^* R_t u_t)dt + G x_T^2\right], \tag{3.4}$$

where $x_\cdot$ is the solution of (3.3) under the control $u_\cdot$ or, equivalently, $K_\cdot$ (note that the unique solvability of (3.3) under a given $K_\cdot$ is ensured by Theorem 2.3), Q_t and R_t are given essentially bounded deterministic functions in t, and G is a given deterministic scalar. Our optimal stochastic control problem is to minimize the cost functional (3.4), for each given x_0, over the set of all admissible Markovian linear feedback controls.

THEOREM 3.1. *Assume that for a.e.* $t \in [0,T]$, $D_t = 0$, $Q_t \geq 0$, *and* $R_t \succ \delta I$ *for some given* $\delta > 0$ *and* $G \geq 0$. *Then the Riccati equation*

$$\begin{cases} \dot{p}_t + 2p_t[A_t + C_t \int_0^t \phi(t,s)C_s ds] + Q_t - B_t R_t^{-1} B_t^* p_t^2 = 0, \\ \\ p_T = G \end{cases} \tag{3.5}$$

admits a unique solution $p_\cdot$ *over* $[0,T]$ *with* $p_t \geq 0$ *for all* $t \in [0,T]$. *Moreover, the optimal Markovian linear feedback control for problem* (3.3)–(3.4) *is given by*

$$\hat{u}_t = \hat{K}_t x_t \quad \textit{with} \quad \hat{K}_t = -R_t^{-1} B_t^* p_t. \tag{3.6}$$

Finally, the optimal value of (3.4) *is* $p_0 x_0^2$.

Proof. The unique solvability of the (classical) Riccati equation (3.5) was proved in, e.g., [29, Corollary 2.10 p. 297]. Next, for any admissible control $u_t = K_t x_t$, applying Lemma 2.3 to (3.3) with $D_t = 0$, we get

$$d(p_t x_t^2) = x_t^2 \left[\dot{p}_t + 2p_t(A_t + B_t K_t) + 2p_t C_t \int_0^t \phi(t,s)C_s ds\right] dt + 2x_t^2 C_t p_t dW_t^H.$$

Taking integration from 0 to T, we get

$$\begin{aligned} p_T x_T^2 = p_0 x_0^2 &+ \int_0^T x_t^2 \left[\dot{p}_t + 2p_t(A_t + B_t K_t) + 2p_t C_t \int_0^t \phi(t,s)C_s ds\right] \cdot dt \\ &+ 2\int_0^T x_t^2 C_t p_t dW_t^H. \end{aligned}$$

Denote $f_t := x_t^2 C_t p_t$. It follows from (2.21) that

$$\int_0^T \int_0^T \phi(s,t)\mathbb{E}\,|f_s f_t| ds dt < \infty.$$

On the other hand, $D_t^H x_t = x_t \int_0^t \phi(t,s) C_s ds$ by virtue of Lemma 2.2. So again by (2.21) we have

$$\sup_{0\le t\le T} \mathbb{E}\,|D_t^H x_t|^p < \infty \quad \forall\, p \ge 1.$$

This implies that f_t satisfies the condition of [9, Theorem 3.9] leading to $\mathbb{E}\int_0^T f_t dW_t^H = 0$. Hence

$$\mathbb{E}\,\left[p_T x_T^2\right] = p_0 x_0^2 + \mathbb{E}\,\int_0^T x_t^2 \Big[\dot{p}_t + 2p_t(A_t + B_t K_t) + 2p_t C_t \int_0^t \phi(t,s) C_s ds\Big] dt.$$

Since $p_T = G$, we obtain

$$\begin{aligned} J(x_0, K.) &= p_0 x_0^2 + \mathbb{E}\,\int_0^T x_t^2 \Bigg[\dot{p}_t + 2p_t(A_t + B_t K_t) + (Q_t + K_t^* R_t K_t) \\ &\qquad + 2p_t C_t \int_0^t \phi(t,s) C_s ds\Bigg] dt \\ &= p_0 x_0^2 + \mathbb{E}\,\int_0^T x_t^2 \Bigg[\dot{p}_t + 2p_t A_t + 2p_t C_t \int_0^t \phi(t,s) C_s ds + Q_t \\ &\qquad + (K_t + R_t^{-1} B_t^* p_t)^* R_t (K_t + R_t^{-1} B_t^* p_t) - B_t R_t^{-1} B_t^* p_t^2\Bigg] dt \qquad (3.7) \\ &= p_0 x_0^2 + \mathbb{E}\,\int_0^T (K_t + R_t^{-1} B_t^* p_t)^* R_t (K_t + R_t^{-1} B_t^* p_t) dt, \qquad (3.8) \end{aligned}$$

where the last equality was due to the Riccati equation (3.5). Equation (3.8) shows that the cost function achieves its minimum when $\hat{K}_t = -R_t^{-1} B_t^* p_t$, with the minimum value being $p_0 x_0^2$. This proves the theorem. □

Remark 3.1. It is interesting to note that the Riccati equation (3.5) corresponds to the following LQ control problem with (normal) Brownian motion:

$$\text{Minimize} \quad (3.4)$$

$$\text{subject to} \quad \begin{cases} dx_t = (A_t x_t + B_t u_t) dt + \tilde{C}_t x_t dW_t, \\ x_0 \in \mathbb{R} \qquad \text{given and deterministic,} \end{cases} \qquad (3.9)$$

where $\tilde{C}_t = \sqrt{2C_t \int_0^t \phi(t,s) C_s ds}$, assuming that $C_t \int_0^t \phi(t,s) C_s ds \ge 0$ for all $t \ge 0$; see [4]. This suggests that, in the current setting, the LQ control problem with FBM is equivalent (in the sense of sharing the same optimal feedback control and optimal value) to an LQ control problem with Brownian motion, where the diffusion coefficient of the state is properly modified.

4. Control model 2: Scalar state, power cost, and Markovian control. In this section we consider a more general model (than the one tackled in the previous section) where the dynamics is given by (3.1) but the cost functional is of the form

(4.1)

$$J(x_0, u.) \equiv J(x_0, u(\cdot)) = \mathbb{E}\,\left\{\int_0^T \left[Q_t x_t^\alpha + \sum_{i,j=1}^m R_{ij}(t) u_i(t)^{\beta_i} u_j(t)^{\beta_j}\right] dt + G x_T^\gamma\right\},$$

where α, β_i $(i = 1, 2, \dots, m)$, $\gamma \in \mathbb{R}$ are given, Q_t and $R_t \equiv R(t) = (R_{ij}(t))_{1\le i,j\le m}$ are given essentially bounded deterministic functions in t, and G is a given deterministic scalar. Moreover, we assume that $(R_{ij}(t))_{1\le i,j\le m}$ is positive definite for each $t \in [0, T]$. Here we interchangeably denote a control variable as

$$u_t \equiv u(t) = (u_t^1, \dots, u_t^m)^* \equiv (u_1(t), \dots, u_m(t))^*$$

for notational convenience. The problem is to find the *extreme* or *optimum* (i.e., minimum or maximum) of $J(x_0, u.)$, for each fixed $x_0 \in \mathbb{R}$, over the set of all admissible Markovian linear feedback controls of the form $u_t = K_t x_t$, subject to (3.1). With this type of control we can write $J(x_0, u.) = J(x_0, K.)$ as in the previous section. Note that whether the cost functional (4.1) assumes a minimum or maximum (or both) depends on the problem data, in particular, the ranges of α, β, and γ. The objective of this section is to derive *necessary conditions* for a solution to the extreme problem.

LEMMA 4.1. *For any admissible Markovian control $u_t = K_t x_t$, where $K_t \equiv K(t) = (K_1(t), \dots, K_m(t))^*$, the corresponding cost has the following representation:*

$$J(x_0, K.) = \int_0^T \left[Q_t \Theta_{\alpha,t} + \sum_{i,j=1}^m R_{ij}(t) K_i(t)^{\beta_i} K_j(t)^{\beta_j} \Theta_{\beta_i+\beta_j,t} \right] dt + G\Theta_{\gamma,T}, \tag{4.2}$$

where for any $\kappa \in \mathbb{R}$,

$$\begin{aligned} \Theta_{\kappa,t} &\equiv \Theta_{\kappa,t}(x_0, K.) \\ &:= x_0^\kappa \mathbb{E}\ \exp\left\{ \kappa \int_0^t (A_s + B_s K_s) ds \right. \\ &\qquad \left. + \frac{\kappa^2 - \kappa}{2} \int_0^t \int_0^t \phi(s, s')(C_s + D_s K_s)(C_{s'} + D_{s'} K_{s'}) ds ds' \right\}. \end{aligned} \tag{4.3}$$

Proof. Substituting the feedback control $u_t = K_t x_t$ into (3.1), we obtain

$$dx_t = (A_t + B_t K_t) x_t dt + (C_t + D_t K_t) x_t dW_t^H. \tag{4.4}$$

By Theorem 2.5, this equation can be solved explicitly as follows:

$$\begin{aligned} x_t = x_0 \exp\left\{ \int_0^t (C_s + D_s K_s) dW_s^H + \int_0^t (A_s + B_s K_s) ds \right. \\ \left. - \frac{1}{2} \int_0^t \int_0^t \phi(s, s')(C_s + D_s K_s)(C_{s'} + D_{s'} K_{s'}) ds ds' \right\}. \end{aligned} \tag{4.5}$$

Let $\kappa \in \mathbb{R}$. Since $\kappa \int_0^t (C_s + D_s K_s) dW_s^H$ is Gaussian, we have

$$\begin{aligned} &\mathbb{E}\ \exp\left\{ \kappa \int_0^t (C_s + D_s K_s) dW_s^H \right\} \\ &\quad = \exp\left\{ \frac{1}{2} \kappa^2 \mathbb{E} \left| \int_0^t (C_s + D_s K_s) dW_s^H \right|^2 \right\} \\ &\quad = \exp\left\{ \frac{1}{2} \kappa^2 \int_0^t \int_0^t \phi(s, s')(C_s + D_s K_s)(C_{s'} + D_{s'} K_{s'}) ds ds' \right\}, \end{aligned}$$

where the last equality is due to [9, Theorem 3.10]. It then follows from (4.5) that

$$\mathbb{E}\, x_t^\kappa = x_0^\kappa \mathbb{E}\, \exp\left\{\kappa \int_0^t (A_s + B_s K_s) ds \right. \tag{4.6}$$
$$\left. + \frac{\kappa^2 - \kappa}{2} \int_0^t \int_0^t \phi(s, s')(C_s + D_s K_s)(C_{s'} + D_{s'} K_{s'}) ds ds' \right\}$$
$$= \Theta_{\kappa,t}(x_0, K_\cdot) \equiv \Theta_{\kappa,t}.$$

The desired representation (4.2) thus follows immediately. □

THEOREM 4.1. *If $K_t \equiv K(t) = (K_1(t), \ldots, K_m(t))^*$ achieves an extreme for problems* (3.1) *and* (4.1)*, then it must satisfy*

$$\begin{aligned} &\alpha B_s \int_s^T Q_t \Theta_{\alpha,t} dt + (\alpha^2 - \alpha) \int_s^T \int_0^t Q_t \Theta_{\alpha,t} \phi(s', s) C_{s'} D_s ds' dt \\ &+ (\alpha^2 - \alpha) \int_s^T \int_0^t Q_t \Theta_{\alpha,t} \phi(s', s) D_{s'} K_{s'} D_s ds' dt + 2\tilde{R}(s) \\ &+ \sum_{i,j=1}^m \int_s^T R_{ij}(t) K_i(t)^{\beta_i} K_j(t)^{\beta_j} \Theta_{\beta_{ij},t} \left[\beta_{ij} B_s + (\beta_{ij}^2 - \beta_{ij}) \int_0^t \phi(s', s) C_{s'} D_s ds' \right. \\ &\qquad \left. + (\beta_{ij}^2 - \beta_{ij}) \int_0^t \phi(s', s) D_{s'} K_{s'} D_s ds' \right] dt \\ &+ \gamma G \Theta_{\gamma,T} B_s + (\gamma^2 - \gamma) G \Theta_{\gamma,T} \int_0^T \phi(s', s) C_{s'} D_s ds' \\ &+ (\gamma^2 - \gamma) G \Theta_{\gamma,T} \int_0^T \phi(s', s) D_{s'} K_{s'} D_s ds' = 0 \quad \text{a.e. } s \in [0, T], \end{aligned} \tag{4.7}$$

where

$$\beta_{ij} := \beta_i + \beta_j \tag{4.8}$$

and

$$\tilde{R}(t) := \left(\sum_{j=1}^m \beta_1 R_{1j}(t) K_1(t)^{\beta_1 - 1} K_j(t)^{\beta_j} \Theta_{\beta_{1j},t}, \right. \tag{4.9}$$
$$\left. \cdots, \sum_{j=1}^m \beta_m R_{mj}(t) K_m(t)^{\beta_m - 1} K_j(t)^{\beta_j} \Theta_{\beta_{mj},t} \right).$$

Proof. Since $K_\cdot$ attains the extreme value, it necessarily holds that

$$\left. \frac{d}{d\varepsilon} J(K_\cdot + \varepsilon \xi_\cdot) \right|_{\varepsilon = 0} = 0 \tag{4.10}$$

for all $\xi_\cdot \in C^\infty(0, T; \mathbb{R}^m)$, where $C^\infty(0, T; \mathbb{R}^m)$ denotes the set of all smooth ($\mathbb{R}^m$-valued) functions on $[0, T]$.

To compute derivative (4.10) we need to first compute $\frac{d}{d\varepsilon}\Theta_{\alpha,t}(x_0, K_\cdot + \varepsilon\xi_\cdot)|_{\varepsilon=0}$, which is given by

$$\begin{aligned}(4.11)\qquad &\frac{d}{d\varepsilon}\Theta_{\alpha,t}(x_0, K_\cdot + \varepsilon\xi_\cdot)\bigg|_{\varepsilon=0}\\ &\quad= \Theta_{\alpha,t}\bigg\{\alpha\int_0^t B_s\xi_s ds + (\alpha^2-\alpha)\int_0^t\int_0^t \phi(s,s')D_sK_sD_{s'}\xi_{s'}dsds'\\ &\qquad + (\alpha^2-\alpha)\int_0^t\int_0^t \phi(s,s')C_sD_{s'}\xi_{s'}dsds'\bigg\}.\end{aligned}$$

Consequently,

$$\begin{aligned}\frac{d}{d\varepsilon}J(\hat{K}_\cdot + \varepsilon\xi_\cdot)\bigg|_{\varepsilon=0} &= \int_0^T Q_t\Theta_{\alpha,t}\bigg\{\alpha\int_0^t B_s\xi_s ds + (\alpha^2-\alpha)\int_0^t\int_0^t \phi(s,s')C_sD_{s'}\xi_{s'}dsds'\\ &\quad + (\alpha^2-\alpha)\int_0^t\int_0^t \phi(s,s')D_sK_sD_{s'}\xi_{s'}dsds'\bigg\}dt\\ (4.12)\qquad &\quad + 2\sum_{i,j=1}^m\int_0^T \beta_i R_{ij}(t)K_i(t)^{\beta_i-1}K_j(t)^{\beta_j}\xi_i(t)\Theta_{\beta_{ij},t}dt\\ &\quad + \sum_{i,j=1}^m\int_0^T R_{ij}(t)K_i(t)^{\beta_i}K_j(t)^{\beta_j}\Theta_{\beta_{ij},t}\\ &\quad \cdot\bigg\{\beta_{ij}\int_0^t B_s\xi_s ds + (\beta_{ij}^2-\beta_{ij})\int_0^t\int_0^t \phi(s,s')C_sD_{s'}\xi_{s'}dsds'\\ &\quad + (\beta_{ij}^2-\beta_{ij})\int_0^t\int_0^t \phi(s,s')D_sK_sD_{s'}\xi_{s'}dsds'\bigg\}dt\\ &\quad + H\Theta_{\gamma,T}\bigg\{\gamma\int_0^T B_s\xi_s ds + (\gamma^2-\gamma)\int_0^T\int_0^T \phi(s,s')C_sD_{s'}\xi_{s'}dsds'\\ (4.13)\qquad &\quad + (\gamma^2-\gamma)\int_0^T\int_0^T \phi(s,s')D_sK_sD_{s'}\xi_{s'}dsds'\bigg\}\\ &= \int_0^T Q_t\Theta_{\alpha,t}\bigg\{\alpha\int_0^t B_s\xi_s ds + (\alpha^2-\alpha)\int_0^t\int_0^t \phi(s,s')C_sD_{s'}\xi_{s'}dsds'\\ &\quad + (\alpha^2-\alpha)\int_0^t\int_0^t \phi(s,s')D_sK_sD_{s'}\xi_{s'}dsds'\bigg\}dt\\ (4.14)\qquad &\quad + 2\int_0^T \tilde{R}(t)\xi_t dt\\ &\quad + \sum_{i,j=1}^m\int_0^T R_{ij}(t)K_i(t)^{\beta_i}K_j(t)^{\beta_j}\Theta_{\beta_{ij},t}\\ &\quad \cdot\bigg\{\beta_{ij}\int_0^t B_s\xi_s ds + (\beta_{ij}^2-\beta_{ij})\int_0^t\int_0^t \phi(s,s')C_sD_{s'}\xi_{s'}dsds'\end{aligned}$$

$$
\begin{aligned}
&+(\beta_{ij}^2-\beta_{ij})\int_0^t\int_0^t \phi(s,s')D_sK_sD_{s'}\xi_{s'}dsds'\Big\}dt\\
&+G\Theta_{\gamma,T}\Big\{\gamma\int_0^T B_s\xi_s ds+(\gamma^2-\gamma)\int_0^T\int_0^T \phi(s,s')C_sD_{s'}\xi_{s'}dsds'\\
&+(\gamma^2-\gamma)\int_0^T\int_0^T \phi(s,s')D_sK_sD_{s'}\xi_{s'}dsds'\Big\}. \qquad (4.15)
\end{aligned}
$$

Using the Fubini theorem

$$
\int_0^T\int_0^t\int_0^t g(t,s,s')dsds'dt=\int_0^T\int_s^T\int_0^t g(t,s',s)ds'dtds, \qquad (4.16)
$$

we obtain

$$
\begin{aligned}
&\frac{d}{d\varepsilon}J(K_\cdot+\varepsilon\xi_\cdot)\Big|_{\varepsilon=0}\\
&=\int_0^T\Big\{\alpha B_s\int_s^T Q_t\Theta_{\alpha,t}dt+(\alpha^2-\alpha)\int_s^T\int_0^t Q_t\Theta_{\alpha,t}\phi(s',s)C_{s'}D_sds'dt\\
&+(\alpha^2-\alpha)\int_s^T\int_0^t Q_t\Theta_{\alpha,t}\phi(s',s)D_{s'}K_{s'}D_sds'dt+2\tilde{R}(s)\\
&+\sum_{i,j=1}^m\int_s^T R_{ij}(t)K_i(t)^{\beta_i}K_j(t)^{\beta_j}\Theta_{\beta_i+\beta_j,t}\\
&\cdot\Big[\beta_{ij}B_s+(\beta_{ij}^2-\beta_{ij})\int_0^t\phi(s',s)C_{s'}D_sds'+(\beta_{ij}^2-\beta_{ij})\int_0^t\phi(s',s)D_{s'}K_{s'}D_sds'\Big]dt\\
&+\gamma G\Theta_{\gamma,T}B_s+(\gamma^2-\gamma)G\Theta_{\gamma,T}\int_0^T\phi(s',s)C_{s'}D_sds'\\
&+(\gamma^2-\gamma)G\Theta_{\gamma,T}\int_0^T\phi(s',s)D_{s'}K_{s'}D_sds'\Big\}\xi_s ds=0. \qquad (4.17)
\end{aligned}
$$

Since ξ_s is arbitrary, we obtain the desired equation, (4.7). □

The necessary condition (4.7), albeit complicated, is very general. Let us now discuss two special (yet interesting) cases, where this condition can be greatly simplified.

Case 1. Suppose there is no running cost, i.e., $Q_t\equiv 0$ and $R_t\equiv 0$. However, $G\neq 0$ and $\gamma\neq 0$ (otherwise the problem would become trivial). This kind of situation arises often in financial portfolio selection models, where only terminal wealth is of concern (see, e.g., [23, 29]). In this case condition (4.7) reduces to

$$
(\gamma-1)\int_0^T\phi(s',s)D_{s'}K_{s'}ds'\ D_s+(\gamma-1)\int_0^T\phi(s',s)C_{s'}ds'\ D_s+B_s=0\ \text{ a.e. } s\in[0,T]. \qquad (4.18)
$$

Equation (4.18) (with $K_\cdot$ being the unknown) is reduced to the *Carleman-type equation* in the following way. Rewrite (4.18) as

$$
\left[(\gamma-1)\int_0^T\phi(s',s)K_{s'}D_{s'}ds'+(\gamma-1)\int_0^T\phi(s',s)C_{s'}ds'\right]D_s=-B_s.
$$

Since the term inside the bracket is a real-valued function, a necessary condition that (4.18) has a solution is

$$\int_0^T \phi(s',s)K_{s'}D_{s'}ds' + \int_0^T \phi(s',s)C_{s'}ds' = \frac{\lambda_s}{1-\gamma}, \tag{4.19}$$

where λ_s is a real-valued function satisfying

$$\lambda_s D_s = B_s.$$

Equation (4.19) is equivalent to (noting that $\phi(s,t) \equiv \phi(t,s)$)

$$\int_0^T \phi(s,t)K_t D_t dt = \frac{\lambda_s}{1-\gamma} - \int_0^T \phi(t,s)C_t dt. \tag{4.20}$$

This equation, with $K.D.$ being the unknown, is of the following general Carleman type:

$$\int_0^T \phi(x,y)h(y)dy = g(x),$$

where g is a given function, h is the unknown function to be determined by this equation, and $\phi(x,y) = H(2H-1)|x-y|^{2H-2}$. The solution of this equation can be expressed explicitly as

$$h(x) = -a_H x^{1/2-H}\frac{d}{dx}\int_x^T \Big[w^{2H-1}(w-x)^{1/2-H}\frac{d}{dw}\int_0^w z^{1/2-H}(w-z)^{1/2-H}g(z)dz\Big]dw,$$

where

$$a_H := \frac{\Gamma(2-2H)}{2H\Gamma(\frac{1}{2}+H)\Gamma(\frac{3}{2}-H)^3};$$

see [11], [12], [13]. Applying this general formula to (4.20) we obtain $K.D.$. Note that, in general, $K.$ may not be unique if its dimension is strictly greater than 1.

A particular case for (4.18) is when $\gamma = 1$ *or* $D_s = 0$ a.e. $s \in [0,T]$. In this case, (4.18) implies it is necessary that $B_s = 0$ a.e. $s \in [0,T]$ (for there to be an extreme achieved by a Markovian control); hence system (3.1) becomes uncontrolled and the cost functional (4.1) is constant valued.

Case 2. Suppose $\alpha = \beta_{ij} = \gamma \neq 0$ for all i,j. Then (4.7) specializes to

(4.21)

$$\begin{aligned}
&\frac{2}{\alpha}\tilde{R}(s) + B_s\int_s^T \Theta_{\alpha,t}\left(Q_t + \sum_{i,j=1}^m R_{ij}(t)K_i(t)^{\beta_i}K_j(t)^{\beta_j}\right)dt \\
&\quad + (\alpha-1)D_s\int_s^T\int_0^t \Theta_{\alpha,t}\phi(s',s)(C_{s'}+D_{s'}K_{s'})\left(Q_t + \sum_{i,j=1}^m R_{ij}(t)K_i(t)^{\beta_i}K_j(t)^{\beta_j}\right)ds'dt \\
&\quad + G\Theta_{\alpha,T}\left[B_s + (\alpha-1)D_s\int_0^T \phi(s',s)(C_{s'}+D_{s'}K_{s'})ds'\right] = 0 \text{ a.e. } s \in [0,T].
\end{aligned}$$

If $\alpha = \gamma = 2$ and $\beta_i = \beta_j = 1$ for all i, j (i.e., the usual LQ case), then the above equation can be further simplified to

$$\begin{aligned} &K_s^* R_s \Theta_s + B_s \int_s^T \Theta_t (Q_t + K_t^* R_t K_t) dt \\ &\quad + D_s \int_s^T \int_0^t \Theta_t \phi(s', s)(C_{s'} + D_{s'} K_{s'})(Q_t + K_t^* R_t K_t) ds' dt \\ &\quad + G\Theta_T \left[B_s + D_s \int_0^T \phi(s', s)(C_{s'} + D_{s'} K_{s'}) ds' \right] = 0 \quad \text{a.e. } s \in [0, T], \end{aligned} \tag{4.22}$$

where $\Theta_t := \Theta_{2,t}$.

Therefore, in the LQ case, where the state is scalar and $D_t \neq 0$, one needs to solve the functional differential equation (4.22) in order to obtain an optimal Markovian feedback control. Recall that in section 3 we derived Riccati equation (3.5) for the LQ control when $D_t \equiv 0$. Let us now show that (3.5) can also be recovered from (4.22) under the assumptions of Theorem 3.1. To this end, first note that (4.22) reduces to the following when $D_t \equiv 0$:

$$K_s^* R_s \Theta_s + B_s \int_s^T \Theta_t (Q_t + K_t^* R_t K_t) dt + G B_s \Theta_T = 0 \quad \text{a.e. } s \in [0, T]. \tag{4.23}$$

Define

$$p_s := \Theta_s^{-1} \left[\int_s^T \Theta_t (Q_t + K_t^* R_t K_t) dt + G\Theta_T \right] \tag{4.24}$$

and

$$K_s := -R_s^{-1} B_s^* p_s. \tag{4.25}$$

It is clear that $K_\cdot$ defined above satisfies the necessary condition (4.23). Now we derive the equation that governs $p_\cdot$. Multiplying (4.24) by Θ_s and then taking derivative in s, we obtain

$$\dot{p}_s \Theta_s + p_s \dot{\Theta}_s + \Theta_s (Q_s + K_s^* R_s K_s) = 0. \tag{4.26}$$

However, (4.6) gives (noting $D_t \equiv 0$)

$$\dot{\Theta}_s = \Theta_s \left[2(A_s + B_s K_s) + 2(C_s + D_s K_s) \int_0^s \phi(s, s')(C_{s'} + D_{s'} K_{s'}) ds' \right]. \tag{4.27}$$

Plugging (4.27) into (4.26), noting (4.25), and going through some manipulation, we get that $p_\cdot$ satisfies Riccati equation (3.5) (that $p_T = G$ is evident from (4.24)).

It should be noted that the above derivation shows only that $K_\cdot$ defined by (4.25), where $p_\cdot$ is the solution to Riccati equation (3.5), satisfies the *necessary condition* of an optimum. Hence, it can by no means supersede Theorem 3.1.

Another interesting case for necessary condition (4.22) is when $H = \frac{1}{2}$ (i.e., the Brownian motion), whereas $D_t \neq 0$. In this case $\phi(s, t) = \delta(s, t)$, the Dirac delta function, and (4.22) becomes

$$\begin{aligned} &K_s^* R_s \Theta_s + [B_s + (C_s + K_s^* D_s^*) D_s] \int_s^T \Theta_t (Q_t + K_t^* R_t K_t) dt \\ &\quad + G\Theta_T [B_s + (C_s + K_s^* D_s^*) D_s] = 0. \end{aligned} \tag{4.28}$$

Let

$$p_s := \Theta_s^{-1}\left[\int_s^T \Theta_t(Q_t + K_t^* R_t K_t)dt + G\Theta_T\right] \tag{4.29}$$

and

$$R_s K_s = -\left[B_s^* + D_s^*(C_s + D_s K_s)\right] p_s, \tag{4.30}$$

or equivalently,

$$K_s = -(R_s + p_s D_s^* D_s)^{-1}[B_s^* + D_s^* C_s]p_s, \tag{4.31}$$

assuming that $R_s + p_s D_s^* D_s$ is invertible. On the other hand, in the present case, Θ_s satisfies

$$\dot{\Theta}_s = \Theta_s\left[2(A_s + B_s K_s) + (C_s + D_s K_s)^2\right]. \tag{4.32}$$

Going through the same argument as above, we can show that $p.$ follows:

(4.33)
$$\begin{cases} \dot{p}_s + (2A_s + C_s^2)p_s + Q_s - (B_s + C_s^* D_s)(R_s + p_s D_s^* D_s)^{-1}(B_s^* + D_s^* C_s)p_s^2 = 0, \\ p_T = G, \\ R_s + p_s D_s^* D_s \succ 0. \end{cases}$$

This is the standard (stochastic) Riccati equation (see [4]).

5. Control model 3: Vector state, quadratic cost, and general control with Stratonovich integral. In the two models previously studied, it is assumed that the state is scalar-valued and only Markovian-type feedback controls are considered. In this section, we investigate a general model, where the state is multidimensional and admissible controls are not necessarily Markovian. In addition, all the coefficients are allowed to be stochastic processes. The system is described by the linear dynamics

$$\begin{cases} dx_t = (A_t x_t + B_t u_t)dt + C_t x_t \circ dW_t^H, \\ x_0 \in \mathbb{R}^n \qquad \text{given and deterministic,} \end{cases} \tag{5.1}$$

where $A_t, C_t, 0 \le t \le T$, are $n \times n$ matrix-valued $\mathcal{F}_t$-adapted stochastic processes, and $B_t, 0 \le t \le T$, is an $n \times m$ matrix-valued $\mathcal{F}_t$-adapted process. An *admissible control* u_t, $0 \le t \le T$, is an $\mathbb{R}^m$-valued $\mathcal{F}_t$-adapted process such that (5.1) has a unique solution. Notice that, unlike in the previous two models, the Stratonovich-type integral is involved in the dynamics.

Let Q_t and $R_t, 0 \le t \le T$, be given $n \times n$ and $m \times m$ matrix-valued $\mathcal{F}_t$-adapted stochastic processes, respectively, and let G be an $n \times n$ $\mathcal{F}_T$-measurable random matrix. Define the cost functional

$$J(x_0, u.) := \mathbb{E}\left[\int_0^T (x_t^* Q_t x_t + u_t^* R_t u_t)dt + x_T^* G x_T\right]. \tag{5.2}$$

The optimal stochastic control problem is to minimize the cost functional (5.2), subject to the dynamics (5.1), over the set of all admissible controls for each given x_0.

5.1. Optimal control. In this subsection we present the solution to the optimal control problem formulated above. Let W_t be the Brownian motion in the representation of the FBM W_t^H as given by Lemma 2.1. We introduce the following matrix-valued backward SDE (BSDE):

(5.3)
$$\begin{cases} dP_t + (A_t^* P_t + P_t A_t + Q_t - P_t B_t R_t^{-1} B_t^* P_t)dt + (C_t^* P_t + P_t C_t) \circ dW_t^H - \Lambda_t dW_t = 0, \\ P_T = G. \end{cases}$$

This is a Riccati-type equation. A *pair* of $n \times n$ matrix-valued $\mathcal{F}_t$-adapted processes, $(P_\cdot, \Lambda_\cdot)$, is called a solution of (5.3) if it satisfies

$$\begin{aligned} P_t = \quad & G + \int_t^T (A_t^* P_t + P_t A_t + Q_t - P_t B_t R_t^{-1} B_t^* P_t)dt \\ & + \int_t^T (C_t^* P_t + P_t C_t) \circ dW_t^H - \int_t^T \Lambda_t dW_t \ \text{ a.e. } t \in [0,T], \end{aligned}$$

with all the integrals above well defined. Notice that, unlike a standard BSDE (see [29]), (5.3) involves simultaneously both the Brownian motion and the FBM.

Now assume that Riccati equation (5.3) admits solution $(P_\cdot, \Lambda_\cdot) \in L^2_{\mathcal{F}}(0,T;\mathbb{R}^{n\times n}) \times L^2_{\mathcal{F}}(0,T;\mathbb{R}^{n\times n})$ with $\Lambda_\cdot$ being a semimartingale of the form

$$\Lambda_t = \Lambda_0 + \int_0^t \Theta_s ds + \int_0^t \Xi_s dW_s, \tag{5.4}$$

where Θ_s and Ξ_s are continuous $\mathcal{F}_t$-adapted processes. Then

$$d\langle \Lambda\,, W\rangle_t = \Xi_t dt,$$

where the bracket $\langle \Lambda\,, W\rangle$ denotes the quadratic variational process corresponding to Λ_t and W_t. Hence, we may rewrite the Riccati (5.3) as

$$\begin{cases} dP_t = -(A_t^* P_t + P_t A_t + Q_t - P_t B_t R_t^{-1} B_t^* P_t)dt, \\ \qquad\qquad -(C_t^* P_t + P_t C_t) \circ dW_t^H + \Lambda_t \circ dW_t - \frac{1}{2}\Xi_t dt, \\ P_T = G. \end{cases} \tag{5.5}$$

Let $u_\cdot$ be any given admissible control and $x_\cdot$ be the corresponding state process. Applying the Itô formula for the Stratonovich differential [9, Theorem 4.7], we have

$$\begin{aligned} d(x_t^* P_t x_t) = {} & (A_t x_t + B_t u_t)^* P_t x_t dt + x_t^* C_t^* P_t x_t \circ dW_t^H \\ & - x_t^* (A_t^* P_t + P_t A_t + Q_t - P_t B_t R_t^{-1} B_t^* P_t) x_t dt - \frac{1}{2} x_t^* \Xi_t x_t dt \\ & - x_t^* (C_t^* P_t + P_t C_t) x_t \circ dW_t^H + x_t^* \Lambda_t x_t \circ dW_t \\ & + x_t^* P_t (A_t x_t + B_t u_t) dt + x_t^* P_t C_t x_t \circ dW_t^H \\ = {} & \left[2u_t^* B_t^* P_t x_t - x_t^* (Q_t - P_t B_t R_t^{-1} B_t^* P_t) x_t \right] dt \\ & - \frac{1}{2} x_t^* \Xi_t x_t dt + x_t^* \Lambda_t x_t \circ dW_t. \end{aligned} \tag{5.6}$$

Integrating from 0 to T and taking expectation, we get

$$\begin{aligned}\mathbb{E}\ [x_T^* P_T x_T] = \quad & x_0^* P_0 x_0 + \mathbb{E}\int_0^T \left[2u_t^* B_t^* P_t x_t - x_t^*(Q_t - P_t B_t R_t^{-1} B_t^* P_t)x_t\right] dt \\ & -\frac{1}{2}\mathbb{E}\int_0^T x_t^* \Xi_t x_t dt + \mathbb{E}\left[\int_0^T x_t^* \Lambda_t x_t \circ dW_t\right].\end{aligned} \tag{5.7}$$

In view of the relation between the Itô integral and the Stratonovich integral [9, Theorem 3.12], we have

$$\begin{aligned}x_t &= x_0 + \int_0^t (A_s x_s + B_s u_s)ds + \int_0^t C_s x_s \circ dW_s^H \\ &= x_0 + \int_0^t [A_s x_s + B_s u_s + D_s^H(C_s x_s)]ds + \int_0^t C_s x_s dW_s^H.\end{aligned}$$

Writing it in component form, we have

$$x_t^i = x_0^i + \int_0^t g_s^i ds + \int_0^t h_s^i dW_s^H, \quad i = 1, 2, \ldots, n,$$

where g_s^i is the ith component of $A_s x_s + B_s u_s + D_s^H(C_s x_s)$ and h_s^i the ith component of $C_s x_s$. By the Itô formula [9, Theorem 4.6], we have

$$\begin{aligned}d(x_t^i x_t^j) &= x_t^i g_t^j dt + x_t^i h_t^j dW_t^H + x_t^j g_t^i dt + x_t^j h_t^i dW_t^H + h_t^j D_t^H x_t^i dt + h_t^i D_t^H x_t^j dt \\ &= g_t^{ij} dt + h_t^{ij} dW_t^H,\end{aligned}$$

where

$$g_t^{ij} := x_t^i g_t^j + x_t^j g_t^i + h_t^j D_t^H x_t^i + h_t^i D_t^H x_t^j$$

and

$$h_t^{ij} := x_t^i h_t^j + x_t^j h_t^i.$$

We also write $d\Lambda_t$ in component form:

$$d\Lambda_t^{ij} = \Theta_t^{ij} dt + \Xi_t^{ij} dW_t, \quad 1 \le i, j \le n.$$

Then taking into account that $dW_t^H dW = 0$ (which can be proved by the same estimate as that in the proof of Theorem 2.1), we have

$$d[\Lambda_t^{ij} x_t^i x_t^j] = \Lambda_t^{ij} g_t^{ij} dt + \Lambda_t^{ij} h_t^{ij} dW_t^H + x_t^i x_t^j \Theta_t^{ij} dt + x_t^i x_t^j \Xi_t^{ij} dW_t.$$

Taking the sum over i and j and then integrating, we obtain

$$x_t^* \Lambda_t x_t = x_0^* \Lambda_0 x_0 + \int_0^t M_s ds + \int_0^t N_s dW_s^H + \int_0^t x_s^* \Xi_s x_s dW_s,$$

where

$$M_s := \sum_{i,j=1}^n (\Lambda_s^{ij} g_s^{ij} + x_s^i x_s^j \Theta_s^{ij}), \quad N_s := \Lambda_s^{ij} h_s^{ij}.$$

We set

$$Y_t := x_0^* \Lambda_0 x_0 + \int_0^t M_s ds + \int_0^t N_s dW_s^H, \tag{5.8}$$

which is of the form (2.9). Assuming that the processes M_s and N_s satisfy conditions (2.10), it then follows from Theorem 2.1 that

$$\mathbb{E} \int_0^T Y_s \circ dW_s = 0.$$

On the other hand, the well-known relation between the Itô and Stratonovich integrals for Brownian motion yields (see [21])

$$\mathbb{E}\left[\int_0^T \left(\int_0^t x_s^* \Xi_s x_s dW_s\right) \circ dW_t\right] = \frac{1}{2}\mathbb{E} \int_0^T x_t^* \Xi_t x_t dt.$$

Therefore we obtain

$$\mathbb{E}\left[\int_0^T x_t^* \Lambda_t x_t \circ dW_t\right] = \frac{1}{2}\mathbb{E} \int_0^T x_t^* \Xi_t x_t dt.$$

It then follows from (5.7) that

$$\mathbb{E}\ [x_T^* P_T x_T] = \mathbb{E}\ \left[x_0^T P_0 x_0\right] + \mathbb{E} \int_0^T \left[2u_t^* B_t^* P_t x_t - x_t^* (Q_t - P_t B_t R_t^{-1} B_t^* P_t) x_t\right] dt.$$

Hence

$$\begin{aligned} J(x_0, u_\cdot) &= x_0^* P_0 x_0 + \mathbb{E} \int_0^T (2u_t^* B_t^* P_t x_t + u_t^* R u_t + x_t^* P_t B_t R_t^{-1} B_t^* P_t x_t) dt \\ &= x_0^* P_0 x_0 + \mathbb{E} \int_0^T \left(u_t + R_t^{-1} B_t^* P_t x_t\right)^* R_t \left(u_t + R_t^{-1} B_t^* P_t x_t\right) dt. \end{aligned}$$

Assuming that $R_t \succ 0$, the minimum of the above functional is achieved when $u_t = -R_t^{-1} B_t^* P_t x_t$ with the minimum value being $x_0^* P_0 x_0$, provided that the control $u_t = -R_t^{-1} B_t^* P_t x_t$ induces an admissible control.

Summarizing, we have obtained the following result.

THEOREM 5.1. *Assume that $R_t \succ 0$ a.e. $t \in [0,T]$ a.s. and that Riccati equation* (5.3) *admits a solution $(P_\cdot, \Lambda_\cdot) \in L^2_{\mathcal{F}}(0,T;\mathbb{R}^{n\times n}) \times L^2_{\mathcal{F}}(0,T;\mathbb{R}^{n\times n})$ with $\Lambda_\cdot$ being a semimartingale of the form* (5.4). *Moreover, assume that the process Y_t defined by* (5.8) *satisfies conditions* (2.10) *and that* (5.6) *holds. Then, the stochastic control problem* (5.1)–(5.2) *is solvable and the optimal control $\hat{u}_\cdot$ must be of Markovian feedback type, given explicitly by*

$$\hat{u}_t = -R_t^{-1} B_t^* P_t x_t, \tag{5.9}$$

provided that the control $\hat{u}_\cdot$ (5.9) *is admissible. Moreover, the minimum value is achieved as $J(x_0, \hat{u}_\cdot) = x_0^* P_0 x_0$.*

Remark 5.1. Even in the case when the system is driven by Brownian motion, to the best of our knowledge, equations involving the Stratonovich-type integral (5.1) have not been considered in the literature on stochastic LQ control.

Remark 5.2. In the Brownian motion case, the Riccati equation is a deterministic differential equation when all the coefficients A_t, B_t, and C_t are deterministic (see [4]). However, In the FBM case, Riccati equation (5.3) is still stochastic even when all the coefficients are deterministic.

5.2. Origin of the Riccati equation. In the previous subsection we established the optimal solution to the control problem (5.1)–(5.2) by making use of Riccati equation (5.3). The approach is a standard completion-of-square technique commonly employed for LQ control once the Riccati equation is in place. However, one may be curious as to how (5.3) was derived in the first place. In this subsection we present, in a rather formal way, how we get (5.3).

We first approximate the underlying FBM using an adapted convolution approach of Malliavin (see [2], [18]). Set $W_s^H = 0$ for $s \le 0$. Define

$$W_t^{H,\varepsilon} := \frac{1}{\varepsilon}\int_{t-\varepsilon}^{t} W_s^H ds. \tag{5.10}$$

The derivative of $W_t^{H,\varepsilon}$ is

$$\dot{W}_t^{H,\varepsilon} = \frac{1}{\varepsilon}\left(W_t^H - W_{t-\varepsilon}^H\right). \tag{5.11}$$

We approximate the stochastic differential system (5.1) by

$$\dot{x}_t^\varepsilon = A_t x_t^\varepsilon + B_t u_t + C_t x_t^\varepsilon \dot{W}_t^{H,\varepsilon} = (A_t + C_t \dot{W}_t^{H,\varepsilon}) x_t^\varepsilon + B_t u_t. \tag{5.12}$$

The above can be regarded as a controlled linear system with the normal Brownian motion (but the diffusion coefficient is zero) and *random* coefficient in drift. Consider an LQ control problem with dynamics (5.12) and cost functional (5.2). The corresponding Riccati equation, following [4], is

$$\begin{cases} dP_t^\varepsilon + \big[(A_t + C_t\dot{W}_t^{H,\varepsilon})^* P_t^\varepsilon + P_t^\varepsilon (A_t + C_t \dot{W}_t^{H,\varepsilon}) \\ \qquad\qquad + Q_t - P_t^\varepsilon B_t R_t^{-1} B_t^* P_t^\varepsilon\big] dt - \Lambda_t^\varepsilon dW_t = 0, \\ P_T^\varepsilon = G. \end{cases} \tag{5.13}$$

Arranging this equation properly, we get

$$\begin{cases} dP_t^\varepsilon + (A_t^* P_t^\varepsilon + P_t^\varepsilon A_t + Q_t - P_t^\varepsilon B_t R_t^{-1} B_t^* P_t^\varepsilon)dt - \Lambda_t^\varepsilon dW_t + C_t^* P_t^\varepsilon \dot{W}_t^{H,\varepsilon} dt \\ \qquad + P_t^\varepsilon C_t \dot{W}_t^{H,\varepsilon} dt = 0, \\ P_T^\varepsilon = G. \end{cases} \tag{5.14}$$

A (formal) limiting equation of (5.14) is exactly Riccati equation (5.3) introduced earlier. It should be noted that it is hard to prove the convergence of (5.14). Fortunately, once we have obtained the Riccati equation, albeit formally, we can then use the completion-of-square technique to solve the original control problem, as shown in the previous subsection.

It is also worth noting the difficulty in dealing with the controlled system

$$\begin{cases} dx_t = (A_t x_t + B_t u_t)dt + (C_t x_t + D_t u_t) \circ dW_t^H, \\ x_0 \in \mathbb{R}^n \qquad \text{given and deterministic} \end{cases} \tag{5.15}$$

instead of (5.1). Going through the same approximation procedure as above, we get that (5.13) now becomes

$$\begin{aligned} dP^\varepsilon + \Big[(A_t + C_t\dot{W}_t^{H,\varepsilon})^* P_t^\varepsilon + P_t^\varepsilon (A_t + C_t \dot{W}_t^{H,\varepsilon}) + Q_t\Big] dt \\ - P_t^\varepsilon (B_t + D_t W_t^{H,\varepsilon}) R_t^{-1} (B_t + D_t W_t^{H,\varepsilon})^* P_t^\varepsilon dt - \Lambda_t^\varepsilon dW_t = 0. \end{aligned}$$

Rearranging, we have

$$\begin{aligned} dP_t^\varepsilon &+ \left[A_t^* P_t^\varepsilon + P_t^\varepsilon A_t + Q_t - P_t^\varepsilon (B_t + D_t W_t^{H,\varepsilon}) R_t^{-1} (B_t + D_t W_t^{H,\varepsilon})^* P_t^\varepsilon\right] dt \\ &-\Lambda_t^\varepsilon dW_t + C_t^* P_t^\varepsilon \dot{W}_t^{H,\varepsilon} dt + P_t^\varepsilon C_t \dot{W}_t^{H,\varepsilon} dt \\ &-P_t^\varepsilon D_t R_t B_t^* P_t^\varepsilon \dot{W}_t^{H,\varepsilon} - P_t^\varepsilon B_t R_t D_t^* P_t^\varepsilon \dot{W}_t^{H,\varepsilon} - P_t^\varepsilon D_t R_t^{-1} D_t P_t^\varepsilon (\dot{W}_t^{H,\varepsilon})^2 dt = 0. \end{aligned}$$

When $D_t \not\equiv 0$, the term $(\dot{W}_t^{H,\varepsilon})^2 dt$ presents, whose limit does not exist. This explains the difficulty of studying the case when $D_t \not\equiv 0$.

Let us conclude this section by remarking again that the discussion in this section, to some extent, is rather formal, in the sense that Theorem 5.1 depends upon the existence of a solution to Riccati equation (5.3) along with some required (very technical) conditions of the solution. Nevertheless, the highlight of this section is indeed the introduction of the new equation (5.3), which is a highly unconventional BSDE involving stochastic integrals with respect to both the normal and fractional Brownian motions. Research on this type of equation, we believe, will prove to be very interesting and challenging from both the BSDE and fractional analysis points of view.

6. Concluding remarks. In this paper we have attempted to systematically tackle an important class of stochastic control problems with linear dynamic systems involving FBM. Three specific control models have been studied. It should be noted that the results in this paper are far from complete. Indeed the paper is intended to be more inspirational—in the sense that it will inspire more research along the same line—than exhaustive and conclusive. Many interesting and challenging problems remain open. The foremost is the optimal control over the set of general controls—not necessarily that of the Markovian controls. The reason is that the FBM has long-range dependence; hence, it is natural to expect that the optimal control would not only depend on the present time, but on the past as well. In fact, we conjecture that the optimal control is generally linear but non-Markovian (in a special case, this fact has been proved in a recent paper [22]). In particular, it is interesting to seek the optimal control in the class of the controls

$$u_t = \int_0^t K(t,s) x_s ds.$$

We will study such problems in forthcoming papers.

Acknowledgments. We thank the two anonymous referees for their careful reading of an earlier version of the paper and for their constructive comments that led to an improved version. In particular, we are grateful to one of the referees for bringing the paper [22] to our attention.

REFERENCES

[1] M. Ait Rami, J. B. Moore, and X. Y. Zhou, *Indefinite stochastic linear quadratic control and generalized differential Riccati equation*, SIAM J. Control Optim., 40 (2001), pp. 1296–1311.

[2] J.-M. Bismut, *Mécanique aléatoire*, Lecture Notes in Math. 866, Springer-Verlag, Berlin, New York, 1981.

[3] F. Biagini, Y. Hu, B. Øksendal, and A. Sulem, *A stochastic maximum principle for processes driven by fractional Brownian motion*, Stochastic Process. Appl., 100 (2002), pp. 233–253.

[4] S. Chen, X. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[5] S. Chen and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs.* II, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.

[6] W. Dai and C. C. Heyde, *Itô formula with respect to fractional Brownian motion and its applications*, J. Appl. Math. Stochastic Anal., 9 (1996), pp. 439–448.

[7] L. Decreusefond and A. S. Üstünel, *Stochastic analysis of the fractional Brownian motion*, Potential Anal., 10 (1999), pp. 177–214.

[8] C. Dellacherie, P.-A. Meyer, and B. Maisonneuve, *Probabilités et Potentiel*, V. Hermann, Paris, 1992.

[9] T. E. Duncan, Y. Hu, and B. Pasik-Duncan, *Stochastic calculus for fractional Brownian motion.* I. *Theory*, SIAM J. Control Optim., 38 (2000), pp. 582–612.

[10] X. Fernique, *Regularité de fonctions aléatoires non gaussiennes*, in Proceedings of the Eleventh Saint Flour Probability Summer School (Saint Flour, 1981), Lecture Notes in Math. 976, Springer, Berlin, 1983, pp. 1–74.

[11] Y. Z. Hu, *Probability structure preserving and absolute continuity*, Ann. Inst. H. Poincaré, Probab. Statist. 38 (2002), pp. 557–580.

[12] Y. Z. Hu, *Prediction and translation of fractional Brownian motions*, in Stochastics in Finite and Infinite Dimensions, T. Hida et al., eds., Trends Math., Birkhäuser, Boston, 2001, pp. 153–171.

[13] Y. Z. Hu, *Integral transformation and anticipative calculus for fractional Brownian motions*, Mem. Amer. Math. Soc., to appear.

[14] Y. Z. Hu, *A unified approach to several inequalities for Gaussian and diffusion measures*, in Séminaire de Probabilités XXXIV, J. Azéma et al., eds., Lecture Notes in Math. 1729, Springer, Berlin, 2000, pp. 329–335.

[15] Y. Z. Hu, B. Øksendal, and A. Sulem, *Optimal consumption and portfolio in a Black–Scholes market driven by fractional Brownian motion*, Infin. Dimens. Anal. Quantum Probab. Relat. Top., 6 (2003), pp. 519–536.

[16] Y. Z. Hu and P. A. Meyer, *Chaos de Wiener et intégrales de Feynman*, in Séminaire de Probabilités XXII, J. Azema, P. A. Meyer, and M. Yor, eds., Lecture Notes in Math. 1321, Springer-Verlag, Berlin, 1988, pp. 51–71.

[17] Y. Z. Hu and P. A. Meyer, *Sur les intégrales multiples de Stratonovich*, in Séminaire de Probabilités XXVI, J. Azema, P. A. Meyer, and M. Yor, eds., Lecture Notes in Math. 1321, Springer-Verlag, Berlin, 1988, pp. 72–81.

[18] Y. Z. Hu and P. A. Meyer, *On the approximation of Stratonovich multiple integrals*, in Stochastic Processes, a Festschrift in Honor of G. Kallianpur, S. Cambanis et al., eds., Springer, Berlin, 1993, pp. 141–147.

[19] Y. Z. Hu and B. Øksendal, *Fractional white noise calculus and applications to finance*, Infin. Dimens. Anal. Quantum Probab. Relat. Top., 6 (2003), pp. 1–32.

[20] Y. Hu, B. Øksendal, and A. Sulem, *Optimal portfolio in a fractional Black and Scholes market*, in Mathematical Physics and Stochastic Analysis: Essays in Honor of Ludwig Streit, S. Albeverio et al., eds., World Scientific, River Edge, NJ, 2000, pp. 267–279.

[21] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, 2nd ed., North-Holland Math. Library 24, North-Holland, Amsterdam; Kodansha, Ltd., Tokyo, 1989.

[22] M. L. Kleptsyna, A. Le Breton, and M. Viot, *About the linear-quadratic regulator problem under a fractional Brownian perturbation*, ESAIM Probab. Statist., 7 (2003), pp. 161–170.

[23] M. Kohlmann and X. Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach*, SIAM J. Control Optim., 38 (2000), pp. 1392–1407.

[24] J. A. León and C. Tudor, *Semilinear fractional stochastic differential equations*, Bol. Soc. Mat. Mexicana (3), 8 (2002), pp. 205–226.

[25] S. J. Lin, *Stochastic analysis of fractional Brownian motions*, Stoch. Stoch. Rep., 55 (1995), pp. 121–140.

[26] I. Norros, E. Valkeila, and J. Virtamo, *An elementary approach to a Girsanov formula and other analytic results on fractional Brownian motions*, Bernoulli, 5 (1999), pp. 571–587.

[27] C. Tudor, *Fractional bilinear stochastic differential equations with the drift in the first fractional chaos*, Stoch. Anal. Appl., 22 (2004), pp. 1209–1233.

[28] D. D. Yao, S. Zhang, and X. Y. Zhou, *Stochastic linear-quadratic control via semidefinite programming*, SIAM J. Control Optim., 40 (2001), pp. 801–823.

[29] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Appl. Math., 43, Springer-Verlag, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 43, No. 6, pp. 2278–2295

STABILITY RADII OF POSITIVE LINEAR FUNCTIONAL DIFFERENTIAL EQUATIONS UNDER MULTI-PERTURBATIONS*

PHAM HUU ANH NGOC† AND NGUYEN KHOA SON‡

Abstract. We study stability radii of linear retarded systems described by general linear functional differential equations. A lower and an upper bound for the complex stability radius with respect to multi-perturbations are given. Furthermore, in some special cases concerning the structure matrices, the complex stability radius can precisely be computed via the associated transfer function. Then, the class of positive linear retarded systems is studied in detail. It is shown that for this class, complex, real and positive stability radius under multi-perturbations or multi-affine perturbations coincide and can be computed by simple formulae expressed in terms of the system matrices.

Key words. functional differential equation, retarded system, positive system, stability radius, multi-perturbation

AMS subject classifications. 39A11, 93D09

DOI. 10.1137/S0363012903434789

1. Introduction.

Motivated by many application in control engineering, problems of robust stability of dynamical systems have attracted a lot of attention of researchers during the last 20 years. In the study of these problems, the notion of stability radius was proved to be a very effective tool. By definition, the stability radius of a given asymptotically stable system $\dot{x}(t) = Ax(t)$ is the *maximal* $\gamma > 0$ for which all the systems of the form

$$\dot{x}(t) = (A + D\Delta E)x(t), \qquad \|\Delta\| < \gamma$$

are asymptotically stable. Here, Δ is an unknown disturbance matrix, and D and E are given matrices defining the structure of the perturbations. Depending upon whether complex or real disturbances Δ are considered, this maximal γ is called a complex or real stability radius, respectively. The basic problem in the study of robustness of stability of the system is to characterize and compute these radii in terms of given matrices A, D, E. It is important to note that these two stability radii are in general distinct. The analysis and computation of the complex stability radius for systems under structured perturbations has been done first in [6] and extended later in many subsequent papers (see [7] for a survey up to 1990) while the computation of the real stability radius, being a much more difficult problem, has been solved only recently; see, e.g., [17].

The situation is much simpler for the class of positive systems. It has been shown in [19] that if A is a Metzler matrix (i.e., all off-diagonal entries of A are nonnegative) and D, E are nonnegative matrices, then the complex and the real stability radii coincide and can be computed directly by a simple formula. These results have been extended quite recently to infinite-dimensional positive systems in [3], [4], to positive

*Received by the editors September 15, 2003; accepted for publication (in revised form) March 18, 2005; published electronically May 27, 2005.
http://www.siam.org/journals/sicon/43-6/43478.html

†Department of Mathematics, University of Hue, 32 Le Loi Str., Hue City, Vietnam (phanhngoc@yahoo.com).

‡Institute of Mathematics, P.O. Box 631, Bo Ho, Hanoi, Vietnam.

continuous time-delay systems in [21], [22], [23] and to discrete time-delay systems in [15], [10].

It is worth noticing that the notion of stability radius can be extended to various perturbation types [7]. Among perturbation types, two of the following perturbation types

$$A \to A + \sum_{i=1}^{N} D_i \Delta_i E_i \qquad \text{(multi-perturbation)}, \tag{1}$$

$$A \to A + \sum_{i=1}^{N} \delta_i A_i \qquad \text{(multi-affine perturbation)} \tag{2}$$

are most well known in control theory and include perturbation types studied in the literature. However, so far, the problem of finding computable formulas for stability radii of the linear systems $\dot{x}(t) = Ax(t),\ t \geq 0;\ x(k+1) = Ax(k),\ k \in \mathbb{N}$, under multi-perturbations still remains open; see, e.g., [8], [9]. In these papers, for multi-perturbations (1), only lower bounds for the real stability radius were obtained. In the literature, there are no results available on the computation of stability radii for this case.

The main purpose of this paper is to develop a framework for studying stability radii of general retarded systems described by linear functional differential equations of the form

$$\dot{x}(t) = A_0 x(t) + \int_{-h}^{0} d[\eta(\theta)] x(t+\theta)$$

under multi-perturbations (1), multi-affine perturbations (2), and to establish the previous results in this more general situation. Formulae will be derived for the stability radii of these general systems in terms of coefficient matrices appearing in the initial system equations.

The organization of the paper is as follows. In the next section, we summarize some notations and preliminary results which will be used in what follows. In section 3, a lower and an upper bound for the complex stability radius with respect to multi-perturbations are given. Furthermore, in some special cases concerning the structure matrices, the complex stability radius can precisely be computed via the associated transfer function. Then, for the class of positive functional differential equations it will be shown that the complex, real and positive stability radius under multi-perturbations coincide and a simple formula for their computation is established. In section 4, the class of multi-affine perturbations (which enables us to deal with diagonal disturbances) is considered and an explicit formula for computing the real stability radius of positive retarded systems is given. Two examples illustrate the obtained results.

2. Preliminaries. In this section we shall define some notations and recall some well known results which will be used in the subsequent sections. Let $\mathbb{K} = \mathbb{C}$ or $\mathbb{R}$, where $\mathbb{C}$ and $\mathbb{R}$ denote the sets of all complex and all real numbers, respectively. For an integer $l, q \geq 1, \mathbb{K}^l$ denotes the l-dimensional vector space over $\mathbb{K}$, $(\mathbb{K}^l)^*$ is its dual, and $\mathbb{K}^{l\times q}$ stands for the set of all $l \times q$-matrices with entries in $\mathbb{K}$. Inequalities between real matrices or vectors will be understood componentwise, i.e., for two real matrices $A = (a_{ij})$ and $B = (b_{ij})$ in $\mathbb{R}^{l\times q}$, we write $A \geq B$ if and only if $a_{ij} \geq b_{ij}$ for $i = 1, \ldots, l$, $j = 1, \ldots, q$. We denote by $\mathbb{R}_+^{l\times q}$ the set of all nonnegative matrices $A \geq 0$. Similar notations are adopted for vectors. For $x \in \mathbb{K}^n$ and $P \in \mathbb{K}^{l\times q}$

we define $|x| = (|x_i|)$ and $|P| = (|p_{ij}|)$. For any matrix $A \in \mathbb{K}^{n\times n}$ the *spectral radius* and *spectral abscissa* of A are denoted by $\rho(A) = \max\{|\lambda|; \lambda \in \sigma(A)\}$ and $\mu(A) = \max\{\mathrm{Re}\lambda; \lambda \in \sigma(A)\}$, where $\sigma(A) := \{s \in \mathbb{C}; \det(sI_n - A) = 0\}$ is the spectrum of A. $A \in \mathbb{R}^{n\times n}$ is called a *Metzler matrix* if all off-diagonal elements of A are nonnegative or, equivalently, $tI_n + A \geq 0$ for some $t \geq 0$. It is clear that any $A \in \mathbb{R}^{n\times n}_+$ is a Metzler matrix and, moreover, $\rho(A) = \mu(A)$.

A norm $\|\cdot\|$ on $\mathbb{K}^n$ is said to be *monotonic* if $\|x\| = \| \, |x| \, \|$ for all $x \in \mathbb{K}^n$. Every p-norm on $\mathbb{K}^n, 1 \leq p \leq \infty$, is monotonic. Throughout the paper, if not otherwise stated, the norm of a matrix $P \in \mathbb{K}^{l\times q}$ is understood as its operator norm associated with a given pair of monotonic vector norms on $\mathbb{K}^l$ and $\mathbb{K}^q$, that is, $\|P\| = \max\{\|Py\|; \ \|y\| = 1\}$. We note that the operator norm is in general not monotonic norm on $\mathbb{K}^{l\times q}$ even if $\mathbb{K}^l, \mathbb{K}^q$ are provided with monotonic norms. However, such monotonicity holds for nonnegative matrices. Moreover, we have (see, e.g., [19])

$$P \in \mathbb{K}^{l\times q}, \quad Q \in \mathbb{R}^{l\times q}_+, \quad |P| \leq Q \quad \Rightarrow \quad \|P\| \leq \| \, |P| \, \| \leq \|Q\|. \tag{3}$$

The matrix function $\eta(.) : \ [-h, 0] \to \mathbb{R}^{l\times q}$ is called a increasing matrix function if

$$\eta(\theta_2) \geq \eta(\theta_1) \quad \text{for} \quad -h \leq \theta_1 \leq \theta_2 \leq 0.$$

To make the presentation self-contained we now summarize in the following theorem some existing results on properties of Metzler matrices which will be used in what follows (see, e.g., [2], [19]).

THEOREM 2.1. *Suppose that $A \in \mathbb{R}^{n\times n}$ is a Metzler matrix. Then*

(i) (Perron–Frobenius) *$\mu(A)$ is an eigenvalue of A and there exists a nonnegative eigenvector $x \geq 0$, $x \neq 0$ such that $Ax = \mu(A)x$;*

(ii) *given $\alpha \in \mathbb{R}$, there exists a nonzero vector $x \geq 0$ such that $Ax \geq \alpha x$ if and only if $\mu(A) \geq \alpha$;*

(iii) *$(tI_n - A)^{-1}$ exists and is nonnegative if and only if $t > \mu(A)$;*

(iv) *given $B \in \mathbb{R}^{n\times n}_+$, $C \in \mathbb{C}^{n\times n}$, then*

$$|C| \leq B \quad \Longrightarrow \quad \mu(A + C) \leq \mu(A + B). \tag{4}$$

A matrix function $\eta(\cdot) : [\alpha, \beta] \to \mathbb{K}^{m\times n}$ is said to be of bounded variation if

$$\mathrm{Var}(\eta; \alpha, \beta) := \sup_{P[\alpha,\beta]} \sum_k \|\eta(\theta_k) - \eta(\theta_{k-1})\| < +\infty, \tag{5}$$

where the supremum is taken over the set of all finite partitions of the interval $[\alpha, \beta]$. The set $BV([\alpha, \beta], \mathbb{K}^{m\times n})$ of all matrix functions $\eta(\cdot)$ of bounded variation on $[\alpha, \beta]$ satisfying $\eta(\alpha) = 0$ is a Banach space endowed with the norm $\|\eta\| = \mathrm{Var}(\eta; \alpha, \beta)$. Since all matrix norms on $\mathbb{K}^{m\times n}$ are equivalent, it follows that the matrix function $\eta(\cdot) = (\eta_{ij}(\cdot)) \in \mathbb{K}^{m\times n}$ is of bounded variation if and only if each $\eta_{ij}(\cdot)$ is of bounded variation. Moreover, it is easy to show that if $\mathbb{K}^{m\times n}$ is provided with ∞-norm, then

$$\max_{1\leq i\leq m} \sum_{j=1}^{n} \mathrm{Var}(\eta_{ij}; \alpha, \beta) \leq \|\eta\| = \mathrm{Var}(\eta; \alpha, \beta) \leq \sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{Var}(\eta_{ij}; \alpha, \beta). \tag{6}$$

Given $\eta(\cdot) \in BV([\alpha,\beta],\mathbb{K}^{m\times n})$ then for any continuous functions $\gamma \in C([\alpha,\beta],\mathbb{K})$ and $\phi \in C([\alpha,\beta],\mathbb{K}^n)$, the integrals

$$\int_\alpha^\beta \gamma(\theta)d[\eta(\theta)] \quad \text{and} \quad \int_\alpha^\beta d[\eta(\theta)]\phi(\theta)$$

exist and are defined, respectively, as the limits of $S_1(P) := \sum_{k=1}^p \gamma(\zeta_k)(\eta(\theta_k) - \eta(\theta_{k-1}))$ and $S_2(P) := \sum_{k=1}^p (\eta(\theta_k) - \eta(\theta_{k-1}))\phi(\zeta_k)$ as $d(P) := \max_k |\theta_k - \theta_{k-1}| \to 0$, where $P = \{\theta_1 = \alpha \le \theta_2 \le \cdots \le \theta_p = \beta\}$ is any finite partition of the interval $[\alpha,\beta]$ and $\zeta_k \in [\theta_{k-1},\theta_k]$. It is immediate from the definition that

$$\begin{aligned} \left\| \int_\alpha^\beta \gamma(\theta)d[\eta(\theta)] \right\| &\le \max_{\theta\in[\alpha,\beta]} |\gamma(\theta)|\,\|\eta\|, \\ \left\| \int_\alpha^\beta d[\eta(\theta)]\phi(\theta) \right\| &\le \max_{\theta\in[\alpha,\beta]} \|\phi(\theta)\|\,\|\eta\|. \end{aligned} \tag{7}$$

Let $\mathbb{K}^n$ be endowed with a vector norm $\|\cdot\|$ and $C([-h,0],\mathbb{K}^n)$ be a Banach space of all continuous functions on $[-h,0]$ with values in $\mathbb{K}^n$ normed by the maximum norm $\|\phi\| = \max_{\theta\in[-h,0]} \|\phi(\theta)\|$. Let $L : C([-h,0],\mathbb{K}^n) \to \mathbb{K}^n$ be a linear bounded operator. Then, by the Riesz representation theorem, there exists unique matrix function $\eta = (\eta_{ij}(\cdot)) \in BV([-h,0],\mathbb{K}^{n\times n})$ which is *continuous from the left* (or briefly c.f.l.) on $(-h,0)$ such that

$$L\phi = \int_{-h}^0 d[\eta(\theta)]\phi(\theta) \qquad \forall \phi \in C([-h,0],\mathbb{K}^n). \tag{8}$$

For any vector norm on $\mathbb{K}^n$, we have by (7), $\|L\| \le \|\eta\|$. Moreover, if $\mathbb{K}^n$ is provided with the ∞-norm so that for all $x \in \mathbb{K}^n$ and $\theta \in [-h,0]$,

$$\|x\| = \max_{1\le i\le n} |x_i| \quad \text{and} \quad \|\eta(\theta)\| = \max_{1\le i\le n} \sum_{j=1}^n |\eta_{ij}(\theta)|,$$

then it can be shown immediately that $\|L\| = \|\eta\|$. In the subsequent sections the following subspace of $BV([-h,0],\mathbb{K}^{m\times n})$ will be used frequently :

$$NBV([-h,0],\mathbb{K}^{m\times n}) := \{\eta \in BV([-h,0],\mathbb{K}^{m\times n}); \eta(-h) = 0, \eta \text{ is c.f.l. on } [-h,0]\}. \tag{9}$$

It is clear that $NBV([-h,0],\mathbb{K}^{m\times n})$ is closed in $BV([-h,0],\mathbb{K}^{m\times n})$ and thus it is a Banach space with the norm $\|\delta\| = \mathrm{Var}(\delta;-h,0)$.

3. Stability radii of linear functional differential equations under multi-perturbations. Consider a linear retarded system described by the following general functional differential equation:

$$\begin{aligned} \dot{x}(t) &= A_0x(t) + Lx_t, \quad t \ge 0, \quad x(t) \in \mathbb{R}^n, \\ x(\theta) &= \phi_0(\theta), \qquad \theta \in [-h,0], \end{aligned} \tag{10}$$

where, for each $t \ge 0, x_t \in C([-h,0],\mathbb{R}^n)$ is defined by $x_t(\theta) = x(t+\theta), \theta \in [-h,0], A_0 \in \mathbb{R}^{n\times n}$ is a given matrix and $L : C([-h,0],\mathbb{R}^n) \to \mathbb{R}^n$ is a linear bounded operator defined by

$$L\phi = \int_{-h}^0 d[\eta(\theta)]\phi(\theta), \quad t \ge 0, \quad \phi \in C([-h,0],\mathbb{R}^n), \tag{11}$$

where $\eta(\cdot) \in NBV([-h,0],\mathbb{R}^{n\times n})$ is given (real $n\times n$-matrix function of bounded variation on $[-h,0]$) such that η vanishes at $-h$ and is c.f.l. on $[-h,0]$. We shall extend the definition of η to $\mathbb{R}$ by setting $\eta(\theta)=\eta(-h)=0$ for all $\theta\le -h, \eta(\theta)=\eta(0)$ for all $\theta\ge 0$.

It is well known that, for any given $\phi_0\in C:=C([-h,0],\mathbb{R}^n)$, the system (10)–(11) has a unique solution $x(\phi_0,\cdot)$ defined and continuous on $[-h,\infty)$ (see, e.g. [5]).

The system (10)–(11) is said to be exponentially asymptotically stable or, more simply, *Hurwitz stable*, if there are constants $c>0,\alpha>0$ such that for all $\phi\in C$, the solution $x(\phi,\cdot)$ of (10)–(11) satisfies

$$\|x(\phi,t)\|\le ce^{-\alpha t}\|\phi\|, \qquad t\ge 0.$$

Denote by $H(s)$ the characteristic quasi-polynomial of the system (10)–(11), that is,

$$H(s)=sI_n-A_0-\int_{-h}^{0}e^{s\theta}d[\eta(\theta)]. \tag{12}$$

Then the necessary and sufficient condition for the system (10)–(11) to be Hurwitz stable is

$$\sigma(A_0,\eta)\subset\mathbb{C}^-:=\{s\in\mathbb{C}:\operatorname{Re}\, s<0\}, \tag{13}$$

where $\sigma(A_0,\eta)$ denotes the set of all roots of the characteristic equation of the system (10)–(11):

$$\sigma(A_0,\eta):=\{s\in\mathbb{C}:\det H(s)=0\}. \tag{14}$$

By the property of $\sigma(A_0,\eta)$ it can be shown that (13) is equivalent to the condition

$$\mu(A_0,\eta):=\max\{\operatorname{Re} s: s\in\sigma(A_0,\eta)\}<0.$$

The number $\mu(A_0,\eta)$ is called the *spectral abscissa* of the retarded system (10)–(11).

Assume that the retarded system (10)–(11) is Hurwitz stable and subjected to multi-perturbations of the type

$$\begin{aligned} A_0\to A_{0\Delta}&=A_0+\sum_{j=1}^{N}D_{0j}\Delta_jE_{0j},\quad \Delta_j\in\mathbb{C}^{l_{0j}\times q_{0j}},\quad j\in\underline{N}:=\{1,2,\dots,N\},\\ \eta\to\eta_\delta&=\eta+\sum_{j=1}^{N}D_{1j}\delta_jE_{1j},\quad \delta_j\in NBV([-h,0],C^{l_{1j}\times q_{1j}}),\quad j\in\underline{N}, \end{aligned} \tag{15}$$

and thus the perturbed system is described by

$$\begin{aligned} \dot x(t)=&\left(A_0+\sum_{j=1}^{N}D_{0j}\Delta_jE_{0j}\right)x(t)+\int_{-h}^{0}d\left[\eta(\theta)+\sum_{j=1}^{N}D_{1j}\delta_j(\theta)E_{1j}\right]x(t+\theta),\quad t\ge 0,\\ &x(\theta)=\phi_0(\theta),\qquad \theta\in[-h,0]. \end{aligned} \tag{16}$$

Here $D_{ij}\in\mathbb{C}^{n\times l_{ij}}, E_{ij}\in\mathbb{C}^{q_{ij}\times n}, i\in I:=\{0,1\}, j\in\underline{N}$ are given matrices determining the *structure* of perturbations, Δ_j and $\delta_j(\cdot), j\in\underline{N}$ are unknown disturbances. We always assume that each disturbance matrix function $\delta_j(\cdot)$ is extended to the whole

$\mathbb{R}$ by $\delta_j(\theta) = \delta_j(-h) = 0$ for $\theta \le -h$, $\delta_j(\theta) = \delta_j(0)$ for $\theta > 0$. We shall measure the size of each perturbation $\tilde{\Delta} := [\Delta, \delta]$, where

$$\Delta := (\Delta_1, \dots, \Delta_N),$$
$$\delta := (\delta_1, \dots, \delta_N), \quad \Delta_j \in \mathbb{C}^{l_{0j} \times q_{0j}}, \quad \delta_j \in NBV([-h, 0], \mathbb{C}^{l_{1j} \times q_{1j}}), \quad j \in \underline{N},$$

by the norm

$$\|\tilde{\Delta}\| := \sum_{j=1}^{N} \|\Delta_j\| + \sum_{j=1}^{N} \|\delta_j\|, \quad \|\delta_j\| := \operatorname{Var}(\delta_j; -h, 0), \quad j \in \underline{N}. \tag{17}$$

Set

$$\mathcal{D}_{\mathbb{C}} := \{\tilde{\Delta} = [\Delta, \delta] : \Delta_j \in \mathbb{C}^{l_{0j} \times q_{0j}}, \quad \delta_j \in NBV([-h, 0], \mathbb{C}^{l_{1j} \times q_{1j}}), \quad j \in \underline{N}\},$$
$$\mathcal{D}_{\mathbb{R}} := \{\tilde{\Delta} = [\Delta, \delta] : \Delta_j \in \mathbb{R}^{l_{0j} \times q_{0j}}, \quad \delta_j \in NBV([-h, 0], \mathbb{R}^{l_{1j} \times q_{1j}}), \quad j \in \underline{N}\},$$
$$\mathcal{D}_{+} := \{\tilde{\Delta} = [\Delta, \delta] : \Delta_j \in \mathbb{R}_{+}^{l_{0j} \times q_{0j}}, \quad \delta_j \in NBV([-h, 0], \mathbb{R}^{l_{1j} \times q_{1j}}) \text{ is increasing}, \ j \in \underline{N}\}.$$

Then $\mathcal{D}_{\mathbb{C}}, \mathcal{D}_{\mathbb{R}}, \mathcal{D}_{+}$ are called, respectively, the *class of complex, real, nonnegative perturbations*. In what follows, we always define $\inf \emptyset = +\infty, 0^{-1} = +\infty$. To study robustness of stability of the retarded system (10)–(11) we introduce the following.

DEFINITION 3.1. *Let the retarded system* (10)–(11) *be Hurwitz stable. The complex, real and positive stability radius of the system with respect to perturbations of the form* (15)*, measured by the norm* (17)*, is defined, respectively, by*

$$r_{\mathbb{C}} = \inf\{\|\tilde{\Delta}\| \ : \ \tilde{\Delta} \in \mathcal{D}_{\mathbb{C}}, \ \sigma(A_{0\Delta}, \eta_\delta) \not\subset \mathbb{C}^-\}, \tag{18}$$

$$r_{\mathbb{R}} = \inf\{\|\tilde{\Delta}\| \ : \ \tilde{\Delta} \in \mathcal{D}_{\mathbb{R}}, \ \sigma(A_{0\Delta}, \eta_\delta) \not\subset \mathbb{C}^-\}, \tag{19}$$

and

$$r_{+} = \inf\{\|\tilde{\Delta}\| \ : \ \tilde{\Delta} \in \mathcal{D}_{+}, \ \sigma(A_{0\Delta}, \eta_\delta) \not\subset \mathbb{C}^-\}. \tag{20}$$

REMARK 3.2. (a) *It is important to note that we can consider the above problem with respect to the multi-perturbations of the form*

$$A_0 \to A_{0\Delta} = A_0 + \sum_{j=1}^{N_1} D_{0j} \Delta_j E_{0j}, \quad \Delta_j \in \mathbb{C}^{l_{0j} \times q_{0j}}, \quad j \in \underline{N}_1 := \{1, \dots, N_1\}, \tag{21}$$

$$\eta \to \eta_\delta = \eta + \sum_{j=1}^{N_2} D_{1j} \delta_j E_{1j}, \quad \delta_j \in NBV([-h, 0], \mathbb{C}^{l_{1j} \times q_{1j}}), \quad j \in \underline{N}_2 := \{1, \dots, N_2\}.$$

However, if, for instance, $N_1 > N_2$, then the parameter perturbations (21) *can be represented in the form* (15) *by setting D_{1j} or E_{1j} being equal to zero matrix for every $j \in \underline{N}_1, j > N_2$.*

(b) *In the particular case, if*

$$\eta(\theta) = \begin{cases} 0 & \text{if} \quad \theta = -h, \\ A_1 & \text{if} \quad \theta \in (-h, 0], \end{cases} \qquad \delta_j(\theta) = \begin{cases} 0 & \text{if} \quad \theta = -h, \\ \Delta_{1j} & \text{if} \quad \theta \in (-h, 0], \end{cases}$$

where A_1 is a given matrix and $\Delta_{1j} \in \mathbb{K}^{l_{1j}\times q_{1j}}, j \in \underline{N}, \mathbb{K} = \mathbb{R}, \mathbb{C}$ are unknown matrices then the problem of computing the stability radii of the system (10)–(11) *under multi-perturbations* (15) *returns the problem of computing the stability radii of delay-time system of the form*

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-h), \qquad t \geq 0,$$

under multi-perturbations of the form

$$A_0 \to A_{0\Delta} = A_0 + \sum_{j=1}^{N} D_{0j}\Delta_{0j}E_{0j}, \quad \Delta_{0j} \in \mathbb{C}^{l_{0j}\times q_{0j}}, \quad j \in \underline{N},$$

$$A_1 \to A_{1\Delta} = A_1 + \sum_{j=1}^{N} D_{1j}\Delta_{1j}E_{1j}, \quad \Delta_{1j} \in \mathbb{C}^{l_{1j}\times q_{1j}}, \quad j \in \underline{N}, \tag{22}$$

and it has been just solved in [16].

We define the associated transfer functions $G_{(i,j;\ u,v)} : \mathbb{C}\backslash\sigma(A_0,\eta) \to \mathbb{C}^{q_{ij}\times l_{uv}}$ of the perturbed system (16) by setting

$$G_{(i,j;\ u,v)}(s) = E_{ij}H(s)^{-1}D_{uv}, \qquad i,u \in I; \quad j,v \in \underline{N}. \tag{23}$$

We need the following technical lemmas.

LEMMA 3.3 (see [23]). *Let the retarded system* (10)–(11) *be Hurwitz stable. Then, for $i,u \in I;\ j,v \in \underline{N}$, the function $s \to \|G_{(i,j;\ u,v)}(s)\|$ $(s \in \mathbb{C}, \mathrm{Re\ s} \geq 0)$ attains the supremum on $\imath\mathbb{R}$ and moreover,*

$$\max_{s\in\mathbb{C},\ \mathrm{Re}\, s\geq 0} \|G_{(i,j;\ u,v)}(s)\| = \max_{\omega\in\mathbb{R}} \|G_{(i,j;\ u,v)}(\imath\omega)\|. \tag{24}$$

LEMMA 3.4. *Let the retarded system* (10)–(11) *be Hurwitz stable. Then, for every $s_0 \in \mathbb{C}, \mathrm{Re}\ s_0 = 0$, $\max_{i\in I; j\in\underline{N}} \|G_{(i,j;\ i,j)}(s_0)\| \neq 0$, there exists a complex perturbation $\tilde{\Delta} := [\Delta, \delta]$, where $\Delta := (\Delta_1, \ldots, \Delta_N)$, $\delta := (\delta_1, \ldots, \delta_N)$, $\Delta_j \in \mathbb{C}^{l_{0j}\times q_{0j}}$, $\delta_j \in NBV([-h,0], \mathbb{C}^{l_{1j}\times q_{1j}})$, $j \in \underline{N}$ such that*

$$\|\tilde{\Delta}\| = \frac{1}{\max_{i\in I; j\in\underline{N}} \|G_{(i,j;\ i,j)}(s_0)\|} \quad \text{and} \quad \mu(A_{0\Delta}, \eta_\delta) \geq 0. \tag{25}$$

Moreover, if $G_{(i,j;\ i,j)}(0) \in {\mathbb{R}_+}^{q_{ij}\times l_{ij}}$ for every $i \in I, j \in \underline{N}$ then there exists a nonnegative perturbation $\tilde{\Delta} \in \mathcal{D}_+$ satisfying (25) *with $s_0 = 0$.*

Proof. Assume that $\max_{i\in I;\ j\in\underline{N}} \|G_{(i,j;\ i,j)}(s_0)\| = \|G_{(i_1,j_1;\ i_1,j_1)}(s_0)\|$ for some $i_1 \in I, j_1 \in \underline{N}$. From the definition of $\|G_{(i_1,j_1;\ i_1,j_1)}(s_0)\|$, there exists a vector $u_0 \in \mathbb{C}^{l_{i_1j_1}}, \|u_0\| = 1$ such that $\|G_{(i_1,j_1;\ i_1,j_1)}(s_0)u_0\| = \|G_{(i_1,j_1;\ i_1,j_1)}(s_0)\|$. Then, by the Hahn–Banach theorem, there exist $y^* \in (\mathbb{C}^{q_{i_1j_1}})^*, \|y^*\| = 1$ satisfying

$$y^*(G_{(i_1,j_1;\ i_1,j_1)}(s_0)u_0) = \|G_{(i_1,j_1;\ i_1,j_1)}(s_0)u_0\|.$$

Define $\Delta_{i_1j_1} := \|G_{(i_1,j_1;\ i_1,j_1)}(s_0)u_0\|^{-1}u_0y^* \quad \in \mathbb{C}^{l_{i_1j_1}\times q_{i_1j_1}}$. It is easy to see that $\|\Delta_{i_1j_1}\| = \|G_{(i_1,j_1;\ i_1,j_1)}(s_0)\|^{-1}$. We consider now the following step function:

$$\delta_{i_1j_1}(\theta) = \begin{cases} 0 & \text{if} \quad \theta = -h, \\ \Delta_{i_1j_1}e^{hs_0} & \text{if} \quad \theta \in (-h,0]. \end{cases}$$

It is clear that

$$\delta_{i_1j_1} \in NBV([-h,0],\mathbb{C}^{l_{i_1j_1}\times q_{i_1j_1}}), \qquad \|\delta_{i_1j_1}\| = \operatorname{Var}(\delta_{i_1j_1};-h,0) = \|\Delta_{i_1j_1}\|$$

and $\int_{-h}^{0} e^{s_0\theta} d[\delta_{i_1j_1}(\theta)] = \Delta_{i_1j_1}$. Setting $x := H(s_0)^{-1}D_{i_1j_1}u_0$, we have

$$\int_{-h}^{0} e^{s_0\theta} d[\delta_{i_1j_1}(\theta)]E_{i_1j_1}x = \Delta_{i_1j_1}E_{i_1j_1}x = u_0.$$

Therefore,

$$x = H(s_0)^{-1}D_{i_1j_1}\Delta_{i_1j_1}E_{i_1j_1}x = H(s_0)^{-1}D_{i_1j_1}\int_{-h}^{0} e^{s_0\theta} d[\delta_{i_1j_1}(\theta)]E_{i_1j_1}x \neq 0$$

and hence

$$H(s_0)x = D_{i_1j_1}\Delta_{i_1j_1}E_{i_1j_1}x = D_{i_1j_1}\int_{-h}^{0} e^{s_0\theta} d[\delta_{i_1j_1}(\theta)]E_{i_1j_1}x.$$

This implies

$$\begin{aligned}&\left(A_0 + \int_{-h}^{0} e^{s_0\theta} d[\eta(\theta)] + D_{i_1j_1}\Delta_{i_1j_1}E_{i_1j_1}\right)x\\ &\quad = \left(A_0 + \int_{-h}^{0} e^{s_0\theta} d[\eta + D_{i_1j_1}\delta_{i_1j_1}(\theta)E_{i_1j_1}]\right)x = s_0x.\end{aligned}$$

If $i_1 = 0$, then we define $\tilde{\Delta} := [\Delta,\delta] = ((\Delta_1,\dots,\Delta_N),(\delta_1,\dots,\delta_N))$, where $\Delta_j = \Delta_{i_1j_1}$ if $j = j_1$; otherwise $\Delta_j = 0, j \in \underline{N}$ and $\delta_j = 0$ for all $j \in \underline{N}$. Then the perturbation $\tilde{\Delta}$ satisfies (25). If $i_1 = 1$, then we define $\tilde{\Delta} := [\Delta,\delta] = ((\Delta_1,\dots,\Delta_N),(\delta_1,\dots,\delta_N))$, where $\Delta_j = 0$ for all $j \in \underline{N}$ and $\delta_j = \delta_{i_1j_1}$ if $j = j_1$ otherwise $\delta_j = 0, j \in \underline{N}$. Then we have $\tilde{\Delta}$ satisfying (25).

Moreover, if $G_{(i_1,j_1;\ i_1,j_1)}(s_0) \in \mathbb{R}_+^{q_{i_1j_1}\times l_{i_1j_1}}$ for $s_0 = 0$, then we have $\|G_{(i_1,j_1;\ i_1,j_1)}(0)\| = \max_{u\in\mathbb{R}_+^{l_{i_1j_1}},\|u\|=1}\|G_{(i_1,j_1;\ i_1,j_1)}(0)u\|$; see [11]. Thus we can choose $u_0 \in \mathbb{R}_+^{l_{i_1j_1}}$ such that $\|u_0\| = 1$ and $\|G_{(i_1,j_1;\ i_1,j_1)}(0)u_0\| = \|G_{(i_1,j_1;\ i_1,j_1)}(0)\|$. Since $G_{(i_1,j_1;\ i_1,j_1)}(0)u_0 \geq 0$ there exists by a theorem of Krein and Rutman [12] a positive linear form $y^* \in (\mathbb{C}^{q_{i_1j_1}})^*$ of dual norm $\|y^*\| = 1$ such that $y^*G_{(i_1,j_1;\ i_1,j_1)}(0)u_0 = \|G_{(i_1,j_1;\ i_1,j_1)}(0)u_0\|$. Hence the perturbation $\tilde{\Delta}$ constructed as above is nonnegative. The proof is complete. □

Using the above lemmas we obtain the following estimates for the complex radius.

THEOREM 3.5. *Let the retarded system* (10)–(11) *be Hurwitz stable and be subjected to multi-perturbations of the form* (15). *Then*

(26)
$$\frac{1}{\max_{i,u\in I;j,v\in\underline{N}}\{\max_{\omega\in\mathbb{R}}\|G_{(i,j;\,u,v)}(\imath\omega)\|\}} \leq r_{\mathbb{C}} \leq \frac{1}{\max_{i\in I;j\in\underline{N}}\{\max_{\omega\in\mathbb{R}}\|G_{(i,j;\,i,j)}(\imath\omega)\|\}}.$$

In particular, if $D_{ij} = D$ *for all* $i \in I; j \in \underline{N}$ *or* $E_{ij} = E$ *for all* $i \in I; j \in \underline{N}$, *then*

$$r_{\mathbb{C}} = \frac{1}{\max_{i\in I;j\in\underline{N}}\{\max_{\omega\in\mathbb{R}}\|G_{(i,j;\ i,j)}(\imath\omega)\|\}}. \tag{27}$$

Proof. Obviously, (27) is immediate from (26), if $D_{ij} = D$ for all $i \in I; j \in \underline{N}$ or $E_{ij} = E$ for all $i \in I; j \in \underline{N}$. Let $\tilde{\Delta} := [\Delta, \delta] = ((\Delta_1, \dots, \Delta_N), (\delta_1, \dots, \delta_N)), \Delta_j \in \mathbb{C}^{l_{0j} \times q_{0j}}$, $\delta_j \in NBV([-h, 0], \mathbb{C}^{l_{1j} \times q_{1j}})$, $j \in \underline{N}$ be destabilizing disturbances, so that $\sigma(A_{0\Delta}, \eta_\delta) \not\subset \mathbb{C}^-$. It follows that there exist a nonzero $x_0 \in \mathbb{C}^n$ and $s_0 \in \mathbb{C}, \operatorname{Re} s_0 \geq 0$ such that

$$\left(s_0 I_n - \left(A_0 + \sum_{j=1}^{N} D_{0j}\Delta_j E_{0j}\right) - \int_{-h}^{0} e^{s_0\theta} d\left[\eta(\theta) + \sum_{j=1}^{N} D_{1j}\delta_j(\theta)E_{1j}\right]\right)x_0 = 0.$$

This implies

$$H(s_0)x_0 = \left(\sum_{j=1}^{N} D_{0j}\Delta_j E_{0j} + \int_{-h}^{0} e^{s_0\theta} d\left[\sum_{j=1}^{N} D_{1j}\delta_j(\theta)E_{1j}\right]\right)x_0$$

or, equivalently,

$$x_0 = H(s_0)^{-1}\left(\sum_{j=1}^{N} D_{0j}\Delta_j E_{0j} + D_{1j}\int_{-h}^{0} e^{s_0\theta} d\left[\sum_{j=1}^{N} \delta_j(\theta)\right]E_{1j}\right)x_0.$$

Let $i_0 \in I, j_0 \in \underline{N}$ be indexes such that $\|E_{i_0 j_0} x_0\| = \max\{\|E_{ij}x_0\| : i \in I, j \in \underline{N}\}$ then from the last equality it follows that $E_{i_0 j_0} x_0 \neq 0$. Multiplying the last equation with $E_{i_0 j_0}$ from the left we can deduce

$$\sum_{j=1}^{N}\left(\|E_{i_0 j_0}H(s_0)^{-1}D_{0j}\|\|\Delta_j\| + \|E_{i_0 j_0}H(s_0)^{-1}D_{1j}\|\|\int_{-h}^{0} e^{s_0\theta} d[\delta_j(\theta)]\|\right) \geq 1. \tag{28}$$

Since, by (7), $\|\int_{-h}^{0} e^{s_0\theta} d[\delta_j(\theta)]\| \leq \|\delta_j\|, j \in \underline{N}$, (28) yields

$$\begin{aligned}&(\max_{i,u\in I;\ j,v\in\underline{N}} \|G_{(i,j;\ u,v)}(s_0)\|)\sum_{j=1}^{N}(\|\Delta_j\| + \|\delta_j\|)\\ &\quad = (\max_{i,u\in I;\ j,v\in\underline{N}} \|G_{(i,j;\ u,v)}(s_0)\|)\|\tilde{\Delta}\| \geq 1,\end{aligned}$$

and thus

$$\|\tilde{\Delta}\| \geq \frac{1}{\max_{i,u\in I;\, j,v\in\underline{N}} \|G_{(i,j;\,u,v)}(s_0)\|} \geq \frac{1}{\max_{i,u\in I;\, j,v\in\underline{N}}\{\max_{s\in\mathbb{C},\operatorname{Re} s\geq 0} \|G_{(i,j;\,u,v)}(s)\|\}}.$$

Therefore, by Lemma 3.3 and Definition 3.5, we have

$$r_{\mathbb{C}} \geq \frac{1}{\max_{i,u\in I;j,v\in\underline{N}}\{\max_{\omega\in\mathbb{R}} \|G_{(i,j;\ u,v)}(\imath\omega)\|\}}.$$

It remains to prove that

$$r_{\mathbb{C}} \leq \frac{1}{\max_{i,u\in I;j,v\in\underline{N}}\{\max_{\omega\in\mathbb{R}} \|G_{(i,j;\ u,v)}(\imath\omega)\|\}}.$$

However, this inequality follows from Lemma 3.4 and the definition of $r_{\mathbb{C}}$. This completes our proof. ∎

The above theorem includes, as a particular case, the results of Theorem 1 in [20] and of Theorem 3.3 in [23] where the formulae for complex stability radii of the linear retarded systems under single perturbations (i.e., $N = 1$) are established. The problem of computation of the real stability radius $r_{\mathbb{R}}$ is much more difficult and has been so far solved only for ordinary linear systems of the form $\dot{x}(t) = A_0x(t),\ t \geq 0$, where the system matrix A_0 is subjected to single perturbations; see, e.g., [17]. We note that, by definition, $r_{\mathbb{C}} \leq r_{\mathbb{R}}$, so $r_{\mathbb{C}}$ can be accepted as the lower bound for $r_{\mathbb{R}}$. Unfortunately, as shown in many previous papers (see, e.g., [7]) these two stability radii can be arbitrarily distinct. Therefore, it is an interesting problem to find classes of systems of practical interest for which these two stability radii coincide. Motivated by the results of [8], [9], [21], [22], [23], in this section we show that for the class of positive functional differential equations, the positive, real and complex stability radius under multi-perturbations coincide and can be computed by a simple formula.

Consider a linear system described by the functional differential equation of the form

$$\begin{aligned} \dot{x}(t) &= A_0x(t) + \int_{-h}^{0} d[\eta(\theta)]x(t+\theta) \quad t \geq 0, \quad x(t) \in \mathbb{R}^n, \\ x(\theta) &= \phi_0(\theta), \qquad \theta \in [-h, 0], \end{aligned} \tag{29}$$

where $\phi_0(\cdot) \in C([-h, 0], \mathbb{R}^n)$ is given and $\eta \in NBV([-h, 0], \mathbb{R}^{n\times n})$ being a given function of bounded variation such that $\eta(-h) = 0$ and η is c.f.l. on $[-h, 0]$. The solution of the system (29) will be denoted by $x(\phi_0, \cdot)$.

Definition 3.6. *System* (29) *is called positive if for every nonnegative initial function* $\phi_0 \in C([-h, 0], \mathbb{R}^n_+)$, *the corresponding solution* $x(\phi_0, \cdot)$ *satisfies* $x(\phi_0, t) \in \mathbb{R}^n_+$ *for every* $t \geq 0$.

The following theorem is a corollary of Proposition 3.5 in [14, p. 224].

Theorem 3.7 (see [23]). *The system* (29) *is positive if and only if* $A_0 \in \mathbb{R}^{n\times n}$ *is a Metzler matrix and* $\eta(.) \in NBV([-h, 0], \mathbb{R}^{n\times n})$ *is an increasing matrix function.*

From Theorem 3.7 it is easy to see that the linear retarded system of the form

$$\dot{x}(t) = A_0x(t) + A_1x(t-h), \qquad t \geq 0$$

is positive if and only if A_0 is a Metzler matrix and $A_1 \geq 0$.

In order to give characterization of the real stability radius of the positive systems, we need some technical lemmas. Define, for each $s \in \mathbb{C}$,

$$Q(s) = \int_{-h}^{0} e^{s\theta} d[\eta(\theta)].$$

Then, it follows from $\eta(.)$ to be an increasing matrix function that, for any $s \in \mathbb{C}$ and $x \in \mathbb{C}^n$,

$$|Q(s)x| = \left|\left(\int_{-h}^{0} e^{s\theta} d[\eta(\theta)]\right)x\right| \leq \left(\int_{-h}^{0} e^{\theta \operatorname{Re} s} d[\eta(\theta)]\right)|x| = Q(\operatorname{Re} s)|x|. \tag{30}$$

Lemma 3.8 (see [23]). *Assume* $A_0 \in \mathbb{R}^{n\times n}$ *a Metzler matrix and* $\eta(.) \in NBV([-h, 0], \mathbb{R}^{n\times n})$ *is an increasing matrix function. Let* $\mu_0 := \mu(A_0, \eta)$ *be the spectral abscissa of the delay system* (29) *and* $\mu_1 := \mu(A_0 + Q(\mu_0))$ *be the spectral abscissa of the matrix* $A_0 + Q(\mu_0)$. *Then we have*

$$\mu_0 = \mu_1.$$

LEMMA 3.9 (see [23]). *Let A_0 be a Metzler matrix and $\eta(\cdot) \in NBV([-1,0],\mathbb{R}^{n\times n})$ be an increasing matrix function. Then*

(i) *with the characteristic quasi-polynomial of the system* (29) *$H(s)$ defined by* (12)*, we have*

$$H(t_1)^{-1} \geq H(t_2)^{-1} \geq 0 \quad for \quad t_2 > t_1 > \mu(A_0,\eta); \tag{31}$$

(ii) *assume that the linear system* (29) *is Hurwitz stable and $D_{uv} \in \mathbb{R}_+^{n\times l_{uv}}, E_{ij} \in \mathbb{R}_+^{q_{ij}\times n}$ for some $i,u \in I; j,v \in \underline{N}$ and the transfer function $G_{(i,j;\ u,v)}(\cdot)$ is defined by* (23)*. Then*

$$\max_{\omega\in\mathbb{R}} \|G_{(i,j;\ u,v)}(\imath\omega)\| = \|G_{(i,j;\ u,v)}(0)\|. \tag{32}$$

We are now in position to prove the main result of this section.

THEOREM 3.10. *Let the linear retarded system* (29) *be positive and Hurwitz stable. Assume A_0,η are subjected to multi-perturbation of the form* (15)*, where $D_{ij} \in \mathbb{R}_+^{n\times l_{ij}}$, $E_{ij} \in \mathbb{R}_+^{q_{ij}\times n}$, $i \in I, j \in \underline{N}$. If $D_{ij} = D$ for all $i \in I, j \in \underline{N}$ or $E_{ij} = E$ for all $i \in I, j \in \underline{N}$, then we have*

$$r_{\mathbb{C}} = r_{\mathbb{R}} = r_+ = \frac{1}{\max_{i\in I, j\in\underline{N}} \|G_{(i,j;\ i,j)}(0)\|}.$$

Proof. From (27) and (32),

$$r_{\mathbb{C}} = \frac{1}{\max_{i\in I, j\in\underline{N}} \|G_{(i,j;\ i,j)}(0)\|}. \tag{33}$$

Since the system (29) is Hurwitz stable, positive, and $D_{ij} \in \mathbb{R}_+^{n\times l_{ij}}, E_{ij} \in \mathbb{R}_+^{q_{ij}\times n}, i \in I, j \in \underline{N}$, it follows from (31) that $G_{(i,j;\ i,j)}(0) \in \mathbb{R}_+^{q_{ij}\times l_{ij}}$ for all $i \in I, j \in \underline{N}$. By definition of r_+ and Lemma 3.4, we get

$$r_+ \leq \frac{1}{\max_{i\in I, j\in\underline{N}} \|G_{(i,j;\ i,j)}(0)\|}. \tag{34}$$

It follows from (33), (34), and the inequalities $r_{\mathbb{C}} \leq r_{\mathbb{R}} \leq r_+$ that

$$r_{\mathbb{C}} = r_{\mathbb{R}} = r_+ = \frac{1}{\max_{i\in I, j\in\underline{N}} \|G_{(i,j;\ i,j)}(0)\|}. \qquad \square$$

We illustrate the above result by two following simple examples.

Example 3.11. Consider a positive linear time-delay system described by the following scalar equation:

$$\dot{x}(t) = -x(t) + \int_{-1}^{0} e^{\theta} x(t+\theta) d\theta \quad t \geq 0, \quad x(t) \in \mathbb{R}. \tag{35}$$

It is easy to see that (35) can be represented of the form (29), with $\eta = e^{\theta} - e^{-1}$. Clearly, $\eta \in NBV([-1,0],\mathbb{R})$ and it is an increasing function on $[-1,0]$. Then, the characteristic quasi-polynomial of (35) is given by

$$H(s) = s + 1 - \int_{-1}^{0} e^{s\theta} d[\eta(\theta)] = s + 1 - \int_{-1}^{0} e^{(s+1)\theta} d\theta. \tag{36}$$

Therefore, $H(s) = 0$ if and only if $(s^2 + 2s)e^s + e^{-1} = 0$. By Theorem 13.9 in [1], it is easy to verify that all roots of this equation have negative real parts. Thus, the system (35) is Hurwitz stable. Assume the system (35) is perturbed as follows:

$$\dot{x}(t) = (-1 + \delta_0)x(t) + \int_{-1}^{0} (e^{\theta} + 2003\Delta_1(\theta) + 2002\Delta_2(\theta))x(t+\theta)d\theta, \tag{37}$$

where $\delta_0 \in \mathbb{R}$ is an unknown parameter scalar and $\Delta_1(\theta), \Delta_2(\theta)$ are unknown integrable functions on $[-1, 0]$. This perturbed system can be rewritten in the form

$$\dot{x}(t) = (-1 + \delta_0)x(t) + \int_{-1}^{0} d[\eta(\theta) + 2003\delta_1(\theta) + 2002\delta_2(\theta)]x(t+\theta), \tag{38}$$

where

$$\delta_1(\theta) = \int_{-1}^{\theta} \Delta_1(\tau)d\tau, \quad \delta_2(\theta) = \int_{-1}^{\theta} \Delta_2(\tau)d\tau, \quad \theta \in [-1, 0].$$

By Theorem 3.10, we conclude that the perturbed system (37) is Hurwitz stable for all $\delta_0 \in \mathbb{R},\ \Delta_1(\cdot),\ \Delta_2(\cdot) \in L_1([-1, 0], \mathbb{R})$ satisfying

$$|\delta_0| + V_{-1}^0(\delta_1) + V_{-1}^0(\delta_2) = |\delta_0| + \int_{-1}^{0} |\Delta_1(\theta)|d\theta + \int_{-1}^{0} |\Delta_2(\theta)|d\theta < r_{\mathbb{R}} = \frac{1}{2003e}.$$

On the other hand, if we take the perturbation $\delta_0 = 0,\ \Delta_1(\theta) = \frac{1-e^{\theta}}{2003},\ \Delta_2(\theta) = 0,\ \theta \in [-1, 0]$, then the perturbed system becomes

$$\dot{x}(t) = -x(t) + \int_{-1}^{0} x(t+\theta)d\theta, \quad t \geq 0, \quad x(t) \in \mathbb{R}$$

for which the characteristic quasi-polynomial is

$$H(s) = s + 1 - \int_{-1}^{0} e^{s\theta}d\theta.$$

Since, clearly, $H(0) = 0$ it follows that the perturbed system is not Hurwitz stable. Note that

$$|\delta_0| + V_{-1}^0(\delta_1) + V_{-1}^0(\delta_2) = \int_{-1}^{0} |\Delta_1(\theta)|d\theta = \frac{1}{2003e} = r_{\mathbb{R}}.$$

Example 3.12. Consider the positive linear delay system

$$\dot{x}(t) = A_0x(t) + A_1x(t-1), \qquad t \geq 0,$$

where

$$A_0 = \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix}, \qquad A_1 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Then the characteristic equation of the system is

$$e^s(s^2 + 3s + 2) - 1 = 0.$$

By Theorem 3.9 in [1] this equation has only roots with negative real parts. Therefore, the above delay system is Hurwitz stable. Suppose the system matrices A_0, A_1 are subjected to parameter perturbations of the form

$$A_0 \to A_0 + D_{01}\Delta_{01}E_{01} + D_{02}\Delta_{02}E_{02},$$
$$A_1 \to A_1 + D_{11}\Delta_{11}E_{11} + D_{12}\Delta_{12}E_{12},$$

where

$$E_{01} = E_{02} = E_{11} = E_{12} = I_2,$$

$$D_{01} = \begin{bmatrix}1\\0\end{bmatrix}, \quad D_{02} = \begin{bmatrix}1 & 1\\0 & 1\end{bmatrix}, \quad D_{11} = \begin{bmatrix}0\\1\end{bmatrix}, \quad D_{12} = \begin{bmatrix}1 & 0\\1 & 1\end{bmatrix},$$

$\Delta_{01},\ \Delta_{11} \in \mathbb{K}^{1\times 2},\ \Delta_{02},\ \Delta_{12} \in \mathbb{K}^{2\times 2},\ \mathbb{K} = \mathbb{R}, \mathbb{C}.$

Then we get

$$G_{(0,1;\ 0,1)}(0) = (-A_0 - A_1)^{-1} D_{01} = \begin{bmatrix}2 & 1\\1 & 1\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}2\\1\end{bmatrix},$$

$$G_{(0,2;\ 0,2)}(0) = (-A_0 - A_1)^{-1} D_{02} = \begin{bmatrix}2 & 1\\1 & 1\end{bmatrix}\begin{bmatrix}1 & 1\\0 & 1\end{bmatrix} = \begin{bmatrix}2 & 3\\1 & 2\end{bmatrix},$$

$$G_{(1,1;\ 1,1)}(0) = (-A_0 - A_1)^{-1} D_{11} = \begin{bmatrix}2 & 1\\1 & 1\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix},$$

$$G_{(1,2;\ 1,2)}(0) = (-A_0 - A_1)^{-1} D_{12} = \begin{bmatrix}2 & 1\\1 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\1 & 1\end{bmatrix} = \begin{bmatrix}3 & 1\\2 & 1\end{bmatrix}.$$

Therefore, if $\mathbb{R}^2$ is endowed with, respectively, with the 1-norm, 2-norm, ∞-norm, then by Theorem 3.10, the perturbed system

$$\dot{x}(t) = \left(A_0 + \sum_{j=1}^{2} D_{0j}\Delta_{0j}E_{0j}\right)x(t) + \left(A_1 + \sum_{j=1}^{2} D_{1j}\Delta_{1j}E_{1j}\right)x(t-1), \qquad t \geq 0,$$

remains Hurwitz stable for all $\Delta_{ij}, i \in \{0,1\}; j \in \{1,2\}$ satisfying

$$\sum_{i=0}^{1}\sum_{j=1}^{2} \|\Delta_{ij}\| < \frac{1}{5}, \quad \frac{1}{\sqrt{9+\sqrt{80}}}, \quad \frac{1}{5}, \text{ respectively.}$$

On the other hand, if, for instance, $\mathbb{R}^2$ is endowed with 1-norm then the destabilizing perturbation of minimal size $\sum_{i=0}^{1}\sum_{j=1}^{2} \|\Delta_{ij}\| = \frac{1}{5}$ is given by $\Delta := (\Delta_{01}, \Delta_{02}, \Delta_{11}, \Delta_{12})$, where $\Delta_{01} = (0,0), \Delta_{02} = \left[\begin{smallmatrix}0 & 0\\ \frac{1}{5} & \frac{1}{5}\end{smallmatrix}\right], \Delta_{11} = (0,0), \Delta_{12} = \left[\begin{smallmatrix}0 & 0\\ 0 & 0\end{smallmatrix}\right]$.

4. Stability radii of positive linear functional differential equations under multi-affine perturbations. In this section we study the robust stability of the positive system (29) where A_0, η are subjected to multi-affine perturbations of the following form:

$$\begin{aligned} A_0 &\to A_0 + \sum_{i=1}^{N_0} \alpha_i A_i, \\ \eta &\to \eta + \sum_{j=1}^{N_1} \beta_j \eta_j, \end{aligned} \tag{39}$$

where $A_i \in \mathbb{R}^{n\times n}, \eta_j \in NBV([-h,0],\mathbb{R}^{n\times n})$ are given and $\alpha_i, \beta_j \in \mathbb{C}$ are unknown scalar parameters. We note that the class of multi-affine perturbations of this type has been considered first in [8], [9], [18] for positive linear systems with no time delays and has just extended to positive time-delay systems in [22], [10]. In this section, we extend these results to general functional differential equations.

As in the previous section, we assume that A_0 is a Metzler matrix and $\eta \in NBV([-h,0],\mathbb{R}^{n\times n})$ is an increasing matrix function under which the linear retarded system (29) is positive. In view of positivity of the system under consideration, we assume that $A_i \in \mathbb{R}_+^{n\times n}, i \in \underline{N}_0 := \{1,2,\dots,N_0\}$, and $\eta_j \in NBV([-h,0],\mathbb{R}^{n\times n})$ are increasing, for all $j \in \underline{N}_1 := \{1,2,\dots,N_1\}$. We define the complex and the real stability radii of the linear retarded system (29) under affine parameter perturbations (39) by setting, for $\mathbb{K} = \mathbb{C}$ and, respectively, $\mathbb{K} = \mathbb{R}$,

$$r_{\mathbb{K}}^a = \inf\left\{ \max(\max_{i\in\underline{N}_0}|\alpha_i|;\ \max_{j\in\underline{N}_1}|\beta_j|) : \alpha_i,\beta_j \in \mathbb{K},\ \mu\Big(A_0 + \sum_{i=1}^{N_0}\alpha_i A_i, \eta + \sum_{j=1}^{N_1}\beta_j\eta_j\Big) \ge 0\right\}. \tag{40}$$

Similarly, the positive stability radius r_+^a is obtained by restricting, in the above definition, the disturbances $(\alpha,\beta) := ((\alpha_i)_{i\in\underline{N}_0}, (\beta_j)_{j\in\underline{N}_1})$ to be nonnegative.

It is clear that $r_{\mathbb{C}}^a \le r_{\mathbb{R}}^a \le r_+^a$. The following theorem gives us computable formulae of stability radii $r_{\mathbb{K}}^a, r_+^a, (\mathbb{K} = \mathbb{C}, \mathbb{R})$. For the simplicity of presentation, as in the previous section, let us define, for $s \in \mathbb{R}$ and $j \in \underline{N}_1$,

$$Q(s) = \int_{-h}^{0} e^{s\theta} d[\eta(\theta)]; \qquad Q_j(s) = \int_{-h}^{0} e^{s\theta} d[\eta_j(\theta)]. \tag{41}$$

By the above assumption, for all $s \ge 0, j \in \underline{N}_1$,

$$Q(0) \ge Q(s) \ge 0, \qquad Q_j(0) \ge Q_j(s) \ge 0. \tag{42}$$

THEOREM 4.1. *Suppose the positive linear retarded system* (29) *is Hurwitz stable and A_0, η are subjected to multi-affine perturbations of the form* (39)*. If the stability radii of the system are given by* (40)*, then*

$$r_{\mathbb{C}}^a = r_{\mathbb{R}}^a = r_+^a = \frac{1}{\mu[(-A_0 - Q(0))^{-1}(\sum_{i=1}^{N_0} A_i + \sum_{j=1}^{N_1} Q_j(0))]}. \tag{43}$$

Proof. Let $(\alpha,\beta) = ((\alpha_i)_{i\in\underline{N}_0}, (\beta_j)_{j\in\underline{N}_1})$ be an arbitrary nonnegative destabilizing perturbation, that is,

$$\mu_1 := \mu\Big(A_0 + \sum_{i=1}^{N_0}\alpha_i A_i, \eta + \sum_{j=1}^{N_1}\eta_j\Big) \ge 0.$$

Using Lemma 3.8 and Theorem 2.1(i) (Perron–Frobenius), we have

$$\Big(A_0+\sum_{i=1}^{N_0}\alpha_iA_i+\int_{-h}^{0}e^{\mu_1\theta}d\Big[\eta+\sum_{j=1}^{N_1}\beta_j\eta_j(\theta)\Big]\Big)y_0=\mu_1y_0$$

for some $y_0\in\mathbb{R}^n_+, y_0\neq 0$. Since the system (29) is Hurwitz stable, $\mu_0:=\mu(A_0,\eta)<0\leq\mu_1$. This yields to

$$H(\mu_1)^{-1}\Big(\sum_{i=1}^{N_0}\alpha_iA_i+\sum_{j=1}^{N_1}\beta_jQ_j(\mu_1)\Big)y_0=y_0, \tag{44}$$

where $H(s)=sI_n-A_0-Q(s)$ is the characteristic quasi-polynomial of the system (29). Let $\gamma=\max\{\max_{i\in\underline{N}_0}\alpha_i,\max_{j\in\underline{N}_1}\beta_j\}$. From $\mu_1\geq 0>\mu_0$, applying the Lemma 3.9(i), we have $H(0)^{-1}\geq H(\mu_1)^{-1}\geq 0$. Therefore, from (44)

$$\gamma\Big(H^{-1}(0)\Big(\sum_{i=1}^{N_0}A_i+\sum_{j=1}^{N_1}Q_j(0)\Big)\Big)y_0=\gamma By_0\geq y_0,$$

where $B:=H(0)^{-1}(\sum_{i=1}^{N_0}A_i+\sum_{j=1}^{N_1}Q_j(0))$. By Theorem 2.1(ii),

$$\mu(B)\geq\frac{1}{\gamma}>0.$$

Hence, $\gamma\geq\frac{1}{\mu(B)}$. Since this holds for arbitrary nonnegative perturbation (α,β), we conclude $r^a_+\geq\frac{1}{\mu(B)}$.

We shall prove that the converse inequality holds true. Since $B\geq 0$, by Theorem 2.1(i) (Perron–Frobenius), there exists a nonzero vector $y\in\mathbb{R}^n_+$ such that $By=\mu(B)y$. This implies

$$H(0)y=\Big(\sum_{i=1}^{N_0}\frac{1}{\mu(B)}A_i+\sum_{j=1}^{N_1}\frac{1}{\mu(B)}Q_j(0)\Big)y,$$

or, equivalently,

$$\Big(\Big(A_0+\sum_{i=1}^{N_0}\frac{1}{\mu(B)}A_i\Big)+\int_{-h}^{0}d\Big[\eta+\sum_{j=1}^{N_1}\frac{1}{\mu(B)}\eta_j(\theta)\Big]\Big)y=0.$$

It means that nonnegative perturbation (α^*,β^*) defined by $\alpha^*_i=1/\mu(B), i\in\underline{N}_0, \beta^*_j=1/\mu(B), j\in\underline{N}_1$ is destabilizing. By the definition of r^a_+, we have

$$r^a_+\leq\frac{1}{\mu(B)}.$$

Thus, we obtain

$$r^a_+=\frac{1}{\mu(B)}=\frac{1}{\mu[(-A_0-Q(0))^{-1}(\sum_{i=1}^{N_0}A_i+\sum_{j=1}^{N_1}Q_j(0))]}.$$

We are now ready to show that $r^a_{\mathbb{C}} = r^a_{\mathbb{R}} = r^a_+$. Suppose that $(\alpha, \beta) = ((\alpha_i)_{i\in\underline{N}_0}, (\beta_j)_{j\in\underline{N}_1})$ is an arbitrary complex destabilizing perturbation. Set

$$\mu_c := \mu\Big(A_0 + \sum_{i=1}^{N_0} \alpha_i A_i, \eta + \sum_{j=1}^{N_1} \beta_j \eta_j\Big) \geq 0,$$

$$\mu_+ := \mu\Big(A_0 + \sum_{i=1}^{N_0} |\alpha_i| A_i, \eta + \sum_{j=1}^{N_1} |\beta_j| \eta_j\Big). \tag{45}$$

By Lemma 3.8,

$$\mu_+ = \mu\Big(A_0 + \sum_{i=1}^{N_0} |\alpha_i| A_i + \int_{-h}^{0} e^{\mu_+ \theta} d\Big[\eta(\theta) + \sum_{j=1}^{N_1} |\beta_j| \eta_j(\theta)\Big]\Big). \tag{46}$$

On the other hand, from the definition of μ_c there exist a nonzero x and $\omega \in \mathbb{R}$ such that

$$\Big(A_0 + \sum_{i=1}^{N_0} \alpha_i A_i + \int_{-h}^{0} e^{(\mu_c + \imath\omega)\theta} d\Big[\eta(\theta) + \sum_{j=1}^{N_1} \beta_j \eta_j(\theta)\Big]\Big) x = (\mu_c + \imath\omega) x. \tag{47}$$

Since A_0 is a Metzler matrix, it follows that $t_0 I_n + A_0 \in \mathbb{R}^{n\times n}_+$, for some $t_0 \in \mathbb{R}$. From (47) we derive

$$\Big((t_0 I_n + A_0) + \sum_{i=1}^{N_0} \alpha_i A_i + \sum_{j=1}^{N_1} \beta_j \int_{-h}^{0} e^{(\mu_c + \imath\omega)\theta} d[\eta(\theta) + \eta_j(\theta)]\Big) x = ((t_0 + \mu_c) + \imath\omega) x.$$

It follows that

$$\begin{aligned}(t_0 + \mu_c)|x| &\leq |((t_0 + \mu_c) + \imath\omega)x| \\ &\leq \Big(|t_0 I_n + A_0| + \sum_{i=1}^{N_0} |\alpha_i| A_i + \sum_{j=1}^{N_1} |\beta_j| \int_{-h}^{0} |e^{(\mu_c + \imath\omega)\theta}| d[\eta(\theta) + \eta_j(\theta)]\Big)|x|\end{aligned}$$

and hence

$$\Big(A_0 + \sum_{i=1}^{N_0} |\alpha_i| A_i + \int_{-h}^{0} e^{\mu_c \theta} d\Big[\eta(\theta) + \sum_{j=1}^{N_1} |\beta_j| \eta_j(\theta)\Big]\Big)|x| \geq \mu_c |x|.$$

This yields, by virtue of Theorem 2.1(ii),

$$\mu\Big(A_0 + \sum_{i=1}^{N_0} |\alpha_i| A_i + \int_{-h}^{0} e^{\mu_c \theta} d\Big[\eta(\theta) + \sum_{j=1}^{N_1} |\beta_j| \eta_j(\theta)\Big]\Big) \geq \mu_c. \tag{48}$$

Combining (46) and (48), taking Theorem 2.1(iv) into account, we get $\mu_+ \geq \mu_c \geq 0$, which means that $(|\alpha|, |\beta|) := ((|\alpha_i|)_{i\in\underline{N}_0}, (|\beta_j|)_{j\in\underline{N}_1})$ is a nonnegative destabilizing perturbation. From the definition of $r^a_{\mathbb{C}}, r^a_+$, we conclude $r^a_{\mathbb{C}} \geq r^a_+$. In combining with the inequalities $r^a_{\mathbb{C}} \leq r^a_{\mathbb{R}} \leq r^a_+$, this implies $r^a_{\mathbb{C}} = r^a_{\mathbb{R}} = r^a_+$. In addition, from the

above arguments, we observe that $r^a_{\mathbb{C}} = r^a_{\mathbb{R}} = r^a_+ = \infty$ if and only if $\mu(B) = 0$. This completes our proof. □

Note that in the formula (43) μ can be replaced by ρ because for nonnegative matrices the spectral abscissa and the spectral radius coincide.

As an example, let, for $j \in \underline{N}_1$,

$$\eta(\theta) = \begin{cases} 0 & \text{if} \quad \theta = -h, \\ B_0 & \text{if} \quad \theta \in (-h, 0], \end{cases} \qquad \eta_j(\theta) = \begin{cases} 0 & \text{if} \quad \theta = -h, \\ B_j & \text{if} \quad \theta \in (-h, 0], \end{cases} \tag{49}$$

where $B_0, B_j \in \mathbb{R}^{n\times n}_+, j \in \underline{N}_1$. Then, by the above theorem, the real stability radius of the positive delay system $\dot{x}(t) = A_0x(t) + B_0x(t-h),\ t \geq 0$, under the affine parameter perturbations

$$\begin{aligned} A_0 &\to A_0 + \textstyle\sum_{i=1}^{N_0} \alpha_i A_i, \\ B_0 &\to B_0 + \textstyle\sum_{j=1}^{N_1} \beta_j B_j \end{aligned} \tag{50}$$

(with $A_i \geq 0, B_j \geq 0, i \in \underline{N}_0, j \in \underline{N}_1$) is given by

$$r^a_{\mathbb{C}} = r^a_{\mathbb{R}} = r^a_+ = \frac{1}{\rho[(-A_0 - B_0)^{-1}(\sum_{i=1}^{N_0} A_i + \sum_{j=1}^{N_1} B_j)]};$$

compare with [13], [22].

REFERENCES

[1] R. Bellman and K.L. Cooke, *Differential-Difference Equations,* Academic Press, New York, 1963.

[2] A. Berman and R.J. Plemmons, *Nonnegative Matrices in Mathematical Sciences,* Academic Press, New York, 1979.

[3] A. Fischer, *Stability radii of infinite-dimensional positive systems,* Math. Control Signals Systems, 10 (1997), pp. 223–236.

[4] A. Fischer, D. Hinrichsen, and N.K. Son, *Stability radii of Metzler operators,* Vietnam J. Math., 26 (1998), pp. 147–163.

[5] J. Hale, *Theory of Functional Differential Equations,* Academic Press, New York, 1977.

[6] D. Hinrichsen and A.J. Pritchard, *Stability radius for structured perturbations and the algebraic Riccati equation,* Systems Control Lett., 8 (1986), pp. 105–113.

[7] D. Hinrichsen and A.J. Pritchard, *Real and complex stability radii: A survey,* in Control of Uncertain Systems, D. Hinrichsen and B. Mårtensson, eds., Progr. Systems Control Theory 6, Birkhäuser, Basel, 1990, pp. 119–162.

[8] D. Hinrichsen and N.K. Son, *μ-Analysis and robust stability of positive linear systems,* Appl. Math. Comput. Sci., 8 (1998), pp. 253–268.

[9] D. Hinrichsen and N.K. Son, *Stability radii of positive discrete-time systems under affine parameter perturbations,* Internat. J. Robust Nonlinear Control, 8 (1998), pp. 1169–1188.

[10] D. Hinrichsen, N.K. Son, and P.H.A. Ngoc, *Stability radii of positive higher order difference systems,* Systems Control Lett., 49 (2003), pp. 377–388.

[11] R.A. Horn and Ch.R. Johnson, *Matrix Analysis,* Cambridge University Press, Cambridge, UK, 1993.

[12] M.G. Krein and M.A. Rutman, *Linear operators leaving invariant a cone in a Banach space,* Amer. Math. Soc. Transl., 10 (1948), pp. 199–325.

[13] J.S. Lou, A. Johnson, and P.P.J. Van Den Bosch, *Delay-independent robust stability of uncertain linear systems,* Systems Control Lett., 24 (1995), pp. 33–39.

[14] R. Nagel, ed., *One-Parameter Semigroups of Positive Operators,* Springer-Verlag, Berlin, 1986.

[15] P.H.A. Ngoc and N.K. Son, *Stability radii of positive linear difference equations under affine parameter perturbations,* Appl. Math. Comput., 134 (2003), pp. 577–594.

[16] P.H.A. Ngoc and N.K. Son, *Stability radii of linear systems under multi-perturbations,* Numer. Funct. Anal. Optim., 25 (2004), pp. 221–238.

[17] L. Qiu, B. Bernhardsson, A. Rantzer, E.J. Davison, P.M. Young, and J.C. Doyle, *A formula for computation of the real stability radius,* Automatica, 31 (1995), pp. 879–890.

[18] B. Shafai, J. Chen, and M. Kothandaraman, *Explicit formulas for stability radii of nonnegative and Metzlerian matrices,* IEEE Trans. Automat. Control, 42 (1997), pp. 265–269.

[19] N.K. Son and D. Hinrichsen, *Robust stability of positive continuous-time systems,* Numer. Funct. Anal. Optim., 17 (1996), pp. 649–659.

[20] N.K. Son and P.H.A. Ngoc, *Complex stability radius of linear time-delay systems,* Vietnam J. Math., 26 (1998), pp. 379–383.

[21] N.K. Son and P.H.A. Ngoc, *Stability radius of linear delay systems,* in Proceedings of the American Control Conference, San Diego, California, 1999, pp. 815–816.

[22] N.K. Son and P.H.A. Ngoc, *Robust stability of positive linear time delay systems under affine parameter perturbations,* Acta Math. Vietnam., 24 (1999), pp. 353–372.

[23] N.K. Son and P.H.A. Ngoc, *Robust stability of linear functional differential equations,* Adv. Stud. Contemp. Math., 3(2) (2001), pp. 43–59.